FROM APOLLONIUS TO ZAREMBA: LOCAL-GLOBAL PHENOMENA IN THIN ORBITS

ALEX KONTOROVICH

Abstract. We discuss a number of natural problems in arithmetic, arising in completely unrelated settings, which turn out to have a common formulation involving "thin" orbits. These include the local-global problem for integral Apollonian gaskets and Zaremba's Conjecture on finite continued fractions with absolutely bounded partial quotients. Though these problems could have been posed by the ancient Greeks, recent progress comes from a pleasant synthesis of modern techniques from a variety of fields, including harmonic analysis, algebra, geometry, combinatorics, and dynamics. We describe the problems, partial progress, and some of the tools alluded to above.

Contents

Date: November 15, 2012.

2010 *Mathematics Subject Classification.* 11F41, 11J70, 11P55, 20H10, 22E40.

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

1. Introduction

In this article we will discuss recent developments on several seemingly unrelated arithmetic problems, which each boil down to the same issue of proving a "local-global principle for thin orbits". In each of these problems, we study the *orbit*

$$\mathcal{O} = \Gamma \cdot \mathbf{v}_0,$$

of some given vector $\mathbf{v}_0 \in \mathbb{Z}^d$, under the action of some given group or semigroup, Γ, (under multiplication) of d-by-d integer matrices. It will turn out that the orbits arising naturally in our problems are "thin"; roughly speaking, this means that each orbit is "degenerate" in its algebro-geometric closure, containing relatively very few points.

Each of the problems then takes another vector $\mathbf{w}_0 \in \mathbb{Z}^d$, and for the standard inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^d$, forms the set

$$\mathscr{S} := \langle \mathbf{w}_0, \mathcal{O} \rangle \subset \mathbb{Z}$$

of integers, asking what numbers are in $\mathscr{S}$. For an integer $q \geq 1$, the projection map

$$\mathbb{Z} \to \mathbb{Z}/q\mathbb{Z}$$

can give an obvious obstruction to membership. Let $\mathscr{S}(\mathrm{mod}\, q)$ be the image of this projection,

$$\mathscr{S}(\mathrm{mod}\, q) := \{s(\mathrm{mod}\, q) : s \in \mathscr{S}\} \subset \mathbb{Z}/q\mathbb{Z}.$$

For example, suppose that any number in $\mathscr{S}$ leaves a remainder of $1, 2$ or 3 when divided by 4, that is, $\mathscr{S}(\mathrm{mod}\, 4) = \{1, 2, 3\}$. Then one can conclude, without any further consideration, that $10^{10^{10}} \notin \mathscr{S}$, since $10^{10^{10}} \equiv 0 (\mathrm{mod}\, 4)$. This is called a *local* obstruction. Call n *admissible* if it avoids all local obstructions,

$$n \in \mathscr{S}(\mathrm{mod}\, q), \qquad \text{for all } q \geq 1.$$

In many applications, the set $\mathscr{S}(\mathrm{mod}\, q)$ is significantly easier to analyze than the set $\mathscr{S}$ itself. But a local to global phenomenon predicts that, if n is admissible, then in fact $n \in \mathscr{S}$, thereby reducing the seemingly more difficult problem to the easier one.

It is the combination of these concepts, (i) thin orbits, and (ii) local-global phenomena, which will turn out to be the "beef" of the problems we intend to discuss.

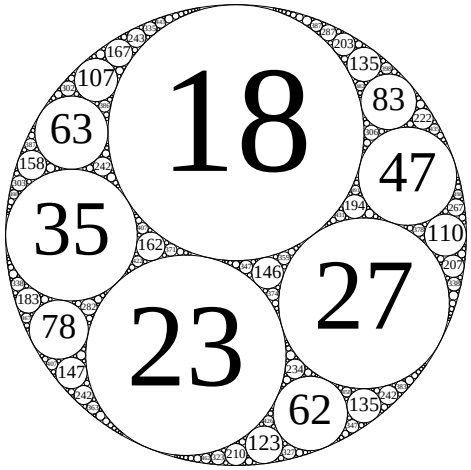


FIGURE 1. An integral Apollonian gasket.

1.1. **Outline.**

We begin in §2 with Zaremba's Conjecture. We will explain how this problem arose naturally in the study of "good lattice points" for quasi-Monte Carlo methods in multi-dimensional numerical integration, and how it also has applications to the linear congruential method for pseudo-random number generators. But the assertion of the conjecture is a statement about continued fraction expansions of rational numbers, and as such is so elementary that Euclid himself could have posed it. We will discuss recent progress by Bourgain and the author, proving a density version of the conjecture.

We change our focus in §3 to the ancient geometer Apollonius of Perga. As we will explain, his straight-edge and compass construction of tangent circles, when iterated *ad infinitum*, gives rise to a beautiful fractal circle packing in the plane, such as that shown in Figure 1. Recall that the curvature of a circle is just one over its radius. For special configurations, all the curvatures of circles in the given packing turn out to be integers; these are the numbers shown in Figure 1. We will present in §3 progress on the problem: which integers appear? It was recently proved by Bourgain and the author that almost every admissible number appears.

In §4, also stemming from Greek mathematics, we describe a local-global problem for a thin orbit of Pythagorean triples, as will be defined there. This problem is a variant of the so-called Affine Sieve, recently introduced by Bourgain, Gamburd, and Sarnak. We will explain an "almost" local-global theorem in this context due to Bourgain and the author.

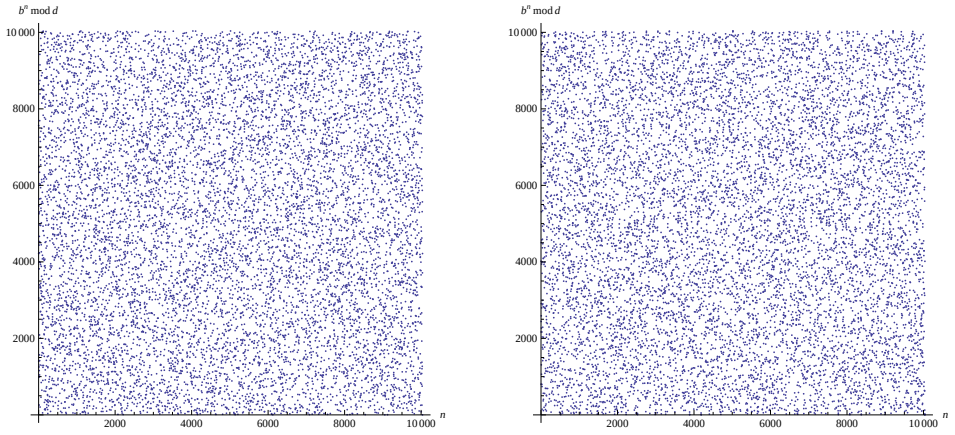
Finally, these three problems are reformulated to the aforementioned common umbrella in §5, where some of the ingredients of the proofs are sketched. The problems do not naturally fit in an established area of research, having no L-functions or Hecke theory (though they are unquestionably problems about whole numbers), being not part of the Langlands Program (though involving automorphic forms and representations), nor falling under the purview of the classical circle method or sieve, which attempt to solve equations or produce primes in polynomials (here it is not polynomials that generate points, but the aforementioned matrix actions). Instead the proofs borrow bits and pieces from these fields and others, the major tools including: analysis (the circle method, exponential sum bounds, infinite volume spectral theory), algebra (strong approximation, Zariski density, spin and orthogonal groups associated to quadratic forms, representation theory), geometry (hyperbolic manifolds, circle packings, diophantine approximation), combinatorics (sum-product, expander graphs, spectral gaps), and dynamics (ergodic theory, mixing rates, the thermodynamic formalism). We aim to highlight some of these ingredients throughout.

1.2. **Notation.**

We use the following standard notation. A quantity is defined via the symbol ":=", and a concept being defined is italicized. Write $f \sim g$ for $f/g \to 1$, $f = o(g)$ for $f/g \to 0$, and $f = O(g)$ or $f \ll g$ for $f \leq Cg$. Here $C > 0$ is called an implied constant, and is absolute unless otherwise specified. Moreover, $f \asymp g$ means $f \ll g \ll f$. We use $e(x) = e^{2\pi i x}$. The cardinality of a finite set S is written as $|S|$ or $\#S$. The transpose of a vector $\mathbf{v}$ is written $\mathbf{v}^t$. The meaning of algebraic symbols can change from section to section; for example the (semi)group Γ and quadratic form Q will vary depending on the context.

Acknowledgements.

I wish to thank Andrew Granville for encouraging me to pen these notes and for his insightful and detailed input on various drafts. Thanks to Mel Nathanson for inviting me to give a mini-course at CANT 2012, as a result of which these notes were finally assembled. I am grateful to Peter Sarnak for introducing me to Apollonian gaskets and infinite volume spectral methods, to Hee Oh for introducing me to homogeneous dynamics, and to Dorian Goldfeld for his constant support and advice. Thanks to Elena Fuchs, Aryeh Kontorovich, Sam Payne, and especially the referee for detailed comments on an earlier draft. Most of all, I owe a huge debt of gratitude to Jean Bourgain for his generous tutelage and collaboration.



(a) Multiplier $b = 4217$. (b) Multiplier $b = 4015$.

FIGURE 2. Graphs of the map (2.2) with prime modulus $d = 10037$, and multiplier b as shown.

2. Zaremba's Conjecture

Countless applications require *pseudo-random numbers*: deterministic algorithms which "behave randomly." Probably the simplest, oldest, and best known among these is the so-called *linear congruential method*: For some starting seed x_0, iterate the map

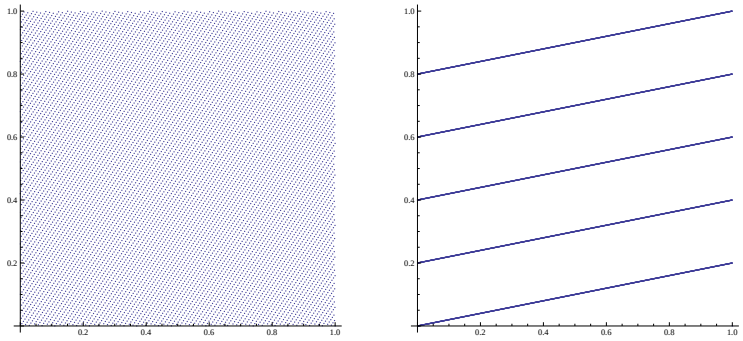
$$x \mapsto bx + c \pmod{d}. \tag{2.1}$$

Here b is called the multiplier, c the shift, and d the modulus. For simplicity, we consider the homogeneous case $c = 0$. To have as long a sequence as possible, take d to be prime, and b a primitive root mod d, that is, a generator of the cyclic group $(\mathbb{Z}/d\mathbb{Z})^\times$. In this case we may as well start with the seed $x_0 = 1$; then the iterates of (2.1) are nothing more than the map

$$n \mapsto b^n \pmod{d}. \tag{2.2}$$

We show graphs of this map in Figure 2 for the prime $d = 10037$, with two choices of roots $b = 4217$ and $b = 4015$. In both cases, the graphs "look" random, in that, given b and n, it is hard to guess where $b^n \pmod{d}$ will lie (without just computing). Similarly, given b and $b^n \pmod{d}$, it is typically difficult to determine n; this is the classical problem of computing a discrete logarithm.

A slightly more rigorous statistical test for randomness is the serial correlation of pairs: how well can we guess where b^{n+1} is, knowing b^n? To this end, we plot in Figure 3 these pairs, or what is the same, the



(a) Multiplier $b = 4217$. (b) Multiplier $b = 4015$.

FIGURE 3. Plots of the points (2.3) for the same choices of modulus $d = 10037$ and multipliers as in Figure 2.

pairs

$$\left\{ \left(\frac{b^n}{d}, \frac{b^{n+1}}{d} \right) (\mathrm{mod}\, 1) \right\}_{n=1}^{d} \subset \mathbb{R}^2/\mathbb{Z}^2 \tag{2.3}$$

in the unit square, with the previous choices of modulus and multiplier. Focus first on Figure 3a: it looks like a fantastically equidistributed grid. Keep in mind that the mesh in each coordinate is of size $1/d \approx 1/10000$, so we have $(10000)^2$ points from which to choose, yet we are only plotting 10000 points, square-root the total number of options.

On the other hand, look at Figure 3b: these parameters make a terrible random number generator! Given the first few terms in this sequence $(x_1, x_2, x_3, \ldots, x_k)$, with $x_n = b^n/d (\mathrm{mod}\, 1)$, we simply plot the pairs $(x_1, x_2), (x_2, x_3), \ldots, (x_{k-1}, x_k)$, and then have a $1:5$ guess for where x_{k+1} will be.

A related phenomenon also appears in two-dimensional numerical integration: Suppose that you wish to integrate a "nice" function f on $\mathbb{R}^2/\mathbb{Z}^2 \cong [0,1) \times [0,1)$, say of finite variation, $V(f) < \infty$, where

$$V(f) := \int_0^1 \int_0^1 \left(|f| + \left| \frac{\partial}{\partial x} f \right| + \left| \frac{\partial}{\partial y} f \right| + \left| \frac{\partial^2}{\partial x \partial y} f \right| \right) dx dy.$$

The idea is to take a large sample of points $\mathcal{Z}$ in $\mathbb{R}^2/\mathbb{Z}^2$, and approximate the integral by the average of $f(z)$, $z \in \mathcal{Z}$. For this to be a good approximation one obviously needs that f does not vary much in a small ball, and that the points of $\mathcal{Z}$ are well-distributed throughout $\mathbb{R}^2/\mathbb{Z}^2$. In fact, the famous Koksma-Hlawka inequality (see [Nie78, p. 966]) states, rather beautifully, that this is all that one needs to take

into account:

$$\left| \int_0^1 \int_0^1 f(x,y)dx\,dy - \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} f(z) \right| \le C \cdot V(f) \cdot \mathrm{Disc}(\mathcal{Z}).$$

Here $C > 0$ is an absolute constant, and Disc is the *discrepancy* of the set $\mathcal{Z}$, defined as follows. Take a rectangle $R = [a,b] \times [c,d] \subset \mathbb{R}^2/\mathbb{Z}^2$. One would like the fraction of points in R to be close to its area, so set

$$\mathrm{Disc}(\mathcal{Z}) := \sup_{R\subset\mathbb{R}^2/\mathbb{Z}^2} \left| \frac{\#(\mathcal{Z}\cap R)}{\#\mathcal{Z}} - \mathrm{Area}(R) \right|.$$

It is elementary that for a growing family $\mathcal{Z}^{(k)} \subset \mathbb{R}^2/\mathbb{Z}^2$, $|\mathcal{Z}^{(k)}| \to \infty$, the discrepancy $\mathrm{Disc}(\mathcal{Z}^{(k)})$ decays to 0 if and only if $\mathcal{Z}^{(k)}$ becomes equidistributed in $\mathbb{R}^2/\mathbb{Z}^2$. But more than just indicating equidistribution, the discrepancy measures the rate. For example, observe that for any finite sample set $\mathcal{Z}$, we have the lower bound $\mathrm{Disc}(\mathcal{Z}) \ge 1/|\mathcal{Z}|$. Indeed, take a family of rectangles R zooming in on a single point in $\mathcal{Z}$; the proportion of points in R is always $1/|\mathcal{Z}|$, while the area of R can be made arbitrarily small. It turns out there is a sharpest possible lower bound, due to Schmidt [Sch72]:

$$\text{For any finite } \mathcal{Z} \subset S, \qquad \mathrm{Disc}(\mathcal{Z}) \gg \frac{\log|\mathcal{Z}|}{|\mathcal{Z}|}. \tag{2.4}$$

Standard Monte Carlo integration is the process of computing the integral of f by just sampling $z \in \mathcal{Z}$ according to the uniform measure; the Central Limit Theorem then predicts that

$$\mathrm{Disc}(\mathcal{Z}) \approx \frac{1}{|\mathcal{Z}|^{1/2}}, \tag{2.5}$$

ignoring $\log\log$ factors. So comparing (2.5) to (2.4), it is clear that uniformly sampled sequences are far from optimal in numerical integration. Alternatively, one could take $\mathcal{Z}$ to be an evenly spaced d-by-d grid,

$$\mathcal{Z} = \{(i/d, j/d) : 0 \le i, j < d\},$$

with $|\mathcal{Z}| = d^2$. But then the rectangle $[\varepsilon, 1/d - \varepsilon] \times [0,1]$ contains no grid points while its area is almost $1/d = 1/|\mathcal{Z}|^{1/2}$, again giving (2.5).

In the *quasi* Monte Carlo method, rather than sampling uniformly, one tries to find a special sample set $\mathcal{Z}$ to come as close as possible to the optimal discrepancy (2.4). Ideally, such a set $\mathcal{Z}$ would also be quickly and easily constructible by a computer algorithm. Not surprisingly, the set $\mathcal{Z}$ illustrated in Figure 3a makes an excellent sample set. It was this problem which led Zaremba to his theorem and conjecture,

described below.

Returning to our initial discussion, observe that the sequence (2.3) is essentially (since b is a generator) the same as

$$\mathcal{Z}_{b,d} := \left\{ \left(\frac{n}{d}, \frac{bn}{d} \right) \right\}_{n=1}^{d} \pmod 1. \tag{2.6}$$

And this is nothing more than a graph of our first map (2.1). Now it is clear that both Figures 3a and 3b are "lines", but the first must be "close to a line with irrational slope," causing the equidistribution. This Diophantine property is best described in terms of continued fractions, as follows.

For $x \in (0,1)$, we use the notation

$$x = [a_1, a_2, \dots]$$

for the continued fraction expansion

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

The integers $a_j \geq 1$ are called *partial quotients* of x. Rational numbers have finite continued fraction expansions.

