Thin groups and superstrong approximation
MSRI Publications
Volume **61**, 2013

The orbital circle method

ALEX KONTOROVICH

We discuss several recent applications of the circle method to the study of integers represented by thin orbits.

1. "Almost every" local-global theorems in thin orbits

A number of articles in these proceedings will discuss the *affine sieve* method which in general produces some (usually unspecified) number of almost-primes in an orbit of affine linear maps [Bourgain et al. 2010a; Salehi Golsefidy and Sarnak 2011]. Our goal in these notes is to highlight several occasions in which actual primes can be produced in the thin setting.[1] In the results surveyed below, these come from the stronger statement of an "almost every" local-to-global phenomenon, which has recently been established in several a priori unrelated settings by Bourgain and the author. The key idea is to replace sieving with the circle method, and develop the latter in the setting of thin orbits. We now state these results.

1A. ***Representing numbers by thin subgroups of*** $\mathbf{SL_2}(\mathbb{Z})$**.** In [Bourgain and Kontorovich 2010] we studied integers represented by certain "thin" subgroups $\Gamma < \mathrm{SL}(2,\mathbb{Z})$, where thin means that Γ is of infinite index in $\mathrm{SL}(2,\mathbb{Z})$ (equivalently, the hyperbolic surface $\Gamma\backslash\mathbb{H}$ has infinite volume). Fix vectors $\boldsymbol{v}_0, \boldsymbol{w}_0 \in \mathbb{Z}^2$, assumed to be primitive (their coordinates are coprime). We shall say that $n \in \mathbb{Z}$ is *represented* by the triple $(\Gamma, \boldsymbol{v}_0, \boldsymbol{w}_0)$ if there is a $\gamma \in \Gamma$ such that $n = \langle \boldsymbol{v}_0\gamma, \boldsymbol{w}_0\rangle$. Here the inner product is the standard one on $\mathbb{R}^2$. Let S be the set of represented integers,

$$S = \langle \boldsymbol{v}_0 \cdot \Gamma, \boldsymbol{w}_0\rangle. \tag{1.1}$$

Partially supported by NSF grants DMS-1209373, DMS-1064214 and DMS-1001252.

[1] For our purposes, "thin" means the following. Let $\Gamma \subset \mathrm{GL}_n(\mathbb{Z})$ be a finitely generated semigroup and for $\boldsymbol{v}_0 \in \mathbb{Z}^n$, let $\mathcal{O} := \boldsymbol{v}_0 \cdot \Gamma$ be a Γ-orbit. Fix a norm $\|\cdot\|$ on $\mathbb{R}^n$, and let $\bar{\mathcal{O}}$ be the Zariski closure of $\mathcal{O}$. We call the orbit $\mathcal{O}$ *thin* if there is an $\eta > 0$ such that, for all T large,

$$|\{\boldsymbol{v} \in \mathcal{O} : \|\boldsymbol{v}\| < T\}| < |\{\boldsymbol{v} \in \bar{\mathcal{O}}(\mathbb{Z}) : \|\boldsymbol{v}\| < T\}|^{1-\eta}.$$

Other authors can simply define a Γ-orbit to be thin if Γ has infinite covolume in its Zariski closure, but this definition only makes sense when Γ is a group. Since our applications require us to consider semigroups Γ, we reformulate thinness as above.

For example, if $\boldsymbol{v}_0 = \boldsymbol{w}_0 = (0, 1)$, then $n \in S$ exactly when it is the bottom right entry of some matrix in Γ. We will see in Section 1B–1C that this setting is, roughly speaking, the starting point, and our other examples will boil down to similar problems. It was shown in [Bourgain and Kontorovich 2010], under some conditions on Γ, that S satisfies an "almost every" local-global principle. First we explain the conditions on Γ, before stating the theorem precisely.

Assume that Γ is free, finitely generated with no parabolic elements, and assume that Γ, while thin, is "not too thin." To explain what this last condition means, recall that one can attach a Poincare series $\mathcal{P}_\Gamma(s)$ to Γ, given by

$$\mathcal{P}_\Gamma(s) := \sum_{\gamma \in \Gamma} \|\gamma\|^{-2s}, \tag{1.2}$$

with $s \in \mathbb{C}$ and $\|\gamma\|$ the Euclidean norm of the matrix entries of γ. This series converges in some half-plane $\operatorname{Re} s \geq C$, and its abscissa of convergence is called the *critical exponent* $\delta = \delta(\Gamma)$. For example, if $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ or any lattice, then $\delta(\Gamma) = 1$, while for Γ thin as above, it is known that $\delta < 1$. The condition that Γ is not too thin is then encoded in the requirement that

$$\delta > 1 - \varepsilon_0, \tag{1.3}$$

for some explicit $\varepsilon_0 > 0$; in particular $\varepsilon_0 = 5 \times 10^{-5}$ suffices.

We now explain the "almost every" local-global principle. Call $n \in \mathbb{Z}$ *admissible* if it is everywhere locally represented, meaning $n \in S \bmod q$ for all q. Let $\mathscr{A}$ denote the set of all admissible numbers. Though it is not *a priori* obvious, it follows from strong approximation that only finitely many q need be checked for admission to $\mathscr{A}$.

