

# Gaps in $\sqrt{n}$ mod 1 and ergodic theory

Noam D. Elkies and Curtis T. McMullen

November 17, 2005

## Contents

1	Introduction . . . . .	1
2	Ergodic theory . . . . .	8
2.1	The affine group of the plane . . . . .	9
2.2	Mixing of the Teichmüller flow . . . . .	10
2.3	Uniform distribution of horocycle sections . . . . .	11
2.4	Limits of measures . . . . .	13
2.5	Unipotent invariance . . . . .	14
2.6	Ratner's theorem . . . . .	16
2.7	Nonlinearity . . . . .	19
3	Distribution of gaps . . . . .	22
3.1	Gap-counting functions . . . . .	22
3.2	From gaps to lattice translates in $\mathbb{R}^2$ . . . . .	23
3.3	Consequences of ergodic theory . . . . .	29
3.4	Formulas for the gap distribution . . . . .	33
3.5	Generalizations . . . . .	38
3.6	Open questions . . . . .	45

---

Supported in part by the Packard Foundation (NDE) and the NSF (CTM).

2000 Mathematics Subject Classification: Primary 11J71, 22E40; Secondary 37A17, 37A25.

### Abstract

Cut the unit circle  $S^1 = \mathbb{R}/\mathbb{Z}$  at the points  $\{\sqrt{1}\}, \{\sqrt{2}\}, \dots, \{\sqrt{N}\}$ , where  $\{x\} = x \bmod 1$ , and let  $J_1, \dots, J_N$  denote the complementary intervals, or *gaps*, that remain. We show that, in contrast to the case of random points (whose gaps are exponentially distributed), the lengths  $|J_i|/N$  are governed by an explicit piecewise real-analytic distribution  $F(t) dt$  with phase transitions at  $t = 1/2$  and  $t = 2$ .

The gap distribution is related to the probability  $p(t)$  that a random unimodular lattice translate  $\Lambda \subset \mathbb{R}^2$  meets a fixed triangle  $S_t$  of area  $t$ ; in fact  $p''(t) = -F(t)$ . The proof uses ergodic theory on the universal elliptic curve

$$E = (\mathrm{SL}_2(\mathbb{R}) \ltimes \mathbb{R}^2) / (\mathrm{SL}_2(\mathbb{Z}) \ltimes \mathbb{Z}^2)$$

and Ratner's theorem on unipotent invariant measures.

# 1 Introduction

For any real number  $x$ , let

$$\{x\} = x \bmod 1 \in S^1 = \mathbb{R}/\mathbb{Z}$$

denote the fractional part of  $x$ . In this paper we determine the distribution of *gaps* in  $\{\sqrt{n}\}$ , the sequence of fractional parts of square-roots of whole numbers  $n > 0$ .

The theory of distribution mod 1 has a long history. Kronecker proved that for any irrational number  $\theta$ , the fractional parts  $\{n\theta\}$  are dense in  $S^1$ . Weyl proved the same sequence is *uniformly distributed*, meaning

$$\frac{\#\{0 < n \leq N : \{n\theta\} \in I\}}{N} \rightarrow \frac{|I|}{|S^1|}$$

as  $N \rightarrow \infty$ , for any interval  $I \subset S^1$ . Many other sequences have been studied; for example,  $\{\theta^n\}$  is known to be uniformly distributed on  $S^1$  for almost every  $\theta > 1$ , while the distribution of specific sequences such as  $\{(3/2)^n\}$  is an open problem.

Now consider the sequence  $\{n^\alpha\}$  for  $0 < \alpha < 1$ . It is easy to see  $\{n^\alpha\}$  is uniformly distributed on  $S^1$ , using the fact that  $(n+1)^\alpha - n^\alpha \rightarrow 0$ .

To explore the distribution of  $\{n^\alpha\}$  in more detail, we study the lengths of the complementary intervals or *gaps*  $\mathcal{J}(N) = \{J_1, \dots, J_N\}$  left over when the circle is cut at the points  $\{1^\alpha\}, \{2^\alpha\}, \dots, \{N^\alpha\}$ . The average gap length,  $(1/N) \sum |J_i|$ , is clearly  $1/N$ , so it is natural to study the ratio of the gap lengths to  $1/N$ .

The gap distribution provides a test of the ‘randomness’ of the points  $\{n^\alpha\}$ . When the circle is cut at a sequence of random points, the resulting gaps are exponentially distributed: that is, we have

$$\frac{\#\{J \in \mathcal{J}(N) : |J| \in [a/N, b/N]\}}{N} \rightarrow \int_a^b e^{-t} dt \quad (1.1)$$

almost surely as  $N \rightarrow \infty$  (compare [Fe, p. 158]). Experiments suggest that for most values of  $\alpha$ , the gaps for  $\{n^\alpha\}$  are also exponentially distributed; in fact, (1.1) appears to hold for all values of  $\alpha \neq 1/2$ .

The gap distribution for  $\{\sqrt{n}\}$  is radically different. Our main result is the following.

**Theorem 1.1** *The gap distribution for the sequence  $\{\sqrt{n}\}$  is given by a continuous function*

$$F(t) = \begin{cases} 6/\pi^2 & t \in [0, 1/2], \\ F_2(t) & t \in [1/2, 2], \text{ and} \\ F_3(t) & t \in [2, \infty), \end{cases}$$

where  $F_2(t)$  and  $F_3(t)$  are explicit real-analytic functions. That is, for any interval  $[a, b] \subset [0, \infty)$  we have

$$\frac{\#\{J \in \mathcal{J}(N) : |J| \in [a/N, b/N]\}}{N} \rightarrow \int_a^b F(t) dt$$

as  $N \rightarrow \infty$ .

Explicit formulas for  $F_2$  and  $F_3$  are given in equations (3.54) and (3.56) below. Figure 1 compares the experimental gap distribution at a finite value of  $N$  with the limiting distribution  $F(t)$ .

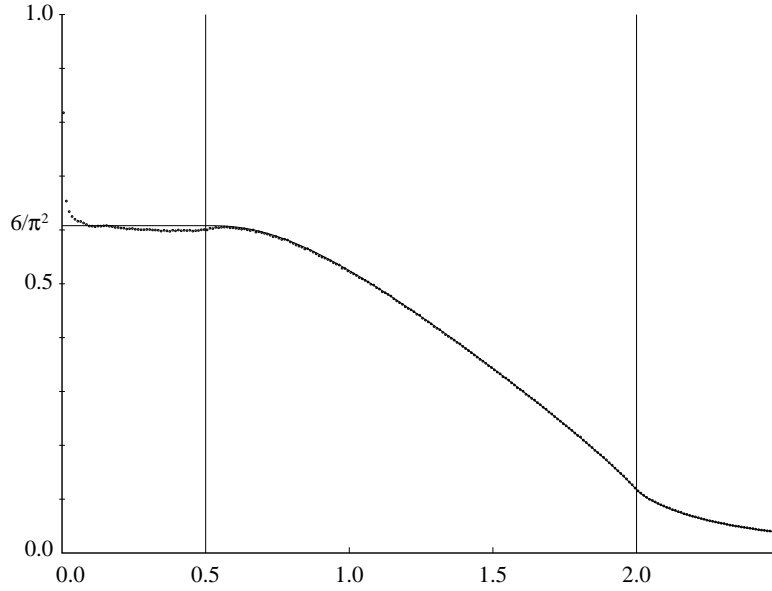


Figure 1. Gaps in  $\{\sqrt{n}\}_1^N$ ,  $N = 2.5 \times 10^7$ , together with the graph of the limiting gap distribution  $y = F(t)$ .

We emphasize that the function  $F(t)$  is *not* analytic or even  $C^3$  at the points  $t = 1/2$  and  $t = 2$ . The gap distribution has genuine phase transitions at these two critical points. Moreover, the tail of the distribution is not exponential; instead, we have  $F(t) \sim (3/\pi^2)t^{-3}$  as  $t \rightarrow \infty$ . Thus large gaps are much more likely for  $\{\sqrt{n}\}_1^N$  than for  $N$  random points (although both are rare events).

**From gaps to lattices.** The distribution of gaps in  $\{\sqrt{n}\}$  is related, via ergodic theory, to the probability  $p(t)$  that a random lattice translate  $\Lambda \subset \mathbb{R}^2$  meets a given triangle  $S_t$  of area  $t$ .

To explain this relation, we first discuss spaces of lattices, their natural measures and the dynamical systems they support. Recall that a *lattice*  $\Lambda^0 \subset \mathbb{R}^2$  is a discrete subgroup isomorphic to  $\mathbb{Z}^2$ ; it is *unimodular* if the quotient torus  $\mathbb{R}^2/\Lambda^0$  has area one. A lattice *translate* is simply a coset  $\Lambda = v + \Lambda^0 \subset \mathbb{R}^2$ .

The space of all translates of unimodular lattices in  $\mathbb{R}^2$  can be naturally identified with the homogeneous space

$$E = \mathrm{ASL}_2(\mathbb{R})/\mathrm{ASL}_2(\mathbb{Z}). \quad (1.2)$$

Here  $\mathrm{ASL}_2(\mathbb{R})$  is the group of area-preserving affine maps  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  of the form  $g(v) = Av + b$  with  $\det A = 1$ , and  $\mathrm{ASL}_2(\mathbb{Z})$  is the discrete subgroup with  $A \in \mathrm{SL}_2(\mathbb{Z})$  and  $b \in \mathbb{Z}^2$ . (To see the identification, just note that  $\mathrm{ASL}_2(\mathbb{R})$  acts transitively on the set of lattice translates, and  $\mathrm{ASL}_2(\mathbb{Z})$  is the stabilizer of  $\Lambda = \mathbb{Z}^2$ .)

The space  $E$  carries a unique probability measure  $\mu_E$  invariant under the left action of  $\mathrm{ASL}_2(\mathbb{R})$ . Using this measure, it makes sense to talk about a ‘random lattice translate’  $\Lambda \subset \mathbb{R}^2$ .

Now fix  $t > 0$ , and for  $N \gg 0$  consider an interval  $I = [x, x+t/N] \subset \mathbb{R}/\mathbb{Z}$  with  $x \in [0, 1]$  chosen at random (with respect to uniform measure). To determine the gap distribution for  $\sqrt{n} \bmod 1$ , it suffices to estimate the probability  $P_N(t)$  that  $I$  contains  $\{\sqrt{n}\}$  for some integer  $n \in [0, N]$ . On the other hand, we have

$$\begin{aligned} \{\sqrt{n}\} \in I &\iff \sqrt{n} \in I + a, \quad \text{some } a \in \mathbb{Z}, \\ &\iff n \in (I + a)^2. \end{aligned}$$

In §3 we will show that  $(I + a)^2$  can be replaced by the linear approximation

$$\begin{aligned} (I + a)^2 &\approx (a + x)^2 + 2(a + x)(I - x) \\ &= a^2 - x^2 + 2(a + x)I \end{aligned}$$

without changing the asymptotics of the gap distribution. We can also assume that  $N$  is a square. Once this is done, we observe that the condition

$$n \in a^2 - x^2 + 2(a+x)I$$

holds for integers  $a, n$ ,  $n \in [0, N]$ , if and only if

$$(\mathbb{Z} + x^2) \cap 2(a+x)I \neq \emptyset$$

for some  $a$  with  $a+x \in [0, \sqrt{N}]$ ; equivalently, if and only if

$$T \cap \mathbb{Z}^2 \neq \emptyset,$$

where  $T \subset \mathbb{R}^2$  is the triangle of area  $t$  given by

$$T = \{(a, b) : b + x^2 \in 2(a+x)I \text{ and } a+x \in [0, \sqrt{N}]\}. \quad (1.3)$$

See Figure 2.

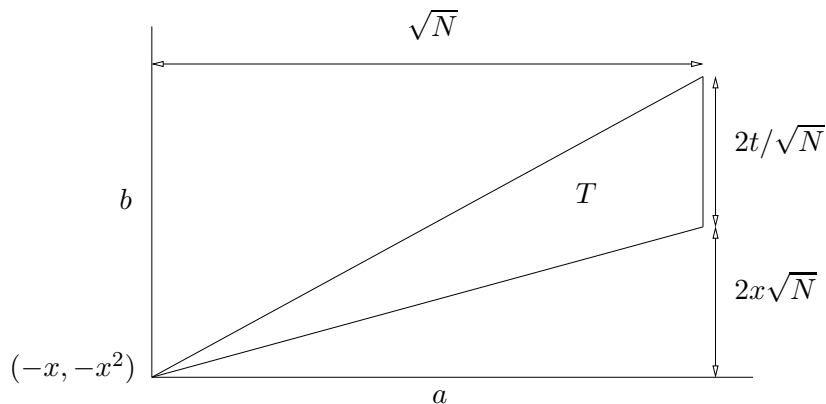


Figure 2. The triangle determined by  $I = [x, x + t/N]$ .

Let  $S_t$  be a standard triangle of area  $t$  with vertices  $(0, 0)$ ,  $(1, 0)$  and  $(0, 2t)$ . Let  $g \in \text{ASL}_2(\mathbb{R})$  be the unique affine map such that  $g(T) = S_t$  and  $g(-x, -x^2) = (0, 0)$ . Summing up the preceding discussion, we find that the probability  $P_N(t)$  that

- $I = [x, x + t/N]$  contains  $\sqrt{n} \bmod 1$  for some  $n \leq N$

is essentially the same as the probability that

- the lattice translate  $\Lambda_N(x) = g(\mathbb{Z}^2)$  meets the standard triangle  $S_t$ .

On the other hand, for  $N \gg 0$  one might expect that  $\Lambda_N(x) \in E$  behaves like a random lattice translate. In fact, in §2 we will use ergodic theory to show:

**Theorem 1.2** *The lattice translates  $\Lambda_N(x)$  are uniformly distributed on  $E$  as  $N \rightarrow \infty$ . That is, for any  $f \in C_0(E)$  we have*

$$\int_0^1 f(\Lambda_N(x)) dx \rightarrow \int_E f(\Lambda) d\mu_E(\Lambda).$$

Here  $C_0(E)$  denotes the space of compactly-supported continuous functions on  $E$ .

Because of this uniform distribution, we find that  $P_N(t)$  converges to  $p(t)$ , the probability that a random  $\Lambda \in E$  meets  $S_t$ . Converting back to the gap distribution, we have:

**Corollary 1.3** *The probability  $p(t)$  that a random unimodular lattice translate  $\Lambda$  meets a given triangle  $S_t$  of area  $t$  satisfies*

$$p''(t) = -F(t),$$

where  $F(t)$  is the gap distribution for  $\{\sqrt{n}\}$ .

The proof of Theorem 1.1 is completed by computing  $p''(t)$ , using explicit formulas for the natural invariant measure  $\mu_E$ . Since  $p(t) = t - \int_0^t \int_0^s F(u) du ds$ , our explicit formula for  $F(t)$  also leads to one for  $p(t)$ .

In summary, we find that the uniform distribution of lattices explains the exotic distribution of gaps.

**Hyperbolic geometry.** We now turn to the ergodic theory side of the argument, to indicate the proof of the uniform distribution of  $\langle \Lambda_N(x) \rangle$  on  $E$ .