One is then immediately prompted to study the continued fraction expansions of the "slopes" b/d in Figure 3:

$$\begin{aligned} 4217/10037 &= [2,2,1,1,1,2,2,2,1,2,2,1,2], \\ 4015/10037 &= [2,2,2007]. \end{aligned}$$

Note the gigantic partial quotient 2007 in the second expression, while the partial quotients in the first are all ones and twos. Observations of this kind naturally led Zaremba to the following

Theorem 2.7 (Zaremba 1966 [Zar66, Corollary 5.2])**.** *Fix $(b,d) = 1$ with $b/d = [a_1, a_2, \dots, a_k]$ and let $A := \max a_j$. Then for $\mathcal{Z}_{b,d}$ given in (2.6),*

$$\operatorname{Disc}(\mathcal{Z}_{b,d}) \leq \left(\frac{4A}{\log(A+1)} + \frac{4A+1}{\log d} \right) \frac{\log d}{d}. \tag{2.8}$$

Since $|\mathcal{Z}_{b,d}| = d$, comparing the upper bound (2.8) to Schmidt's lower bound (2.4) shows that the sequences (2.6) are essentially best possible, up to the "constant" A (and this optimal equidistribution is precisely what we observe visually in Figure 3a). But the previous sentence is

complete nonsense: A is not constant at all; it depends on d,[1] and Figure 3b perfectly illustrates what can go wrong.

With this motivation, Zaremba predicted that in fact A *can* be taken constant:

Conjecture Z (Zaremba 1972 [Zar72, p. 76])**.** *Every natural number is the denominator of a reduced fraction whose partial quotients are absolutely bounded.*

That is, there exists some absolute $A > 1$ so that for each $d \geq 1$, there is some $(b, d) = 1$, so that $b/d = [a_1, \ldots, a_k]$ with $\max a_j \leq A$.

Zaremba even suggested a sufficient value for A, namely $A = 5$. So this is really a problem that could have been posed in Book VII of the *Elements* (after Euclid's algorithm): using the partial quotients $a_j \in \{1, \ldots, 5\}$, does the set of (reduced) fractions with expansion $[a_1, \ldots, a_k]$ contain every integer as a denominator? The reason for Zaremba's guess $A = 5$ is simply that it is false for $A = 4$, as we now explain. First some more notation.

Let $\mathscr{R}_A$ be the set of rationals with the desired property that all partial quotients are at most A:

$$\mathscr{R}_A := \left\{ \frac{b}{d} = [a_1, \ldots, a_k] : (b, d) = 1, \text{ and } a_j \leq A, \ \forall j \right\},$$

and let $\mathscr{D}_A$ be the set of denominators which arise:

$$\mathscr{D}_A := \left\{ d : \exists (b, d) = 1 \text{ with } \frac{b}{d} \in \mathscr{R}_A \right\}.$$

Then Zaremba's conjecture is that $\mathscr{D}_5 = \mathbb{N}$, and we claim that this is false for $\mathscr{D}_4$. Indeed, $6 \notin \mathscr{D}_4$: the only numerators to try are 1 and 5, but the continued fraction expansion of $1/6$ is just $[6]$, and $5/6 = [1, 5]$, so the largest partial quotient in both is too big.

That said, there are only two other numbers, 54 and 150, known to be missing from $\mathscr{D}_4$ (see [OEI]), leading one to ask what happens if a finite number of exceptions is permitted. Indeed, Niederreiter [Nie78, p. 990] conjectured in 1978 that for $A = 3$, $\mathscr{D}_3$ already contains every sufficiently large number; we write this as

$$\mathscr{D}_3 \supset \mathbb{N}_{\gg 1}.$$

With lots more computational capacity and evidence, Hensley almost 20 years later [Hen96] conjectured even more boldly that the same holds

[1]The value A also depends on b, but the important variable for applications is $|\mathcal{Z}_{b,d}| = d$.

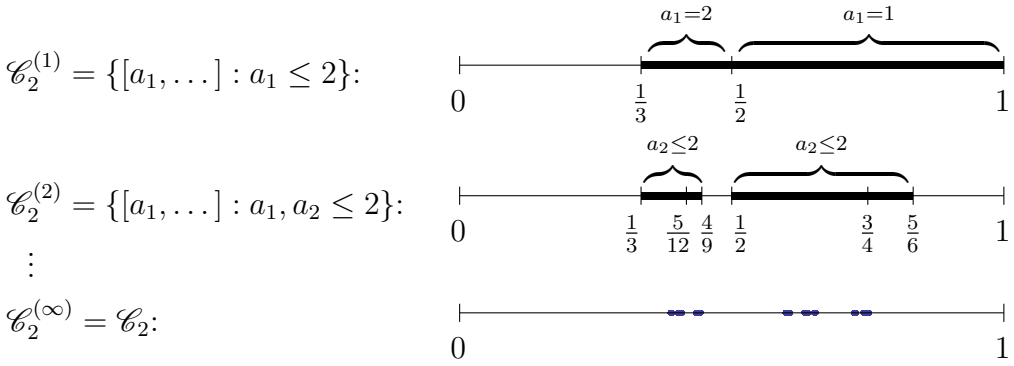


FIGURE 4. The Cantor set $\mathscr{C}_2 = \bigcap_{k=1}^{\infty} \mathscr{C}_2^{(k)}$, where $\mathscr{C}_2^{(k)} = \{[a_1, \dots, a_j, \dots, a_k, \dots] : a_j \le A \text{ for all } 1 \le j \le k\}$ restricts only the first k partial quotients.

already for $A = 2$:

$$\mathscr{D}_2 \supset \mathbb{N}_{\gg 1}. \tag{2.9}$$

Lest the reader be tempted to one-up them all, let us consider the case $A = 1$. Here $\mathscr{R}_1$ contains only continued fractions of the form $[1, \dots, 1]$, and these are quotients of consecutive Fibonacci numbers F_n,

$$\mathscr{R}_1 = \{F_n/F_{n+1}\}.$$

So $\mathscr{D}_1 = \{F_n\}$ is just the Fibonacci numbers, and this is an exponentially thin sequence.

In fact, Hensley conjectured something much stronger than (2.9). First some more notation. Let $\mathscr{C}_A$ be the set of limit points of $\mathscr{R}_A$,

$$\mathscr{C}_A := \{[a_1, a_2, \dots] : a_j \le A, \ \forall j\}.$$

To get our bearings, consider again the case $A = 1$. Then $\mathscr{C}_1 = \{1/\varphi\}$ is just the singleton consisting of the reciprocal of the golden mean.

Now take $A = 2$. Consider the unit interval $[0, 1]$. The numbers in the range $(1/2, 1]$ have first partial quotient $a_1 = 1$, and those in $(1/3, 1/2]$ have first partial quotient $a_1 = 2$. The remaining interval $[0, 1/3]$ has numbers whose first partial quotient is already too big, and thus is cut out. We repeat in this way, cutting out intervals for each partial quotient, and arriving at $\mathscr{C}_2$; see Figure 4.

For any $A \ge 1$, the Cantor-like set $\mathscr{C}_A$ has some Hausdorff dimension

$$\delta_A := \dim(\mathscr{C}_A), \tag{2.10}$$

which recall is defined as the infimum of all $s \geq 0$ for which

$$\inf_{\bigcup_j B_j \supset \mathscr{C}_A} \left\{ \sum_j r(B_j)^s \right\} \tag{2.11}$$

vanishes. The infimum in (2.11) is over collections $\{B_j\}_j$ of open balls (intervals) which cover $\mathscr{C}_A$, and $r(B_j)$ is the radius of B_j (half the length of the interval).

Clearly $\delta_1 = 0$, since $\mathscr{C}_1$ is a single point. There is a substantial literature estimating the dimension δ_2 which we will not survey, but the current record is due to Jenkinson-Pollicott [JP01], whose super-exponential algorithm estimates

$$\delta_2 = 0.531280506277205141624468 6\ldots \tag{2.12}$$

If we relax the bound A, the Cantor sets increase, as do their dimensions. In fact, Hensley [Hen92] determined the asymptotic expansion, which to first order is

$$\delta_A = 1 - \frac{6}{\pi^2 A} + o\left(\frac{1}{A}\right), \tag{2.13}$$

as $A \to \infty$. In particular, the dimension can be made arbitrarily close to 1 by taking A large.

We can now explain Hensley's stronger conjecture. His observation is that one need not only consider restricting the partial quotients a_j to the full interval $[1, A]$; one can allow more flexibility by fixing any finite "alphabet" $\mathcal{A} \subset \mathbb{N}$, and restricting the partial quotients to the "letters" in this alphabet. To this end, let $\mathscr{C}_{\mathcal{A}}$ be the Cantor set

$$\mathscr{C}_{\mathcal{A}} := \{[a_1, a_2, \ldots] : a_j \in \mathcal{A},\ \forall j \geq 1\},$$

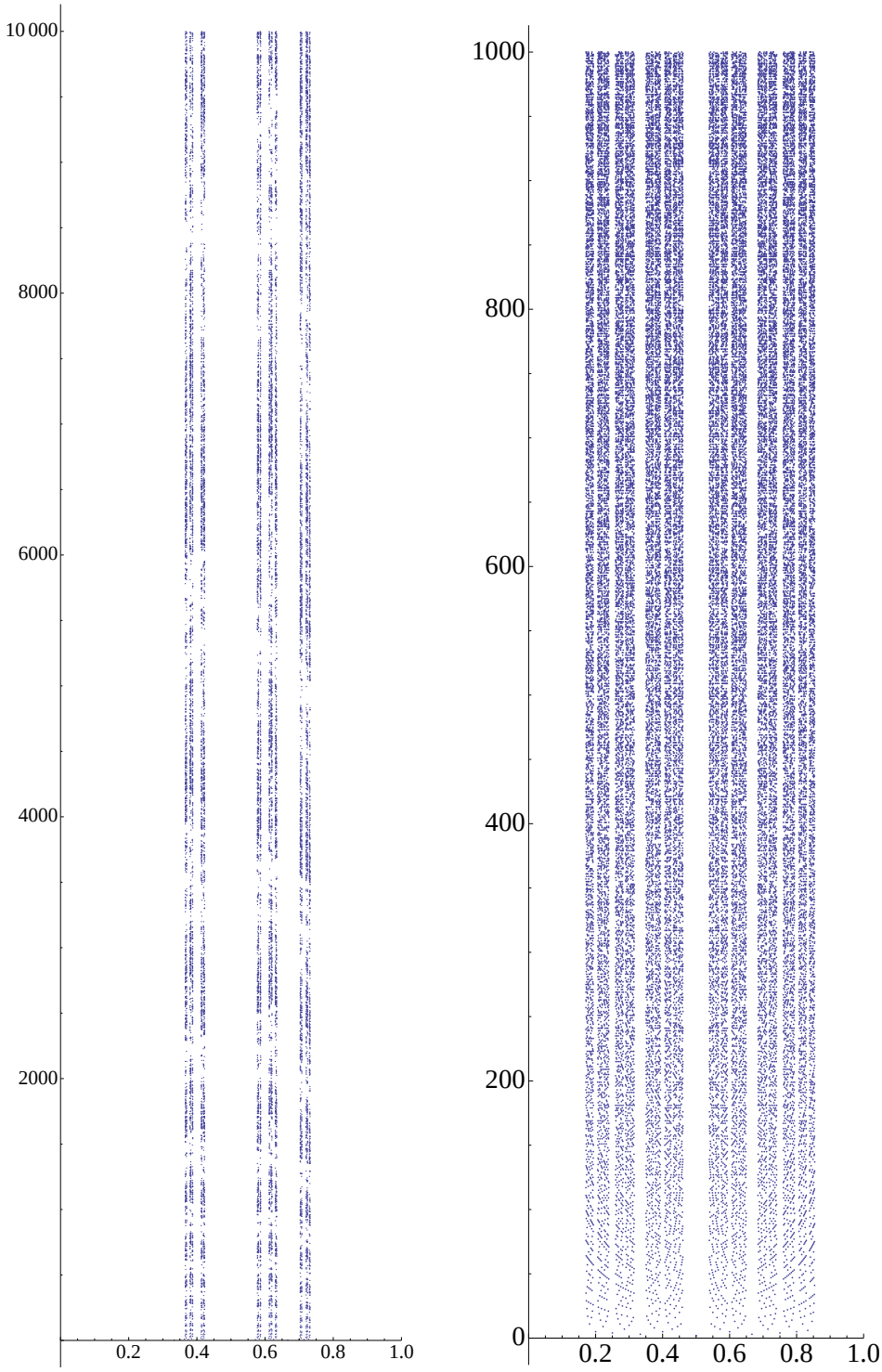
and similarly let $\mathscr{R}_{\mathcal{A}}$ be the partial convergents to $\mathscr{C}_{\mathcal{A}}$, $\mathscr{D}_{\mathcal{A}}$ the denominators of $\mathscr{R}_{\mathcal{A}}$, and $\delta_{\mathcal{A}}$ the Hausdorff dimension of $\mathscr{C}_{\mathcal{A}}$. Then Hensley's elegant claim is the following

Conjecture 2.14 (Hensley 1996 [Hen96, Conjecture 3, p. 16])**.**

$$\mathscr{D}_{\mathcal{A}} \supset \mathbb{N}_{\gg 1} \qquad \Longleftrightarrow \qquad \delta_{\mathcal{A}} > 1/2. \tag{2.15}$$

Observe in particular that δ_2 in (2.12) exceeds $1/2$, and hence Hensley's full conjecture (2.15) implies the special case $A = 2$ in (2.9).

Here is some heuristic evidence in favor of (2.15). Let us visualize the set $\mathscr{R}_{\mathcal{A}}$ of rationals, by grading each fraction according to the denominator. That is, plot each fraction b/d at height d, showing the



(a) $A = 2$, $N = 10000$. (b) $A = 5$, $N = 1000$.

FIGURE 5. For each $b/d \in \mathscr{R}_A(N)$, plot b/d versus d, with A and truncation parameter N as shown.

set

$$\left\{ \left(\frac{b}{d} \ , \ d \right) : \ \frac{b}{d} \in \mathscr{R}_{\mathcal{A}}, \quad (b,d) = 1 \right\}. \tag{2.16}$$

We show this plot in Figure 5a for $\mathcal{A} = \{1,2\}$ truncated at height $N = 10000$, and in Figure 5b for $\mathcal{A} = \{1,2,3,4,5\}$ truncated at height $N = 1000$. We give a name to this truncation, defining

$$\mathscr{R}_{\mathcal{A}}(N) := \left\{ \frac{b}{d} \in \mathscr{R}_{\mathcal{A}} : (b,d) = 1, \ 1 \le b < d < N \right\}.$$

Observe that the "vertical tentacles" in Figure 5 emanate from points on the x-axis lying in the Cantor sets $\mathscr{C}_{\mathcal{A}}$; compare Figures 5a and 4. Moreover, note that if at least one point has been placed at height d, then $d \in \mathscr{D}_{\mathcal{A}}$. That is, the "beef" of this problem boils down to: what are the projections of the plots in Figure 5 to the y-axis? In particular, does every (sufficiently large) integer appear?

The first question to address is: how big is $|\mathscr{R}_{\mathcal{A}}(N)|$, that is, how many points are being plotted in Figures 5a and 5b? Hensley [Hen89] showed that, as $N \to \infty$,

$$\#\mathscr{R}_{\mathcal{A}}(N) \quad \asymp \quad N^{2\delta_{\mathcal{A}}}, \tag{2.17}$$

where the implied constant can depend on $\mathcal{A}$. (Hensley proved this for the alphabet $\mathcal{A} = \{1,2,\dots,A\}$, but the same proof works for an arbitrary finite $\mathcal{A}$.)

Now, the $\Longrightarrow$ direction of (2.15) is trivial. Indeed, let

$$\mathscr{D}_{\mathcal{A}}(N) := \mathscr{D}_{\mathcal{A}} \cap [1,N],$$

so that the left hand side of (2.15) is equivalent to

$$\#\mathscr{D}_{\mathcal{A}}(N) = N + O(1), \qquad \text{as } N \to \infty. \tag{2.18}$$

Then it is clear that $\#\mathscr{R}_{\mathcal{A}}(N)$ counts d's with multiplicity, whereas $\#\mathscr{D}_{\mathcal{A}}(N)$ counts each appearing d only once; hence

$$\#\mathscr{D}_{\mathcal{A}}(N) \le \#\mathscr{R}_{\mathcal{A}}(N) \overset{(2.17)}{\ll} N^{2\delta_{\mathcal{A}}}. \tag{2.19}$$

So if (2.18) holds, then (2.19) implies that $2\delta_{\mathcal{A}}$ must be at least 1.

A caveat: we do not know how to verify (2.18) for a single alphabet! Nevertheless the content of Hensley's Conjecture is clearly the opposite $\Longleftarrow$ direction. Here is some evidence in favor of this claim.

An old theorem of Marstrand's [Mar54] states the following. Let $E \subset [0,1] \times [0,1]$ be a Hausdorff measurable set having Hausdorff dimension $\alpha > 1$. Then the projection of E into a line of slope $\tan\theta$ is "large," for Lebesgue-almost every $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Here "large" means of positive Lebesgue measure. One may thus heuristically think of (2.16)

as E above, with (2.17) suggesting the "dimension" $\alpha = 2\delta_{\mathcal{A}}$. Then $\mathscr{D}_{\mathcal{A}}$ is the projection of this E to the y-axis, and it should be "large" according to the analogy. Marstrand's theorem says nothing about an individual line, and does not apply to the countable set (2.16), so the analogy cannot be furthered in any meaningful way. Nevertheless, we see the condition $\alpha > 1$ is converted into $2\delta_{\mathcal{A}} > 1$, giving evidence for the $\Longleftarrow$ direction of (2.15).

For another heuristic, if one uniformly samples $N^{2\delta}$ pairs (b, d) out of the integers up to N, a given d is expected to appear with multiplicity roughly $N^{2\delta-1}$. For $\delta > 1/2$ and N growing, this multiplicity will be positive with probability tending to 1.