Theorem 1.4 [Bourgain and Kontorovich 2010]. *Let Γ be as above. For almost every n (in the sense of natural density),*

$$n \text{ is represented if and only if } n \text{ is admissible.} \tag{1.5}$$

Quantitatively, the number of exceptions to (1.5) *up to N, is at most $O(N^{1-\eta_0})$, for some $\eta_0 > 0$, that is,*

$$\frac{\#(S \cap [1, N])}{\#(\mathscr{A} \cap [1, N])} = 1 + O(N^{-\eta_0}) \quad \text{as } N \to \infty.$$

In particular, almost every admissible prime is produced, since up to N, the exceptional set of order $O(N^{1-\eta_0})$ is too small to contain even a positive proportion of primes, of cardinality $N/\log N$.

On probabilistic grounds, one may expect something like (1.5) to hold with a more relaxed condition on δ in (1.3). Indeed, if $\|\gamma\|$ is about T, then so is $n = \langle \boldsymbol{v}_0 \cdot \gamma, \boldsymbol{w}_0 \rangle$, roughly speaking. The number of such γ is on the order of $T^{2\delta}$,

so conspiracies notwithstanding, n is expected to occur with multiplicity $T^{2\delta-1}$. As long as $\delta > \frac{1}{2}$, this multiplicity should grow, giving heuristic evidence that (1.3) can be relaxed to any $\varepsilon_0 < \frac{1}{2}$. If δ falls below $\frac{1}{2}$, then certainly (1.5) is false; the set S, even counted with multiplicity, is already too thin to be even a positive proportion of the integers.

1B. ***Zaremba's conjecture.*** If $x \in (0, 1)$ has the continued fraction expansion

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

we write $x = [a_1, a_2, \dots]$. The numbers $a_j \in \mathbb{N}$ are called the *partial quotients* of x. For $A \geq 1$, let $\mathfrak{R}_A$ denote by $\mathfrak{R}_A$ the set of all rational numbers b/d, $(b, d) = 1$, whose partial quotients are bounded by A:

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

Let $\mathfrak{D}_A$ be the set of denominators appearing in $\mathfrak{R}_A$:

$$\mathfrak{D}_A := \left\{ d \geq 1 : \text{there exists } b \text{ such that } (b, d) = 1 \text{ and } \frac{b}{d} \in \mathfrak{R}_A \right\}.$$

Conjecture 1.6 [Zaremba 1972, p. 76]. Every number is the denominator of a reduced fraction whose partial quotients are absolutely bounded. That is, there is some absolute $A > 1$ (possibly $A = 5$ suffices) such that

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

This seemingly innocuous conjecture has important applications to the theory of good lattice points in multidimensional quasi-Monte Carlo numerical integration, as well as to the linear congruential method for generating pseudorandom numbers; see, for example, [Niederreiter 1978].

The conjecture is known to be true for a set of density one:

Theorem 1.7 [Bourgain and Kontorovich 2011]. *Almost every number is the denominator of a reduced fraction whose partial quotients are bounded by* 50. *That is, for* $A = 50$,

$$\frac{\#(\mathfrak{D}_A \cap [1, N])}{N} \to 1 \quad \text{as } N \to \infty. \tag{1.8}$$

The best previously known bound, due to Hensley [2006, Theorem 3.2], states that, for any fixed $\varepsilon > 0$, there is some large $A = A(\varepsilon)$ such that

$$\#(\mathfrak{D}_A \cap [1, N]) > C_\varepsilon \cdot N^{1-\varepsilon}.$$

This problem turns out to be nearly identical to that in Section 1A. Observe that $b/d = [a_1, a_2, \ldots, a_k]$ if and only if

$$\begin{pmatrix} * & b \\ * & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_2 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_k \end{pmatrix}. \tag{1.9}$$

Hence we set Γ_A to be the semigroup generated by the matrices of the above type:

$$\Gamma_A := \left\langle \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} : a \le A \right\rangle^+. \tag{1.10}$$

(The superscript plus sign denotes generation as a semigroup.) Then from (1.9), the set $\mathfrak{D}_A$ of denominators is nothing more than the set of bottom right entries,

$$\mathfrak{D}_A = \langle \boldsymbol{v}_0 \cdot \Gamma_A, \boldsymbol{w}_0 \rangle,$$

with $\boldsymbol{v}_0 = \boldsymbol{w}_0 = (0, 1)$. That is, this set is precisely of the same form as (1.1). One can again attach a Poincaré series (1.2) to Γ_A; it has some critical exponent δ_A, which approaches 1 as $A \to \infty$ [Hensley 1992]. In fact, the condition $A = 50$ in (1.8) is derived from the condition

$$\delta_A > 1 - \tfrac{5}{312}, \tag{1.11}$$

as in (1.3).