We begin with dynamics on the simpler space

$$B = \mathrm{SL}_2(\mathbb{R}) / \mathrm{SL}_2(\mathbb{Z}).$$

The space  $B$  classifies unimodular lattices  $\Lambda^0 \subset \mathbb{R}^2$  and carries a natural invariant probability measure  $\mu_B$ .

Geometrically,  $\mathrm{SL}_2(\mathbb{R}) / (\pm I)$  can be identified with the unit tangent bundle  $T_1(\mathbb{H})$  of the hyperbolic plane  $\mathbb{H}$ . Similarly,  $B$  can be identified with  $T_1(\mathcal{M}_1)$ , the unit tangent bundle to the moduli space

$$\mathcal{M}_1 = \mathbb{H} / \mathrm{SL}_2(\mathbb{Z})$$

of Riemann surfaces of genus 1.

The space  $\mathcal{M}_1$  is a hyperbolic orbifold of finite volume with a unique cusp. Under the identification  $B = T_1(\mathcal{M}_1)$ , the measure  $\mu_B$  agrees with Liouville measure for the hyperbolic metric, which is preserved by the geodesic and horocycle flows

$$g_s, h_s : T_1(\mathcal{M}_1) \rightarrow T_1(\mathcal{M}_1).$$

Since  $\mathcal{M}_1$  has finite volume, these flows are ergodic and mixing.

Of special importance for us are the *closed horocycles*  $H_y \subset T_1(\mathcal{M}_1)$ , i.e. the closed orbits for the horocycle flow. When projected to  $\mathcal{M}_1$ , these horocycles are loops of length  $1/y$  around the cusp; they are the images of the lines  $\text{Im } z = y$  in  $\mathbb{H}$ . The geodesic flow expands the closed horocycles, pushing them away from the cusp; indeed, we have

$$g_s(H_y) = H_{e^{-s}y}.$$

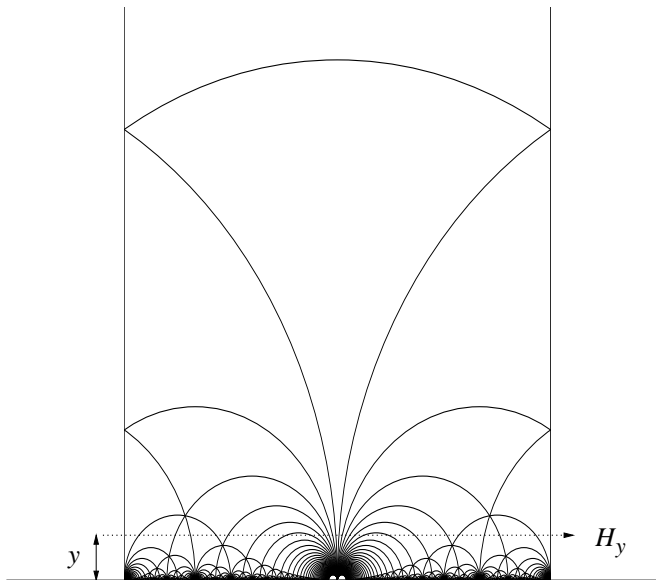


Figure 3. A long closed horocycle  $H_y$  passes randomly through many fundamental domains for  $\text{SL}_2(\mathbb{Z})$ .

**Random elliptic curves.** Using mixing of the geodesic flow, it is not hard to show that  $H_y$  is uniformly distributed on  $B = T_1(\mathcal{M}_1)$  as  $y \rightarrow 0$  (see e.g.

[EsM, §7]). That is, uniform measure along  $H_y$  converges to the invariant measure  $\mu_B$  on  $B$  as the length of  $H_y$  tends to infinity. See Figure 3.

Because of this uniform distribution, one can construct a ‘nearly random’ Riemann surface  $X$  of genus one as follows: pick  $y > 0$  very small, pick  $x \in [0, 1]$  at random, let  $\tau = x + iy$  and set  $X = \mathbb{C}/\mathbb{Z} \oplus \mathbb{Z}\tau$ . As  $y \rightarrow 0$ , the distribution of  $X$  on  $\mathcal{M}_1$  converges to hyperbolic area measure.

**$E$  as a torus bundle.** We now return to the space of lattice translates

$$E = \mathrm{ASL}_2(\mathbb{R})/\mathrm{ASL}_2(\mathbb{Z}).$$

The projection  $\mathrm{ASL}_2(\mathbb{R}) \rightarrow \mathrm{SL}_2(\mathbb{R})$  (sending  $Ax + b$  to  $A$ ) makes the space of lattice translates into a torus bundle

$$E \xrightarrow{D} B.$$

The fiber over  $\Lambda^0$  is the torus  $\mathbb{R}^2/\Lambda^0$ . Moreover  $E$  carries a canonical connection, sending fibers to fibers by group isomorphism. Using this connection, the geodesic flow  $g_s : B \rightarrow B$  lifts to the fiber-preserving *Teichmüller flow*

$$A_s : E \rightarrow E.$$

(The terminology is suggested by the fact that the induced maps between fibers, regarded as Riemann surfaces, are Teichmüller mappings.)

A *horocycle section* is a smooth loop

$$\sigma : S^1 \rightarrow E$$

such that  $D \circ \sigma : S^1 \rightarrow B$  travels with constant speed along a closed horocycle  $H_y \subset B$ . It is not hard to see that  $\sigma_N(x) = \Lambda_N(x)$  is a horocycle section, because the triangle  $T$  shears as  $x$  increases. Moreover, the loops  $\sigma_N$  are simply translates of a single loop under the Teichmüller flow; that is, we have

$$\sigma_N = A_{s(N)} \cdot \sigma_1$$

where  $s(N)$  tends to infinity as  $N$  does. Finally, the section  $\sigma_1$  turns out to be *nonlinear* (see §2.3 for a precise definition). Thus the uniform distribution of  $\langle \Lambda_N(x) \rangle$  on  $E$  follows from:

**Theorem 1.4** *For any nonlinear horocycle section  $\sigma : S^1 \rightarrow E$ , the loops  $\sigma_s = A_s \cdot \sigma$  are uniformly distributed on  $E$  as  $s \rightarrow \infty$ .*

Here is a sketch of the proof (§2). Let  $\mu$  be any measure on  $E$  that arises as a limit of the uniform measures  $m(\sigma_s)$  as  $s \rightarrow \infty$ . By assumption,  $D \circ \sigma_s$  covers a horocycle  $H_{e^{-s}y}$  whose length is tending to infinity. Since these horocycles are uniformly distributed, we have  $D_*(\mu) = \mu_B$ .

In addition,  $\mu$  is invariant under a certain unipotent subgroup  $N(\mathbb{R}) \subset \text{ASL}_2(\mathbb{R})$ . Using a powerful result of Ratner (1991), we can classify the possible ergodic components  $\nu$  of  $\mu$ : either

- $\nu = \mu_E$ , or
- $\nu$  is supported on  $E[n] \subset E$ ,

where  $E[n]$  is the bundle of points of order  $n$  on the torus fibers of  $E$ . The condition that  $\sigma$  is nonlinear rules out the second possibility. Thus  $\nu = \mu = \mu_E$ , and therefore the loops  $\sigma_s$  are uniformly distributed on  $E$ .

In the case at hand, the nonlinearity of the horocycle section  $\Lambda_N(x)$  comes from the fact that the triangle  $T$  has one vertex at  $(0, x^2)$  — and  $x^2$  is a nonlinear function of  $x$ . Thus  $\Lambda_N(x)$  is uniformly distributed on  $E$ , validating our calculation of the gap distribution  $F(t)$ .

**Remarks and references.** The unusual gap distribution for  $\sqrt{n} \bmod 1$  was observed experimentally by M. Boshernitzan in 1993 and communicated to us by Z. Rudnick. See [Sw], [Sos], [RS] and [Bo] for related work on gaps and uniform distribution.

The idea of relating gaps to lattices, as above, is a variation on the method used in [El] to find small nonzero values of  $|x^3 - y^2|$  ( $x, y \in \mathbb{Z}$ ) via lattice reduction.

It is not hard to evaluate the error term in the uniform distribution of  $\{n^\alpha\}$ ,  $0 < \alpha < 1$ : we have

$$\frac{\#\{0 < n \leq N : \{n\theta\} \in I\}}{N} = \frac{|I|}{|S^1|} + O(N^{-\alpha}),$$

and this estimate is sharp (errors of size comparable to  $N^{-\alpha}$  actually occur, when  $0 < |I| < 1$ .)

For more on distribution of sequences modulo 1, see, for example, [We], [HW, Ch. XXIII], [Sa] and [KN].

## 2 Ergodic theory

In this section we prove a general form of Theorem 1.4 on the uniform distribution of horocycle sections  $\sigma : S^1 \rightarrow E$ . This ergodic-theoretic result

will allow us to relate the gap distribution for  $\{\sqrt{n}\}$  to random lattices, as sketched in the Introduction.

## 2.1 The affine group of the plane

All the Lie groups we will consider reside inside the *special affine group*  $G(\mathbb{R}) = \text{ASL}_2(\mathbb{R})$  of the plane, defined by

$$G(\mathbb{R}) = \left\{ \begin{pmatrix} a & b & x \\ c & d & y \\ 0 & 0 & 1 \end{pmatrix} : ad - bc = 1 \right\} \subset \text{SL}_3(\mathbb{R}).$$

This group acts on  $\mathbb{R}^2$  by the area-preserving affine transformations

$$\begin{pmatrix} X \\ Y \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} x \\ y \end{pmatrix}.$$

The affine group is a semidirect product  $G(\mathbb{R}) = \text{SL}_2(\mathbb{R}) \ltimes \text{V}_2(\mathbb{R})$ , where

$$\text{SL}_2(\mathbb{R}) = \left\{ \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\} \quad \text{and} \quad \text{V}_2(\mathbb{R}) = \left\{ \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \right\} \cong \mathbb{R}^2.$$

There is a natural exact sequence

$$0 \rightarrow \text{V}_2(\mathbb{R}) \rightarrow G(\mathbb{R}) \xrightarrow{D} \text{SL}_2(\mathbb{R}) \rightarrow 0,$$

where  $D(g)$  records the linear part of  $g$ .

Within  $\text{SL}_2(\mathbb{R})$  we have the 1-parameter subgroups

$$A(\mathbb{R}) = \left\{ \begin{pmatrix} s & 0 & 0 \\ 0 & 1/s & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\} \quad \text{and} \quad N(\mathbb{R}) = \left\{ \begin{pmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\}; \quad (2.1)$$

we denote their typical elements by  $A_s (s \in \mathbb{R}^*)$  and  $N_t (t \in \mathbb{R})$ .

**Lattice translates.** Let  $G(\mathbb{Z}) \subset G(\mathbb{R})$  denote the arithmetic subgroup of matrices with integral entries. As we remarked in the Introduction, the coset space  $G(\mathbb{R})/G(\mathbb{Z})$  can be identified with the moduli space of translates of unimodular lattices in  $\mathbb{R}^2$ . This identification can be made explicit by taking

the lattice  $\mathbb{Z}^2 \subset \mathbb{R}^2$  as our basepoint: then we may associate to any  $g \in G(\mathbb{R})$  the lattice translate

$$\Lambda(g) = \left\{ (w_1, w_2) \in \mathbb{R}^2 : \begin{pmatrix} w_1 \\ w_2 \\ 1 \end{pmatrix} \in g \begin{pmatrix} \mathbb{Z} \\ \mathbb{Z} \\ 1 \end{pmatrix} \right\}. \quad (2.2)$$

Every unimodular lattice translate can be obtained in this way, and  $\Lambda(g) = \Lambda(h)$  if and only if we have  $g \in h \cdot G(\mathbb{Z})$ .

## 2.2 Mixing of the Teichmüller flow

Let  $\Gamma \subset G(\mathbb{Z})$  be a subgroup of finite index in the integral points of  $G(\mathbb{R})$ . From  $D$  we obtain a fibration

$$\begin{array}{ccc} F = \mathbb{R}^2 / (\Gamma \cap V_2(\mathbb{Z})) & \longrightarrow & E = G(\mathbb{R}) / \Gamma \\ & & \downarrow D \\ & & B = \mathrm{SL}_2(\mathbb{R}) / D(\Gamma). \end{array}$$

In the special case  $\Gamma = G(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z}) \times \mathbb{Z}^2$ , we can regard:

- $B = \mathrm{SL}_2(\mathbb{R}) / \mathrm{SL}_2(\mathbb{Z})$  as the unit tangent bundle  $T_1(\mathcal{M}_1)$  to the moduli space of curves of genus 1,
- $E = G(\mathbb{R}) / G(\mathbb{Z})$  as the pullback to  $T_1(\mathcal{M}_1)$  of the universal elliptic curve  $\mathcal{E} \rightarrow \mathcal{M}_1$ ; and
- $F = \mathbb{R}^2 / \mathbb{Z}^2$ , the fiber over the identity, as the square torus.

For a general subgroup  $\Gamma$ , we obtain a finite cover of the case above. The general case can also be interpreted as a bundle of elliptic curves with basepoints, as follows.

First, note there is an action of  $G(\mathbb{R})$  on  $E$  by left multiplication, and the fibers of  $E \rightarrow B$  are just the orbits of  $V_2(\mathbb{R})$ . Fixing the usual identification  $V_2(\mathbb{R}) = \mathbb{C}$  with complex coordinate  $z = x + iy$ , each fiber obtains a natural complex structure and a natural Euclidean metric (coming from  $|z|$ ). Each fiber meets the submanifold  $\mathrm{SL}_2(\mathbb{R}) / D(\Gamma) \subset E$  in a single point, providing a natural basepoint and making the fibers into groups. Explicitly, for  $g \in \mathrm{SL}_2(\mathbb{R})$  the fiber  $F_g = D^{-1}(g)$  is isomorphic to  $\mathbb{C} / \Lambda_g$  where

$$\Lambda_g = g(\Gamma \cap V_2(\mathbb{R})) \subset V_2(\mathbb{R}) = \mathbb{C}.$$

There is a natural flat *Teichmüller connection* on the fibration  $E \rightarrow B$  that locally identifies fibers via group isomorphisms. (These group isomorphisms are also extremal quasiconformal maps, hence the terminology.) The orbits of  $\mathrm{SL}_2(\mathbb{R})$  on  $E$  are the horizontal submanifolds for this connection.

**The Teichmüller flow.** The group  $\mathrm{SL}_2(\mathbb{R})$  also has a left action on the base  $B$ , compatible with its action on  $E$ . Under the identification  $B = T_1(\widetilde{\mathcal{M}}_1)$  where  $\widetilde{\mathcal{M}}_1 = \mathbb{H}/D(\Gamma)$ , the action of the diagonal subgroup  $A \subset \mathrm{SL}_2(\mathbb{R})$  on  $B$  is identified with the geodesic flow for the hyperbolic metric. The action of  $A$  on  $E \rightarrow B$  is just the lift of the geodesic flow via the Teichmüller connection. Since it sends fibers to fibers by Teichmüller maps, we refer to the action of  $A$  on  $E$  as the *Teichmüller flow*.