This heuristic does not rule out the possible conspiracy that only very few (about $N^{2\delta-1}$) d's actually appear, each with very high (about N) multiplicity. But such an argument in reverse leads to another bit of evidence towards (2.15): since the multiplicity of any $d < N$ is at most N, we have the elementary lower bound

$$\#\mathscr{D}_{\mathcal{A}}(N) \geq \frac{1}{N}\#\mathscr{R}_{\mathcal{A}}(N) \overset{(2.17)}{\gg} \frac{1}{N}N^{2\delta_{\mathcal{A}}} = N^{2\delta_{\mathcal{A}}-1}.$$

So if $\delta_{\mathcal{A}} > 1/2$, then the set $\mathscr{D}_{\mathcal{A}}$ already grows at least at a power rate. Furthermore, for any fixed $\varepsilon > 0$, one can take some $\mathcal{A} = \mathcal{A}(\varepsilon)$ sufficiently large so that $2\delta_{\mathcal{A}} - 1 > 1 - \varepsilon$. For example, using (2.13), we can take $\mathcal{A} = \{1, 2, \dots, A\}$ where

$$A > \frac{12}{\pi^2\varepsilon}(1 + o(1)).$$

Here $o(1) \to 0$ as $\varepsilon \to 0$. Hence one can produce $N^{1-\varepsilon}$ points in $\mathscr{D}_{\mathcal{A}}(N)$, which is already substantial progress towards (2.18).

But unfortunately, Hensley's conjecture (2.15), as stated, is false.

Lemma 2.20 (Bourgain-K. 2011 [BK11, Lemma 1.19])**.** *The alphabet* $\mathcal{A} = \{2, 4, 6, 8, 10\}$ *has dimension* $\delta_{\mathcal{A}} = 0.517\dots$, *which exceeds* $1/2$, *but does not contain every sufficiently large number.*

Proof. The dimension can be computed by the Jenkinson-Pollicott algorithm used to establish (2.12). It is an elementary calculation from the definitions to show for this alphabet that every fraction in $\mathscr{R}_{\mathcal{A}}$ is of the form $2m/(4n+1)$ or $(4n+1)/(2m)$, and so $\mathscr{D}_{\mathcal{A}} \equiv \{0, 1, 2\} (\mathrm{mod}\, 4)$. Hence $\mathscr{D}_{\mathcal{A}}$ does not contain every sufficiently large number. □

That is, there can be congruence obstructions, in addition to the condition on dimension. This suggests instead a closer analogy with Hilbert's 11th problem, which asks: what numbers are represented by a given integral (or rational) quadratic form? According to this analogy, we make the following

Definition 2.21. Call d *represented* by the given alphabet $\mathcal{A}$ if $d \in \mathscr{D}_{\mathcal{A}}$. Also, call d *admissible* for the alphabet $\mathcal{A}$ if it is everywhere locally represented, meaning that $d \in \mathscr{D}_{\mathcal{A}} (\mathrm{mod}\, q)$ for all $q \geq 1$.

One can then modify Hensley's conjecture to state that, if $\delta_{\mathcal{A}}$ exceeds $1/2$ (an archimedean condition), then every sufficiently large admissible number is represented, akin to Hasse's local-to-global principle.

Remark 2.22. We will explain in §2.2 that the alphabet $\mathcal{A} = \{1, 2\}$ has no local obstructions, so Hensley's first conjecture (2.9) is still plausible.

Here is some progress towards the conjecture.

Theorem Z (Bourgain-K. 2011 [BK11])**.** *Almost every natural number is the denominator of a reduced fraction whose partial quotients are bounded by* 50.

Here "almost every" is in the sense of density: for $\mathcal{A} = \{1, 2, \dots, 50\}$,

$$\frac{1}{N} \#(\mathscr{D}_{\mathcal{A}} \cap [1, N]) \to 1,$$

as $N \to \infty$. The proof in fact shows that for any alphabet $\mathcal{A}$ having sufficiently large dimension

$$\delta_{\mathcal{A}} > \delta_0, \tag{2.23}$$

almost every admissible number is represented, where the value

$$\delta_0 = 1 - 5/312 \approx 0.98 \tag{2.24}$$

is sufficient. Using refined versions of Hensley's asymptotic expansion (2.13), the value $A = 50$ seems to satisfy (2.23). The reason Theorem Z needs no mention of admissibility is that any alphabet $\mathcal{A}$ with such a large dimension (2.24) must already contain both 1 and 2; missing even one of these letters will drop the dimension by too much. Hence there are actually no local obstructions in the theorem, cf. Remark 2.22.

To explain the source of this progress, we reformulate Zaremba's problem in a way that highlights the role of the hitherto unmentioned "thin orbit" lurking underneath.

2.1. **Reformulation.**

The key to the above progress is the old and elementary observation that

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

is equivalent to

$$\begin{pmatrix} * & b \\ * & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_k \end{pmatrix}. \tag{2.25}$$

With this observation, it is natural to introduce the semigroup generated by matrices of the above form with partial quotients restricted to the given alphabet. Let

$$\Gamma = \Gamma_{\mathcal{A}} := \left\langle \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} : a \in \mathcal{A} \right\rangle^{+}, \tag{2.26}$$

where the superscript "+" denotes generation as a semigroup (no inverse matrices). Then the orbit

$$\mathcal{O} = \mathcal{O}_{\mathcal{A}} := \Gamma \cdot \mathbf{v}_0 \tag{2.27}$$

with

$$\mathbf{v}_0 = (0,1)^t \tag{2.28}$$

isolates the set of second columns in Γ, and from (2.25) is hence in bijection with the set $\mathscr{R}_{\mathcal{A}}$. The "thinness" of the orbit is explained by Hensley's counting statement (2.17), which implies that

$$\#\{\mathbf{v} \in \mathcal{O} : \|\mathbf{v}\| < N\} \asymp N^{2\delta_{\mathcal{A}}},$$

as $N \to \infty$. If $\mathcal{O}$ consisted of all integer pairs $(b,d)^t$, the above count would be replaced by N^2, ignoring constants. So this is the reason we call $\mathcal{O}$ thin: it contains many fewer points than the ambient set in which it naturally sits.

From (2.25) again, the set $\mathscr{D}_{\mathcal{A}}$ is nothing more than the set of bottom right entries of matrices in $\Gamma_{\mathcal{A}}$. This can be isolated via:

$$\langle \mathbf{v}_0, \mathcal{O} \rangle = \langle \mathbf{v}_0, \Gamma \cdot \mathbf{v}_0 \rangle = \mathscr{D}_{\mathcal{A}}, \tag{2.29}$$

where the inner product is the standard one on $\mathbb{R}^2$. Thus d is represented if and only if there is a $\gamma \in \Gamma$ so that

$$d = \langle \mathbf{v}_0, \gamma \cdot \mathbf{v}_0 \rangle, \tag{2.30}$$

with $\mathbf{v}_0$ given in (2.28).

2.2. **Local Obstructions.**

One can now easily understand Remark 2.22, and the source of any potential local obstructions. The key observation, via (2.29), is that to understand $\mathscr{D}_{\mathcal{A}}(\bmod q)$, one needs only to understand the reduction of $\Gamma(\bmod q)$. And the latter can be analyzed by some algebra, namely the so-called *strong approximation* property; see e.g. [Rap12] for a comprehensive survey. As we will see below, this is a property which determines when the reduction mod q map is onto. For general

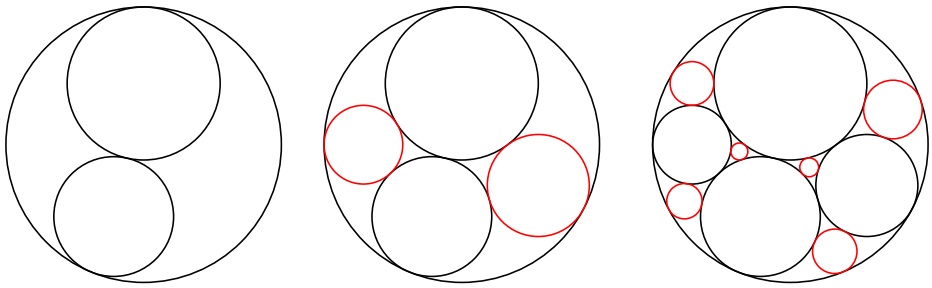
algebraic groups this is a deep theory, the first proof [MVW84] using the classification of finite simple groups. But for SL_2, the proofs are elementary, see e.g. [DSV03].

First observe that Γ sits inside the integer points of the *algebraic* group GL_2, meaning that any solution in $\mathbb{Z}$ to the polynomial equation $(ad-bc)m = 1$ gives an element $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z})$, and vice-versa. Actually GL_2 does not have strong approximation, (e.g. the determinant in $\mathrm{GL}_2(\mathbb{Z})$ can only be ± 1, while in $\mathrm{GL}_2(\mathbb{Z}/5\mathbb{Z})$ it is $1, 2, 3$ or 4; hence the reduction map cannot be onto). So we first pass to SL_2, as follows. The generators in (2.26) all have determinant -1, so the product of any two has determinant $+1$. We make these products the generators for a subsemigroup $\tilde{\Gamma}$ of Γ, that is, set $\tilde{\Gamma} := \Gamma \cap \mathrm{SL}_2$. We recover the original Γ-orbit $\mathcal{O}$ in (2.27) by a finite union of $\tilde{\Gamma}$-orbits. The limiting Cantor set and its Hausdorff dimension are unaffected.

Then strong approximation says essentially that for p a sufficiently large prime, and $q = p^e$ any p power, the reduction of $\tilde{\Gamma}$ mod q is all of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$. (It does not matter that $\tilde{\Gamma}$ is only a semigroup; upon reduction mod q, it becomes a group.) Moreover for *ramified* primes p (those for which the reduction mod p is not onto), the reduction mod sufficiently large powers of p *stabilizes* after some finite height. This means that there is some power $e_0 = e_0(p, \tilde{\Gamma})$ so that the following holds. For any higher power $e > e_0$, if $M \in \mathrm{SL}_2(\mathbb{Z}/p^e\mathbb{Z})$ is such that its reduction is in $\tilde{\Gamma} (\mathrm{mod}\, p^{e_0})$, then M is also in $\tilde{\Gamma} (\mathrm{mod}\, p^e)$. (These statements are best made in the language of p-adic numbers, which we avoid here.) A key ingredient is that, while $\tilde{\Gamma}$ is some strange subset of $\mathrm{SL}_2(\mathbb{Z})$, it is nevertheless *Zariski dense* in SL_2. This means that if $P(a, b, c, d)$ is a polynomial which vanishes for every $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \tilde{\Gamma}$, then P also vanishes on all matrices in SL_2 with entries in $\mathbb{C}$.

In the above, "sufficiently large," both for primes p to be unramified, and the stabilizing powers e_0 of ramified primes, can be effectively computed in terms of the generators. Then for an arbitrary modulus $q = p_1^{e_1} \cdots p_k^{e_k}$, the reduction mod q can be pieced together from those mod $p_j^{e_j}$ using a type of Chinese Remainder Theorem for groups called Goursat's Lemma. This leaves some finite group theory to determine completely the reduction of $\tilde{\Gamma}$ mod any q, and hence explains all local obstructions via (2.29).

We now leave Zaremba's problem, and return to sketch a proof of Theorem Z in §5.



(a) Three mutually tangent circles (b) Two more tangent circles (c) Six more tangent circles

FIGURE 6. Tangent circles

3. Integral Apollonian Gaskets

Apollonius of Perga (ca 262-190 BC) wrote a two-volume book on Tangencies, solving in every conceivable configuration the following general problem: Given three circles in the plane, any of which may have radius zero (a point) or infinity (a line), construct a circle tangent to the given ones. The volumes were lost but the statements survived via a survey of the work by Pappus. In the special case when the given three circles are themselves mutually tangent with disjoint points of tangency (Figure 6a), Apollonius proved that

$$\text{there are exactly two solutions} \tag{3.1}$$

to his problem (Figure 6b). Adding these new circles to the configuration, one has many other triples of tangent circles, and Apollonius's construction can be applied to them (Figure 6c). Iterating in this way *ad infinitum*, as apparently was first done in Leibniz's notebook, gives rise to a circle packing, the closure of which has become known in the last century as an *Apollonian gasket*. We restrict our discussion henceforth to bounded gaskets, such as that illustrated in Figure 1; there the number shown inside a circle is its curvature, that is, one over its radius. Such pictures have received considerable attention recently, see e.g. [LMW02, GLM+03, GLM+05, GLM+06a, GLM+06b, EL07, Sar07, Sar08, BGS10, KO11, Oh10, BF11, Sar11, Fuc11, FS11, OS12, Vin12, LO12, BK12]. We will focus our discussion on the following two problems:

(1) The Counting Problem: For a fixed gasket $\mathscr{G}$, how quickly do the circles shrink, or alternatively, how many circles are there in $\mathscr{G}$ with curvature bounded by a growing parameter T?

(2) The Local-Global Problem: Suppose $\mathscr{G}$ is furthermore *integral*, meaning that its circles all have integer curvatures, such as the gasket in Figure 1. How many distinct integers appear up to a growing parameter N? That is, count curvatures up to N, but without multiplicity.

Problem (2) does not yet look like a local-global question, but will soon turn into one. We first address Problem (1) in more detail.

3.1. **The Counting Problem.**

3.1.1. *Preliminaries.*

Some notation: for a typical circle C in a fixed bounded gasket $\mathscr{G}$, let $r(C)$ be its radius and

$$\kappa(C) = 1/r(C)$$

its curvature. Let

$$\mathcal{N}_{\mathscr{G}}(T) := \#\{C \in \mathscr{G} : \kappa(C) < T\} \tag{3.2}$$

be the desired counting function. To study this quantity, one might introduce an "L-function":

$$\mathcal{L}_{\mathscr{G}}(s) := \sum_{C \in \mathscr{G}} \frac{1}{\kappa(C)^s} = \sum_{C \in \mathscr{G}} r(C)^s. \tag{3.3}$$

Since the sum of the areas of inside circles in $\mathscr{G}$ yields the area of the bounding circle, the series $\mathcal{L}_{\mathscr{G}}$ converges for $\mathfrak{Re}(s) \geq 2$. It has some *abscissa of convergence* δ, meaning $\mathcal{L}_{\mathscr{G}}$ converges for $\mathfrak{Re}(s) > \delta$ and diverges for $\mathfrak{Re}(s) < \delta$. Boyd [Boy73] proved that this abscissa δ is none other than the Hausdorff dimension of the gasket $\mathscr{G}$, as should not be too surprising, comparing (3.3) with the definition (see (2.11)). In fact, Apollonian gaskets are rigid, in the sense that one can be mapped to any other by Möbius transformations. The latter are conformal (angle preserving) motions of the complex plane, sending $z \mapsto (az+b)/(cz+d)$, $ad - bc = 1$. Hence δ is a universal constant; McMullen [McM98] estimates that

$$\delta = 1.30568\ldots \tag{3.4}$$

From such considerations, Boyd [Boy82] was able to conclude that

$$\frac{\log \mathcal{N}_{\mathscr{G}}(T)}{\log T} \to \delta,$$

as $T \to \infty$.

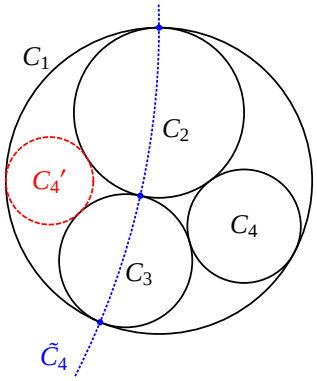


FIGURE 7. Generation from a root quadruple

To refine this crude estimate to an asymptotic formula for $\mathcal{N}_{\mathscr{G}}(T)$, the author and Oh [KO11] established a "spectral interpretation" for $\mathcal{L}_{\mathscr{G}}$, proving:

$$\mathcal{N}_{\mathscr{G}}(T) \sim c \cdot T^{\delta}, \tag{3.5}$$

for some $c = c(\mathscr{G}) > 0$, as $T \to \infty$. (This asymptotic was recently refined further in Vinogradov's thesis [Vin12] and independently by Lee-Oh [LO12], giving lower order error terms.) The remainder of this subsection is devoted to explaining this spectral interpretation and highlighting some of the ideas going into the proof of (3.5).

3.1.2. *Root quadruples and generation by reflection.*

It is easy to see [GLM+03, p. 14] that each such gasket $\mathscr{G}$ contains a *root configuration* $\mathcal{C} = \mathcal{C}(\mathscr{G}) := (C_1, C_2, C_3, C_4)$ of four largest mutually tangent circles in $\mathscr{G}$. Let

$$\mathbf{v}_0 = \mathbf{v}_0(\mathscr{G}) = (\kappa_1, \kappa_2, \kappa_3, \kappa_4)^t \tag{3.6}$$

with $\kappa_j = \kappa(C_j)$ be the *root quadruple* of corresponding curvatures. The bounding circle, being internally tangent to the others, is given opposite orientation to make all interiors disjoint; this is accounted for by giving it negative curvature. For example in Figure 1, the root quadruple is

$$\mathbf{v}_0 = (-10, 18, 23, 27)^t, \tag{3.7}$$

where the bounding circle has radius $1/10$.

Three tangent circles, say C_1, C_2, C_3 have three points of tangency, and determine a *dual circle* $\tilde{C}_4$ passing through these points, see Figure 7. Thus the root configuration $\mathcal{C}$ determines a *dual configuration* $\tilde{\mathcal{C}} = (\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4)$ of four mutually tangent circles, orthogonal to those in $\mathcal{C}$, see Figure 8. Reflection through $\tilde{C}_4$ fixes C_1, C_2, and C_3,

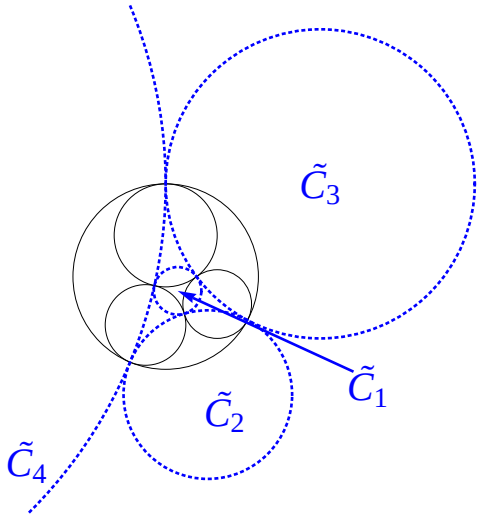


FIGURE 8. Root and dual configurations

and sends C_4 to C_4', the other solution to Apollonius's problem (3.1), see Figure 7. Starting with the root configuration, repeated reflections through the dual circles give the whole circle packing.