1C. ***The local-global conjecture for Apollonian gaskets.*** In Figure 1 we see a bounded integral Apollonian gasket $\mathcal{G}$. This is constructed by starting with three tangent circles in the plane, and iteratively packing new circles into the

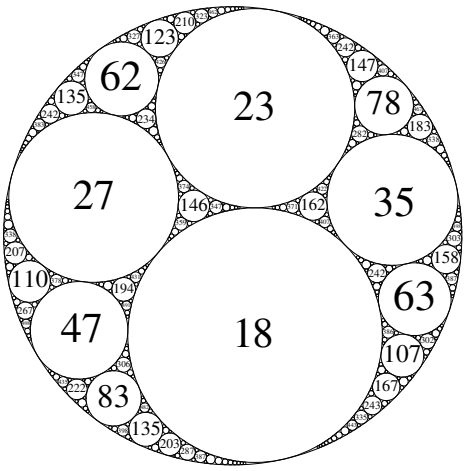

Figure 1. An integral Apollonian gasket.

complementary interstices. The closure of this circle packing in the plane is a fractal-like gasket of Hausdorff dimension

$$\delta \approx 1.3. \tag{1.12}$$

In the special configuration shown, all circles in the gasket have integral curvatures (one over their radii); these are the numbers illustrated. The author and Oh [2011] proved an asymptotic formula for the number of circles $\mathcal{C}$ in such a gasket $\mathcal{G}$ whose curvatures $b(\mathcal{C})$ are bounded by a growing parameter T: There is a constant $c > 0$ such that

$$\#\{\mathcal{C} \in \mathcal{G} : b(\mathcal{C}) < T\} = c\ T^{\delta}(1 + o(1)) \quad \text{as } T \to \infty. \tag{1.13}$$

This result was made effective in Vinogradov's thesis [2012], and independently by Lee and Oh [2013], replacing the error $o(1)$ in (1.13) by an effective rate $O(T^{-\varepsilon_0})$ with $\varepsilon_0 > 0$; see also [Sarnak 2011; Oh and Shah 2013].

In a lovely series of papers on Apollonian packings and generalizations, Graham et al. [2003] posed a different question: what numbers appear in Figure 1? Let $\mathcal{B} = \mathcal{B}(\mathcal{G})$ be the set of curvatures in $\mathcal{G}$,

$$\mathcal{B} := \{n \in \mathbb{Z} : \text{the exists } \mathcal{C} \in \mathcal{G} \text{ with } b(\mathcal{C}) = n\}.$$

As before, we say n is *represented* if $n \in \mathcal{B}$. A moment's reflection reveals that there are local obstructions, for example in the gasket $\mathcal{G}$ of Figure 1, every number is

$$\equiv 2, 3, 6, 11, 14, 15, 18, \text{ or } 23 \bmod 24. \tag{1.14}$$

As before, call such a number *admissible*, and let $\mathcal{A} = \mathcal{A}(\mathcal{G})$ be the set of admissible numbers. These obstructions, observed empirically in [Graham et al. 2003], are determined rigorously by Fuchs [2011], who proved that for any gasket, admissibility is a condition mod 24.

Conjecture 1.15 [Graham et al. 2003; Fuchs and Sanden 2011]. Every sufficiently large admissible number is represented.

This is again an instance of a local-global problem in thin orbits. (The orbit and thinness will be revealed below.)

Theorem 1.16 [Bourgain and Kontorovich 2013]. *Almost every admissible number is represented. That is*,

$$\frac{\#(\mathcal{B} \cap [1, N])}{\#(\mathcal{A} \cap [1, N])} \to 1 \quad \textit{as } N \to \infty.$$

The best previously known result, due to Bourgain and Fuchs [2011], is the

so-called positive density conjecture, which states that

$$\#(\mathcal{B} \cap [1, N]) \gg N.$$

To relate this problem to the previous two, we assume that the reader is familiar with the Apollonian group and the root quadruple of a packing; see for example [Fuchs 2014] in this volume. Specifically, the Apollonian group Γ is the integer matrix group

$$\Gamma := \langle S_1, S_2, S_3, S_4 \rangle \tag{1.17}$$

generated by the following four explicit matrices:

$$S_1 = \begin{pmatrix} -1 & & & \\ 2 & 1 & & \\ 2 & & 1 & \\ 2 & & & 1 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 1 & 2 & & \\ & -1 & & \\ & 2 & 1 & \\ & 2 & & 1 \end{pmatrix}, \quad S_3 = \begin{pmatrix} 1 & & 2 & \\ & 1 & 2 & \\ & & -1 & \\ & & 2 & 1 \end{pmatrix}, \quad S_4 = \begin{pmatrix} 1 & & & 2 \\ & 1 & & 2 \\ & & 1 & 2 \\ & & & -1 \end{pmatrix}.$$

The Apollonian group sits inside the orthogonal group of integer matrices preserving a certain quadratic form Q of signature $(3, 1)$:

$$\Gamma < O_Q(\mathbb{Z}). \tag{1.18}$$

The index of Γ in $O_Q(\mathbb{Z})$ is infinite, but nevertheless Γ is Zariski dense in O_Q.