**Theorem 2.1** *The Teichmüller flow on the bundle of elliptic curves  $E$  is ergodic and mixing.*

**Proof.** First we observe that the action of  $\mathrm{SL}_2(\mathbb{R})$  on  $E$  is ergodic. Since  $\mathrm{SL}_2(\mathbb{R})$  acts transitively on  $B$ , its ergodicity on  $E$  is equivalent to the ergodicity of  $D(\Gamma)$  on the fiber  $F$ . In the case  $\Gamma = G(\mathbb{Z})$ , we need to show  $\mathrm{SL}_2(\mathbb{Z})$  acts ergodically on  $\mathbb{R}^2/\mathbb{Z}^2$ , and this is easily established using Fourier series. The general case is similar.

Since  $\mathrm{SL}_2(\mathbb{R})$  acts ergodically, the trivial representation is absent from its unitary action on  $L_0^2(E)$ , the square-integrable functions of mean zero. By a general result of Howe and Moore, the matrix coefficients of such a representation vanish at infinity [Zim, Theorem 2.2.20]; that is, for any sequence  $g_n$  tending to infinity in  $\mathrm{SL}_2(\mathbb{R})$  and any  $f_1, f_2 \in L_0^2(E)$ , we have  $\langle g_n \cdot f_1, f_2 \rangle \rightarrow 0$ . Restricting to the 1-parameter group  $A$ , we conclude that the Teichmüller flow is mixing (and hence ergodic). ■

Let  $\mu_E$  denote unique  $G(\mathbb{R})$ -invariant probability measure on  $E$ . Mixing of the Teichmüller flow means that for any set  $X \subset E$  of positive measure and  $f \in C_0(E)$ , we have

$$\lim_{s \rightarrow \infty} \frac{1}{\mu_E(X)} \int_X f(A_s \cdot x) d\mu_E = \int_E f(x) d\mu_E.$$

In other words,  $A_s \cdot X$  is uniformly distributed in  $E$  as  $s \rightarrow \infty$ .

### 2.3 Uniform distribution of horocycle sections

Our main result shows certain loops in  $E$  are also uniformly distributed under the Teichmüller flow.

**Horocycle sections.** We define a *horocycle section*  $\sigma : \mathbb{R} \rightarrow G(\mathbb{R})$  to be a smooth map of the form

$$\sigma(t) = \left\{ \begin{pmatrix} 1 & t & x(t) \\ 0 & 1 & y(t) \\ 0 & 0 & 1 \end{pmatrix} \right\} \quad (2.3)$$

satisfying, for some integer  $p_0 > 0$  and  $\gamma_0 \in G(\mathbb{Z})$ ,

$$\sigma(t + p_0) = \sigma(t)\gamma_0.$$

Note that  $D \circ \sigma(t) = N_t \in \mathrm{SL}_2(\mathbb{R})$ .

Since  $\Gamma$  has finite index in  $G(\mathbb{Z})$ , there is also a minimal  $p > 0$  and  $\gamma \in \Gamma$  such that  $\sigma(t + p) = \sigma(t)\gamma$ . We refer to  $p$  as the *period* of  $\sigma$  on  $E = G(\mathbb{R})/\Gamma$ .

Passing to the quotient space, a horocycle section gives a loop  $\sigma : \mathbb{R}/p\mathbb{Z} \rightarrow E$ . In the case where  $N(p\mathbb{Z}) = N(\mathbb{R}) \cap \Gamma$  (which can always be arranged by passing to a subgroup of finite index), we obtain a bijection

$$D \circ \sigma : \mathbb{R}/p\mathbb{Z} \rightarrow N(\mathbb{R})/N(p\mathbb{Z})$$

between the domain of  $\sigma$  and a standard closed horocycle  $H = N(\mathbb{R})/N(p\mathbb{Z})$  around a cusp of  $B$ . Then  $\sigma$  can be regarded as a section of the universal elliptic curve  $E \rightarrow B$  over  $H \subset B$  — hence the terminology.

**Nonlinearity.** We say a horocycle section  $\sigma$  as in (2.3) is (rationally) *linear* if for some  $\alpha, \beta \in \mathbb{Q}$ , we have:

$$m\{t \in [0, p] : x(t) = \alpha t + \beta\} > 0;$$

in other words, if the graph of  $x(t)$  meets a rational line in a set of positive measure. Otherwise  $\sigma$  is *nonlinear*. (The behavior of  $y(t)$  plays no role in this definition.) A real-analytic section is linear iff  $x(t)$  is *identically* of the form  $\alpha t + \beta$ ,  $\alpha, \beta \in \mathbb{Q}$ .

**Smoothness.** The assumption that  $\sigma(t)$  is smooth ( $C^\infty$ ) is only for convenience; in fact the proof of uniform distribution works so long as  $y(t)$  is continuous and  $x(t)$  is Lipschitz on  $[0, p]$ .

With these definitions in place, we can now state our main result, which generalizes Theorem 1.4.

**Theorem 2.2 (Uniform distribution)** *Let  $\sigma : \mathbb{R} \rightarrow G(\mathbb{R})$  be a nonlinear horocycle section of period  $p$ . Then as  $s \rightarrow \infty$ , the loops  $\sigma_s = A_s \cdot \sigma$  are uniformly distributed in  $E = G(\mathbb{R})/\Gamma$ . That is, for any  $f \in C_0(E)$  we have*

$$\lim_{s \rightarrow \infty} \frac{1}{p} \int_0^p f(A_s \cdot \sigma(t)) dt = \int_E f(x) d\mu_E.$$

As a special case, let  $\Gamma = G(\mathbb{Z})$  and consider the unipotent subgroup

$$U(\mathbb{R}) = \left\{ \begin{pmatrix} 1 & -t & -t^2/4 \\ 0 & 1 & t/2 \\ 0 & 0 & 1 \end{pmatrix} : t \in \mathbb{R} \right\} \subset G(\mathbb{R}). \quad (2.4)$$

Then  $U(\mathbb{R}) \cap \Gamma = U(2\mathbb{Z})$ , so we can regard  $S = U(\mathbb{R})/U(2\mathbb{Z}) \subset E$  as the image of a nonlinear horocycle section with  $x(t) = t^2/4$  and with period  $p = 2$ . Let  $\mu_S$  denote the  $U$ -invariant probability measure on  $S$ . Then we have:

**Corollary 2.3** *The loops  $A_s \cdot S$  are uniformly distributed in  $G(\mathbb{R})/G(\mathbb{Z})$  as  $s \rightarrow \infty$ . That is, for all  $f \in C_0(E)$  we have*

$$\lim_{s \rightarrow \infty} \int_S f(A_s \cdot x) d\mu_S = \int_E f(x) d\mu_E.$$

## 2.4 Limits of measures

Given a loop  $\sigma : \mathbb{R}/p\mathbb{Z} \rightarrow E$ , let  $m(\sigma)$  denote the natural probability measure supported along the image of  $\sigma$ , satisfying

$$\int_E f(x) dm(\sigma) = \frac{1}{p} \int_0^p f(\sigma(t)) dt$$

for all  $f \in C_0(E)$ .

Now let  $\sigma : \mathbb{R}/p\mathbb{Z} \rightarrow E$  be a nonlinear horocycle section of period  $p$ , and let  $\sigma_s = A_s \cdot \sigma$ . To prove Theorem 2.2 (Uniform distribution), we must show that

$$m(\sigma_s) = (A_s)_* m(\sigma) \rightarrow \mu_E$$

as  $s \rightarrow \infty$ . Here convergence takes place in the weak\* topology on the space of measures  $M(E) = C_0(E)^*$ , the dual to the space of compactly-supported continuous functions.

By compactness of the unit ball in the weak\* topology (Alaoglu's theorem), we can pass to a subsequence such that  $m(\sigma_s) \rightarrow \mu \in M(E)$ . Our task is to show that for any such subsequence,  $\mu = \mu_E$ .

We begin by observing that the projection of  $\mu$  to  $B$  is correct. Let

$$\mu_B = D_* \mu_E$$

be the unique  $\mathrm{SL}_2(\mathbb{R})$ -invariant probability measure on  $B$ .

**Theorem 2.4** *We have  $D_*\mu = \mu_B$ .*

**Proof.** As a loop in  $B = T_1(\widetilde{\mathcal{M}}_1)$ , the image  $H$  of  $D \circ \sigma$  represents the outward-pointing normals to a closed horocycle around a cusp of  $\widetilde{\mathcal{M}}_1$ . Under the geodesic flow,  $H$  is pushed away from the cusp and becomes uniformly distributed in  $T_1(\widetilde{\mathcal{M}}_1)$ . (This uniform distribution follows easily from mixing of the geodesic flow by slightly thickening  $H$ ; see [EsM, §7].) Since  $D$  transports the Teichmüller flow on  $E$  to the geodesic flow on  $B$ , and transports  $m(\sigma)$  to the  $N(\mathbb{R})$ -invariant probability measure  $\mu_H$  along  $H$ , we have  $D_*\mu = \lim(A_s)_*\mu_H = \mu_B$ . ■

**Conservation of mass.** Note that the preceding result implies  $\mu(E) = 1$ ; the mass of the probability measures  $m(\sigma_s)$  is conserved under passage to the limit. (In principle mass could be lost because  $E$  is noncompact.)

**Program for the proof.** To show  $A_s \cdot \sigma$  is uniformly distributed in  $E$ , we must prove  $\mu = \mu_E$ . Here are the main steps in the argument.

1. We first show  $\mu$  is invariant under the unipotent subgroup  $N(\mathbb{R}) \subset \mathrm{SL}_2(\mathbb{R})$ .
2. Let  $\nu$  be an ergodic component of  $\mu$ , and let  $J \subset G(\mathbb{R})$  be the largest subgroup leaving  $\nu$  invariant. By Ratner's theorem on unipotent orbits,  $\nu$  is supported on a single orbit of  $J$ .
3. From the fact that  $D_*\mu = \mu_B$ , we conclude that  $J = G(\mathbb{R})$  or  $J = \mathrm{SL}_2(\mathbb{R})$ . Then  $\nu = \mu_E$  or  $\mathrm{supp} \nu \subset E[n]$ , the bundle of points of order  $n$  on each elliptic curve.
4. Using nonlinearity of  $\sigma$ , we show  $\mu(E[n]) = 0$ , and thus  $\mu = \mu_E$ .

## 2.5 Unipotent invariance

In this section we establish:

**Theorem 2.5** *The measure  $\mu$  is  $N(\mathbb{R})$ -invariant.*

Recall that each fiber of the bundle  $E \rightarrow B$  carries a natural Euclidean metric, making it into a flat torus of area one. Suppose  $\sigma_i : \mathbb{R}/p\mathbb{Z} \rightarrow E$ ,  $i = 1, 2$  are a pair of loops satisfying  $D \circ \sigma_1 = D \circ \sigma_2$ . Then the points  $\sigma_1(t)$  and  $\sigma_2(t)$  reside in the same fiber for every  $t$ , and we can measure the distance between these loops by the quantity

$$d(\sigma_1, \sigma_2) = \sup d(\sigma_1(t), \sigma_2(t)).$$

More precisely, suppose the loops are specified upstairs by paths  $\sigma_i : \mathbb{R} \rightarrow G(\mathbb{R})$  of the form

$$\sigma_i(t) = \begin{pmatrix} a(t) & b(t) & x_i(t) \\ c(t) & d(t) & y_i(t) \\ 0 & 0 & 1 \end{pmatrix};$$

then we define

$$d(\sigma_1, \sigma_2) = \sup_{t \in [0, p]} |z_1(t) - z_2(t)| \quad (2.5)$$

where  $z_i(t) = x_i(t) + iy_i(t)$ . (The lifts of  $\sigma_i$  to maps  $\mathbb{R} \rightarrow G$  determine a homotopy from  $\sigma_1$  to  $\sigma_2$ , and  $d(\sigma_1(t), \sigma_2(t))$  is measured using the geodesic on the torus in the given homotopy class.)

Now suppose we have two sequences of loops  $\sigma_1^k, \sigma_2^k$ , with  $d(\sigma_1^k, \sigma_2^k) \rightarrow 0$  and  $m(\sigma_1^k) \rightarrow \nu$ . Then  $m(\sigma_2^k) \rightarrow \nu$  as well. Indeed, any  $f \in C_0(E)$  is uniformly continuous in the fiber direction, so  $|\int f dm(\sigma_1^k) - \int f dm(\sigma_2^k)| \rightarrow 0$ .

**Proof of Theorem 2.5.** Suppose  $\sigma(t)$  is given by (2.3). Then we have

$$\sigma_s(t) = A_s \cdot \sigma(t) = \begin{pmatrix} s & st & sx(t) \\ 0 & s^{-1} & s^{-1}y(t) \\ 0 & 0 & 1 \end{pmatrix}.$$

To test  $N(\mathbb{R})$ -invariance, fix  $\tau \in \mathbb{R}$  and consider the section

$$\eta_s(t) = N_\tau \cdot \sigma_s(t) = \begin{pmatrix} s & st + s^{-1}\tau & sx(t) + s^{-1}\tau y(t) \\ 0 & s^{-1} & s^{-1}y(t) \\ 0 & 0 & 1 \end{pmatrix}.$$

Let  $u = s^{-2}\tau$ . Changing variables, we obtain the section

$$\rho_s(t) = \eta_s(t - u) = \begin{pmatrix} s & st & sx(t - u) + s^{-1}\tau y(t - u) \\ 0 & s^{-1} & s^{-1}y(t - u) \\ 0 & 0 & 1 \end{pmatrix}.$$

Restricting to the subsequence of  $s \rightarrow \infty$  along which  $m(\sigma_s) \rightarrow \mu$ , we have

$$m(\rho_s) = m(\eta_s) = (N_\tau)_* m(\sigma_s) \rightarrow (N_\tau)_* \mu. \quad (2.6)$$

Since the upper-left  $2 \times 2$  submatrices of  $\rho_s(t)$  and  $\sigma_s(t)$  agree, we have  $D \circ \rho_s = D \circ \sigma_s$ , so we can measure their distance:

$$d(\rho_s, \sigma_s) \leq \sup_{t \in [0, p]} |sx(t) - sx(t - u)| + |s^{-1}\tau y(t - u)| + |s^{-1}y(t)| + |s^{-1}y(t - u)|.$$

As  $s \rightarrow \infty$ , the term involving  $x$  satisfies

$$|sx(t) - sx(t - u)| = O(su) = O(s^{-1}) \rightarrow 0$$

since  $x(t)$  is smooth. The terms involving  $y$  go to zero because  $y(t)$  is bounded on  $[0, p]$ . Thus  $d(\rho_s, \sigma_s) \rightarrow 0$ , and therefore  $m(\rho_s)$  and  $m(\sigma_s)$  have the same limit, namely  $\mu$ , by the observations preceding the proof. From (2.6) we obtain  $(N_\tau)_*\mu = \mu$ .  $\blacksquare$

Geometrically, the  $N(\mathbb{R})$ -invariance of  $\mu$  comes from the fact that  $A_s$  stretches more horizontally than along the fibers, and hence the loop  $A_s \cdot \sigma$ ,  $s \gg 1$  is nearly horizontal (flat for the Teichmüller connection) when  $s \gg 1$ .