3.1.3. *Hyperbolic space and the group $\mathcal{A}$.*

Following Poincaré, we extend these circle reflections to the hyperbolic upper half space,

$$\mathbb{H}^3 := \{(x_1, x_2, y) : x_1, x_2 \in \mathbb{R}, y > 0\}, \tag{3.8}$$

replacing the action of the dual circle $\tilde{C}_j$ by a reflection through a (hemi)sphere $\mathfrak{s}_j$ whose equator is $\tilde{C}_j$ (with $j = 1, \ldots, 4$). We abuse notation, writing $\mathfrak{s}_j$ for both the hemisphere and the conformal map reflecting through $\mathfrak{s}_j$. The group

$$\mathcal{A} := \langle \mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_3, \mathfrak{s}_4 \rangle < \mathrm{Isom}(\mathbb{H}^3), \tag{3.9}$$

generated by these reflections acts discretely on $\mathbb{H}^3$; it is a so-called *Schottky* group, in that the four generating spheres have disjoint interiors.

The $\mathcal{A}$-orbit of any fixed base point $p_0 \in \mathbb{H}^3$ has a limit set in the boundary $\partial\mathbb{H}^3$, which is easily seen to be the original gasket, see Figure 9. A *fundamental domain* for an action is a region

$$\Omega \subset \mathbb{H}^3 \tag{3.10}$$

so that any point in $\mathbb{H}^3$ can be sent to Ω in an essentially unique way; for the action of $\mathcal{A}$, one can take Ω to be the exterior of the four

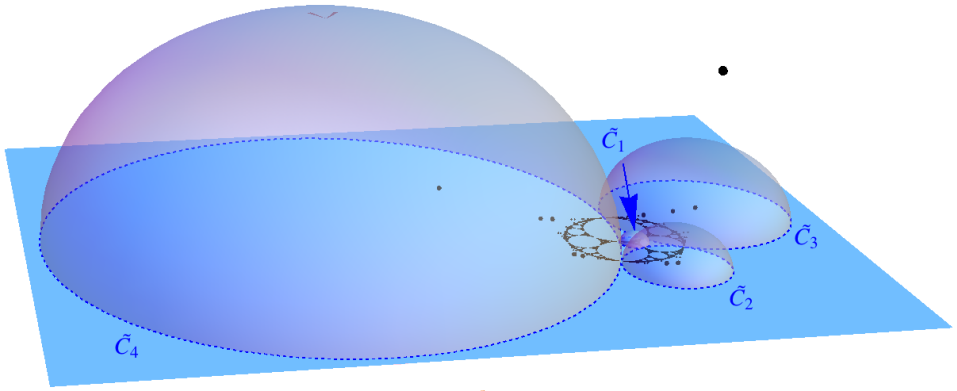


FIGURE 9. Poincaré extension: an $\mathcal{A}$-orbit in $\mathbb{H}^3$

hemispheres. To see this, observe that if a point $p = (x_1, x_2, y) \in \mathbb{H}^3$ is inside one of the spheres $\mathfrak{s}_j$, then its reflection $\mathfrak{s}_j(p)$ is outside of $\mathfrak{s}_j$ and has a strictly larger y-value. This does not guarantee that $\mathfrak{s}_j(p)$ is outside all of the other spheres, but if it is inside some $\mathfrak{s}_k$, then reflection through $\mathfrak{s}_k$ will again have even higher y-value. This procedure must halt after finitely many iterations, since the only limit points of $\mathcal{A}$ are in the boundary $\partial\mathbb{H}^3$ where $y = 0$. And it halts only when the image is outside of the four geodesic hemispheres. Uniqueness follows since any reflection $\mathfrak{s}_j$ takes a point in Ω to a point inside $\mathfrak{s}_j$, that is, not in Ω.

Two facts are evident from the above: first of all, $\mathcal{A}$ is *geometrically finite*, meaning it has a fundamental domain bounded by a finite number (here it is four) of geodesic[2] hemispheres; on the other hand, $\mathcal{A}$ has *infinite* co-volume, that is, any fundamental domain has infinite volume with respect to the hyperbolic measure

$$y^{-3} dx_1 dx_2 dy$$

in the coordinates (3.8). Note moreover that $\mathcal{A}$ has the structure of a *Coxeter group*, being free save the relations $\mathfrak{s}_j^2 = I$ for the generators. It is also the symmetry group of all Möbius transformations fixing $\mathscr{G}$.

3.1.4. *Descartes' Circle Theorem and integral gaskets.*

Next we need an observation due to Descartes in the year 1643 [Des01, pp. 37-50] (though his proof had a gap [Cox68]), that a quadruple $\mathbf{v} = (b_1, b_2, b_3, b_4)^t$ of signed curvatures of four mutually tangent circles lies on the cone

$$Q(\mathbf{v}) = 0, \tag{3.11}$$

[2]A geodesic in hyperbolic space is a straight vertical line or a semicircle orthogonal to the boundary $\partial\mathbb{H}^3$.

where Q is the so-called "Descartes quadratic form"

$$Q(\mathbf{v}) := 2\left(\kappa_1^2+\kappa_2^2+\kappa_3^2+\kappa_4^2\right) - \left(\kappa_1+\kappa_2+\kappa_3+\kappa_4\right)^2. \tag{3.12}$$

By a real linear change of variables, Q can be diagonalized to the form

$$x^2+y^2+z^2-w^2,$$

that is, it has signature $(3,1)$. Arguably the most beautiful formulation of Descartes' Theorem (rediscovered on many separate occasions) is the following excerpt from Soddy's 1936 *Nature* poem [Sod36]:

> Four circles to the kissing come. / The smaller are the bender. /
> The bend is just the inverse of / The distance from the center. /
> Though their intrigue left Euclid dumb / There's now no need for rule of thumb. /
> Since zero bend's a dead straight line / And concave bends have minus sign, /
> The sum of the squares of all four bends / Is half the square of their sum.

If κ_1, κ_2 and κ_3 are given, then (3.11) is a quadratic equation in κ_4 with two solutions, κ_4 and κ_4', say; this is an algebraic proof of Apollonius's theorem (3.1). It is then an elementary exercise to see that

$$\kappa_4+\kappa_4' = 2(\kappa_1+\kappa_2+\kappa_3).$$

In other words, if the quadruple $(\kappa_1,\kappa_2,\kappa_3,\kappa_4)^t$ is given, then one obtains the quadruple with κ_4 replaced by κ_4' via a linear action:

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 2 & 2 & 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \\ \kappa_4 \end{pmatrix} = \begin{pmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \\ \kappa_4' \end{pmatrix}.$$

Hence we have given an algebraic realization to the geometric action of $\tilde{C}_4$ (or $\mathfrak{s}_4$) on the root quadruple, see again Figure 7. Call the above 4×4 matrix S_4. Of course one could also send other κ_j to κ_j' keeping the three complementary curvatures fixed, via the matrices

$$S_1 = \begin{pmatrix} -1 & 2 & 2 & 2 \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, S_2 = \begin{pmatrix} 1 & & & \\ 2 & -1 & 2 & 2 \\ & & 1 & \\ & & & 1 \end{pmatrix}, S_3 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ 2 & 2 & -1 & 2 \\ & & & 1 \end{pmatrix}. \tag{3.13}$$

Moreover one can iterate these actions, so we introduce the so-called *Apollonian group* Γ, isomorphic to $\mathcal{A}$, generated by the S_j:

$$\Gamma := \langle S_1, S_2, S_3, S_4 \rangle. \tag{3.14}$$

Then the orbit

$$\mathcal{O} := \Gamma \cdot \mathbf{v}_0 \tag{3.15}$$

of the root quadruple $\mathbf{v}_0$ under the Apollonian group Γ consists of all quadruples corresponding to curvatures of four mutually tangent circles in the gasket $\mathscr{G}$. We can now explain the integrality of all curvatures in Figure 1: the group Γ has only integer matrices, so if the root quadruple $\mathbf{v}_0$ (or for that matter any four curvatures of mutually tangent circles in $\mathscr{G}$) is integral, then all curvatures in $\mathscr{G}$ are integers! This fact seems to have been first observed by Soddy [Sod37].

3.1.5. *Reformulating the counting statement, and the thin orbit.*

Moreover, note that starting with $\mathbf{v}_0$, any new circle generated by a reflection is the smallest in its configuration, and hence has largest curvature. That is, for $\mathbf{v} = \gamma \cdot \mathbf{v}_0 \in \mathcal{O}$, writing $\gamma \in \Gamma$ as a reduced word in the generators $\gamma = S_{i_k} \cdots S_{i_1}$, the last multiplication by S_{i_k} changes one entry, which is the largest entry in $\mathbf{v}$. Hence, setting $\|\mathbf{v}\|_\infty$ to be the max-norm, and for T large, we can rewrite $\mathcal{N}_{\mathscr{G}}(T)$ in (3.2) as

$$\mathcal{N}_{\mathscr{G}}(T) = 4 + \#\left\{\mathbf{v} \in \mathcal{O} : \mathbf{v} \neq \mathbf{v}_0,\ \|\mathbf{v}\|_\infty < T\right\}. \tag{3.16}$$

Here the first "4" accounts for the root quadruple $\mathbf{v}_0$.

We have thus converted the circle counting problem into something seemingly more tractable: the counting problem for a Γ-orbit. That said, we clearly need a better understanding of the group Γ. Returning to the Descartes form Q in (3.12), we have by construction (and one can check directly) that for each $j = 1, \ldots, 4$,

$$Q(S_j \cdot \mathbf{v}) = Q(\mathbf{v}),$$

for any $\mathbf{v}$. That is, each generator S_j lies in the so-called *orthogonal group* preserving the quadratic form Q,

$$O_Q := \left\{g \in \mathrm{GL}_4 : Q(g \cdot \mathbf{v}) = Q(\mathbf{v}),\ \forall \mathbf{v}\right\}.$$

Hence Γ also sits inside O_Q, and moreover inside $O_Q(\mathbb{Z})$, the group of matrices in O_Q with integer entries. The latter is a well understood *algebraic* group, again meaning that any solution to a certain set of polynomial equations gives an element in O_Q, and vice-versa. But Γ is quite a mysterious group, in particular having infinite index in $O_Q(\mathbb{Z})$ (this fact is equivalent to $\mathcal{A}$ having infinite co-volume). It is also worth noting here that the general membership problem in a group is known to be undecidable [Nov55], so presenting a matrix group via its generators leaves much to be desired.[3]

Just as in Zaremba's problem, we can now again call this orbit $\mathcal{O}$ *thin*; indeed, for the counting problem with Γ replaced by the full

[3]That said, for our particular group Γ, one can use a reduction algorithm to root quadruples to determine membership.

group $O_Q(\mathbb{Z})$ (which is an example of what's called an "arithmetic lattice"), standard arguments in automorphic forms or ergodic theory [DRS93, EM93] show that

$$\#\{\mathbf{v} \in O_Q(\mathbb{Z}) \cdot \mathbf{v}_0 : \|\mathbf{v}\|_\infty < T\} \sim c\, T^2, \qquad \text{as } T \to \infty, \tag{3.17}$$

for some $c > 0$. So comparing (3.17) to (3.16), (3.5) and (3.4), where the power drops from T^2 to T^δ with $\delta < 2$, we see that the Γ orbit is quite degenerate, having many fewer points.

3.1.6. *Sketch of the counting statement.*

Finally, we explain the aforementioned spectral interpretation, by first giving an analogous elementary example of a counting statement in another discrete group: the integers. Let us spectrally count the number of integers of size at most T:

$$\mathcal{N}_{\mathbb{Z}}(T) := \#\{n \in \mathbb{Z} : |n| < T\}.$$

Of course this is a trivial problem,

$$\mathcal{N}_{\mathbb{Z}}(T) = \lfloor 2T + 1 \rfloor = 2T + O(1), \tag{3.18}$$

but it will be instructive to analyze it by harmonic analysis. To this end, let

$$f(x) := \mathbf{1}_{\{|x|<1\}},$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Scale f to

$$f_T(x) := f(x/T) = \mathbf{1}_{\{|x|<T\}},$$

and periodize it with respect to the discrete group $\mathbb{Z}$:

$$F_T(x) := \sum_{n \in \mathbb{Z}} f_T(n + x). \tag{3.19}$$

Then we have

$$F_T(0) = \sum_{n \in \mathbb{Z}} \mathbf{1}_{\{|n|<T\}} = \mathcal{N}_{\mathbb{Z}}(T). \tag{3.20}$$

By construction, $F_T(x) = F_T(x + 1)$, that is, it takes values on the circle $X := \mathbb{Z}\backslash\mathbb{R}$, and is square-integrable, $F_T \in L^2(X)$. The Laplace operator

$$\Delta := -\operatorname{div} \circ \operatorname{grad} = -\frac{\partial^2}{\partial x^2}$$

on smooth functions can be extended to act on the whole Hilbert space $L^2(X)$ and is self-adjoint and positive definite (by our choice of sign) with respect to the standard inner product

$$\langle F, G \rangle = \int_X F(x)\bar{G}(x)dx.$$

(Proof: partial integration.) Its *spectrum* $\mathrm{Spec}(\Delta)$ is just the set of its eigenvalues, with multiplicity. Elementary Fourier analysis shows that eigenfunctions of Δ invariant under $\mathbb{Z}$-translations are scalar multiples of

$$\varphi_m : x \mapsto e^{2\pi i m x}$$

for $m \in \mathbb{Z}$. This function has Laplace eigenvalue

$$\lambda_m = 4\pi^2 m^2,$$

and hence these numbers λ_m completely exhaust the spectrum (they have multiplicity two, except when $m = 0$). Expanding spectrally gives

$$F_T(x) = \sum_{\lambda_m \in \mathrm{Spec}(\Delta)} \langle F_T, \varphi_m \rangle \, \varphi_m(x), \tag{3.21}$$

where equality is in the L^2-sense. (Note that the φ_m are already scaled to have unit L^2-norm.) The bottom of the spectrum $\lambda_0 = 0$ corresponds to the constant function $\varphi_0(x) = 1$, and contributes the entire "main term" in (3.18) to (3.21):

$$\langle F_T, \varphi_0 \rangle \cdot \varphi_0 = \left(\int_{\mathbb{Z}\backslash\mathbb{R}} \sum_{n \in \mathbb{Z}} f_T(n + x) \cdot 1 \, dx \right) \cdot 1 = T \int_{\mathbb{R}} f(x) dx = 2T,$$

after inserting (3.19), a change of variables, and "unfolding" $\int_{\mathbb{Z}\backslash\mathbb{R}} \sum_{\mathbb{Z}}$ to just $\int_{\mathbb{R}}$. That said, the equality (3.21) is in the L^2 sense, not pointwise (we cannot evaluate (3.21) at the point $x = 0$, as needed in (3.20)). Moreover, the rest of the spectrum in (3.21), if bounded in absolute value,

$$\sum_{\substack{\lambda_m \in \mathrm{Spec}(\Delta) \\ \lambda_m \neq \lambda_0}} \left| \langle F_T, \varphi_m \rangle \, \varphi_m \right|,$$

does not converge, the mth term being of size $1/m$. (Exercise.) But there are standard methods (smoothing and later unsmoothing) which overcome these technical irritants.

A version of the above works with the Apollonian group Γ in place of $\mathbb{Z}$, once one overcomes a number of further technical obstructions. The reader may wish to omit the following paragraph on the first pass; it is not essential to the sequel.

We now need non-abelian harmonic analysis on the space $L^2(X)$ with

$$X := \mathcal{A}\backslash\mathbb{H}^3,$$

the hyperbolic 3-fold in Figure 9. The (positive definite) hyperbolic Laplacian is

$$\Delta = -y^2\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y^2}\right) + y\frac{\partial}{\partial y}$$

in the coordinates (3.8). The spectrum in this setting, as studied by Lax-Phillips [LP82], has both continuous and discrete components (though only a finite number of the latter). As X has *infinite* volume, the constant function is no longer square-integrable, and the bottom eigenvalue λ_0 is strictly positive. A beautiful result in Patterson-Sullivan theory [Pat76, Sul84] relates this eigenvalue to the Hausdorff dimension of the limiting gasket $\mathscr{G}$, namely

$$\lambda_0 = \delta(2-\delta).$$

The corresponding base eigenfunction φ_0 replaces the role of the constant function. Here we have used crucially that $\mathcal{A}$ is geometrically finite, and that $\delta > 1$, see (3.4). Even this is insufficient: because of the non-Euclidean norm $\|\cdot\|_\infty$ in (3.16), one must work not on X but its unit tangent bundle $Y := T^1(X)$. And moreover we do not know how to handle the continuous spectrum directly, applying instead general results in the representation theory of semisimple groups about ergodic properties of flows on Y. At this point, we will not say more about the proof, inviting the interested reader to consult the original references [KO11, Vin12, LO12].

3.2. **The Local-Global Problem.**

Assume now that $\mathscr{G}$ is not only bounded but also integral (recall that this means it has only integer curvatures). If the curvatures are all even, say, then we can stretch the gasket by a factor of two, doubling the radii and halving the (still integral) curvatures. In this way, we can rescale an integral gasket to make it *primitive*, meaning that there is no number other than ± 1 dividing all of the curvatures. In fact, all of the salient features of the problem persist if we fix $\mathscr{G}$ to be the packing shown in Figure 1, and we do so henceforth. Recall that the problem we wish to now address is: How many curvatures are there up to some parameter N, counting without multiplicity, that is, counting only distinct curvatures?