The root quadruple $\boldsymbol{v}_0 \in \mathbb{Z}^4$ is the set of curvatures corresponding to a configuration of four mutually tangent circles in $\mathcal{G}$ of largest radius (smallest curvature). For the packing $\mathcal{G}$ in Figure 1, the root quadruple is $\boldsymbol{v}_0 = (-10, 18, 23, 27)$. Here the negative sign attached to the outermost circle's curvature accounts for its opposite orientation, being internally tangent to the other circles. It is the only circle with negative curvature.

It is a consequence of Descartes' kissing circles theorem that the orbit

$$\mathcal{O} := \boldsymbol{v}_0 \cdot \Gamma$$

of the root quadruple under the Apollonian group consists of all quadruples of curvatures corresponding to four mutually tangent circles in $\mathcal{G}$. (Note that this orbit is thin, since Γ has infinite index in the ambient group $O_Q(\mathbb{Z})$.) In particular, the entries of $\mathcal{O}$ contain all curvatures in $\mathcal{G}$, which one can recover taking the union of sets of the form

$$\langle \boldsymbol{v}_0 \cdot \Gamma, \boldsymbol{w}_0 \rangle,$$

as $\boldsymbol{w}_0$ ranges through the standard basis vectors

$$\boldsymbol{w}_0 \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_4\}.$$

Here $\boldsymbol{e}_1 = (1,0,0,0), \dots, \boldsymbol{e}_4 = (0,0,0,1)$. So this problem again turns into the form (1.1).

2. Sketches of proofs

2A. ***The circle method.*** We now describe some of the ingredients going into the proofs of Theorems 1.4, 1.7, and 1.16. Specifically, we will highlight the following tools:

- the circle method; decomposition into major and minor arcs;
- infinite volume spectral and representation theory; Lax–Phillips and Patterson–Sullivan theory;
- the thermodynamic formalism of Ruelle transfer operators and their "congruence" versions;
- expansion and uniform spectral gaps;
- strong approximation, and explicit versions thereof;
- thin bisector counting with effective rates;
- Vinogradov's method of estimating bilinear forms;
- estimates of Gauss sums and a Kloosterman refinement.

Recall that in each of our applications, the set of represented numbers is of the form (1.1). We begin with the Hardy–Littlewood circle method, looking at the exponential sum

$$\mathcal{S}_N(\theta) := \sum_{\gamma \in \Omega_N} e(\theta \langle \boldsymbol{v}_0 \gamma, \boldsymbol{w}_0 \rangle), \tag{2.1}$$

where $\theta \in \mathbb{R}/\mathbb{Z}$, $e(x) := e^{2\pi i x}$, and

$$\Omega_N \subset \{\gamma \in \Gamma : \|\gamma\| < N\}$$

is a certain carefully chosen ensemble. For now, one can think of it as the whole Γ-ball of radius N. Observe that n is certainly represented if the n-th Fourier coefficient of $\mathcal{S}_N$,

$$\widehat{\mathcal{S}_N}(n) = \int_0^1 \mathcal{S}_N(\theta) e(-n\theta)\, d\theta = \sum_{\gamma \in \Omega_N} \mathbf{1}_{\{n = \langle \boldsymbol{v}_0 \gamma, \boldsymbol{w}_0 \rangle\}}$$

is nonzero. Here $\mathbf{1}_{\{\cdot\}}$ is the indicator function. According to the circle method, one breaks the range $[0,1]$ of integration into the major arcs $\mathfrak{M}$ (where $\theta \approx a/q$ with q small) and the complementary minor arcs $\mathfrak{m} = [0,1] \setminus \mathfrak{M}$, writing

$$\widehat{\mathcal{S}_N}(n) = \mathcal{M}_N(n) + \mathcal{E}_N(n). \tag{2.2}$$

Here

$$\mathcal{M}_N(n) := \int_{\mathfrak{M}} \mathcal{S}_N(\theta) e(-n\theta)\, d\theta$$

should be the *main term* and

$$\mathcal{E}_N(n) := \int_{\mathfrak{m}} \mathcal{S}_N(\theta) e(-n\theta)\, d\theta \tag{2.3}$$

is the *error*. For the main term $\mathcal{M}_N$, one attempts to prove an asymptotic formula, or at least a good lower bound. In particular, since every admissible number is expected to be represented, one would like to show, say, for $N/2 < n < N$, that

$$\mathcal{M}_N(n) \gg \mathfrak{S}(n) \frac{\#\Omega_N}{N}. \tag{2.4}$$

Here $\mathfrak{S}(n) \geq 0$ is a certain product of local densities called the *singular series*. In particular, $\mathfrak{S}(n)$ vanishes if n is not admissible, and otherwise is $\gg N^{-\varepsilon}$ for any $\varepsilon > 0$. The singular series is well-understood, and for ease of exposition, we pretend henceforth that every n is admissible and

$$\mathfrak{S}(n) = 1. \tag{2.5}$$

For the minor arcs integral $\mathcal{E}_N$, one wishes to give an upper bound, asymptotically smaller than the lower bound given for $\mathcal{M}_N$. If one is able to do this at the level of individual n, then one can show that every sufficiently large admissible n is represented. At present, we do not know how to give such strong upper bounds, instead settling for a sharp upper bound on average, in particular in L^2-norm, as follows. From the definition (2.3), Parseval's theorem states that

$$\sum_n |\mathcal{E}_N(n)|^2 = \int_{\mathfrak{m}} |\mathcal{S}_N(\theta)|^2\, d\theta. \tag{2.6}$$