## 2.6 Ratner's theorem

The main tool in our proof of equidistribution is the following theorem [Rat]:

**Theorem 2.6 (Ratner)** *Let  $\Gamma \subset G$  be a discrete subgroup of a connected Lie group  $G$ , and let  $N \subset G$  be a unipotent subgroup. Let  $\nu$  be an ergodic  $N$ -invariant probability measure on  $G/\Gamma$ , and let  $J \subset G$  be the largest subgroup leaving  $\nu$  invariant. Then there is an  $x \in G/\Gamma$  such that  $\nu(J \cdot x) = 1$ .*

Here the group  $N$  is *unipotent* if for any  $g \in N$ , the eigenvalues of the adjoint action of  $g$  on the Lie algebra of  $G$  are all 1. The measure  $\nu$  is *ergodic* if any  $N$ -invariant measurable set  $X \subset G/\Gamma$  satisfies  $\nu(X) = 0$  or  $\nu(X) = 1$ . The *support* of  $\nu$ , denoted  $\text{supp } \nu$ , is the smallest closed set with  $\nu(\text{supp } \nu) = 1$ .

The Theorem implies that:

- $J/(J \cap x\Gamma x^{-1})$  has finite volume,
- $\nu$  coincides with normalized Haar measure on  $J \cdot x \subset G/\Gamma$ , and
- $J \cdot x$  is closed in  $G/\Gamma$  (cf. [Rat, Prop. 1.4]), so
- $J \cdot x = \text{supp}(\nu)$ .

**Concentration on torsion points.** Each fiber  $F$  of the bundle  $E \rightarrow B$  has the structure of a complex torus  $\mathbb{C}/\Lambda$ . For any integer  $n \geq 1$ , we let  $F[n] = (n^{-1}\Lambda)/\Lambda \subset F$  denote the points of order  $n$  with respect to the group law on  $F$ , and let  $E[n] \subset E$  be the bundle whose fibers are  $F[n]$ . Then  $\bigcup E[n]$  is the set of *torsion points* in  $E$ .

The projection  $E[n] \rightarrow B$  is a covering map of degree  $n^2$ , and  $E[n]$  is the union of a finite number of  $\mathrm{SL}_2(\mathbb{R})$ -orbits on  $E$ . Every closed  $\mathrm{SL}_2(\mathbb{R})$  orbit is contained in  $E[n]$  for some  $n$ .

Let us denote the group horizontal translations of  $\mathbb{R}^2$  by:

$$H(\mathbb{R}) = \left\{ \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} : x \in \mathbb{R} \right\} \subset G(\mathbb{R}).$$

This group commutes with  $N(\mathbb{R})$ . Applying Ratner's theorem, we will show that the sections  $\sigma_s$  are either uniformly distributed, or they concentrate on horizontal translates of the torsion points in  $E$ .

**Theorem 2.7 (Torsion alternative)** *Either  $\mu = \mu_E$  or  $\mu(H(\mathbb{R}) \cdot E[n]) > 0$  for some  $n \geq 1$ .*

To apply Theorem 2.6 to the case at hand, we first note that  $G = \mathrm{ASL}_2(\mathbb{R})$  is connected and  $N = N(\mathbb{R})$  is unipotent. However, we do not yet know if  $\mu$  is ergodic.

Nevertheless, a basic result of ergodic theory furnishes a canonical decomposition

$$\mu = \int \nu dP(\nu)$$

of  $\mu$  into a convex combination of ergodic,  $N(\mathbb{R})$ -invariant probability measures  $\nu$  [Wa, p. 153], [Ph, §10]. Here  $dP$  is itself a probability measure on the set of ergodic  $\nu$ , which form the extreme points of the unit ball in  $M(E)^{N(\mathbb{R})}$ . (The traditional setting for the ergodic decomposition is dynamics on a compact metric space, which can be obtained by replacing  $E$  with its one-point compactification.)

Given an ergodic,  $N(\mathbb{R})$ -invariant probability measure  $\nu$  on  $E$ , let

$$J(\nu) = \{g \in G(\mathbb{R}) : g_*\nu = \nu\}$$

be the largest subgroup of  $G(\mathbb{R})$  leaving  $\nu$  invariant. Then  $J(\nu)$  is closed and we have  $N(\mathbb{R}) \subset J(\nu)$ .

**Theorem 2.8** *Almost every  $\nu$  occurring in the ergodic decomposition of  $\mu$  satisfies  $D_*\nu = \mu_B$  and  $D(J(\nu)) = \mathrm{SL}_2(\mathbb{R})$ .*

**Proof.** The action of  $N(\mathbb{R})$  on  $(B, \mu_B)$  is ergodic, since it coincides with the horocycle flow on  $T_1(\widetilde{\mathcal{M}}_1)$ . Thus in the decomposition  $\mu_B = D_*\mu = \int D_*\nu dP(\nu)$  we must have  $D_*\nu = \mu_B$  for almost every  $\nu$ .

By Ratner's theorem,  $\nu$  is supported on the single orbit  $J(\nu) \cdot x \subset E$ . Since  $E \rightarrow B$  has compact fibers, we have

$$D(J(\nu)) \cdot D(x) = D(\mathrm{supp} \nu) = \mathrm{supp} D_*\nu = B = \mathrm{SL}_2(\mathbb{R})/D(\Gamma).$$

Thus  $D(J(\nu)) = \mathrm{SL}_2(\mathbb{R})$ . ■

**Fixed points.** To determine the possibilities for  $J(\nu)$ , we will use:

**Lemma 2.9** *Any affine action of  $\mathrm{SL}_2(\mathbb{R})$  on  $\mathbb{R}^k$  has a fixed point.*

**Proof.** Weyl's unitary trick [Kn, Prop. 2.1] allows one to extend the affine action of  $\mathrm{SL}_2(\mathbb{R})$  on  $\mathbb{R}^k$  to an affine action of  $\mathrm{SL}_2(\mathbb{C})$  on  $\mathbb{C}^k$ . Then a fixed point  $p \in \mathbb{C}^k$  for the compact group  $\mathrm{SU}(2, \mathbb{C})$  can be constructed by averaging. Since  $\mathbb{C} \cdot \mathfrak{su}_2 = \mathfrak{sl}_2(\mathbb{C})$ , the point  $p$  is also fixed by  $\mathrm{SL}_2(\mathbb{C})$ , and hence by  $\mathrm{SL}_2(\mathbb{R})$ . The real part of  $p$  then gives a fixed point in  $\mathbb{R}^k$ . ■

**Corollary 2.10** *Let  $H \subset G(\mathbb{R})$  be a subgroup with  $D(H) = \mathrm{SL}_2(\mathbb{R})$ . Then either  $H = G(\mathbb{R})$  or  $H$  is a conjugate of  $\mathrm{SL}_2(\mathbb{R}) \subset G(\mathbb{R})$ .*

**Proof.** Since  $D(H) = \mathrm{SL}_2(\mathbb{R})$ , the kernel  $K = \mathrm{Ker}(D : H \rightarrow \mathrm{SL}_2(\mathbb{R}))$  is an  $\mathrm{SL}_2(\mathbb{R})$ -invariant subgroup of  $V_2(\mathbb{R}) \cong \mathbb{R}^2$ , so  $K = V_2(\mathbb{R})$  or  $K = \{e\}$ . In the former case  $H = G(\mathbb{R})$ , while in the latter case

$$D^{-1} : \mathrm{SL}_2(\mathbb{R}) \rightarrow H \subset G(\mathbb{R}) = \mathrm{ASL}_2(\mathbb{R})$$

gives an affine action of  $\mathrm{SL}_2(\mathbb{R})$  on  $\mathbb{R}^2$ . By the preceding Lemma, this action has a fixed point. After conjugating by an element of  $V_2(\mathbb{R})$ , we can assume the fixed point is the origin, and then  $H = \mathrm{SL}_2(\mathbb{R})$ . ■

Since  $D(J(\nu)) = \mathrm{SL}_2(\mathbb{R})$  and  $N(\mathbb{R}) \subset J(\nu)$ , we have:

**Corollary 2.11** *Either  $J(\nu) = G(\mathbb{R})$  or  $J(\nu) = g\mathrm{SL}_2(\mathbb{R})g^{-1}$  for some  $g \in H(\mathbb{R})$ .*

**Corollary 2.12** *Either  $\nu = \mu_E$  or  $\mathrm{supp}(\nu) \subset g \cdot E[n]$  for some  $n > 0$  and  $g \in H(\mathbb{R})$ .*

**Proof.** If  $J(\nu) = G(\mathbb{R})$  then  $\nu = \mu_E$  by  $J(\nu)$ -invariance. Otherwise, there exists a  $g \in H(\mathbb{R})$  such that  $g^{-1} \cdot \mathrm{supp} \nu$  is a closed  $\mathrm{SL}_2(\mathbb{R})$  orbit in  $E$ , and any such orbit is contained in  $E[n]$  for some  $n$ . ■

**Proof of Theorem 2.7 (Torsion alternative).** Let  $\mu = \int \nu dP(\nu)$  be the ergodic decomposition of  $\mu$ . By the preceding Corollary, for almost every ergodic component  $\nu$  of  $\mu$ , either  $\nu = \mu_E$  or  $\mathrm{supp} \nu \subset H(\mathbb{R}) \cdot E[n]$  for some  $n$ . Thus we can write  $\mu = a_0\mu_E + \sum_1^\infty a_n\mu_n$  where  $\sum_0^\infty a_n = 1$  and  $\mathrm{supp} \mu_n \subset H(\mathbb{R}) \cdot E[n]$ . If  $\mu \neq \mu_E$ , then  $a_n \neq 0$  for some  $n > 0$  and  $\mu(H(\mathbb{R}) \cdot E[n]) > 0$ . ■

## 2.7 Nonlinearity

In this section we use the nonlinearity of  $\sigma$  to show its limit under the Teichmüller flow does not concentrate on the torsion points of  $E$ .

**Theorem 2.13** *For any  $n \geq 1$  and any accumulation point  $\mu$  of the measures  $m(A_s \cdot \sigma)$  as  $s \rightarrow \infty$ , we have  $\mu(H(\mathbb{R}) \cdot E[n]) = 0$ .*

**Proof.** Since the torsion points of  $G/\Gamma$  lie over those of  $G/G(\mathbb{Z})$ , it suffices to treat the case where  $\Gamma = G(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z}) \times \mathbb{Z}^2$ .

Let  $H(r, \epsilon) \subset V_2(\mathbb{R})$  denote the set of translations by  $(x, y) \in \mathbb{R}^2$  with  $|x| < r$ ,  $|y| < \epsilon$ , and let  $H(r) = H(r, 0)$ . Since  $H(\mathbb{R}) = \bigcup_1^\infty H(r)$ , to prove the theorem it suffices to show  $\mu(H(r) \cdot E[n]) = 0$  for each fixed  $r > 0$ .

Consider the open set  $U = H(r, \epsilon) \cdot E[n]$ . Let  $\sigma_s = A_s \cdot \sigma$ , let  $m_s = m(\sigma_s)$  and let

$$T_s = \{t \in [0, p] : \sigma_s(t) \in U\}.$$

We will show that the linear measure of  $T_s$  satisfies

$$\limsup m(T_s) = O(\epsilon). \tag{2.7}$$

(The implied constant may depend on  $r, n$ .) Since  $m_s(U) = m(T_s)/p$ , it will follow that  $\mu(H(r) \cdot E[n]) \leq \limsup m_s(U) = O(\epsilon^2)$  and hence  $\mu(H(r) \cdot E[n]) = 0$  as desired.

To estimate  $m(T_s)$ , it is useful to pass to the universal cover  $G$  of  $E = G/G(\mathbb{Z})$ . Then we can regard  $\sigma_s$  as the map  $\sigma_s : [0, p] \rightarrow G$  given by

$$\sigma_s(t) = \begin{pmatrix} s & st & sx(t) \\ 0 & s^{-1} & s^{-1}y(t) \\ 0 & 0 & 1 \end{pmatrix}.$$

The set  $E[n] \subset E$  is covered by the  $\mathrm{SL}_2(\mathbb{R})$ -orbits  $G[n] = \bigcup G[n]^{ij} \subset G$ ,  $i, j \in \mathbb{Z}$ , given by

$$G[n]^{ij} = \left\{ \begin{pmatrix} a & b & (i/n)a + (j/n)b \\ c & d & (i/n)c + (j/n)d \\ 0 & 0 & 1 \end{pmatrix}, \quad ad - bc = 1 \right\}.$$

Thus the points of  $G[n]$  in the same fiber as  $\sigma_s(t)$  are given by

$$\rho_s^{ij}(t) = \begin{pmatrix} s & st & (i/n)s + (j/n)st \\ 0 & s^{-1} & s^{-1}(j/n) \\ 0 & 0 & 1 \end{pmatrix}, \quad i, j \in \mathbb{Z}.$$

Taking the difference between the final columns of  $\rho_s^{ij}$  and  $\sigma_s$ , we find  $T_s = \bigcup T_s^{ij}$  where

$$T_s^{ij} = \left\{ t : \left| \begin{pmatrix} sx(t) \\ s^{-1}y(t) \end{pmatrix} - \begin{pmatrix} (i/n)s + (j/n)st \\ s^{-1}(j/n) \end{pmatrix} \right| < \begin{pmatrix} r \\ \epsilon \end{pmatrix} \right\},$$

and the vector inequality is taken componentwise. Equivalently, we have  $T_s^{ij} = X_s^{ij} \cap Y_s^{ij}$ , where

$$\begin{aligned} X_s^{ij} &= \{t : |x(t) - (i/n) - (j/n)t| < s^{-1}r\} \quad \text{and} \\ Y_s^{ij} &= \{t : |y(t) - (j/n)| < s\epsilon\}. \end{aligned}$$

Since  $\sigma$  is nonlinear, the set of  $t$  such that  $x(t) = (i/n) + (j/n)t$  has measure zero; therefore, for each fixed  $i, j$  we have

$$\lim_{s \rightarrow \infty} m(X_s^{ij}) = 0. \tag{2.8}$$

Also, when  $j$  is large we have

$$m(X_s^{ij}) = O(s^{-1}/|j|),$$

since  $X_s^{ij}$  is the preimage of an interval of length  $s^{-1}r$  under a map with derivative approximately  $-j/n$ . More precisely, this bound on  $m(X_s^{ij})$  holds so long as  $|j| > M$ , where  $M = 2n \sup_{[0,p]} |x'(t)|$ .

We also have  $Y_s^{ij} = \emptyset$  unless

$$|j| < J_s = n \left( s\epsilon + \sup_{[0,p]} |y(t)| \right) = O(s\epsilon)$$

for large  $s$ . Similarly, we have  $X_s^{ij} = \emptyset$  unless

$$|i| < I_s(j) = n \left( s^{-1}r + |j/n| + \sup_{[0,p]} |x(t)| \right) = O(|j| + 1).$$

Therefore we have:

$$\begin{aligned} m(T_s) &\leq \sum_{|j| < J_s} \sum_{|i| < I_s(j)} m(X_s^{ij}) \\ &\leq \sum_{M < |j| < J_s} \sum_{|i| < I_s(j)} O(s^{-1}/|j|) + \sum_{|j| \leq M} \sum_{|i| < I_s(M)} m(X_s^{ij}). \end{aligned}$$

Since the second sum is over a finite set of  $i, j$ , it tends to zero as  $s \rightarrow \infty$  by (2.8). Since  $I_s(j) = O(|j| + 1)$ , the first sum is  $O(|J_s|s^{-1}) = O(\epsilon)$ , establishing (2.7).  $\blacksquare$

**Completion of the proof of Theorem 2.2 (Uniform distribution).**

Let  $\mu$  be any accumulation point of the measures  $m(A_s \cdot \sigma)$  as  $s \rightarrow \infty$ . By the preceding result,  $\mu$  assigns zero mass to the torsion points  $H(\mathbb{R}) \cdot E[n] \subset E$ , so by Theorem 2.7 (Torsion alternative), we have  $\mu = \mu_E$ . Since the accumulation point is unique,  $\mu_E$  is actually the limit of  $m(A_s \cdot \sigma)$ , and thus the loops  $A_s \cdot \sigma$  are uniformly distributed in  $E$  as  $s \rightarrow \infty$ .  $\blacksquare$

**Sharpness.** In conclusion we remark that the converse to Theorem 2.2 also holds:  $A_s \cdot \sigma$  is never uniformly distributed when  $\sigma$  is *linear*. Indeed, if  $x(t) = (i/n) + (j/n)t$  on a set of positive measure, then we have  $\mu(E(n)) > 0$ , so  $m(A_s \cdot \sigma)$  cannot converge to the uniform measure  $\mu_E$ .