First some more notation: let $\mathscr{K} = \mathscr{K}(\mathscr{G})$ be the set of all curvatures of circles C in the gasket $\mathscr{G}$,

$$\mathscr{K} := \{n \in \mathbb{Z} : \exists C \in \mathscr{G} \text{ with } \kappa(C) = n\},$$

and call n *represented* if $n \in \mathscr{K}$. Staring at Figure 1 for a moment or two, one might observe that every curvature in our $\mathscr{G}$ is

$$\equiv 2, 3, 6, 11, 14, 15, 18, \text{ or } 23 \pmod{24}. \tag{3.22}$$

These are the local obstructions for $\mathscr{G}$; accordingly, we call n *admissible* if it satisfies (3.22), and set $\mathscr{A} = \mathscr{A}(\mathscr{G})$ to be the set of admissible numbers. In general, one calls n admissible if, as before, it is everywhere locally represented,

$$n \in \mathscr{K} \pmod{q}, \ \forall q \geq 1. \tag{3.23}$$

It cannot be the case that $\mathscr{A} = \mathscr{K}$, since, for example, $n = 15$ is admissible, but a circle of radius $1/15$ does not appear in our gasket. Nevertheless, as in Zaremba's problem, we have the following

Conjecture A. *Every sufficiently large admissible number is the curvature of some circle in $\mathscr{G}$.*

This conjecture is stated by Graham-Lagarias-Mallows-Wilks-Yan [GLM+03, p. 37], in the first of a lovely series of papers on Apollonian gaskets and generalizations. They observe empirically that congruence obstructions for any integral gasket seem to be to the modulus 24, and this is completely clarified (as we explain below) by Fuchs [Fuc11] in her thesis. Further convincing numerical evidence towards the conjecture is given in Fuchs-Sanden [FS11]. Here is some recent progress.

Theorem A (Bourgain-K. 2012 [BK12])**.** *Almost every admissible number is the curvature of some circle in $\mathscr{G}$.*

Again, "almost every" is in the sense of density, that

$$\frac{\#(\mathscr{K} \cap [1, N])}{\#(\mathscr{A} \cap [1, N])} \to 1, \tag{3.24}$$

as $N \to \infty$. It follows from the congruence restrictions (3.22) that for N large, $\#(\mathscr{A} \cap [1, N])$ is about $N/3$ (there are 8 admissible residue classes mod 24), so (3.24) is equivalent to

$$\#(\mathscr{K} \cap [1, N]) \sim \frac{N}{3}.$$

Some history on this problem: Graham *et al* [GLM+03] already made the first progress, proving that

$$\#(\mathscr{K} \cap [1, N]) \gg N^{1/2}. \tag{3.25}$$

Then Sarnak [Sar07] showed

$$\#(\mathscr{K} \cap [1, N]) \gg \frac{N}{\sqrt{\log N}}, \tag{3.26}$$

before Bourgain-Fuchs [BF11] settled the so-called Positive Density Conjecture, that

$$\#(\mathscr{K} \cap [1, N]) \gg N. \tag{3.27}$$

A key observation in the proof of Theorem A is that the problem is nearly identical to Zaremba's, in the following sense. Recall from (3.15) that the orbit $\mathcal{O} = \Gamma \cdot \mathbf{v}_0$ of the root quadruple $\mathbf{v}_0$ under the Apollonian group Γ contains all quadruples of curvatures, and in particular its entries consist of all curvatures in $\mathscr{G}$. Hence the set $\mathscr{K}$ of all curvatures is simply the finite union of sets of the form

$$\langle \mathbf{w}_0, \mathcal{O} \rangle = \langle \mathbf{w}_0, \Gamma \cdot \mathbf{v}_0 \rangle , \tag{3.28}$$

as $\mathbf{w}_0$ ranges through the standard basis vectors $\mathbf{e}_1 = (1, 0, 0, 0)^t, \ldots,$ $\mathbf{e}_4 = (0, 0, 0, 1)^t$, each picking off one entry of $\mathcal{O}$. A heuristic analogy between Zaremba and the Apollonian problem is actually already given in [GLM+03, p. 37], but it is crucial for us that both problems are exactly of the form (3.28); compare to (2.29). That is, n is represented if and only if there is a γ in the Apollonian group Γ and some $\mathbf{w}_0 \in \{\mathbf{e}_1, \ldots, \mathbf{e}_4\}$ so that

$$n = \langle \mathbf{w}_0, \gamma \cdot \mathbf{v}_0 \rangle . \tag{3.29}$$

Before saying more about the proof of Theorem A, we first discuss admissibility in greater detail.

3.2.1. *Local obstructions.*

Through (3.28), the admissibility condition (3.23) is again reduced to the study of the projection of Γ modulo q. An important feature here is that, like in the Zaremba case, the group Γ is Zariski dense in O_Q. Recall that this means: if $P(\gamma)$ is a polynomial in the entries of a 4×4 matrix γ which vanishes for every $\gamma \in \Gamma$, then P also vanishes on all complex matrices in O_Q.

We would like again to exploit strong approximation, but neither O_Q nor its orientation preserving subgroup $\mathrm{SO}_Q := O_Q \cap \mathrm{SL}_4$ have this property (being not even connected). But there is a standard method of applying strong approximation anyway, by first passing to a certain cover, as we now describe.

From the theory of rational quadratic forms [Cas78], special orthogonal groups are covered by so-called *spin groups*, and it is a pleasant accident that, since Q has signature $(3, 1)$, the spin group of $\mathrm{SO}_Q(\mathbb{R})$ is isomorphic to $\mathrm{SL}_2(\mathbb{C})$; let us explain this covering map. The formulae are nicer if we first change variables (over $\mathbb{Q}$) from our quadratic form Q to the equivalent form

$$\tilde{Q}(x, y, z, w) := xw + y^2 + z^2.$$

Observe that the matrix

$$M := \begin{pmatrix} -x & y+iz \\ y-iz & w \end{pmatrix}$$

has determinant equal to $-\tilde{Q}$ and is Hermitian, that is, fixed under transpose-conjugation. The group $\mathrm{SL}_2(\mathbb{C})$, consisting of 2×2 complex matrices of determinant one, acts on M by

$$\mathrm{SL}_2(\mathbb{C}) \ni g : M \mapsto g\cdot M\cdot \bar{g}^t =: M' = \begin{pmatrix} -x' & y'+iz' \\ y'-iz' & w' \end{pmatrix},$$

with M' also Hermitian and of determinant $-\tilde{Q}$. Then it is easy to see that $(x',y',z',w')^t$ is a linear change of variables from $(x,y,z,w)^t$, via left multiplication by a matrix whose entries are quadratic in the entries of g. Explicitly, if

$$g = \begin{pmatrix} a+\alpha i & b+\beta i \\ c+\gamma i & d+\delta i \end{pmatrix}, \tag{3.30}$$

then the change of variables matrix is

$$\frac{1}{|\det(g)|^2}\begin{pmatrix} a^2+\alpha^2 & 2(ac+\alpha\gamma) & 2(c\alpha-a\gamma) & -c^2-\gamma^2 \\ ab+\alpha\beta & bc+ad+\beta\gamma+\alpha\delta & d\alpha+c\beta-b\gamma-a\delta & -cd-\gamma\delta \\ a\beta-b\alpha & -d\alpha+c\beta-b\gamma+a\delta & -bc+ad-\beta\gamma+\alpha\delta & d\gamma-c\delta \\ -b^2-\beta^2 & -2(bd+\beta\delta) & 2(b\delta-d\beta) & d^2+\delta^2 \end{pmatrix}. \tag{3.31}$$

Let $\tilde{\rho}$ be the (rational) map from $\mathrm{SL}_2(\mathbb{C})$ to $\mathrm{GL}_4(\mathbb{R})$, sending (3.30) to (3.31); then by construction (again one can verify directly) the image is in $\mathrm{SO}_{\tilde{Q}}(\mathbb{R})$. (Some minor technical points: Being quadratic in the entries, $\tilde{\rho}$ is a double cover, with $\pm I$ having the same image. Moreover, $\mathrm{SL}_2(\mathbb{C})$ is connected while $\mathrm{SO}_{\tilde{Q}}(\mathbb{R})$ has two connected components, so $\tilde{\rho}$ only maps onto the identity component $\mathrm{SO}^\circ_{\tilde{Q}}$.) Then changing variables from $\tilde{Q}$ back to the Descartes form Q by a conjugation, one gets the desired map

$$\rho : \mathrm{SL}_2(\mathbb{C}) \to \mathrm{SO}_Q(\mathbb{R}).$$

It is straightforward then to compute the pullback of $\Gamma\cap\mathrm{SO}_Q$ under ρ (see [GLM+05, Fuc11]), the answer being the following

Lemma 3.32. *There is*[4] *a homomorphism $\rho : \mathrm{SL}_2(\mathbb{C}) \to \mathrm{SO}_Q(\mathbb{R})$ so that the group $\tilde{\Gamma} := \rho^{-1}(\Gamma\cap\mathrm{SO}_Q)$ sits in $\mathrm{SL}_2(\mathbb{Z}[i])$ and is generated by*

$$\tilde{\Gamma} = \left\langle \pm\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \pm\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \pm\begin{pmatrix} 1+2i & -2 \\ -2 & 1-2i \end{pmatrix} \right\rangle. \tag{3.33}$$

[4]And one can easily write it down explicitly: it is a conjugate of (3.31), but much messier and not particularly enlightening. We spare the reader.

Moreover, recalling the generators S_j for Γ in (3.13), one can arrange ρ so that $\rho : \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right) \mapsto S_2S_3$, and $\rho : \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right) \mapsto S_4S_3$.

In fact, we have just realized a conjugate of the group $\mathcal{A}$ (or rather its index-two orientation preserving subgroup) explicitly in terms of matrices in $\mathrm{PSL}(2,\mathbb{C}) \cong \mathrm{Isom}^+(\mathbb{H}^3)$.

From here, one follows the strategy outlined in §2.2. Fuchs [Fuc11] proved an explicit version of strong approximation for $\tilde{\Gamma} < \mathrm{SL}_2(\mathbb{Z}[i])$ (one considers reduction mod principal ideals (q)) via Goursat's Lemma, some finite group theory, and other ingredients, enabling her to determine completely the reduction of Γ modulo any q, and hence explaining all local obstructions. The answer is that all primes other than 2 and 3 are *unramified*, meaning, as in §2.2, that for $(q, 6) = 1$,

$$\Gamma \cap \mathrm{SO}_Q \pmod q \quad = \quad \mathrm{SO}_Q(\mathbb{Z}/q\mathbb{Z}).$$

Recall again that the right hand side above is a well-understood group. And moreover, the prime 2 stabilizes (with the same meaning as §2.2) at the power $e_0(2) = 3$, that is at 8, and the prime 3 stabilizes immediately at $e_0(3) = 1$. Then reducing Γ modulo $2^3 \cdot 3 = 24$, one obtains some explicit finite group, and looking at all the values of (3.28) for the given root quadruple $\mathbf{v}_0(\mathscr{G})$, one immediately sees all admissible residue classes.

3.2.2. *Partial Progress.*

Lemma 3.32 can already be quite useful; in particular, it easily implies (3.25) and (3.26), as follows.

The Apollonian group Γ contains the matrix S_4S_3, which by Lemma 3.32 is the image under ρ of $\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$. The latter (and hence the former) is a *unipotent* matrix, meaning that all its eigenvalues are equal to 1. These have the important property that they grow only polynomially under exponentiation; in particular, $\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)^k = \left(\begin{smallmatrix} 1 & 0 \\ 2k & 1 \end{smallmatrix}\right)$, and one can check directly from the definitions (3.13) that

$$(S_4S_3)^k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 4k^2 - 2k & 4k^2 - 2k & 1 - 2k & 2k \\ 4k^2 + 2k & 4k^2 + 2k & -2k & 2k + 1 \end{pmatrix}.$$

Put the above matrix into (3.29) with the root quadruple $\mathbf{v}_0$ for our fixed gasket from (3.7), and take $\mathbf{w}_0 = \mathbf{e}_4$, say. Then for any $k \in \mathbb{Z}$, the number

$$\left\langle \mathbf{e}_4 \ , \ (S_4S_3)^k \cdot \mathbf{v}_0 \right\rangle = 32k^2 + 24k + 27 \tag{3.34}$$

is represented. That is, the set of represented numbers contains the values of this quadratic polynomial. From this observation, made in [GLM+03], it is immediate that (3.25) holds. Geometrically, these curvatures correspond to circles in the packing tangent to C_1 and C_2, since these are fixed by the corresponding reflections through $\tilde{C}_4$ and $\tilde{C}_3$. For example, the values $k = -2, -1, 0, 1, 2$ in (3.34) give curvatures $107, 35, 27, 83, 203$, respectively. These are visible in Figure 1; they are all tangent to the circles of curvature -10 (the bounding circle) and 18, skipping every other such circle. Using $\mathbf{w}_0 = \mathbf{e}_3$ instead of $\mathbf{e}_4$ in (3.34) gives the polynomial $32k^2 - 8k + 23$, the values of which correspond to the skipped circles.

To prove (3.26), we make the following observation, due to Sarnak [Sar07]. It is well known that the matrices $\pm\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$ and $\pm\left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ (which map under ρ to S_4S_3 and S_2S_3, respectively) generate the group

$$\Lambda(2) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : \begin{array}{c} a \equiv d \equiv 1 (\mathrm{mod}\, 2) \\ b \equiv c \equiv 0 (\mathrm{mod}\, 2) \end{array} \right\}. \tag{3.35}$$

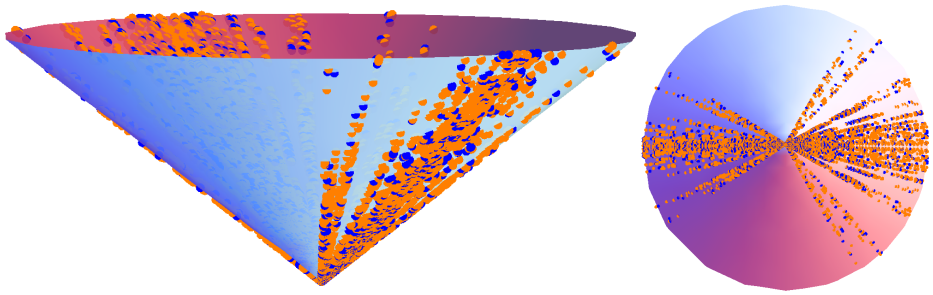
This is the so-called level-2 *principal congruence* subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Hence by Lemma 3.32, the group Γ contains

$$\Xi := \langle\ S_2S_3\ ,\ S_4S_3\ \rangle = \rho\left(\Lambda(2)\right). \tag{3.36}$$

The point is that $\Lambda(2)$ is *arithmetic*, being defined in (3.35) by congruences. Then for any integer ℓ coprime to $2k$, there is a matrix $\left(\begin{smallmatrix} * & * \\ 2k & \ell \end{smallmatrix}\right)$ in $\Lambda(2)$. One can work out, with the same $\mathbf{v}_0$ and $\mathbf{w}_0$ as above, that

$$\left\langle \mathbf{e}_4\ ,\ \rho\begin{pmatrix} * & * \\ 2k & \ell \end{pmatrix} \cdot \mathbf{v}_0 \right\rangle = 32k^2 + 24k\ell + 17\ell^2 + 10. \tag{3.37}$$

For example, the choices $(2k, \ell) = (4, -3), (2, -1), (4, -1)$, and $(6, -1)$ give curvatures $147, 35, 107$, and 243, respectively, visible up the left side of Figure 1, all tangent to the bounding circle (since Ξ in (3.36) fixes C_1). Observe also that setting $\ell = 1$ in (3.37) recovers (3.34). In this way, Sarnak [Sar07] proved that the set $\mathscr{K}$ of represented numbers contains all *primitive* (meaning with $2k$ and ℓ coprime) values of the shifted binary quadratic form in (3.37). Note that the quadratic form has discriminant $24^2 - 4 \cdot 32 \cdot 17 = -1600$, and so (3.37) is definite, taking only positive values. The number of distinct primitive values of (3.37) up to N was determined by Landau [Lan08]: it is asymptotic to a constant times $N/\sqrt{\log N}$, thereby proving (3.26). A much more delicate and clever but still "elementary" (no automorphic forms are harmed) argument goes into the proof of the Positive Density Conjecture (3.27), using an ensemble of such shifted binary quadratic forms.



(a) View from the side (b) View from below

FIGURE 10. The thin Pythagorean orbit $\mathcal{O}$ in (4.9). Points are marked according to whether the hypotenuse is prime (●) or composite (●).

For Theorem A, one needs the theory of automorphic representations for the full Apollonian group, as hinted to at the end of §3.1.6.

We now leave the discussion of the Apollonian problem, returning to it again in §5.

4. The Thin Pythagorean Problem

A *Pythagorean triple* $\mathbf{x} = (x, y, z)^t$ is a point on the cone

$$Q(\mathbf{x}) = 0, \tag{4.1}$$

where Q is the "Pythagorean quadratic form"

$$Q(\mathbf{x}) := x^2 + y^2 - z^2.$$

Throughout we consider only *integral* triples, $\mathbf{x} \in \mathbb{Z}^3$, and assume that x, y, and z are coprime; such a triple is called *primitive*. Elementary considerations then force the hypotenuse z to be odd, and x and y to be of opposite parity; we assume henceforth that x is odd and y is even. The cone has a singularity at the origin, so we only consider its top half, assuming subsequently that the hypotenuse is positive, $z > 0$.

Diophantus (and likely the Babylonians [Pli], who preceded him by about as much as he precedes us) knew how to parametrize Pythagorean triples: Given $\mathbf{x}$, there is a pair $\mathbf{v} = (u, v)$ of coprime integers of opposite parity so that

$$\begin{cases} x &= u^2 - v^2 \\ y &= 2uv \\ z &= u^2 + v^2. \end{cases} \tag{4.2}$$

That the converse is true is elementary algebra: any such pair $\mathbf{v}$ inserted into (4.2) gives rise to a triple $\mathbf{x}$ satisfying (4.1). For example, it is easy to see that the triple

$$\mathbf{x}_0 = (3, 4, 5)^t \tag{4.3}$$

corresponds to the pair

$$\mathbf{v}_0 = (2, 1)^t. \tag{4.4}$$

4.1. **Orbits and the Spin Representation.**

As in the Apollonian case, the Pythagorean form Q has a *special* (determinant one) orthogonal group preserving it:

$$\mathrm{SO}_Q := \{g \in \mathrm{SL}_3 : Q(g \cdot \mathbf{x}) = Q(\mathbf{x})\}. \tag{4.5}$$

And as before, this group is also better understood by passing to its spin cover. Since the Pythagorean form Q has signature $(2, 1)$, there is an accidental isomorphism between its spin group and $\mathrm{SL}_2(\mathbb{R})$, given explicitly as follows.