A trivial bound for $\mathcal{S}_N(\theta)$ in (2.1) is just $\leq |\Omega_N|$, giving a trivial bound of $\leq |\Omega_N|^2$ for (2.6). We claim that it suffices for our applications to save a little more than $\sqrt{N}$ on average over $\mathfrak{m}$ off of each term $\mathcal{S}_N$ in (2.6), that is, we seek the bound

$$\int_{\mathfrak{m}} |\mathcal{S}_N(\theta)|^2\, d\theta = o\left(\frac{|\Omega_N|^2}{N}\right). \tag{2.7}$$

First we claim this suffices. Let $\mathfrak{E}(N)$ be the set of exceptional n up to N, that is, those that are admissible but not represented. Certainly the n-th Fourier coefficient (2.2) is nonzero if $\mathcal{M}_N(n) > |\mathcal{E}_N(n)|$, so the number of exceptions is bounded by

$$\#\mathfrak{E}(N) \leq \sum_{\substack{|n|<N \\ n \text{ is admissible}}} \mathbf{1}_{\{|\mathcal{E}_N(n)| \geq \mathcal{M}_N(n)\}}.$$

For admissible n, we have (2.4) and (2.5), thus

$$\#\mathfrak{E}(N) \le \sum_n \mathbf{1}_{\{|\mathcal{E}_N(n)| \gg |\Omega_N|/N\}} \ll \frac{N^2}{|\Omega_N|^2} \sum_n |\mathcal{E}_N(n)|^2.$$

Now applying (2.6) and (2.7) gives

$$\#\mathfrak{E}(N) = o\bigg(\frac{N^2}{|\Omega_N|^2} \frac{|\Omega_N|^2}{N}\bigg) = o(N),$$

so almost every admissible number is represented.

In the next two subsections, we focus individually on the tools needed to establish the major arcs bound (2.4) and the error bound (2.7).

2B. ***The major arcs.*** We now sketch an argument leading to (2.4). Recall that $\mathcal{M}_N$ is an integral over the major arcs $\theta \in \mathfrak{M}$, and we may pretend that θ is just equal to a fraction a/q. Then the analysis reduces to evaluating (2.1) at $\theta = a/q$. Pretend now that Ω_N is just the whole Γ-ball:

$$\Omega_N = \{\gamma \in \Gamma : \|\gamma\| < N\}.$$

Then

$$\mathcal{S}_N\bigg(\frac{a}{q}\bigg) = \sum_{\substack{\gamma\in\Gamma \\ \|\gamma\|<N}} e\bigg(\frac{a}{q}\langle \boldsymbol{v}_0 \cdot \gamma, \boldsymbol{w}_0\rangle\bigg).$$

The exponent only depends on the residue class of $\gamma \mod q$. Let $\Gamma_q = \Gamma \mod q$ be the set of such residue classes. We split the sum as

$$\mathcal{S}_N\bigg(\frac{a}{q}\bigg) = \sum_{\gamma_0\in\Gamma_q} e\bigg(\frac{a}{q}\langle \boldsymbol{v}_0\gamma_0, \boldsymbol{w}_0\rangle\bigg)\bigg[\sum_{\substack{\gamma\in\Gamma \\ \|\gamma\|<N}} \mathbf{1}_{\{\gamma\equiv\gamma_0 \bmod q\}}\bigg]. \tag{2.8}$$

Expansion (that certain corresponding Cayley graphs have a uniform spectral gap) is used critically here to estimate the bracketed term. Roughly speaking, expansion ensures that, for q sufficiently small relative to N, the random walk in $\Gamma \bmod q$ is rapidly uniformly mixing, implying that the bracketed term in (2.8) is, up to acceptable error, just

$$\frac{1}{|\Gamma_q|} \sum_{\substack{\gamma\in\Gamma \\ \|\gamma\|<N}} 1.$$

In the group cases (Section 1A and the Apollonian case, Section 1C), one uses uniform spectral gaps [Gamburd 2002; Bourgain et al. 2010a; Bourgain and Varjú 2012; Varjú 2012], while in the semigroup case for Zaremba (Section 1B) one invokes the thermodynamic formalism and "congruence" Ruelle transfer

operators; see [Bourgain et al. 2011].

The rest of the analysis of $\mathcal{S}_N$ reduces to a certain local analysis via explicit versions of strong approximation, Goursat's lemma, and some other ingredients; see [Fuchs 2011; Bourgain and Kontorovich 2010, Section 4.3]. Inserting this estimation of $\mathcal{S}_N$ into $\mathcal{M}_N$, one needs a final ingredient of independent interest to establish (2.4), namely an effective infinite-volume bisector estimate, as follows.