### 3 Distribution of gaps

In this section we establish the relationship between gaps in  $\sqrt{n} \bmod 1$  and the lattice translates  $\Lambda_N(x)$  sketched in the Introduction. We also determine the probability that a random unimodular lattice translate meets a triangle  $T$  as a function of  $\text{area}(T)$ . Using the main ergodic theory result of §2, we then deduce an explicit formula for the limiting gap distribution  $F(t)$ . Finally we discuss generalizations of the main result and open questions.

#### 3.1 Gap-counting functions

For each positive integer  $N$  we define a normalized gap-counting function  $\lambda_N : [0, \infty) \rightarrow [0, 1]$  as follows. Consider the  $N$  numbers  $\sqrt{n} \bmod 1$  in  $\mathbb{R}/\mathbb{Z}$  for  $n = 1, 2, \dots, N$ . These partition  $\mathbb{R}/\mathbb{Z}$  into  $N$  intervals (of which  $\lfloor \sqrt{N} \rfloor - 1$  have length zero). We call the lengths of these intervals the *gaps* in the sequence

$$\{\sqrt{n} \bmod 1 : 1 \leq n \leq N\}. \quad (3.1)$$

The function  $\lambda_N(x)$  is defined to be  $1/N$  times the number of gaps whose length is  $< x/N$ .

Clearly  $\lambda_N$  is a nondecreasing function, continuous from the left, which is constant except for finitely many jumps, and satisfies  $\lambda_N(0) = 0$  and  $\lambda_N(\infty) = 1$ . Moreover, we have

$$\int_0^\infty (1 - \lambda_N(x)) dx = \int_0^\infty x d(\lambda_N(x)) = 1, \quad (3.2)$$

because  $\int_0^\infty x d(\lambda_N(x))$  is the total length of the gaps (the normalizing factors of  $N$  cancel), and  $\int_0^\infty (1 - \lambda_N(x)) dx = \int_0^\infty x d(\lambda_N(x))$  by integration by parts.

We are interested in the asymptotics of  $\lambda_N$  as  $N \rightarrow \infty$ . We shall show that there exists  $\lambda_\infty : [0, \infty) \rightarrow [0, 1)$  such that  $\lambda_N(x) \rightarrow \lambda_\infty(x)$ , uniformly in  $x$ . Moreover we shall write

$$\lambda_\infty(x) = \int_0^x F(\xi) d\xi, \quad (3.3)$$

where  $F : [0, \infty) \rightarrow (0, \infty)$  is a continuous function to be described later. Thus  $F$  is the asymptotic normalized distribution of gaps in  $\{\sqrt{n} \bmod 1 : 1 \leq n \leq N\}$ : for each  $x_1, x_2 \in [0, \infty)$  with  $x_1 < x_2$ , the number of gaps in  $[x_1/N, x_2/N]$  is asymptotic to  $\int_{x_1}^{x_2} F(x) dx$ .

To get at  $\lambda_N$ , define  $L_N : \mathbb{R}/\mathbb{Z} \rightarrow [0, \infty)$  as follows:  $L_N(t)$  is  $N$  times the length of the gap containing  $t$ , unless  $t \equiv \sqrt{n} \pmod{1}$  for some positive integer  $n \leq N$ , in which case we set  $L_N(t) = 0$ . Then

$$I_N(x) := \{t \in \mathbb{R}/\mathbb{Z} : L_N(t) < x\} \quad (3.4)$$

is the union of the gaps of length  $< x$ , and thus has measure

$$|I_N(x)| = \int_0^x \xi d(\lambda_N(\xi)). \quad (3.5)$$

To prove our claim about the asymptotics of  $\lambda_N$ , it will be enough to show that

$$|I_N(x)| \rightarrow \int_0^x \xi F(\xi) d\xi \quad (3.6)$$

as  $N \rightarrow \infty$ , uniformly as  $x$  varies in bounded subsets of  $[0, \infty)$ .

### 3.2 From gaps to lattice translates in $\mathbb{R}^2$

It is time to use specific properties of the sequence  $\{\sqrt{n} \pmod{1} : n \leq N\}$ . As in the Introduction, we shall write each  $n$  uniquely as  $a^2 + b$  with

$$a = \lfloor \sqrt{n} \rfloor = \sqrt{n} - \{\sqrt{n}\},$$

and exploit the special behavior of the function  $(a, b) \mapsto \{\sqrt{a^2 + b}\}$  on the  $(a, b)$ -plane.

It will be convenient to assume that  $N$  is a perfect square greater than 1, say  $N = s^2$  with  $s > 1$ . Once we prove that  $\lambda_N \rightarrow \lambda_\infty$  uniformly as  $N$  increases through perfect squares, the uniform convergence of all  $\lambda_N$  to  $\lambda_\infty$  will follow:

**Lemma 3.1** *Assume that there exists a continuous function  $\lambda_\infty : [0, \infty) \rightarrow [0, \infty)$  such that  $\{\lambda_{s^2} : s = 1, 2, 3, \dots\}$  converges uniformly to  $\lambda_\infty$ . Then  $\{\lambda_N : N = 1, 2, 3, \dots\}$  also converges uniformly to  $\lambda_\infty$ .*

**Proof.** Every integer  $N$  is within  $O(N^{1/2})$  of a perfect square  $s_1^2$ . Replacing  $N$  by  $s_1^2$  changes at most  $3|N - s_1^2| \ll N^{1/2}$  of the gaps, and multiplies the normalizing factors by  $1 + O(N^{-1/2})$ . Under our assumptions that  $\lambda_\infty$  is continuous and  $\lambda_{s_1^2}$  converges uniformly to  $\lambda_\infty$ , it follows that  $\lambda_N$  does as well. ■

Now let  $t \in [0, 1]$ . Then  $L_N(t) = N(t_2 - t_1)$ , where  $t_2$  is the smallest real number  $\geq t$  such that  $(a_2 + t_2)^2 \in \mathbb{Z}$  for some positive integer  $a_2 < s$ , and  $t_1$  is likewise the largest real number  $\leq t$  such that  $(a_1 + t_1)^2 \in \mathbb{Z}$  for some positive integer  $a_1 < s$ . Our first key observation is that in the binomial expansion

$$(a_j + t_j)^2 = a_j^2 + 2a_j t_j + t_j^2 \quad (j = 1, 2) \quad (3.7)$$

the term  $a_j^2$  is always an integer, so the condition  $(a_j + t_j)^2 \in \mathbb{Z}$  is equivalent to

$$2a_j t_j + t_j^2 = b_j, \quad b_j \in \mathbb{Z}. \quad (3.8)$$

Necessarily

$$0 \leq b_j \leq (a_j + 1)^2 - a_j^2 = 2a_j + 1. \quad (3.9)$$

Define a function  $r_t$  by

$$r_t(a, b) := \sqrt{a^2 + b} - a - t. \quad (3.10)$$

Then we may write

$$L_N(t) = N((t_2 - t) - (t_1 - t)) = N\left(\min_{r_t(a,b) \geq 0} r_t(a, b) - \max_{r_t(a,b) \leq 0} r_t(a, b)\right), \quad (3.11)$$

with  $a, b$  ranging over integers such that

$$0 < a < s, \quad 0 \leq b \leq 2a + 1. \quad (3.12)$$

In fact the conditions on  $b$  are superfluous. Since  $a > 0$ , requiring  $b \in [0, 2a + 1]$  is equivalent to demanding that  $r_t(a, b) + t \in [0, 1]$ . The smallest positive and largest negative values of  $r_t(a, b)$  as  $a, b$  vary over integers with  $0 < a < s$  automatically have  $r_t(a, b) + t \in [0, 1]$ , because  $(a, b) = (1, 0)$  and  $(a, b) = (1, 3)$  already give  $r_t(a, b) + t = 0$  and  $r_t(a, b) + t = 1$  respectively.

The next step is to change  $t_j^2$  in equation (3.8) to its linear approximation

$$t^2 + 2t(t_j - t) = 2tt_j - t^2 = t_j^2 - (t - t_j)^2 \quad (3.13)$$

We expect that usually  $t_j = t + O(1/N)$ , and thus that  $t_j$  will be within  $O(1/aN^2)$  of the solution  $\tau_j$  of the equation

$$2(a_j + t)\tau_j - t^2 = b_j. \quad (3.14)$$

Since for each  $\epsilon > 0$  we have  $1/a < \epsilon$  for all but a few  $(a_j, b_j)$  pairs, we thus expect that replacing  $t_j$  by  $\tau_j$  in the definition of  $L_N$  will change most gaps by  $O(\epsilon/N^2)$ , and thus will not affect the asymptotic behavior of  $|I_N|$ . We next establish the estimates that support our expectations.

The solution of equation (3.14) is

$$\tau_j = \frac{b_j + t^2}{2(a_j + t)}. \quad (3.15)$$

Let  $\rho_t$ , then, be the function

$$\rho_t(a, b) := \frac{b + t^2}{2(a + t)} - t = \frac{a^2 + b - (a + t)^2}{2(a + t)}, \quad (3.16)$$

and define  $L'_N : [0, 1] \rightarrow [0, \infty)$  by

$$L'_N(t) := N \left( \min_{\rho_t(a,b) \geq 0} \rho_t(a, b) - \max_{\rho_t(a,b) \leq 0} \rho_t(a, b) \right), \quad (3.17)$$

with  $a, b$  in the same range (3.12) as in our formula (3.11) for  $L_N(t)$ . As with  $L_N$ , the condition  $b \in [0, 2a + 1]$  holds automatically for the minimal and maximal  $(a, b)$ , and thus need not be required explicitly. Analogous to the formula (3.4) for  $|I_N|$  in terms of  $L_N$ , we then define

$$I'_N(x) := \{t \in \mathbb{R}/\mathbb{Z} : L'_N(t) < x\}. \quad (3.18)$$

We shall prove:

**Proposition 3.2** *Suppose that the limiting gap distribution  $F : [0, \infty) \rightarrow [0, \infty)$  is continuous. Then formula (3.6), with convergence uniform in bounded subsets of  $[0, \infty)$ , is equivalent to the same formula with  $I_N$  replaced with  $I'_N$ :*

$$|I'_N(x)| \rightarrow \int_0^x \xi F(\xi) d\xi \quad (3.19)$$

and the same uniformity.

**Proof.** We use the following technical estimates.

**Lemma 3.3** *For all  $t \in [0, 1]$  we have*

$$\frac{3}{4}L_N(t) \leq L'_N(t) \leq \frac{3}{2}L_N(t). \quad (3.20)$$

Moreover, for each  $A = 1, 2, 3, \dots$  the stronger inequality

$$\frac{2A+1}{2A+2}L'_N(t) \leq \frac{2A+1}{2A}L_N(t) \quad (3.21)$$

holds for all  $t \in [0, 1]$  except for a subset of length at most  $(A+2)(A-1)/(s-1)$ .

**Proof.** Comparing equation (3.10) with (3.16), we see that  $r_t(a, b)$  and  $\rho_t(a, b)$  are either both positive, both negative, or both zero. Moreover, unless  $r_t(a, b) = \rho_t(a, b) = 0$ , we have

$$\frac{\rho_t(a, b)}{r_t(a, b)} = \frac{\sqrt{a^2 + b} + a + t}{2(a + t)} \in \left[ \frac{2a + 1}{2a + 2}, \frac{2a + 1}{2a} \right] \quad (3.22)$$

for each  $a, b$  in the range (3.12) and every  $t \in [0, 1]$ . Since  $a \geq 1$ , inequality (3.20) follows. Furthermore, the stronger inequality (3.21) holds unless  $a < A$  for the  $(a, b)$  pair attaining the maximum or minimum in equation (3.11). But there are only  $(A + 2)(A - 1)$  integer pairs  $(a, b)$  with  $0 < a < A$  and  $0 \leq b \leq 2a + 1$ , and each affects only those  $t$  contained in a gap one of whose endpoints is  $\sqrt{a^2 + b} - a$ . Each gap has length less than  $1/2(s - 1)$ , since  $\{\sqrt{(s - 1)^2 + b} - (s - 1) : 0 \leq b < 2s\}$  already partitions  $[0, 1]$  into intervals of length  $< 1/2(s - 1)$ . Thus the exceptional set for equation (3.21) has length at most  $(A + 2)(A - 1)/(s - 1)$ , as claimed. ■

**Corollary 3.4** *For each  $x \in [0, \infty)$  we have*

$$\left| I'_N\left(\frac{3}{4}x\right) \right| \leq |I_N(x)| \leq \left| I'_N\left(\frac{3}{2}x\right) \right|, \quad (3.23)$$

and for each  $A = 1, 2, 3, \dots$  also

$$\left| I'_N\left(\frac{2A + 1}{2A + 2}x\right) \right| - O(A^2/s) \leq |I_N(x)| \leq \left| I'_N\left(\frac{2A + 1}{2A}x\right) \right| + O(A^2/s). \quad (3.24)$$

**Proof.** The estimates (3.23) and (3.24) follow immediately from the corresponding bounds given by equations (3.20) and (3.21) in Lemma 3.3. ■

We complete the proof of Proposition 3.2 by taking  $A = 1 + \lfloor s^{1/3} \rfloor$  (or any increasing function of  $s$  such that  $A > 1$  and  $A^2/s \rightarrow 0$ ) and using the continuity of  $F$ . ■

We have thus reduced the asymptotics of  $\lambda_N$  to the asymptotic distribution of the values of  $L'_N$  as  $N \rightarrow \infty$ . We next describe  $L'_N(t)$  geometrically in terms of the family of triangles  $T$  in the plane, shown in Figure 2 of the Introduction. In the present notation, for each  $s, t$ ,  $T$  is the triangle with one vertex at  $(a, b) = (-t, -t^2)$  and the opposite side contained in the line  $a + t = s$ , which contains the segment of the line  $b = 2ta + t^2$  from the fixed

vertex to  $(a, b) = (s - t, 2st + t^2)$ . Algebraically,  $T$  is the triangle whose interior is given by the inequalities

$$0 < a + t < s, \quad \frac{2c_-}{s^2}(a + t) < b - 2ta - t^2 < \frac{2c_+}{s^2}(a + t), \quad (3.25)$$

for some  $c_-, c_+$  with  $c_- < 0 < c_+$ .