Observe that SL_2 acts on a pair $\mathbf{v}$ by left multiplication; via (4.2), this action then extends to a linear action on $\mathbf{x}$. In coordinates, it is an elementary computation that the action of $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ on $\mathbf{v}$ corresponds to left multiplication on $\mathbf{x}$ by

$$\frac{1}{ad - bc} \begin{pmatrix} \frac{1}{2}(a^2 - b^2 - c^2 + d^2) & ac - bd & \frac{1}{2}(a^2 - b^2 + c^2 - d^2) \\ ab - cd & bc + ad & ab + cd \\ \frac{1}{2}(a^2 + b^2 - c^2 - d^2) & ac + bd & \frac{1}{2}(a^2 + b^2 + c^2 + d^2) \end{pmatrix}. \tag{4.6}$$

One can check directly from the definition (4.5) that (4.6) is an element of SO_Q, in fact of the connected component SO_Q° of the identity, and hence we have explicitly constructed the spin homomorphism

$$\rho : \mathrm{SL}_2(\mathbb{R}) \to \mathrm{SO}_Q(\mathbb{R}) : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto (4.6).$$

Given a Pythagorean triple $\mathbf{x}_0$, such as that in (4.3), the group $\Gamma := \mathrm{SO}_Q^\circ(\mathbb{Z})$ of all *integer* matrices in SO_Q° acts by left multiplication, giving the full orbit $\mathcal{O} = \Gamma \cdot \mathbf{x}_0$ of all Pythagorean triples (with our convention that $z > 0$, x is odd, and y is even).

Via (4.2) again, this SO_Q action on $\mathbf{x}$ is equivalent to the SL_2 action on $\mathbf{v}$. For a primitive $\mathbf{v} \in \mathbb{Z}^2$, both the integrality and primitivity are preserved by restricting the action to just the integral matrices $\mathrm{SL}_2(\mathbb{Z})$. Moreover, one should preserve the parity condition on $\mathbf{v}$ by restricting further to only the principal 2-congruence subgroup

$$\Lambda(2) = \left\{\gamma \in \mathrm{SL}_2(\mathbb{Z}) : \gamma \equiv I (\mathrm{mod}\, 2)\right\} = \left\langle \pm \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \pm \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \right\rangle,$$

which already appeared in §3.2.2. One can check directly that the image (4.6) of any $\gamma \in \Lambda(2)$ is an integral matrix, that is, in $\mathrm{SO}_Q(\mathbb{Z})$. For $\mathbf{v}_0$ corresponding to $\mathbf{x}_0$, the orbit $\tilde{\mathcal{O}} := \tilde{\Gamma} \cdot \mathbf{v}_0$ under the full group $\tilde{\Gamma} := \Lambda(2)$ consists of all coprime (u, v) with u even and v odd.

Prompted by the Affine Sieve[5] [BGS06, BGS10, SGS11] one may wish to study *thin* orbits $\mathcal{O}$ of Pythagorean triples. Here one replaces the full group $\mathrm{SO}_Q(\mathbb{Z})$ by some finitely generated subgroup Γ of infinite index. Equivalently one can consider an orbit $\tilde{\mathcal{O}}$ of $\mathbf{v}_0$ under an infinite index subgroup $\tilde{\Gamma}$ of $\Lambda(2)$. We illustrate the general theory via the following concrete example.

We first give a sample $\tilde{\mathcal{O}}$ orbit: in comparison with the generators of $\Lambda(2)$, let $\tilde{\Gamma}$ be the group generated by the following two matrices

$$\tilde{\Gamma} := \left\langle \pm \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad \pm \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \right\rangle. \tag{4.7}$$

This group clearly sits inside $\Lambda(2)$ but it is not immediately obvious whether it is of finite or infinite index; as we will see later, the index is infinite. Taking the base pair $\mathbf{v}_0$ in (4.4), we form the orbit

$$\tilde{\mathcal{O}} := \tilde{\Gamma} \cdot \mathbf{v}_0. \tag{4.8}$$

Correspondingly, we can take the base triple $\mathbf{x}_0$ in (4.3), and form the orbit

$$\mathcal{O} := \Gamma \cdot \mathbf{x}_0 \tag{4.9}$$

of $\mathbf{x}_0$ under the group

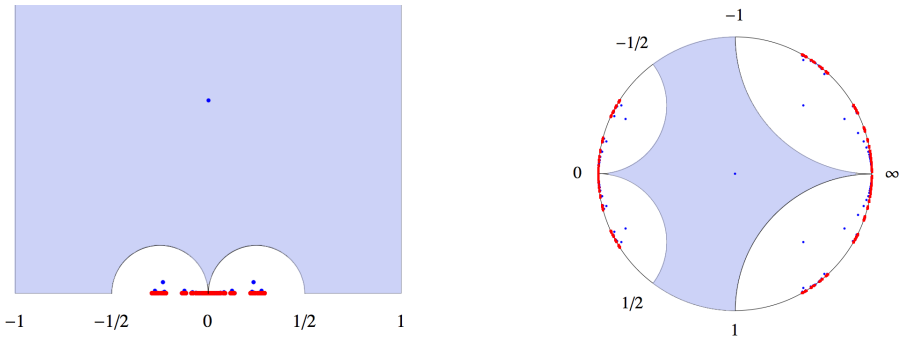
$$\Gamma := \langle M_1, M_2 \rangle, \tag{4.10}$$

where M_1 and M_2 are the images under ρ of the matrices generating $\tilde{\Gamma}$; one can elementarily compute from (4.7) and (4.6) that

$$M_1 := \begin{pmatrix} -1 & -2 & -2 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \end{pmatrix}, \qquad M_2 := \begin{pmatrix} -7 & 4 & 8 \\ -4 & 1 & 4 \\ -8 & 4 & 9 \end{pmatrix}. \tag{4.11}$$

Figure 10 illustrates this orbit $\mathcal{O}$. We can visually verify that the orbit looks thin, and in the next subsection we confirm this rigorously.

[5]We have insufficient room to survey this beautiful theory, for which the reader is directed to any number of excellent surveys, e.g. [SG12].



(a) Upper half plane model (b) Disk model

FIGURE 11. The orbit of $i \in \mathbb{H}$ under $\tilde{\Gamma}$.

4.2. **The Orbit is Thin.**

The group $\mathrm{SL}_2(\mathbb{R})$ also acts on the hyperbolic upper half-plane

$$\mathbb{H} := \{z = x + iy : x \in \mathbb{R}, y > 0\}$$

by fractional linear transformations,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \mapsto \frac{az+b}{cz+d}. \tag{4.12}$$

The action of our group $\tilde{\Gamma}$ in (4.7) on $\mathbb{H}$ has a fundamental domain (the definition is similar to (3.10)) given by

$$\{z \in \mathbb{H} : |\mathfrak{Re}(z)| < 1,\ |z - 1/4| > 1/4,\ |z + 1/4| > 1/4\},$$

where the distances above are Euclidean; see Figure 11a. The hyperbolic measure is $y^{-2}\,dx\,dy$, and hence this region again has *infinite* hyperbolic area. Equivalently, the index of $\tilde{\Gamma}$ in $\Lambda(2)$ is infinite (it is well-known that $\Lambda(2)$ has finite co-area), as claimed.

Any orbit of a fixed base point in $\mathbb{H}$ under $\tilde{\Gamma}$ has some *limit set* $\mathscr{C} = \mathscr{C}(\tilde{\Gamma})$ in the boundary $\partial\mathbb{H}$. A piece of this Cantor-like set can already be seen in Figure 11a. But to see it fully, we show in Figure 11b the same $\tilde{\Gamma}$-orbit in the disk model

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\},$$

by composing the action of $\tilde{\Gamma}$ with the map

$$\mathbb{H} \to \mathbb{D} : z \mapsto \frac{z-i}{z+i}$$

(which encodes the observation that points in the upper half plane are closer to i than they are to $-i$). In the disk model, one more clearly sees the limit set as the set of "directions" in which the orbit $\mathcal{O}$ can

grow – juxtapose Figure 10b on Figure 11b. This limit set $\mathscr{C}$ has some Hausdorff dimension $\delta = \delta(\tilde{\Gamma}) \in [0, 1]$; one can estimate

$$\delta \approx 0.59\ldots \tag{4.13}$$

This dimension (also called the "critical exponent of Γ") is again an important geometric invariant, measuring the "thinness" of Γ, as illustrated in the following counting statement [Kon07, Kon09, KO12]. Let $\|\mathbf{x}\|$ be the Euclidean norm. There is some $c > 0$ so that

$$\#\{\mathbf{x} \in \mathcal{O} : \|\mathbf{x}\| < N\} \sim c\, N^{\delta}, \qquad \text{as } N \to \infty. \tag{4.14}$$

Once again, (4.14) should be compared with the orbit of $\mathbf{x}_0$ under the full ambient group, $\mathrm{SO}_Q(\mathbb{Z})$. Elementary methods show that

$$\#\{\mathbf{x} \in \mathrm{SO}_Q(\mathbb{Z}) \cdot \mathbf{x}_0 : \|\mathbf{x}\| < N\} \sim c\, N.$$

So in passing from the full orbit to $\mathcal{O}$, the asymptotic drops from N to N^{δ}, with $\delta < 1$. Thus the orbit $\mathcal{O}$ is *thin*.

The fact that ρ is a quadratic map in the entries (see (4.6)) implies that the count (4.14) on triples $\mathbf{x} \in \mathcal{O}$ is equivalent to the following asymptotic for the pairs $\mathbf{v} \in \tilde{\mathcal{O}}$:

$$\#\{\mathbf{v} \in \tilde{\mathcal{O}} : \|\mathbf{v}\| < N\} \sim c' \cdot N^{2\delta}, \tag{4.15}$$

as $N \to \infty$. Note that the power of N is now 2δ. This can also be seen immediately from (4.1) and (4.2) that

$$\|\mathbf{x}\| = \sqrt{x^2 + y^2 + z^2} = \sqrt{2}z = \sqrt{2}(u^2 + v^2) = \sqrt{2}\|\mathbf{v}\|^2. \tag{4.16}$$

(Geometrically, the cone (4.1) intersects the sphere of radius N at a circle of radius $N/\sqrt{2}$.) Observe that (4.14) looks like the Apollonian asymptotic (3.5), while (4.15) is more similar to Hensley's estimate (2.17) in Zaremba's problem. This is just a consequence of choosing between working in the orthogonal group or its spin cover.

4.3. **Diophantine Problems.**

One can now pose a variety of Diophantine questions about the values of various functions on such thin orbits. Given an orbit $\mathcal{O} = \Gamma \cdot \mathbf{x}_0$ and a function $f : \mathcal{O} \to \mathbb{Z}$, call

$$\mathscr{P} := f(\mathcal{O}) \qquad \subset \qquad \mathbb{Z} \tag{4.17}$$

the set of *represented* numbers. That is, n is represented by the pair $(\mathcal{O}, f)$ if there is some $\gamma \in \Gamma$ so that $n = f(\gamma \cdot \mathbf{x}_0)$. And as before, we say n is *admissible* if $n \in \mathscr{P} (\mathrm{mod}\, q)$ for all q. For example, if f is the

"hypotenuse" function, $f(\mathbf{x}) = z$, one can ask whether $(\mathcal{O}, f)$ represents infinitely many admissible primes. Evidence to the affirmative is illustrated in Figure 10, where a triple is highlighted if its hypotenuse is prime. Unfortunately this problem on thin orbits[6] seems out of reach of current technology.

But for a restricted class $\mathcal{F}$ of functions f, and orbits $\mathcal{O}$ which are "not too thin," recent progress has been made towards the local-global problem in $\mathscr{P}$. Let $\mathcal{F}$ be the set of functions f which are a linear, not on the triples $\mathbf{x}$, but on the corresponding pairs $\mathbf{v}$. For example, it is not particularly well-know that in a Pythagorean triple, the sum of the hypotenuse z and the even side y is always a perfect square. This follows immediately from the parametrization (4.2); in particular, $y + z = (u + v)^2$. So the function

$$f(\mathbf{x}) = \sqrt{y + z} = u + v \tag{4.18}$$

is integer-valued on $\mathcal{O}$ and linear[7] in $\mathbf{v}$.

Another way of saying this is to pass to the corresponding orbit $\tilde{\mathcal{O}} = \tilde{\Gamma} \cdot \mathbf{v}_0$. Any such linear function on $\mathbf{v}$ is of the form

$$f(\mathbf{v}) = \langle \mathbf{w}_0, \mathbf{v} \rangle , \tag{4.19}$$

for some fixed $\mathbf{w}_0 \in \mathbb{Z}^2$. In the example (4.18), take $\mathbf{w}_0 = (1, 1)^t$. Then $\mathcal{F}$ consists of all functions on $\mathcal{O}$ which, pulled back to $\tilde{\mathcal{O}}$, are of the form (4.19).

Theorem P (Bourgain-K. 2010 [BK10])**.** *Fix any such linear $f \in \mathcal{F}$ and Pythagorean triple $\mathbf{x}_0$. There is some $\delta_0 < 1$ (the value $\delta_0 = 0.99995$ suffices) so that if the orbit $\mathcal{O} = \Gamma \cdot \mathbf{x}_0$ is not too thin, meaning the exponent δ of Γ satisfies*

$$\delta > \delta_0, \tag{4.20}$$

then almost every admissible number is represented in $\mathscr{P} = f(\mathcal{O})$.

We are finally in position to relate this Pythagorean problem to the Apollonian and Zaremba's. Indeed, passing to the corresponding orbit $\tilde{\mathcal{O}} = \tilde{\Gamma} \cdot \mathbf{v}_0$ and fixing the function $f(\mathbf{v}) = \langle \mathbf{w}_0, \mathbf{v} \rangle$, we have that n is represented if there is a $\gamma \in \tilde{\Gamma}$ so that

$$n = \langle \mathbf{w}_0, \gamma \cdot \mathbf{v}_0 \rangle . \tag{4.21}$$

[6]For the full orbit of all Pythagorean triples, infinitely many hypotenuses are prime. This follows from (4.2) that $z = u^2 + v^2$ and Fermat's theorem that all primes $\equiv 1 (\mathrm{mod}\, 4)$ are sums of two squares.

[7]Really we want the values of $|u + v|$, which within the positive integers are the union of the values of $u + v$ and $-u - v$. Alternatively, we can assume that $-I \in \tilde{\Gamma}$, as is the case for (4.7).

That is,

$$\mathscr{P} = \left\langle \mathbf{w}_0, \tilde{\Gamma} \cdot \mathbf{v}_0 \right\rangle, \tag{4.22}$$

which is of the same form as (2.29) and (3.28). The condition of admissibility is analyzed again given the generators of $\tilde{\Gamma}$ by strong approximation, Goursat's Lemma, and finite group theory, as in §2.2.

Note that in light of the asymptotic counting formula (4.15), the minimal dimension δ_0 in (4.20) cannot go below 1/2: the numbers in $\mathscr{P}$ up to N (counted *with* multiplicity) have cardinality roughly $N^{2\delta}$, so if δ is less than 1/2, then certainly a local-global principle fails miserably. (Such a phenomenon appeared already in the context of Hensley's conjecture (2.15) in Zaremba's problem.)

5. The Circle Method: Tools and Proofs

We briefly review the previous three sections, unifying the (re)formulations of the problems. The Apollonian, Pythagorean, and Zaremba Theorems will henceforth be referred to as Theorem X, where

$$X = A, P, \text{ or } Z,$$

respectively. Theorem X concerns the set $\mathscr{S}$ of numbers of the form

$$\mathscr{S} = \langle \mathbf{w}_0, \Gamma \cdot \mathbf{v}_0 \rangle . \tag{5.1}$$

Here

$$\mathscr{S} = \begin{cases} \mathscr{K} = \text{ the set of curvatures (3.28)} & \text{if } X = A, \\ \mathscr{P} = \text{ the set of square-roots of sums of hypotenuses and even sides (4.22), (4.18)} & \text{if } X = P, \\ \mathscr{D}_{\mathcal{A}} = \text{ the set of denominators (2.29)} & \text{if } X = Z, \end{cases}$$

$$\Gamma = \begin{cases} \text{the Apollonian group } \Gamma & \text{if } X = A, \\ \text{an infinite index subgroup } \tilde{\Gamma} < \Lambda(2) & \text{if } X = P, \\ \text{the semigroup } \Gamma_{\mathcal{A}} & \text{if } X = Z, \end{cases}$$

$$\mathbf{v}_0 = \begin{cases} \text{the root quadruple} & \text{if } X = A, \\ \text{any coprime pair of opposite parity} & \text{if } X = P, \\ (0,1)^t & \text{if } X = Z, \end{cases}$$

and

$$\mathbf{w}_0 = \begin{cases} \text{a standard basis vector } \mathbf{e}_j & \text{if } X = A, \\ \text{any fixed pair} & \text{if } X = P, \\ (0,1)^t & \text{if } X = Z. \end{cases}$$

But now we can forget the individual problems and just focus on the general setting (5.1); one need not keep the above taxonomy in one's head throughout.