Considering first the setting of Section 1A, let $G = \mathrm{SL}(2,\mathbb{R})$, and write the Cartan decomposition $G = KA^+K$,

$$g = k_1(g)a(g)k_2(g)^{-1}.$$

Here $k_1, k_2 \in K = \mathrm{SO}(2)$ and

$$a \in A^+ = \left\{ \begin{pmatrix} e^{-t} & 0 \\ 0 & e^t \end{pmatrix} : t \geq 0 \right\}.$$

Given a nonelementary (not virtually abelian), finitely generated, discrete group $\Gamma < G$ of infinite covolume, and two intervals $\mathcal{I}_1, \mathcal{I}_2 \subset K$, a bisector count is an estimate for

$$\mathcal{N}_{\Gamma,\mathcal{I}_1,\mathcal{I}_2}(T) := \#\big\{\gamma \in \Gamma \ : \ \|\gamma\| < T,\ k_1(\gamma) \in \mathcal{I}_1,\ k_2(\gamma) \in \mathcal{I}_2\big\},$$

as $T \to \infty$. In the finite covolume case, such an estimate is given with best known error terms by Good [1983]. To explain the answer in infinite volume, we need to recall a few more notions.

Since Γ acts discontinuously on the hyperbolic plane $\mathbb{H}$, any Γ-orbit has a limit set $\Lambda = \Lambda(\Gamma)$ in the boundary $\partial\mathbb{H} \cong S^1$, and this limit set has some Hausdorff dimension. By Patterson–Sullivan theory [Patterson 1976; Sullivan 1984], this dimension is equal to the critical exponent δ of Γ. It is also connected to the spectral resolution of the hyperbolic Laplacian Δ acting on the Hilbert space of square-integrable functions on $\Gamma\backslash\mathbb{H}$. Namely, it was shown in [Lax and Phillips 1982], under the standard normalization of Δ, that the spectrum below $\frac{1}{4}$ consists of a finite number of discrete eigenvalues, and the spectrum above $\frac{1}{4}$ is purely continuous. Moreover, the discrete spectrum is empty unless $\delta > \frac{1}{2}$, in which case the base eigenvalue is related to the critical exponent by

$$\lambda_0 = \delta(1-\delta).$$

Corresponding to λ_0 is the base eigenfunction, which can be realized explicitly as the integral of a Poisson kernel against the so-called Patterson–Sullivan measure μ on the boundary. Roughly speaking, μ is the weak* limit as $s \to \delta^+$ of the measures

$$\mu_s(x) := \frac{\sum_{\gamma\in\Gamma} \exp(-s\, d(\mathfrak{o}, \gamma\cdot\mathfrak{o}))\mathbf{1}_{x=\gamma\mathfrak{o}}}{\sum_{\gamma\in\Gamma} \exp(-s\, d(\mathfrak{o}, \gamma\cdot\mathfrak{o}))},$$

where $d(\cdot,\cdot)$ is the hyperbolic distance, and $\mathfrak{o}$ is the fixed point of K in $\mathbb{H}$. Then the theorem proved by Bourgain, Kontorovich and Sarnak [2010b] can be stated as follows. Assuming $\delta > \frac{1}{2}$ and $\partial\mathcal{I}_1 \cap \Lambda = \partial\mathcal{I}_2 \cap \Lambda = \varnothing$, there are some $C > 0$ and $\varepsilon_0 > 0$ such that

$$\mathcal{N}_{\Gamma,\mathcal{I}_1,\mathcal{I}_2}(T) = C\ \mu(\mathcal{I}_1)\ \mu(\mathcal{I}_2)\ T^{2\delta} + O(T^{2\delta-\varepsilon_0}), \tag{2.9}$$

as $T \to \infty$. Here $\varepsilon_0 > 0$ depends only on the spectral gap for Γ. Without effective rates, such an asymptotic statement can be proved by other means — see [Sharp 2001; Oh and Shah 2013], for example — but the error terms are used crucially in our applications. The method of proof is a combination of the above-mentioned spectral theory, as well as ergodic and representation theory of $L^2(\Gamma\backslash G)$.

A similar counting statement is established in [Vinogradov 2012] for the Apollonian case (Section 1C) of $G = \mathrm{SL}(2,\mathbb{C})$ (which is the spin double cover of the ambient orthogonal group; see (1.18)). For Zaremba's problem (Section 1B), we have only a *semigroup* Γ_A, so none of the spectral, ergodic, and representation-theoretic tools are available. Instead, one uses the counting technique pioneered by Lalley [1989] (see also [Dolgopyat 1998; Naud 2005]), analytically continuing the resolvent of the transfer operator, and its congruence version studied in [Bourgain et al. 2011].

These techniques and some more standard circle method analysis lead eventually to (2.4).

2C. ***The minor arcs.*** We use different strategies to prove (2.7) for the setting of Sections 1A–1B versus the Apollonian setting of Section 1C, so we present them separately. The details are quite involved and will quickly fall outside the scope of this survey. We will only give a few key ideas, encouraging the interested reader to consult the original references.