**Lemma 3.5** *For each  $N = s^2$  and  $t \in [0, 1]$ , if  $L'_N(t) \neq 0$  then  $L'_N(t)$  is the area  $c_+ - c_-$  of the largest triangle whose interior is given by equation (3.25) such that this interior contains no integer points except possibly  $(0, 0)$  or  $(0, 1)$ . If  $L'_N(t) = 0$  then there is no such triangle because  $0 = b - 2ta - t^2$  has an integer solution  $(a, b)$  with  $0 < a < s$ .*

As  $c_-, c_+$  move away from zero, our triangle expands from the fixed line segment until each of the sides of the triangle through the vertex  $(0, t^2)$  hits a lattice point.

**Lemma 3.6** *When this happens, the triangle has area  $L'_N(t)$ , unless  $L'_N(t) = 0$  when the line  $b = 2ta + t^2$  already contains a lattice point with  $0 < a < s$ .*

**Proof.** That the triangle in fact has area  $c_+ - c_-$  is clear. If  $b = 2ta + t^2$  for some integers  $a, b$  with  $a > 0$  then necessarily  $0 \leq b \leq 2s + 1$  and  $t = \sqrt{a^2 + b} - a$ , so  $L'_N(t) = 0$ , and conversely. We may thus assume that  $L'_N(t) \neq 0$ . In particular,  $0 \neq t \neq 1$ , so, since  $a$  is an integer,  $0 < a + t < s$  if and only if  $0 \leq a < s$ . If  $a = 0$  then  $b = 0$  or  $b = 1$ , since for other choices of  $b$  the triangle also contains  $(1, 0)$  or  $(1, 3)$ . The inequalities on  $b - 2ta - t^2$  in equation (3.25) are equivalent to  $c_- \leq s^2 \rho_t(a, b) \leq c_+$ . Since  $s^2 = N$ , our largest triangle has  $c_+/s^2$  equal to the smallest positive value of  $\rho_t(a, b)$  with  $0 < a < s$ , and  $c_-/s^2$  equal to the largest negative value. Thus  $c_+ - c_- = L'_N(t)$  as claimed.  $\blacksquare$

Fortunately the distracting possibilities  $(a, b) = (0, 0)$  and  $(0, 1)$  do not affect  $L'_N(t)$  except for  $t$  in a subset of  $[0, 1]$  of length  $O(1/s)$ :

**Lemma 3.7** *Allowing  $(a, b) = (0, 0)$  or  $(a, b) = (0, 1)$  in Lemma 3.5 does not change  $L'_N(t)$  unless  $t < 1/(s - 1)$  or  $t^{-1} - t < 1/(s - 1)$ .*

**Proof.** If  $r_t(0, 0) = -t/2$  is smaller than  $c_-$  then there are no lattice points in the triangle interior

$$0 < a + t < s, \quad -t(a + t) < b - 2ta - t^2 < 0. \quad (3.26)$$

But the intersection of this triangle with  $a = s - 1$  is a line segment of length  $> (s - 1)t$ . If  $t \geq 1/(s - 1)$ , this segment has length  $> 1$  and thus contains a lattice point. Likewise, if  $r_t(0, 1) = (1 - t^2)/2t$  then there are no lattice points in the triangle interior

$$0 < a + t < s, \quad 0 < b - 2ta - t^2 < \frac{1 - t^2}{t}(a + t) \quad (3.27)$$

whose intersection with  $a = s - 1$  is a line segment of length  $> (s - 1)(t^{-1} - t)$ . If  $t^{-1} - t \geq 1/(s - 1)$ , this segment has length  $> 1$  and thus contains a lattice point. ■

We may thus ignore  $(0, 0)$  and  $(0, 1)$ , since doing this changes  $L'_N(t)$  by at most  $O(1/s)$  and thus does not affect the asymptotics of  $|I'_N(t)|$ .

We next apply an area-preserving affine linear transformation to  $\mathbb{R}^2$  that maps the fixed vertex  $(-t, -t^2)$  to the origin, and the region defined by (3.25) to a triangle depending only on  $c_-, c_+$  but not on  $s$  and  $t$ . Our transformation is:

$$w_1 = s(b - 2ta - t^2), \quad w_2 = (a + t)/s, \quad (3.28)$$

or in matrix form

$$\begin{pmatrix} w_1 \\ w_2 \\ 1 \end{pmatrix} = \begin{pmatrix} s & -2st & -st^2 \\ 0 & 1/s & t/s \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b \\ a \\ 1 \end{pmatrix} = A_s \sigma(t) \begin{pmatrix} b \\ a \\ 1 \end{pmatrix}, \quad (3.29)$$

where  $A_s = \text{diag}(s, 1/s, 1)$  as in (2.1), and

$$\sigma(t) := U(2t) = \begin{pmatrix} 1 & -2t & -t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} = \exp \begin{pmatrix} 0 & -2t & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{pmatrix} \quad (3.30)$$

is in our unipotent subgroup defined in equation (2.4). Under this transformation, the triangle defined by (3.25) maps to the triangle

$$\Delta_{c_-, c_+} := \{(w_1, w_2) \in \mathbb{R}^2 : 0 < w_2 < 1, 2c_-w_2 < w_1 < 2c_+w_2\}, \quad (3.31)$$

and  $\mathbb{Z}^2$  maps to a translate  $\Lambda_{s^2}(t)$  of a unimodular (covolume-1) lattice:

$$\Lambda_{s^2}(t) := \left\{ (w_1, w_2) \in \mathbb{R}^2 : \begin{pmatrix} w_1 \\ w_2 \\ 1 \end{pmatrix} \in A_s \sigma(t) \begin{pmatrix} \mathbb{Z} \\ \mathbb{Z} \\ 1 \end{pmatrix} \right\}. \quad (3.32)$$

**The maximum area  $L(\Lambda)$  of a triangle disjoint from  $\Lambda$ .** For any lattice translate  $\Lambda$  in the  $(w_1, w_2)$ -plane, we let  $L(\Lambda)$  denote the area  $c_+ - c_-$  of the largest triangle of the form  $\Delta_{c_-, c_+}$  disjoint from  $\Lambda$ . Set  $L(\Lambda) = 0$  if there is no such triangle (because  $\Lambda$  contains the point  $(0, w_2)$  for some  $w_2 \in (0, 1)$ ). If  $\Lambda$  is disjoint from  $\{(w_1, w_2) : 0 < w_2 < 1\}$  then  $c_-$  and  $c_+$  are arbitrary, and we set  $L(\Lambda) = +\infty$ . We then have:

**Proposition 3.8** *For each  $s, x$ , the set of  $t \in [0, 1]$  such that  $L(\Lambda_{s^2}(t)) \leq x$  has length  $|I'_{s^2}(x)| + O(1/s)$ .*

We are thus led to study the function  $L(\cdot)$  on the space, which we shall call  $E$ , of all unimodular lattice translates in  $\mathbb{R}^2$ , and the distribution as  $s \rightarrow \infty$  of  $\{\Lambda_{s^2}(t) : t \in [0, 1]\}$  in that space.

### 3.3 Consequences of ergodic theory

In this section we use the ergodic theory results of §2 to obtain a description of the gap distribution function  $F(x)$  in terms of random lattice translates, establishing Corollary 1.3 of the Introduction.

To apply the results of §2, note that  $\Lambda_{s^2}(t)$ ,  $t \in [0, 1]$ , parameterizes the loop of lattice translates  $A_s \cdot S$  considering in Corollary 2.3. By that Corollary, as  $s \rightarrow \infty$ , the loop  $A_s \cdot S$  becomes uniformly distributed in the space  $E$  of all lattice translates. Combining this result with properties of the function  $L : E \rightarrow \mathbb{R}$  defined above, we obtain:

**Proposition 3.9** *For all  $x \in [0, \infty)$ , we have*

$$|I'_{s^2}(x)| \rightarrow \mu_E(\{\Lambda \in E : L(\Lambda) \leq x\}) \quad (3.33)$$

*as  $s \rightarrow \infty$  through positive integers, and the convergence is uniform on bounded subsets of  $\{x : x \geq 0\}$ . If  $F(\cdot)$  is continuous then*

$$\lim_{N \rightarrow \infty} |I'_N(x)| = \mu_E(\{\Lambda \in E : L(\Lambda) \leq x\}) \quad (3.34)$$

*for all  $x \in [0, \infty)$  and the convergence is uniform in  $x$ .*

**Proof.** To prove the first part, let

$$E_x = \{\Lambda \in E : L(\Lambda) \leq x\}.$$

We claim that  $\mu_E(\partial E_x) = 0$ . Indeed, the function  $L : E \rightarrow [0, \infty]$  is a submersion at most points of  $E$ , so its level sets have measure zero. It fails

to be a submersion only when the lattice translate  $\Lambda$  contains  $(0,0)$  or a point on the horizontal edge  $w_2 = 1$  of its maximal triangle. These lattices meet  $\overline{E_x}$  in a closed set of measure zero, and hence  $\partial E_x$  has measure zero.

Now let  $f : E \rightarrow \{0,1\}$  be the indicator function of  $E_x$  (defined by  $f(\Lambda) = 1$  iff  $\Lambda \in E_x$ .) By Proposition 3.8, to establish equation (3.33) it suffices to show that

$$\int_0^1 f(\Lambda_{s^2}(t)) dt \rightarrow \mu_E(E_x)$$

as  $s \rightarrow \infty$ . This statement would follow immediately from Corollary 2.3 if  $f$  were continuous and compactly supported. In the case at hand,  $f$  is only discontinuous on  $\partial E_x$ , a closed set of measure zero, so the same conclusion readily follows by approximating  $f$  and  $1 - f$  by functions in  $C_0(E)$ .

The first part of the Proposition, equation (3.33), is thus established. For the rest, note that if  $F(\cdot)$  is continuous then so is  $\lambda_\infty(\cdot)$  by equation (3.3). Then equation (3.33) yields (3.34) via Proposition 3.2 and Lemma 3.1. Uniformity over all of  $[0, \infty)$  then follows from the fact that each  $|I_N(x)|$ , and thus necessarily also  $\lim_{s \rightarrow \infty} |I_N(x)|$ , is a nondecreasing function of  $x$  and approaches 1 as  $x \rightarrow \infty$ . ■

**Proof of Corollary 1.3.** While the description (3.34) of  $\lim_{N \rightarrow \infty} |I_N|$  in terms of  $\mu_E$  does give an answer of sorts to the question of the asymptotic distribution of gaps in  $\{\sqrt{n} \bmod 1\}$ , this answer is not yet in a form conducive to either computational or theoretical investigation: we can neither easily compute say  $\lim_{N \rightarrow \infty} |I_N(1)|$ , nor readily deduce the existence and continuity of the asymptotic normalized gap distribution  $F(x)$ . This distribution  $F$  is related with  $\lim_{N \rightarrow \infty} |I_N|$  via equation (3.6), which must be differentiated with respect to  $x$  to recover  $F$ . But it is not yet clear even that the derivative exists. We next reformulate our description of  $\lim_{N \rightarrow \infty} |I_N|$  to make  $F$  visible.

For real  $c_-, c_+$  such that  $c_- < 0 < c_+$ , consider the subset  $S_{c_-, c_+}$  of  $E$  consisting of lattice translates  $\Lambda$  with a point in  $\Delta_{c_-, c_+}$ . The measure of this set depends only on the area  $c_+ - c_-$  of  $\Delta_{c_-, c_+}$ , because all triangles of the same area are equivalent under  $\text{ASL}_2(\mathbb{R})$ , and  $\mu_E$  is invariant under  $\text{ASL}_2(\mathbb{R})$ . We may thus define a nondecreasing function  $p : [0, \infty) \rightarrow [0, 1]$  by

$$p(c_+ - c_-) = \mu_E(S_{c_-, c_+}) \tag{3.35}$$

for all  $c_-, c_+$  such that  $c_- < 0 < c_+$ . We have  $p(0) = 0$  and  $p(+\infty) = 1$ .

We shall presently show that  $p$  has a continuous second derivative. Assuming this for the moment, we can now prove Corollary 1.3 of the Introduction, giving a simple formula for  $F(x)$ :

**Lemma 3.10** *Assume that  $p$  is twice differentiable and that  $p''$  is continuous. Then equation (3.3) holds for*

$$F(x) = -p''(x). \quad (3.36)$$

**Proof.** We rewrite the right-hand side of equation (3.34) in terms of  $p$ :

$$\begin{aligned} \mu_E(\{\Lambda \in E : L(\Lambda) < x\}) &= \mu_E(S_{0,x}) - \\ &\lim_{M \rightarrow \infty} \sum_{j=0}^{M-1} [\mu_E(S_{(j+1)x/M-x, jx/M}) - \mu_E(S_{jx/M-x, jx/M})] \quad (3.37) \\ &= p(x) - \lim_{M \rightarrow \infty} M \left( p(x) - p\left(x - \frac{x}{M}\right) \right). \end{aligned}$$

That is,

$$\mu_E(\{\Lambda \in E : L(\Lambda) < x\}) = p(x) - xp'(x). \quad (3.38)$$

Now  $p(0) = 0$ , so  $p(x) - xp'(x)$  vanishes at zero (this can be seen also from equation (3.38)), and

$$\frac{d}{dx}(p(x) - xp'(x)) = -xp''(x)$$

provided  $p''$  exists. This together with the integral formula (3.6) yields (3.36).  $\blacksquare$

**Integral formula for  $F(x)$ .** To complete the proof of Corollary 1.3, we will show  $p''(x)$  exists and express it as a double integral, leading in the next section to an explicit formula for  $F(x)$ .

We must first recall some facts about primitive lattice vectors and the measure on the space  $B$  of unimodular lattices. A vector  $w$  in some lattice  $\Lambda^0$  is said to be *primitive* if  $w/k \notin \Lambda^0$  for each  $k > 1$ . Equivalently,  $w \in \Lambda^0$  is primitive if and only if there exists  $w' \in \Lambda^0$  such that  $\{w, w'\}$  is a  $\mathbb{Z}$ -basis for  $\Lambda^0$ . For any nonzero  $w \in \mathbb{R}^2$ , the lattices having  $w$  as a primitive vector constitute a circle (a closed horocycle)  $Z_w$  in  $B$ : such a lattice is determined by  $w'$  with  $\det(w, w') = 1$ , and two such  $w'$  determine the same lattice if and only if they differ by an integer multiple of  $w$ . If  $K \subset \mathbb{R}^2$  is a bounded convex set then the area of  $K$  is  $\zeta(2)$  times the integral over  $B$  of the function  $f_K$

taking any lattice  $\Lambda^0$  to the number of primitive vectors in  $K$ . In particular, if  $K$  is small enough that  $f_K(\Lambda^0) \leq 1$  for all  $\Lambda^0 \in B$ , then the lattices with a primitive vector in  $K$  constitute a subset of  $B$  whose measure is  $1/\zeta(2)$  times the area of  $K$ . Moreover, we can recover the measure of any measurable  $B_1 \subset B$  as

$$\mu_B(B_1) = \frac{1}{\zeta(2)} \int_{w \in K} \mu_w(B_1 \cap Z_w). \quad (3.39)$$

Here  $\mu_w$  is the uniform measure on the circle  $Z_w$ , normalized to  $\mu_w(Z_w) = 1$ ; and  $B_1 \cap Z_w$  is a measurable subset of  $Z_w$  for almost all  $w \in K$ .