To study the local-global problem for $\mathscr{S}$, we introduce the representation function

$$\mathcal{R}_N(n) := \sum_{\gamma \in \Omega_N} \mathbf{1}_{\{n = \langle \mathbf{w}_0, \gamma \cdot \mathbf{v}_0 \rangle\}}. \tag{5.2}$$

Here N is a growing parameter, and Ω_N is a certain subset of the radius N ball in Γ,

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

which we will describe in more detail later. For now, one can just think of Ω_N as the whole radius N ball. To get our bearings, let us recall

roughly[8] the size of Ω_N:

$$\#\{\gamma \in \Gamma : \|\gamma\| < N\} \asymp \begin{cases} N^{\delta}, & \text{if } X = A, \text{ see (3.5)} \\ N^{2\delta}, & \text{if } X = P, \text{ see (4.15)} \\ N^{2\delta_{\mathcal{A}}}, & \text{if } X = Z, \text{ see (2.17).} \end{cases}$$

We can write this uniformly by introducing the parameter α, defined by

$$\alpha := \begin{cases} \delta, \text{ the dimension of an Apollonian packing} & \text{if } X = A, \text{ see (3.4)} \\ 2\delta, \text{ where } \delta \text{ is the dimension of } \mathscr{C}(\tilde{\Gamma}) & \text{if } X = P, \text{ see (4.13)} \\ 2\delta_{\mathcal{A}}, \text{ where } \delta_{\mathcal{A}} \text{ is the dimension of } \mathscr{C}_{\mathcal{A}} & \text{if } X = Z, \text{ see (2.10).} \end{cases}$$

In each case α satisfies

$$1 < \alpha < 2. \tag{5.3}$$

Then the cardinality of such a ball Ω_N is roughly

$$|\Omega_N| \asymp N^{\alpha}. \tag{5.4}$$

Returning to (5.2), we see by construction that $\mathcal{R}_N$ is nonnegative. Moreover observe that

$$\text{if } \mathcal{R}_N(n) > 0, \text{ then certainly } n \text{ is represented in } \mathscr{S}. \tag{5.5}$$

Also record that

$$\mathcal{R}_N \text{ is supported on } n \text{ of size } |n| \ll N. \tag{5.6}$$

Recalling the notation $e(x) = e^{2\pi i x}$, the Fourier transform

$$\begin{aligned} \mathcal{S}_N(\theta) &:= \widehat{\mathcal{R}_N}(\theta) = \sum_{n \in \mathbb{Z}} \mathcal{R}_N(n) e(n\theta) \\ &= \sum_{\gamma \in \Omega_N} e(\theta \langle \mathbf{w}_0, \gamma \cdot \mathbf{v}_0 \rangle) \end{aligned} \tag{5.7}$$

is a wildly oscillating exponential sum on the circle $\mathbb{R}/\mathbb{Z} = [0, 1)$, whose graph looks something like Figure 12. One recovers $\mathcal{R}_N$ through elementary Fourier inversion,

$$\mathcal{R}_N(n) = \int_{\mathbb{R}/\mathbb{Z}} \mathcal{S}_N(\theta) e(-n\theta) d\theta, \tag{5.8}$$

but without further ingredients, one is going around in circles (no pun intended).

Hardy and Littlewood had the idea that the bulk of the integral (5.8) could be captured just by integrating over frequencies θ that are very close to rational numbers a/q, $(a, q) = 1$, with very small denominators

[8]Technically the quoted results are about counting in the corresponding orbits $\mathcal{O}$ and not in the groups Γ; but the order of magnitude is the same for both.

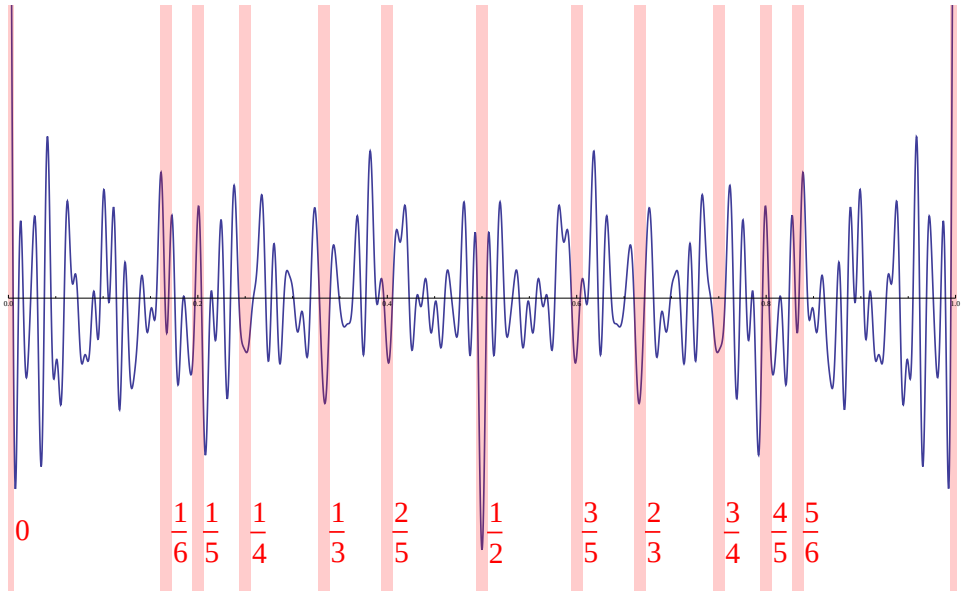


FIGURE 12. The real part of an exponential sum of the form (5.7)

q; some of these intervals are shaded in Figure 12. These are now called the *major arcs* $\mathfrak{M}$; the name refers not to their total length (they comprise a tiny fraction of the circle $\mathbb{R}/\mathbb{Z}$) but to the fact that they are supposed to account for a preponderance of $\mathcal{R}_N(n)$. Accordingly, we decompose (5.8) as

$$\mathcal{R}_N(n) = \mathcal{M}_N(n) + \mathcal{E}_N(n),$$

where the major arc contribution

$$\mathcal{M}_N(n) := \int_{\mathfrak{M}} \mathcal{S}_N(\theta) e(-n\theta) d\theta \tag{5.9}$$

is supposed to give the "main" term, and

$$\mathcal{E}_N(n) := \int_{\mathfrak{m}} \mathcal{S}_N(\theta) e(-n\theta) d\theta \tag{5.10}$$

should be the "error". Here $\mathfrak{m} := [0,1) \setminus \mathfrak{M}$ are the complemenary so-called *minor arcs*. If $\mathcal{M}_N(n)$ is positive and bigger than $|\mathcal{E}_N(n)|$, then certainly

$$\mathcal{R}_N(n) \geq \mathcal{M}_N(n) - |\mathcal{E}_N(n)| > 0, \tag{5.11}$$

so again, n is represented. In practice, one typically tries to prove an asymptotic formula (or at least a lower bound) for $\mathcal{M}_N$, and then give an upper bound for $|\mathcal{E}_N|$.

The reason for this decomposition is that exponential sums such as $\mathcal{S}_N$ should be mostly supported on $\mathfrak{M}$, having their biggest peaks and valleys at (or very near) these frequencies (some of this phenomenon is visible in Figure 12). Indeed, the value $\theta = 0$ is as big as $\mathcal{S}_N$ will ever get,

$$|\mathcal{S}_N(\theta)| \leq \mathcal{S}_N(0) = |\Omega_N|, \tag{5.12}$$

which follows trivially (and is thus called the *trivial bound*) from the triangle inequality: every summand in (5.7) is a complex number of absolute value 1. Also for other $\theta \in \mathfrak{M}$, $\theta \approx a/q$, the summands should all point in a limited number of directions, colluding to give a large contribution to $\mathcal{S}_N$. As we will see later, at these frequencies, one is in a sense measuring the distribution of $\mathscr{S}$ (or equivalently Ω_N) along certain arithmetic progressions. This strategy of coaxing out the (conjectural) main term for $\mathcal{R}_N$ works in surprisingly great generality, but can also give false predictions (even for the Prime Number Theorem, see e.g. [Gra95]).

Having made this decomposition, we should determine what we expect for the main term. From (5.7), we have that

$$\sum_n \mathcal{R}_N(n) = \mathcal{S}_N(0) = |\Omega_N|,$$

so recalling the support (5.2) of $\mathcal{R}_N$, one might expect that an admissible number of size about $n \asymp N$ is represented roughly $|\Omega_N|/N$ times. In particular, since every admissible number is expected to be represented, one would like to show, say, for $N/2 \leq n < N$, that

$$\mathcal{M}_N(n) \gg \mathfrak{S}(n)\frac{|\Omega_N|}{N}. \tag{5.13}$$

Here $\mathfrak{S}(n) \geq 0$ is a certain product of local densities called the *singular series*; it alone is responsible for the notion of admissibility, vanishing on non-admissible n. For admissible n, it typically does not fluctuate too much; crudely one can show in many contexts the lower bound $\gg N^{-\varepsilon}$ for any $\varepsilon > 0$. For ease of exposition, let us just pretend for now that every n is admissible and remove the role of the singular series, allowing ourselves to assume that

$$\mathfrak{S}(n) = 1. \tag{5.14}$$

Observe also that, in light of the cardinality (5.4) of $|\Omega_N|$ and with exponent α ranging in (5.3), the lower bound in (5.13) is of the order $N^{\alpha-1}$, with $\alpha > 1$. That is, there should be quite a lot of representations of an admissible $n \asymp N$ large, giving further indication that every

sufficiently large admissible number may be represented.

One is then left with the problem of estimating away the remainder term $\mathcal{E}_N$, and this is why (as Peter Sarnak likes to say) the circle method is a "method" and not a "theorem": establishing such estimates is much more of an art than a science. The Hardy-Littlewood procedure suggests somehow exploiting the fact that on the minor arc frequencies, $\theta \in \mathfrak{m}$, the exponential sum $\mathcal{S}_N$ in (5.7) should itself already be quite small, being a sum of canceling phases. If one could indeed prove at the level of individual n an upper bound for the error term $\mathcal{E}_N$ which is asymptotically smaller than the lower bound (5.13) for $\mathcal{M}_N$, then one would immediately conclude the full local-global conjecture that every sufficiently large admissible n is represented. Unfortunately, at present we do not know how to give such strong upper bounds on the minor arcs.

Instead, we settle for an "almost" local-global statement, by proving a sharp bound not for individual n, but for n in an average sense, as follows. Parseval's theorem states that the L^2 norm of a function is equal to that of its Fourier transform, that is, the Fourier transform is a unitary operator on these Hilbert spaces. Using the definition (5.10), Parseval's theorem then gives

$$\sum_n |\mathcal{E}_N(n)|^2 = \int_{\mathfrak{m}} |\mathcal{S}_N(\theta)|^2 d\theta. \tag{5.15}$$

Inserting our trivial bound (5.12) for $\mathcal{S}_N$ into the above yields a trivial bound for (5.15) of

$$\int_{\mathfrak{m}} |\mathcal{S}_N(\theta)|^2 d\theta \leq |\Omega_N|^2. \tag{5.16}$$

We claim that it suffices for our applications to establish a bound of the form

$$\int_{\mathfrak{m}} |\mathcal{S}_N(\theta)|^2 d\theta = o\left(\frac{|\Omega_N|^2}{N}\right). \tag{5.17}$$

That is, the above saves a little more than $\sqrt{N}$ on average over $\mathfrak{m}$ off of each term $\mathcal{S}_N$ relative to the trivial bound (5.16). We first explain why this suffices.

5.1. **Proof of Theorem X, Assuming (5.13) and (5.17).**

Let $\mathfrak{E}(N)$ be the set of exceptional n (those that are admissible but not represented) in the range $N/2 \leq n < N$. Recalling the sufficient condition (5.11) for representation, the number of exceptions is

bounded by

$$\#\mathfrak{E}(N) \leq \sum_{\substack{N/2<|n|<N \\ n \text{ is admissible}}} \mathbf{1}_{\{|\mathcal{E}_N(n)|\geq \mathcal{M}_N(n)\}}.$$

For admissible n, we have the supposed major arc lower bound (5.13) and recall our simplifying assumption (5.14) to ignore the singular series; thus

$$\#\mathfrak{E}(N) \leq \sum_n \mathbf{1}_{\{|\mathcal{E}_N(n)|\gg |\Omega_N|/N\}}. \tag{5.18}$$

Here is a pleasant (standard) trick: for those n contributing a 1 rather than 0 to (5.18), we have

$$1 \ll \frac{|\mathcal{E}_N(n)|}{|\Omega_N|/N},$$

both sides of which may be squared. Hence (5.18) implies that

$$\#\mathfrak{E}(N) \ll \frac{N^2}{|\Omega_N|^2} \cdot \sum_n |\mathcal{E}_N(n)|^2.$$

Now we apply Parseval (5.15) and the supposed minor arcs bound (5.17). This gives

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

and thus 100% of the admissible numbers in the range $[N/2, N)$ are represented. Combining such dyadic intervals, we conclude that almost every admissible number is represented.

Now "all" that is left is to establish the major arcs bound (5.13) and the error bound (5.17). In the next two subsections, we focus individually on the tools needed to prove these claims.

5.2. **The Major Arcs.**

Recall that $\mathcal{M}_N$ in (5.9) is an integral over the major arcs $\theta \in \mathfrak{M}$; here θ is very close to a fraction a/q, with q "small" (the meaning of which is explained below). Also let us pretend for now that Ω_N is just the whole Γ-ball,

$$\Omega_N = \{\gamma \in \Gamma : \|\gamma\| < N\}. \tag{5.19}$$

We begin by trying to evaluate (5.7) at $\theta = a/q$:

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

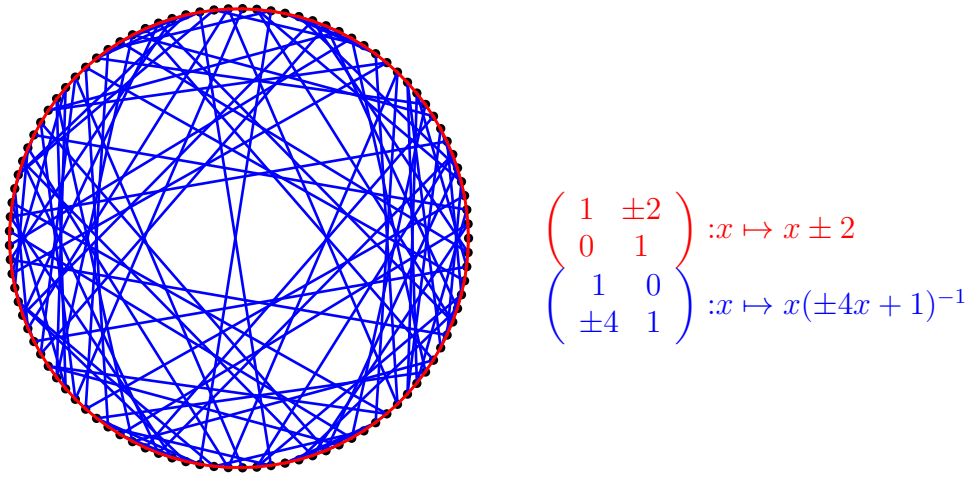


FIGURE 13. An expander; shown with $q = 101$

An important observation in the above is that the summation may be grouped according to the residue class mod q of the integer $\langle \mathbf{w}_0, \gamma \cdot \mathbf{v}_0 \rangle$. Or what is essentially the same, we can decompose the sum according to the the residue class of $\gamma (\mathrm{mod}\, q)$. To this end, let $\Gamma_q = \Gamma (\mathrm{mod}\, q)$ be the set of such residue classes (which we have already studied in the context of admissibility and strong approximation). Then we split the sum as

$$\mathcal{S}_N \left(\frac{a}{q} \right) = |\Omega_N| \sum_{\gamma_0 \in \Gamma_q} e \left(\frac{a}{q} \langle \mathbf{v}_0 \cdot \gamma_0, \mathbf{w}_0 \rangle \right) \cdot \left[\frac{1}{|\Omega_N|} \sum_{\substack{\gamma \in \Gamma \\ \|\gamma\| < N}} \mathbf{1}_{\{\gamma \equiv \gamma_0 (\mathrm{mod}\, q)\}} \right], \tag{5.20}$$

where we have artificially multiplied and divided by the cardinality of Ω_N. Now for γ_0 fixed, the bracketed term is then measuring the "probability" that $\gamma \equiv \gamma_0 (\mathrm{mod}\, q)$. As one may suspect, our groups do not have particular preferences for certain residue classes over others; that is, this probability becomes equidistributed as N grows, with q also allowed to grow, but at a much slower rate. (In fact, this is exactly what we mean by the denominator q being "small" – relative to N – in the major arcs $\mathfrak{M}$.) To explain how this happens, we briefly discuss the crucial notion of an *expander*.

Rather than going into the general theory (for which we refer the reader to the beautiful survey [Lub12]; see also [Sar04]), we content ourselves with but one illustrative example of expansion. Figure 13 shows the following graph. For $q = 101$, say, take the vertices to be the elements of $\mathbb{Z}/q\mathbb{Z}$, organized around the unit circle by placing $x \in \mathbb{Z}/q\mathbb{Z}$

at $e(x/q)$. For the edges, connect each

$$x \text{ to } x \pm 2, \qquad \text{and also} \qquad \text{to } x(\pm 4x+1)^{-1}, \tag{5.21}$$

when inversion $(\mathrm{mod}\, q)$ is possible. This is nothing more than the fractional linear action (see (4.12)) of the generating matrices in (4.7) (and their inverses) on $\mathbb{Z}/q\mathbb{Z}$. We first claim that our graph on q vertices is "sparse". Indeed, the complete graph (connecting any vertex to any other) has on the order of q^2 edges, whereas our graph has only on the order of q edges (since (5.21) implies that any vertex is connected to at most four others). So we have square-root the total number of possible edges, and our graph is indeed quite sparse.

Despite having few edges, it is a fact that this graph is nevertheless highly connected, in the sense that a random walk on it is rapidly mixing. Moreover, this rate of mixing, properly normalized, is independent of the choice of q above. That is, by varying q, we in fact have a whole family of such sparse but highly connected graphs, and with a uniform mixing rate; this is exactly what characterizes an expander.

Proofs of expansion use, among other things, tools from additive combinatorics, in particular, the so-called sum-product [BKT04, Bou08] and triple-product [Hel08, BGT11, PS10] estimates, and quite a lot of other work which we will not survey; see e.g. [SX91, Gam02, BG08, BGS10, Var10, BV11, SGV11]. Once one proves uniform expansion for such finite graphs, the statements must be converted into the archimedean form needed for the bracketed term in (5.20). To handle such counting statements, one uses

$$\begin{cases} \text{infinite volume spectral and representation theory} \\ \text{à la §3.1.6, specifically Vinogradov's thesis [Vin12],} & \text{if } X = A, \\ \text{similar techniques developed by} \\ \text{Bourgain-K.-Sarnak [BKS10],} & \text{if } X = P, \\ \text{the thermodynamic formalism, analytically continuing} \\ \text{certain Ruelle transfer operators [Lal89, Dol98, Nau05]} \\ \text{and their "congruence" extensions; see [BGS11],} & \text{if } X = Z. \end{cases}$$

Without going into details, the upshot is that, up to acceptable errors, the bracketed term in (5.20) is just $1/|\Gamma_q|$, confirming the desired equidistribution. Inserting this estimation into $\mathcal{M}_N$ in (5.9), one uses these techniques and some more standard circle method analysis to eventually conclude (5.13).