The setting of Sections 1A–1B. To handle the minor arcs here, we make the observation that the ensemble Ω_N in the definition of $\mathcal{S}_N$ from (2.1) need not be a full Γ-ball. In fact, the definition of $\mathcal{S}_N$ can be changed to, say,

$$\mathcal{S}_N(\theta) := \sum_{\substack{\gamma_1\in\Gamma \\ \|\gamma_1\|<\sqrt{N}}} \sum_{\substack{\gamma_2\in\Gamma \\ \|\gamma_2\|<\sqrt{N}}} e(\theta\langle \boldsymbol{v}_0\gamma_1\gamma_2, \boldsymbol{w}_0\rangle), \tag{2.10}$$

without damaging the major arcs analysis. This new sum encodes much more of the (semi)group structure of Γ, while preserving the property that a nonvanishing n-th Fourier coefficient implies that n is represented. (In reality, we use an even more complicated exponential sum.) The advantage of (2.10) is that, fixing one of the variables, we can apply the Cauchy–Schwarz inequality in the other, and

begin to exploit the structure of (2.10) à la Vinogradov's method [Vinogradov 1937] for estimating bilinear forms. In so doing, we pass at some point from the thin and mysterious group Γ (or semigroup Γ_A) to the full ambient group $\mathrm{SL}(2,\mathbb{Z})$. This type of perturbation argument only succeeds when δ is near 1, explaining the restrictions (1.3) and (1.11).

The Apollonian case, Section 1C. The above strategy fails for the Apollonian problem, because the Hausdorff dimension (1.12) is a fixed invariant which cannot be adjusted. Instead, we use the observation from [Sarnak 2007] that the Apollonian group, while of infinite index in $O_Q(\mathbb{Z})$ (see (1.18)), contains arithmetic subgroups of $\mathrm{SO}(2,1)$. In particular, consider the group consisting of all even-length words in three of the generators of page 98 — say, S_1, S_2, S_3:

$$\Xi := \langle S_1, S_2, S_3\rangle \cap \mathrm{SO}_Q < \Gamma.$$

Sarnak observed that it is isomorphic to the principal congruence subgroup of level 2:

$$\Lambda(2) := \big\{\gamma \in \mathrm{SL}(2,\mathbb{Z}) : \gamma \equiv I \bmod 2\big\}.$$

Then, like (2.10), we change the definition of the exponential sum to

$$\mathcal{S}_N(\theta) := \sum_{\substack{\xi\in\Xi\\ \|\xi\|<X}} \sum_{\substack{\gamma\in\Gamma\\ \|\gamma\|<T}} e(\theta\langle \boldsymbol{v}_0\cdot\xi\,\gamma, \boldsymbol{w}_0\rangle), \tag{2.11}$$

for certain parameters X and T chosen optimally in relation to N. One uses the full sum over the group Γ to capture the major arcs. For the minor arcs bound, one keeps γ essentially fixed (though at some point even the γ sum is used), and tries to get cancellation from the sum on $\xi \in \Xi$. Since Ξ is an arithmetic group, this sum can be converted into a more classical exponential sum, allowing use of more standard tools (bounds for certain Gauss-like sums, and a Kloosterman-like refinement, among other ingredients). In this way, one gets the requisite cancellation in (2.11) and (2.6) to prove the desired bound (2.7).

Acknowledgements

It is a pleasure to thank MSRI and the organizers for the warm hospitality during the "Hot Topics" workshop. I am grateful to Jean Bourgain for the collaboration, and the referees for comments and corrections.

References

[Bourgain and Fuchs 2011] J. Bourgain and E. Fuchs, "A proof of the positive density conjecture for integer Apollonian circle packings", *J. Amer. Math. Soc.* **24**:4 (2011), 945–967.

[Bourgain and Kontorovich 2010] J. Bourgain and A. Kontorovich, "On representations of integers in thin subgroups of $\mathrm{SL}_2(\mathbb{Z})$", *Geom. Funct. Anal.* **20**:5 (2010), 1144–1174.

[Bourgain and Kontorovich 2011] J. Bourgain and A. Kontorovich, "On Zaremba's conjecture", preprint, 2011. To appear in *Annals Math.* arXiv 1107.3776

[Bourgain and Kontorovich 2013] J. Bourgain and A. Kontorovich, "On the local-global conjecture for integral Apollonian gaskets", *Inventiones mathematicae* (2013).

[Bourgain and Varjú 2012] J. Bourgain and P. P. Varjú, "Expansion in $\mathrm{SL}_d(\mathbf{Z}/q\mathbf{Z})$, q arbitrary", *Invent. Math.* **188**:1 (2012), 151–173.

[Bourgain et al. 2010a] J. Bourgain, A. Gamburd, and P. Sarnak, "Affine linear sieve, expanders, and sum-product", *Invent. Math.* **179**:3 (2010), 559–644.