Now by equation (3.35) the second derivative  $p''(x)$ , if it exists, equals the value of  $(\partial^2/\partial c_- \partial c_+) \mu_E(S_{c_-, c_+})$  at any  $c_-, c_+$  with  $c_- < 0 < c_+$  and  $c_+ - c_- = x$ . Thus  $F(x)$  has the following geometrical interpretation:  $F(c_+ - c_-) dc_- dc_+$  is the measure of the set of lattice translates  $\Lambda \in E$  that intersect  $\Delta_{c_-, c_+}$  in exactly two points, one with  $w_1/2w_2 \in (c_-, c_- + dc_-)$ , the other with  $w_1/2w_2 \in (c_+ - dc_+, c_+)$ .

The difference between these two points of  $\Lambda$  must be a primitive vector, else  $\Lambda$  would contain another point on the line segment joining them, and thus in  $\Delta_{c_-, c_+}$  (because a triangle is convex). Using the formula (3.39) for  $\mu_B$ , we express  $F(x)$  as a double integral over the  $w_2$  coordinates  $v_-, v_+$  of the two vectors where  $\Lambda$  intersects  $\Delta_{c_-, c_+}$ . This lets us complete the proof of Corollary 1.3 by proving the hypotheses of Lemma 3.10.

For  $v_-, v_+ \in (0, 1)$ , let  $w = (c_+ v_+, v_+) - (c_- v_-, v_-)$  be the difference between the two vectors on the boundary of  $\Delta_{c_-, c_+}$ . Then  $Z_w$  parameterizes unimodular lattice translates containing these two vectors. Let  $q_x(v_-, v_+) \in [0, 1]$  be the measure of the subset of  $Z_w$  on which this lattice translate is disjoint from the interior of  $\Delta_{c_-, c_+}$ . We write  $q_x$  rather than  $q_{c_-, c_+}$ , again because it depends only on  $x = c_+ - c_-$ .

**Proposition 3.11** *The function  $(x, v_-, v_+) \mapsto q_x(v_-, v_+)$  on  $[0, \infty) \times (0, 1) \times (0, 1)$  is continuous except on a set contained in  $\{v_- = v_+\}$ . For  $x \in [0, \infty)$  we have*

$$-p''(x) = F(x) = \frac{1}{\zeta(2)} \int_{v_+=0}^1 \int_{v_-=0}^1 4v_- v_+ q_x(v_-, v_+) dv_- dv_+, \quad (3.40)$$

and  $F(x)$  is a continuous function of  $x$ .

**Proof.** The first continuity claim is geometrically evident: our subset of  $Z_w$  varies continuously in  $x, v, v'$  except possibly when  $w$  is horizontal and

thus parallel to the third side of  $\Delta_{c_-,c_+}$ . Since  $q_x(v, v')$  is also bounded, the double integral (3.40) exists and varies continuously with  $x$ . That it equals both  $F(x)$  and  $-p''(x)$  now follows from the geometrical description of  $F(x)$ . The factor  $4v_-v_+$  is the product of the lengths of the line segments

$$\{(w_1, w_2) : w_2 = v_-, 2c_-v_- < w_1 < 2(c_- + dc_-)v_-\}, \quad (3.41)$$

$$\{(w_1, w_2) : w_2 = v_+, 2(c_+ - dc_+)v_+ < w_1 < 2c_+v_+\}, \quad (3.42)$$

on which our vectors lie, and formula (3.39) accounts for the  $q_x(v_-, v_+)/\zeta(2)$  factor.  $\blacksquare$

### 3.4 Formulas for the gap distribution

In this section we complete the proof of Theorem 1.1 by giving, in Theorem 3.14, a closed formula for the gap distribution  $F(x)$ .

We can compute  $q_x$  and the integral in equation (3.40) in closed form. We find:

**Lemma 3.12** *For all  $v, v' \in (0, 1]$  and  $x > 0$  we have*

$$q_x(v, v') = q_x(v', v). \quad (3.43)$$

If  $v \geq v'$  then

$$q_x(v, v') = \max \left( 0, \min \left( 1, \frac{r}{vv'} \right) - \max \left( 0, \frac{v(1-v') - r}{v(v-v')} \right) \right), \quad (3.44)$$

where

$$r := \frac{1}{2x} \quad (3.45)$$

and  $\max(0, (v(1-v') - r)/(v(v-v')))$  is interpreted as  $+\infty$  if  $v = v'$  and  $r < v(1-v')$  and as 0 if  $v = v'$  and  $r \geq v(1-v')$ .

**Proof.** Change coordinates linearly from  $w_1, w_2$  to  $z, z'$ , chosen so that  $\Delta_{c_-,c_+}$  becomes the fixed triangle interior

$$\Delta_0 := \{(z, z') \in \mathbb{R}^2 : z > 0, z' > 0, z + z' < 1\} \quad (3.46)$$

of an isosceles right triangle with unit sides, whose closure intersects the lattice translate  $\Lambda$  at

$$(z, z') = (v, 0), (0, v'). \quad (3.47)$$

Since  $\Delta_0$  has area  $1/2$ , the linear transformation from  $(w_1, w_2)$  to  $(z, z')$  multiplies areas by a factor  $1/2x = r$ . Thus in the  $(z, z')$  plane  $\Lambda$  is a lattice translate of covolume  $r$ . Therefore it is generated by  $(v, 0)$ ,  $(0, v')$  and a third vector on the line

$$vz' + v'z = vv' + r \quad (3.48)$$

through the two points (3.47), determined up to translation by integer multiples of their difference  $(v, -v')$ . The proportion of such lattices disjoint from  $\Delta_0$  is clearly invariant under  $v \leftrightarrow v'$ . This establishes equation (3.43). It follows that the computation of  $q_x(v, v')$  for  $v \geq v'$  will yield  $q_x(v, v')$  for all  $v, v'$ .

Assume, then, that  $v \geq v'$ . The lattice translate  $\Lambda$  is determined by the  $z$ -coordinate of our vector  $(z, z')$  on the line given by (3.48), with two choices of  $z$  yielding the same lattice if and only if they differ by an integer multiple of  $v$ . We may thus parameterize  $\Lambda$  by  $[0, v)$ . We claim that  $\Lambda$  is disjoint from  $\Delta_0$  if and only if neither  $(z, z')$  nor  $(v - z, v' - z')$  is in  $\Delta_0$ . “Only if” is clear because both  $(z, z')$  and  $(v - z, v' - z')$  are in  $\Lambda$ . To prove “if”, observe that the general vector in  $\Lambda$  is

$$(0, v') + m_1(v, -v') + m_2((z, z') - (0, v')) \quad (3.49)$$

for some  $m_1, m_2 \in \mathbb{Z}$ . We are to show that if the intersection  $\Lambda \cap \Delta_0$  is nonempty then it contains the vector given by (3.49) with  $(m_1, m_2) = (0, 1)$  or  $(1, -1)$ . We first reduce to the case  $m_2 = \pm 1$ . If  $m_2 = 0$  then the vector is on the line joining  $(v, 0)$  and  $(0, v')$  and thus cannot be in  $\Delta_0$ . If  $m_2 > 1$  then  $(z, z') + [m_1/m_2](v, -v)$ , of the same form but with  $m_2 = 1$ , is a convex combination of the three vectors (3.47), (3.49), and is thus also contained in  $\Delta_0$ . Likewise if  $m_2 < -1$  we find a vector in  $\Lambda \cap \Delta_0$  with  $m_2 = -1$ . Now if  $m_2 = -1$  then the  $z$ -coordinate  $m_1v - z$  of (3.49) must exceed 0 (because the vector is in  $\Delta_0$ ) but be smaller than  $v$  (since it is on the line  $vz' + v'z = vv'$ , and  $r < vv'$  while  $vz' > 0$ ). Since  $0 \leq z < v$ , this forces  $m_1 = 1$ . If  $m_2 = +1$  then the  $z$ -coordinate is  $m_1v + z$ , and its positivity together with  $z < v$  force  $m_1 \geq 0$ . If  $m_1 > 0$  then  $(z, z')$  itself (with  $m_1 = 0$ ) is in  $\Delta_0$ , because it differs from (3.49) by  $m_1(v, -v')$  and  $v \geq v'$ .

Now, given that  $z \in [0, v)$  and that  $(z, z')$  is on the line (3.48), the condition  $(v - z, v' - z') \notin \Delta_0$  is equivalent to  $z \leq r/v'$ , while  $(z, z') \notin \Delta_0$  becomes  $z \geq (v'(1 - v) - r)/(v - v')$ . (If  $v = v'$  the quotient is interpreted as  $+\infty$  if the numerator is positive,  $-\infty$  if not.) The first condition leaves the interval  $0 \leq z \leq \min(v, r/v')$ , and the second condition leaves

$$\max\left(0, \frac{v'(1 - v) - r}{v - v'}\right) \leq z \leq \min(v, r/v'). \quad (3.50)$$

The length of this interval, divided by the length  $v$  of  $[0, v)$ , is given by the right-hand side of equation (3.44). Since this ratio is  $q_x(v, v')$ , we are done. ■

**Corollary 3.13** *If  $x \leq 1/2$  then  $q_x(v, v') = 1$  for all  $v, v' \in (0, 1]$ . If  $x \geq 2$  then  $q_x(1/2, 1/2) = 0$ . If  $1/2 < x < 2$  then  $q_x(v, v')$  is positive for all  $v, v' \in (0, 1]$  but does not equal 1 identically.*

**Proof.** The three  $x$  ranges are equivalent to  $r \geq 1$ ,  $r \leq 1/4$ , and  $1 > r > 1/4$  respectively. If  $r \geq 1$  then  $r/vv' \geq 1$  and  $v(1 - v') \leq r$ . Thus (3.44) reduces to  $\max(0, 1 - 0) = 1$ . If  $r \leq 1/4$  then  $v(1 - v') > r$  at  $v = v' = 1/2$ , so the formula (3.44) for  $q_x(1/2, 1/2)$  reduces to 0. For any  $r < 1$  we have  $q_x(1, 1) < 1$ . If  $q_x(v, v') = 0$  then  $(v(1 - v') - r)/(v(v - v'))$  is either  $\geq 1$  or  $\geq r/vv'$ . But

$$1 - \frac{v(1 - v') - r}{v(v - v')} = \frac{r - (v - v^2)}{v(v - v')}, \quad (3.51)$$

$$\frac{r}{vv'} - \frac{v(1 - v') - r}{v(v - v')} = \frac{r - (v' - v'^2)}{v'(v - v')}. \quad (3.52)$$

Thus  $r \leq 1/4$  in either case. ■

**Phase transition at  $x = 1/2$ .** The fact that  $q_x(v, v') = 1$  for  $x \leq 1/2$  can also be seen geometrically: if  $\Lambda$  has a point in  $\Delta_0$ , then that point together with  $(v, 0)$  and  $(0, v')$  span a triangle whose area is a positive multiple of  $r/2$ , but strictly less than the area  $1/2$  of  $\Delta_0$ ; therefore  $r < 1$ . This yields the fact that  $F(x) = 1/\zeta(2)$  for  $x \leq 1/2$  (the first part of the following theorem). Once  $x > 1/2$ , it is possible for  $\Lambda$  to meet  $\Delta_0$ , and thus  $q_x(v, v')$  does not equal 1 identically. This explains the phase transition of  $F(x)$  at  $x = 1/2$ .

Lemma 3.12 reduces the computation of  $F(x)$  to a calculus exercise, simplified somewhat by Corollary 3.13. We obtain:

**Theorem 3.14** *i) If  $x \leq 1/2$  then  $F(x) = 1/\zeta(2)$ .  
ii) If  $1/2 \leq x \leq 2$ , let  $r = 1/2x$  as in (3.45) and*

$$\psi(r) = \tan^{-1} \frac{2r - 1}{\sqrt{4r - 1}} - \tan^{-1} \frac{1}{\sqrt{4r - 1}}. \quad (3.53)$$

Then

$$F(x) = \frac{1}{\zeta(2)} \left( \frac{2}{3}(4r - 1)^{3/2} \psi(r) + (1 - 6r) \log r + 2r - 1 \right). \quad (3.54)$$

iii) If  $x \geq 2$ , let  $r = 1/2x$  again and

$$\alpha = \frac{1}{2}(1 - \sqrt{1 - 4r}), \quad (3.55)$$

the smaller root of  $\alpha - \alpha^2 = r$ . Then

$$F(x) = \frac{1}{\zeta(2)} \left( 4(1 - 4\alpha)(1 - \alpha)^2 \log(1 - \alpha) - 2(1 - 2\alpha)^3 \log(1 - 2\alpha) - 2\alpha^2 \right). \quad (3.56)$$

**Remark.** We have

$$\tan \psi(r) = \frac{(r - 1)\sqrt{4r - 1}}{(3r - 1)}, \quad (3.57)$$

by the addition formula for the tangent; but we cannot define  $\psi(r)$  as  $\tan^{-1}((r - 1)\sqrt{4r - 1}/(3r - 1))$ , because  $\psi(r)$  is not the principal value of this arctangent for  $r \leq 1/3$ .

**Proof.** By equations (3.40) and (3.43), we have

$$F(x) = \frac{1}{\zeta(2)} \iint_{0 < v' < v < 1} 8vv'q_x(v, v') dv' dv. \quad (3.58)$$

Case (i) is easy: by Corollary 3.13,  $q_x = 1$  identically, so the double integral is just  $\int_0^1 8v(v^2/2) dv = 1$ .

In case (ii), the last part of Corollary 3.13 simplifies our formula (3.44) to

$$1 - q_x(v, v') = \max\left(0, 1 - \frac{r}{vv'}\right) + \max\left(0, \frac{v(1 - v') - r}{v(v - v')}\right). \quad (3.59)$$

Thus

$$1 - \zeta(2)F(x) = \iint_{0 < v' < v < 1} 8vv' \left[ \max\left(0, 1 - \frac{r}{vv'}\right) + \max\left(0, \frac{v(1 - v') - r}{v(v - v')}\right) \right] dv' dv. \quad (3.60)$$

By symmetry,

$$\iint_{0 < v' < v < 1} 8vv' \max\left(0, 1 - \frac{r}{vv'}\right) dv' dv = \iint_{0 < v, v' < 1} 4vv' \max\left(0, 1 - \frac{r}{vv'}\right) dv' dv. \quad (3.61)$$

This in turn equals

$$\begin{aligned} \int_r^1 4v \left( \int_{r/v}^1 v' \left( 1 - \frac{r}{vv'} \right) dv' \right) dv &= \int_r^1 2(v-r)^4 \frac{dv}{v} \\ &= (3r-1)(r-1) - 2r^2 \log r. \end{aligned} \quad (3.62)$$

For the second term in (3.60) we integrate

$$8vv' \frac{v(1-v')-r}{v(v-v')} = \frac{8v'(v(1-v')-r)}{v-v'} \quad (3.63)$$

over the region  $r < v < 1$ ,  $v(1-v') > r$ . In this region  $v' < v$  holds automatically, else we would have

$$r < v(1-v') < v(1-v) \leq 1/4, \quad (3.64)$$

contradicting  $r \in [1/4, 1]$ . We thus obtain

$$\begin{aligned} &\int_r^1 \left( \int_0^{1-(r/v)} \frac{8v'(v(1-v')-r)}{v-v'} dv' \right) dv \\ &= \int_r^1 \left[ 4v'(2r + v(v' + 2v - 2)) + 8v(r - v + v^2) \log(v - v') \right]_{v'=0}^{1-(r/v)} dv \\ &= \int_r^1 \left( 8v(r - v + v^2) (\log(r - v + v^2) - 2 \log v) + \frac{4(v-r)(r - v + 2v^2)}{v} \right) dv. \end{aligned} \quad (3.65)$$

Evaluating this definite integral, entering its value into equation (3.60), and solving for  $F(x)$  yields (3.54).