5.3. **The Minor Arcs.**

We use different strategies to prove the minor arcs bound (5.17) for the Pythagorean and Zaremba settings $X = P$ or Z, versus the Apollonian setting $X = A$, so we present them individually. As the details quickly become quite technical, we will only scratch the surface, inviting the interested reader to study the original manuscripts [BK10, BK11, BK12]. Hopefully the short sketches below give some indication for the flavor of the arguments involved.

5.3.1. *Pythagorean and Zaremba settings.*

To handle the minor arcs here, we make the observation that the ensemble Ω_N in the definition of $\mathcal{S}_N$ from (5.7) need not be an exact Γ-ball as in (5.19), but can be replaced by, say, a product of two such. That is, the definition of $\mathcal{S}_N$ can be changed to

$$\mathcal{S}_N(\theta) := \sum_{\substack{\gamma_1\in\Gamma\\ \|\gamma_1\|<\sqrt{N}}} \sum_{\substack{\gamma_2\in\Gamma\\ \|\gamma_2\|<\sqrt{N}}} e(\theta\,\langle \mathbf{v}_0\gamma_1\gamma_2, \mathbf{w}_0\rangle), \tag{5.22}$$

without irreparably damaging the major arcs analysis. This new sum encodes much more of the (semi)group structure of Γ, while preserving the "non-vanishing implies represented" property (5.5), where $\mathcal{R}_N$ is redefined by Fourier inversion (5.8). (In reality, we use even more complicated exponential sums.) The advantage of (5.22) is that we can now exploit this structure à la Vinogradov's method [Vin37] for estimating *bilinear forms*: one can think of (5.22) as the sum of all entries in the matrix indexed by γ_1 and γ_2 with entries $e(\theta\,\langle \mathbf{v}_0\gamma_1\gamma_2, \omega_0\rangle)$. Just one standard maneuver in estimating bilinear forms is the following.

Apply the Cauchy-Schwarz inequality to (5.22) in the γ_1 variable:

$$|\mathcal{S}_N(\theta)| \le \left(\sum_{\substack{\gamma_1\in\Gamma\\ \|\gamma_1\|<\sqrt{N}}} 1\right)^{1/2} \left(\sum_{\substack{\gamma_1\in \mathrm{SL}_2(\mathbb{Z})\\ \|\gamma_1\|<\sqrt{N}}} \left|\sum_{\substack{\gamma_2\in\Gamma\\ \|\gamma_2\|<\sqrt{N}}} e(\theta\,\langle \mathbf{v}_0\gamma_1\gamma_2, \mathbf{w}_0\rangle)\right|^2\right)^{1/2}.$$

Notice in the second appearance of a γ_1 sum, we have replaced the thin and mysterious group Γ (or semigroup Γ_A) by the full ambient group $\mathrm{SL}_2(\mathbb{Z})$. On one hand, this allows us to now use more classical tools to get the requisite cancellation (5.17) in the minor arcs integral. On the other hand, this type of perturbation argument only succeeds when δ is near 1, explaining the dimension restrictions (2.23) and (4.20).

5.3.2. *The Apollonian case.*

The above strategy fails for the Apollonian problem, because the Hausdorff dimension (3.4) is a fixed invariant which refuses to be adjusted to suit our needs. Instead, we recall that the Apollonian group Γ contains the special (arithmetic) subgroup Ξ from (3.36). Then, like (5.22), we change the definition of the exponential sum to something of the (again, bilinear) form

$$\mathcal{S}_N(\theta) := \sum_{\substack{\xi\in\Xi\\ \|\xi\|<X}} \sum_{\substack{\gamma\in\Gamma\\ \|\gamma\|<T}} e(\theta\,\langle \mathbf{v}_0\cdot\xi\,\gamma, \mathbf{w}_0\rangle), \tag{5.23}$$

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 and admissibility conditions. For the minor arcs bound, one keeps γ fixed and uses the classical arithmetic group Ξ to get sufficient cancellation to prove the desired bound (5.17). Again, we spare the reader all details.

5.4. **Conclusion.**

Putting together the above-sketched minor arcs upper bound (5.17) with the major arcs lower bound (5.13) discussed in §5.2, we prove the main Theorem X, as explained in §5.1. We end by emphasizing again that, though the problems have nearly identical reformulations, the circle method is only a method and not an applicable theorem: while the idea of breaking the integral (5.8) into major and minor arcs is ubiquitous, the actual execution of this idea is handled by very different tools in each case; see the table below. Besides the circle method, the only other pervasive and critical ingredients are expanders for the major arcs, and bilinear forms for the minor arcs.

Theorem	Tools for Major Arcs	Ingredients for Minor Arcs
A	infinite volume hyperbolic 3-folds, automorphic forms, representations, expanders	that Γ contains the arithmetic subgroup $\Xi \cong \Lambda(2)$, bilinear forms
P	infinite volume hyperbolic 2-folds, automorphic forms, representations, expanders	replacing Γ by $\mathrm{SL}_2(\mathbb{Z})$ for δ near 1, bilinear forms
Z	thermodynamic formalism, congruence transfer operators, expanders	replacing $\Gamma_{\mathcal{A}}$ by $\mathrm{SL}_2(\mathbb{Z})$ for $\delta_{\mathcal{A}}$ near 1, bilinear forms

References

[BF11] Jean Bourgain and Elena Fuchs. A proof of the positive density conjecture for integer Apollonian circle packings. *J. Amer. Math. Soc.*, 24(4):945–967, 2011. 18, 29

[BG08] Jean Bourgain and Alex Gamburd. Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$. *Ann. of Math. (2)*, 167(2):625–642, 2008. 47

[BGS06] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Sieving and expanders. *C. R. Math. Acad. Sci. Paris*, 343(3):155–159, 2006. 35

[BGS10] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Affine linear sieve, expanders, and sum-product. *Invent. Math.*, 179(3):559–644, 2010. 18, 35, 47

[BGS11] J. Bourgain, A. Gamburd, and P. Sarnak. Generalization of Selberg's 3/16th theorem and affine sieve. *Acta Math*, 207:255–290, 2011. 47

[BGT11] Emmanuel Breuillard, Ben Green, and Terence Tao. Approximate subgroups of linear groups. *Geom. Funct. Anal.*, 21(4):774–819, 2011. 47

[BK10] J. Bourgain and A. Kontorovich. On representations of integers in thin subgroups of $\mathrm{SL}(2, \mathbf{Z})$. *GAFA*, 20(5):1144–1174, 2010. 38, 48

[BK11] J. Bourgain and A. Kontorovich. On Zaremba's conjecture, 2011. Preprint, `arXiv:1107.3776`. 14, 15, 48

[BK12] J. Bourgain and A. Kontorovich. On the local-global conjecture for integral Apollonian gaskets, 2012. Preprint, `arXiv:1205.4416v1`. 18, 28, 48

[BKS10] J. Bourgain, A. Kontorovich, and P. Sarnak. Sector estimates for hyperbolic isometries. *GAFA*, 20(5):1175–1200, 2010. 47

[BKT04] J. Bourgain, N. Katz, and T. Tao. A sum-product estimate in finite fields, and applications. *Geom. Funct. Anal.*, 14(1):27–57, 2004. 47

[Bou08] J. Bourgain. The sum-product theorem in $\mathbf{z}_q$ with q arbitrary. *J. Analyse Math.*, 106:1–93, 2008. 47

[Boy73] David W. Boyd. The residual set dimension of the Apollonian packing. *Mathematika*, 20:170–174, 1973. 19

[Boy82] David W. Boyd. The sequence of radii of the Apollonian packing. *Math. Comp.*, 39(159):249–254, 1982. 19

[BV11] J. Bourgain and P. Varju. Expansion in $SL_n(\mathbf{Z}/q\mathbf{Z})$, q arbitrary, 2011. To appear, *Invent. Math.* `arXiv:1006.3365v1`. 47

[Cas78] J. W. S. Cassels. *Rational Quadratic Forms*. Number 13 in London Mathematical Society Monographs. Academic Press, London-New York-San Francisco, 1978. 29

[Cox68] H. S. M. Coxeter. The problem of Apollonius. *Amer. Math. Monthly*, 75:5–15, 1968. 22

[Des01] Rene Descartes. *Œuvres*, volume 4. Paris, 1901. C. Adams and P. Tannery, eds. 22

[Dol98] Dmitry Dolgopyat. On decay of correlations in Anosov flows. *Ann. of Math. (2)*, 147(2):357–390, 1998. 47

[DRS93] W. Duke, Z. Rudnick, and P. Sarnak. Density of integer points on affine homogeneous varieties. *Duke Math. J.*, 71(1):143–179, 1993. 25

[DSV03] G. Davidoff, P. Sarnak, and A. Valette. *Elementary Number Theory, Group Theory and Ramanujan Graphs*, volume 55 of *London Math. Soc., Student Text*. Cambridge University Press, 2003. 17

[EL07] Nicholas Eriksson and Jeffrey C. Lagarias. Apollonian circle packings: number theory. II. Spherical and hyperbolic packings. *Ramanujan J.*, 14(3):437–469, 2007. 18

[EM93] A. Eskin and C. McMullen. Mixing, counting and equidistribution in lie groups. *Duke Math. J.*, 71:143–180, 1993. 25

[FS11] Elena Fuchs and Katherine Sanden. Some experiments with integral Apollonian circle packings. *Exp. Math.*, 20(4):380–399, 2011. 18, 28

[Fuc11] Elena Fuchs. Strong approximation in the Apollonian group. *J. Number Theory*, 131(12):2282–2302, 2011. 18, 28, 30, 31

[Gam02] Alex Gamburd. On the spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbf{Z})$. *Israel J. Math.*, 127:157–200, 2002. 47

[GLM+03] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: number theory. *J. Number Theory*, 100(1):1–45, 2003. 18, 20, 28, 29, 32

[GLM+05] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: geometry and group theory. I. The Apollonian group. *Discrete Comput. Geom.*, 34(4):547–585, 2005. 18, 30

[GLM+06a] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: geometry and group theory. II. Super-Apollonian group and integral packings. *Discrete Comput. Geom.*, 35(1):1–36, 2006. 18

[GLM+06b] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: geometry and group theory. III. Higher dimensions. *Discrete Comput. Geom.*, 35(1):37–72, 2006. 18

[Gra95] Andrew Granville. Harald Cramér and the distribution of prime numbers. *Scand. Actuar. J.*, (1):12–28, 1995. Harald Cramér Symposium (Stockholm, 1993). 43

[Hel08] H. A. Helfgott. Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. *Ann. of Math. (2)*, 167(2):601–623, 2008. 47

[Hen89] Doug Hensley. The distribution of badly approximable numbers and continuants with bounded digits. In *Théorie des nombres (Quebec, PQ, 1987)*, pages 371–385. de Gruyter, Berlin, 1989. 13

[Hen92] Doug Hensley. Continued fraction Cantor sets, Hausdorff dimension, and functional analysis. *J. Number Theory*, 40(3):336–358, 1992. 11

[Hen96] Douglas Hensley. A polynomial time algorithm for the Hausdorff dimension of continued fraction Cantor sets. *J. Number Theory*, 58(1):9–45, 1996. 9, 11

[JP01] Oliver Jenkinson and Mark Pollicott. Computing the dimension of dynamically defined sets: E_2 and bounded continued fractions. *Ergodic Theory Dynam. Systems*, 21(5):1429–1445, 2001. 11

[KO11] A. Kontorovich and H. Oh. Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds. *Journal of the American Mathematical Society*, 24(3):603–648, 2011. 18, 20, 27

[KO12] A. Kontorovich and H. Oh. Almost prime Pythagorean triples in thin orbits. *J. reine angew. Math.*, 667:89–131, 2012. `arXiv:1001.0370`. 37

[Kon07] A. V. Kontorovich. *The Hyperbolic Lattice Point Count in Infinite Volume with Applications to Sieves*. Columbia University Thesis, 2007. 37

[Kon09] A. Kontorovich. The hyperbolic lattice point count in infinite volume with applications to sieves. *Duke J. Math.*, 149(1):1–36, 2009. `arXiv:0712.1391`. 37

[Lal89] Steven P. Lalley. Renewal theorems in symbolic dynamics, with applications to geodesic flows, non-Euclidean tessellations and their fractal limits. *Acta Math.*, 163(1-2):1–55, 1989. 47

[Lan08] E. Landau. Über die Einteilung der positiven ganzen Zahlen in vier Klassen nach der Mindestzahl der zu ihrer additiven Zusammensetzung erforderlichen Quadrate. *Arch. der Math. u. Phys.*, 13(3):305–312, 1908. 32

[LMW02] Jeffrey C. Lagarias, Colin L. Mallows, and Allan R. Wilks. Beyond the Descartes circle theorem. *Amer. Math. Monthly*, 109(4):338–361, 2002. 18

[LO12] M. Lee and H. Oh. Effective circle count for Apollonian packings and closed horospheres, 2012. Preprint, `arXiv:1202.1067`. 18, 20, 27

[LP82] P.D. Lax and R.S. Phillips. The asymptotic distribution of lattice points in Euclidean and non-Euclidean space. *Journal of Functional Analysis*, 46:280–350, 1982. 27

[Lub12] A. Lubotzky. Expander graphs in pure and applied mathematics. *Bull. Amer. Math. Soc.*, 49:113–162, 2012. 46

[Mar54] J. M. Marstrand. Some fundamental geometrical properties of plane sets of fractional dimensions. *Proc. London Math. Soc.*, 4(3):257–302, 1954. 13

[McM98] Curtis T. McMullen. Hausdorff dimension and conformal dynamics. III. Computation of dimension. *Amer. J. Math.*, 120(4):691–721, 1998. 19

[MVW84] C. Matthews, L. Vaserstein, and B. Weisfeiler. Congruence properties of Zariski-dense subgroups. *Proc. London Math. Soc*, 48:514–532, 1984. 17

[Nau05] Frédéric Naud. Expanding maps on Cantor sets and analytic continuation of zeta functions. *Ann. Sci. École Norm. Sup. (4)*, 38(1):116–153, 2005. 47

[Nie78] Harald Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.*, 84(6):957–1041, 1978. 6, 9

[Nov55] P. S. Novikov. *Ob algoritmičeskoĭ nerazrešimosti problemy tožděstva slov v teorii grupp*. Trudy Mat. Inst. im. Steklov. no. 44. Izdat. Akad. Nauk SSSR, Moscow, 1955. 24

[OEI] `http://oeis.org/A195901`. 9

[Oh10] Hee Oh. Dynamics on geometrically finite hyperbolic manifolds with applications to Apollonian circle packings and beyond. In *Proceedings of the International Congress of Mathematicians. Volume III*, pages 1308–1331, New Delhi, 2010. Hindustan Book Agency. 18

[OS12] Hee Oh and Nimish Shah. The asymptotic distribution of circles in the orbits of Kleinian groups. *Invent. Math.*, 187(1):1–35, 2012. 18

[Pat76] S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*, 136:241–273, 1976. 27

[Pli] http://en.wikipedia.org/wiki/Plimpton_322. 33

[PS10] L. Pyber and E. Szabo. Growth in finite simple groups of lie type of bounded rank, 2010. Preprint arXiv:1005.1858. 47

[Rap12] A. Rapinchuk. On strong approximation for algebraic groups, 2012. Preprint arXiv:1207.4425. 16

[Sar04] P. Sarnak. What is...an expander? *Notices Amer. Math. Soc.*, 51(7):762–763, 2004. 46

[Sar07] P. Sarnak. Letter to J. Lagarias, 2007. http://web.math.princeton.edu/sarnak/AppolonianPackings.pdf. 18, 28, 32

[Sar08] Peter Sarnak. Equidistribution and primes. *Astérisque*, (322):225–240, 2008. Géométrie différentielle, physique mathématique, mathématiques et société. II. 18

[Sar11] Peter Sarnak. Integral Apollonian packings. *Amer. Math. Monthly*, 118(4):291–306, 2011. 18

[Sch72] Wolfgang M. Schmidt. Irregularities of distribution. VII. *Acta Arith.*, 21:45–50, 1972. 7

[SG12] A. Salehi Golsefidy. Affine sieve and expanders, 2012. Preprint. 35

[SGS11] A. Salehi Golsefidy and P. Sarnak. Affine sieve, 2011. Preprint. 35

[SGV11] A. Salehi Golsefidy and P. Varju. Expansion in perfect groups, 2011. Preprint. 47

[Sod36] F. Soddy. The kiss precise. *Nature*, 137:1021, 1936. 23

[Sod37] F. Soddy. The bowl of integers and the hexlet. *Nature*, 139:77–79, 1937. 24

[Sul84] D. Sullivan. Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. *Acta Math.*, 153(3-4):259–277, 1984. 27

[SX91] P. Sarnak and X. Xue. Bounds for multiplicities of automorphic representations. *Duke J. Math.*, 64(1):207–227, 1991. 47

[Var10] P. Varju. Expansion in $SL_d(O_K/I)$, I square-free, 2010. arXiv:1001.3664v1. 47

[Vin37] I. M. Vinogradov. Representation of an odd number as a sum of three primes. *Dokl. Akad. Nauk SSSR*, 15:291–294, 1937. 48

[Vin12] I. Vinogradov. Effective bisector estimate with application to Apollonian circle packings, 2012. Princeton University Thesis, arxiv:1204.5498v1. 18, 20, 27, 47

[Zar66] S. C. Zaremba. Good lattice points, discrepancy, and numerical integration. *Ann. Mat. Pura Appl. (4)*, 73:293–317, 1966. 8

[Zar72] S. K. Zaremba. La méthode des "bons treillis" pour le calcul des intégrales multiples. In *Applications of number theory to numerical analysis (Proc. Sympos., Univ. Montreal, Montreal, Que., 1971)*, pages 39–119. Academic Press, New York, 1972. 9

E-mail address: alex.kontorovich@yale.edu

Department of Mathematics, Yale University, New Haven, CT