[Bourgain et al. 2010b] J. Bourgain, A. Kontorovich, and P. Sarnak, "Sector estimates for hyperbolic isometries", *Geom. Funct. Anal.* **20**:5 (2010), 1175–1200.

[Bourgain et al. 2011] J. Bourgain, A. Gamburd, and P. Sarnak, "Generalization of Selberg's $\frac{3}{16}$ theorem and affine sieve", *Acta Math.* **207**:2 (2011), 255–290.

[Dolgopyat 1998] D. Dolgopyat, "On decay of correlations in Anosov flows", *Ann. of Math.* (2) **147**:2 (1998), 357–390.

[Fuchs 2011] E. Fuchs, "Strong approximation in the Apollonian group", *J. Number Theory* **131**:12 (2011), 2282–2302.

[Fuchs 2014] E. Fuchs, "The ubiquity of thin groups", pp. 73–92 in *Thin groups and superstrong approximation*, edited by H. Oh and E. Breuillard, Math. Sci. Res. Inst. Publ. **61**, Cambridge, New York, 2014.

[Fuchs and Sanden 2011] E. Fuchs and K. Sanden, "Some experiments with integral Apollonian circle packings", *Exp. Math.* **20**:4 (2011), 380–399.

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

[Good 1983] A. Good, *Local analysis of Selberg's trace formula*, Lecture Notes in Mathematics **1040**, Springer, Berlin, 1983.

[Graham et al. 2003] R. L. Graham, J. C. Lagarias, C. L. Mallows, A. R. Wilks, and C. H. Yan, "Apollonian circle packings: number theory", *J. Number Theory* **100**:1 (2003), 1–45.

[Hensley 1992] D. Hensley, "Continued fraction Cantor sets, Hausdorff dimension, and functional analysis", *J. Number Theory* **40**:3 (1992), 336–358.

[Hensley 2006] D. Hensley, *Continued fractions*, World Scientific, Hackensack, NJ, 2006.

[Kontorovich and Oh 2011] A. Kontorovich and H. Oh, "Apollonian circle packings and closed horospheres on hyperbolic 3–manifolds", *J. Amer. Math. Soc.* **24**:3 (2011), 603–648.

[Lalley 1989] S. P. Lalley, "Renewal theorems in symbolic dynamics, with applications to geodesic flows, non-Euclidean tessellations and their fractal limits", *Acta Math.* **163**:1-2 (1989), 1–55.

[Lax and Phillips 1982] P. D. Lax and R. S. Phillips, "The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces", *J. Funct. Anal.* **46**:3 (1982), 280–350.

[Lee and Oh 2013] M. Lee and H. Oh, "Effective circle count for Apollonian packings and closed horospheres", *Geom. Funct. Anal.* **23**:2 (2013), 580–621.

[Naud 2005] F. Naud, "Expanding maps on Cantor sets and analytic continuation of zeta functions", *Ann. Sci. École Norm. Sup.* (4) **38**:1 (2005), 116–153.

[Niederreiter 1978] H. Niederreiter, "Quasi-Monte Carlo methods and pseudo-random numbers", *Bull. Amer. Math. Soc.* **84**:6 (1978), 957–1041.

[Oh and Shah 2013] H. Oh and N. A. Shah, "Equidistribution and counting for orbits of geometrically finite hyperbolic groups", *J. Amer. Math. Soc.* **26**:2 (2013), 511–562.

[Patterson 1976] S. J. Patterson, "The limit set of a Fuchsian group", *Acta Math.* **136**:3-4 (1976), 241–273.

[Salehi Golsefidy and Sarnak 2011] A. Salehi Golsefidy and P. Sarnak, "Affine sieve", preprint, 2011. arXiv 1109.6432

[Sarnak 2007] P. Sarnak, Letter to J. Lagarias, 2007, http://tinyurl.com/SarnakToLagarias.

[Sarnak 2011] P. Sarnak, "Integral Apollonian packings", *Amer. Math. Monthly* **118**:4 (2011), 291–306.

[Sharp 2001] R. Sharp, "Sector estimates for Kleinian groups", *Port. Math.* (*N.S.*) **58**:4 (2001), 461–471.

[Sullivan 1984] D. Sullivan, "Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups", *Acta Math.* **153**:3-4 (1984), 259–277.

[Varjú 2012] P. P. Varjú, "Expansion in $\mathrm{SL}_d(\mathcal{O}_K/I)$, I square-free", *J. Eur. Math. Soc.* (*JEMS*) **14**:1 (2012), 273–305.

[Vinogradov 1937] I. M. Vinogradov, "Representation of an odd number as a sum of three primes", *Dokl. Akad. Nauk SSSR* **15** (1937), 291–294. In Russian; translated in *C. R. (Dokl.) Acad. Sci. URSS* **15** (1937), 169–172.

[Vinogradov 2012] I. Vinogradov, *Effective bisector estimate with application to Apollonian circle packings*, Ph.D. thesis, Princeton University, 2012. arXiv 1204.5498v1

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

alex.kontorovich@yale.edu *Department of Mathematics, Yale University, P.O. Box 208283, New Haven, CT 06520-8283, United States*