Finally, in case (iii) we see from equations (3.51) and (3.52) that the integrand of (3.58) is supported on the union of the regions

$$vv' < r \text{ and either } v < \alpha \text{ or } v > 1 - \alpha, \quad (3.66)$$

where it equals  $8v'(r - (v - v^2))/(v - v')$ , and

$$vv' > r \text{ and either } v' < \alpha \text{ or } v' > 1 - \alpha, \quad (3.67)$$

where it equals  $8v(r - (v' - v'^2))/(v - v')$ . The union of these regions consists of: the triangle  $0 < v' < v < \alpha$ , contained in (3.66); the triangle  $1 - \alpha < v' < v < 1$ , contained in (3.67); and the square  $0 < v' < \alpha$ ,  $1 - \alpha < v < 1$ , split between the two regions along the segment  $1 - \alpha < v < 1$  of the hyperbola  $vv' = r$ . Integrating the appropriate integrand over each subregion, adding the results and dividing by  $\zeta(2)$  yields the formula (3.56) for  $F(x)$  when  $x \geq 2$ .  $\blacksquare$

Theorem 3.14 confirms that  $F(x)$  is a continuous function of  $x$ : at the boundary values  $x = 1/2$  and  $x = 2$  the formula (3.54) of (ii) agrees with the values  $F(1/2) = 1/\zeta(2)$ ,  $F(2) = (\log 2 - \frac{1}{2})/\zeta(2)$  given by (i) and (iii) respectively. Moreover,  $F(x)$  is continuously differentiable, but not smooth or even  $C^2$  at  $x = 2$ : the series expansions near  $x = 2$  begin

$$\begin{aligned} \zeta(2)F(2 - \epsilon) &= \log 2 - \frac{1}{2} + 3 \log 2 \frac{\epsilon}{2} - \sqrt{2\pi} \frac{\epsilon^{3/2}}{6} \\ &+ (12 \log 2 + 3) \frac{\epsilon^2}{16} - \sqrt{2\pi} \frac{\epsilon^{5/2}}{8} + (36 \log 2 + 17) \frac{\epsilon^3}{96} \dots \end{aligned} \quad (3.68)$$

for  $x < 2$  and

$$\begin{aligned} \zeta(2)F(2 + \epsilon) &= \log 2 - \frac{1}{2} - 3 \log 2 \frac{\epsilon}{2} - \sqrt{2}(8 - 3 \log(2\epsilon)) \frac{\epsilon^{3/2}}{12} \\ &+ (12 \log 2 + 3) \frac{\epsilon^2}{16} - \sqrt{2}(32 - 15 \log(2\epsilon)) \frac{\epsilon^{5/2}}{80} - (36 \log 2 + 17) \frac{\epsilon^3}{96} \dots \end{aligned} \quad (3.69)$$

for  $x > 2$ , first differing in the  $|x - 2|^{3/2}$  terms. (The coincidence of the coefficients of integral powers of  $x - 2$  persists, as can eventually be shown with some manipulation of the formulas (3.54) and (3.56); it is not clear what significance if any this might have.) At  $x = 1/2$ , the function given by (3.54) first differs from the constant  $1/\zeta(2)$  in the cubic term:

$$\zeta(2)F(2 + \epsilon) = 1 - \frac{16}{3}\epsilon^3 + 24\epsilon^4 - \frac{384}{5}\epsilon^5 + O(\epsilon^6) \quad (3.70)$$

for small  $\epsilon \geq 0$ . Finally, as  $x \rightarrow \infty$  we obtain a Taylor expansion in powers of  $1/x$ :

$$\zeta(2)F(x) = \frac{1}{2x^3} + \frac{9}{16x^4} + \frac{11}{16x^5} + \frac{175}{192x^6} + O(x^{-7}). \quad (3.71)$$

### 3.5 Generalizations

We can generalize our results simultaneously in three directions. First, instead of gaps in  $\{\sqrt{n} \bmod 1 : 1 \leq n \leq N\}$ , we can analyze gaps in the fractional parts of the square roots of the integers in  $(\theta^2 N, N]$  for fixed  $\theta \in (0, 1)$ . Second, we can analyze the joint distribution of two or more consecutive gaps. Third, for each positive rational  $u$ , we can replace  $\{\sqrt{n}\}$  by  $\{\sqrt{un}\}$ . In each case our methods yield answers, sometimes (notably for three or more consecutive gaps) surprising answers. We next briefly discuss each of these generalizations.

**Gaps in  $\{\sqrt{n} \bmod 1 : \theta^2 N < n \leq N\}$ .** This is the easiest generalization: most of the analysis carries through except that the integers  $a$  or  $a_j$  in (3.7) and later must be in the range  $[\theta s, s)$  rather than  $[1, s)$ . (This is why we chose the lower limit  $\theta^2 N$  rather than  $\theta N$ ). This has the effect of replacing the triangles  $\Delta_{c_-, c_+}$  by the truncated triangles (trapezoids)

$$\Delta_{c_-, c_+}^{(\theta)} := \{(w_1, w_2) \in \mathbb{R}^2 : \theta < w_2 < 1, 2c_- w_2 < w_1 < 2c_+ w_2\}. \quad (3.72)$$

We find:

**Theorem 3.15** *Let  $p^{(\theta)}(t)$  be the probability that a random unimodular lattice translate meets a given trapezoid of the form (3.72) with  $c_+ - c_- = t$ . Then  $-p^{(\theta)}$  has a continuous second derivative  $F^{(\theta)}$ , which is the asymptotic gap distribution for  $\{\sqrt{n} \bmod 1 : \theta^2 N < n \leq N\}$ . That is, for any interval  $[t_0, t_1] \subset [0, \infty)$  the number of gaps whose length falls in  $[t_0/N, t_1/N]$  is asymptotic to  $N \int_{t_0}^{t_1} F^{(\theta)}(t) dt$  as  $N \rightarrow \infty$ . Moreover,  $F^{(\theta)}$  is given by a double integral formula (see (3.73) below); it is piecewise analytic on  $[0, \infty)$ , restricts to the constant function  $(1 - \theta^2)^2/\zeta(2)$  on  $[0, 1/(2 - 2\theta)]$ , and is asymptotically proportional to  $t^{-3}$  for large  $t$ .*

**Proof** (sketch). We obtain this in much the same way as we did for  $\theta = 0$ . Let  $x = c_+ - c_-$ , and for  $v_-, v_+ \in (\theta, 1)$  let  $w$  be the difference  $(c_+ v_+, v_+) - (c_- v_-, v_-)$  between two vectors on the boundary of  $\Delta_{c_-, c_+}^{(\theta)}$ . Define  $q_x^{(\theta)}(v_-, v_+)$  to be the measure of the subset of  $Z_w$  that parameterizes unimodular lattice translates containing those two vectors and disjoint from the interior of  $\Delta_{c_-, c_+}^{(\theta)}$ . (As was true for the triangles  $\Delta_{c_-, c_+}$ , all trapezoids  $\Delta_{c_-, c_+}^{(\theta)}$  with the same value of  $x$  are equivalent under  $\text{ASL}_2(\mathbb{R})$ , so the measure of that subset depends only on  $x$ .) We then generalize Proposition (3.11), again with much the same proof, by replacing  $q_x$  with  $q_x^{(\theta)}$  and (3.40) with the double integral formula

$$F^{(\theta)}(x) = \frac{1}{\zeta(2)} \int_{v_+=\theta}^1 \int_{v_-=\theta}^1 4v_- v_+ q_x^{(\theta)}(v_-, v_+) dv_- dv_+. \quad (3.73)$$

Since the largest triangles contained in  $\Delta_{c_-, c_+}^{(\theta)}$  have area  $(1 - \theta)x$ , we find that there exist  $v_-, v_+$  such that  $q_x^{(\theta)}(v_-, v_+) < 1$  if and only if  $2x > 1/(1 - \theta)$ . Hence  $F^{(\theta)}(x) = \frac{1}{\zeta(2)} \int_{v_+=\theta}^1 \int_{v_-=\theta}^1 4v_- v_+ dv_- dv_+ = (1 - \theta^2)^2/\zeta(2)$  for  $x \leq 1/(2 - 2\theta)$ , and  $F^{(\theta)}(x) < (1 - \theta^2)^2/\zeta(2)$  for  $x > 1/(2 - 2\theta)$ . The remaining properties of  $F^{(\theta)}$  also follow from the integral formula (3.73).

We have not attempted to generalize Lemma (3.12) and Theorem (3.14) by computing  $q_x^{(\theta)}(v, v')$  and  $F^{(\theta)}$  explicitly for any  $\theta > 0$ . The symmetry (3.43) does generalize to  $q_x^{(\theta)}(v, v') = q_x^{(\theta)}(v', v)$ , since  $\Delta_{c_-, c_+}^{(\theta)}$  retains the affine bilateral symmetry of  $\Delta_{c_-, c_+}$ . But the general formula for  $q_x^{(\theta)}(v, v')$  will likely be somewhat more complicated than our formula (3.44) for  $q_x(v, v')$ , making the double integral (3.73) even less pleasant to evaluate than it was for  $\theta = 0$ .

**Joint distribution of consecutive gaps.** In Section 1 we introduced the notation  $J_1, \dots, J_N$  for the gaps left over when the circle  $\mathbb{R}/\mathbb{Z}$  is cut at the points  $\{\sqrt{1}\}, \{\sqrt{2}\}, \dots, \{\sqrt{N}\}$ . Let us index these  $J_i$  in their order of appearance on the circle, so that  $J_{i+1}$  is the next gap after  $J_i$ . Theorem 1.1 describes the asymptotic distribution of the normalized gap lengths  $N|J_i|$  in  $[0, \infty)$ . More generally we can study the joint distribution of consecutive gaps: fix a positive integer  $r$ , and consider the  $r$ -tuples  $(N|J_{i+1}|, N|J_{i+2}|, \dots, N|J_{i+r}|) \in [0, \infty)^r$ . Our methods also yield the asymptotic behavior of this joint distribution for each  $r$ . We show:

**Theorem 3.16** *For each  $r = 1, 2, 3, \dots$ , there exists a nonnegative measure  $\Phi$  on  $[0, \infty)^r$  that is an asymptotic joint distribution of  $r$  consecutive gaps in  $\{\{\sqrt{n}\} : 1 \leq n \leq N\}$ . That is, for any box  $B \subset [0, \infty)^r$ , the number of  $r$ -tuples of consecutive gaps whose lengths lie in  $N^{-1}B$  is asymptotic to  $N \int_B d\Phi$  as  $N \rightarrow \infty$ .*

When  $r = 1$  we have  $d\Phi = F(t) dt$  where  $F$  is the distribution function of Theorem 1.1. Once  $r > 1$ , we do not obtain explicit formulas for  $\Phi$ , but can still describe it in terms of lattice translates. Using this description we can show that, as for  $r = 1$ , the distribution differs qualitatively from what one would expect when cutting the circle at  $N$  random points, and indeed the differences become more striking as  $r$  increases.

We already noted that when the circle is cut at  $N$  random points, the resulting gaps are exponentially distributed almost surely as  $N \rightarrow \infty$ . More generally, for each  $r$ , the  $r$ -tuples of consecutive gaps almost surely approach the product of  $r$  exponential distributions. In other words, there is no dependence among nearby gaps. But when the cuts are at  $\{\sqrt{n}\}$  ( $0 < n < N$ ), nearby gaps are markedly dependent, and the dependencies become more pronounced as  $r$  increases. For example, fix  $r$  and a positive real  $T < 1/2$ . Then  $\sum_{j=1}^r N|J_{i+j}| < T$  holds for a positive proportion of  $i \in [0, N - r]$ , and for each of these  $i$ , any two of the  $r$  consecutive normalized gaps  $N|J_{i+j}|$  ( $1 \leq j \leq r$ ) determine all  $r$  of them to within  $O(1/N)$ . In other words, the

simplex

$$\left\{ (t_1, \dots, t_r) : t_j > 0 \text{ (each } j), \sum_{j=1}^r t_j < 1/2 \right\} \quad (3.74)$$

holds a positive fraction of the asymptotic joint distribution of  $r$  consecutive gaps, and this part of the distribution is supported on a 2-dimensional manifold. In particular, this part of the distribution is singular once  $r > 2$ . Moreover, the entire distribution is supported on a union of manifolds of dimension at most 5 in  $[0, \infty)^r$ , and is thus singular once  $r > 5$ .

The number 5 arises as the dimension of the space  $E$  of unimodular lattice translates. Let  $\Lambda$  be a random lattice translate. Our analysis of the case  $r = 1$  led to a relation between the asymptotic distribution of  $N|J_i|$  and the distribution of the largest negative and smallest positive values  $c_-, c_+$  of  $w_1/2w_2$  in the intersection of  $\Lambda$  with the strip  $\{(w_1, w_2) : 0 < w_2 < 1\}$ . More generally, for any  $r$  the distribution of  $N(|J_{i+1}|, |J_{i+2}|, \dots, |J_{i+r}|)$  hinges on the joint distribution of  $c_-$  and the  $r$  smallest positive values of  $w_1/2w_2$ , call them  $c_1, c_2, \dots, c_r$ , in the intersection of  $\Lambda$  with the same strip. Note that  $c_1 = c_+$ ; it will be convenient to also set  $c_0 = c_-$ . We relate the distribution of  $(c_0, \dots, c_r)$  with that of  $N(|J_{i+1}|, |J_{i+2}|, \dots, |J_{i+r}|)$  as follows. Recall that to each  $x \in [0, 1]$  we associated a lattice translate  $\Lambda_{s^2}(x)$ , and proved (Theorem 1.2) that the  $\Lambda_{s^2}(x)$  are uniformly distributed on  $E$  as  $N \rightarrow \infty$ , so we may take  $\Lambda_{s^2}(x)$  to be our random lattice translate  $\Lambda$ . We showed that, for  $x$  outside a subset of  $[0, 1]$  of length  $o(1)$ , the gap containing  $x$  has length  $N^{-1}(c_1 - c_0 + o(1))$ . Let that gap be  $J_1$ . Then our analysis readily generalizes to show that for each  $j = 1, \dots, r$  we have  $N|J_{i+j}| = c_j - c_{j-1} + o(1)$  for almost all  $x \in [0, 1]$ . Since each  $c_j$  is a smooth function on the complement of a measure-zero subset of  $E$ , it follows that the asymptotic distribution  $\Phi$  of  $N(|J_{i+1}|, |J_{i+2}|, \dots, |J_{i+r}|)$  is supported on a countable union of manifolds in  $[0, \infty)^r$  each of dimension at most  $\dim E = 5$ .

Moreover, if  $c_r - c_0 < 1/2$  then the  $r + 1$  points of  $\Lambda \cap \Delta_{c_0, c_r}$  must be collinear, again because any three non-collinear points in  $\Lambda$  are vertices of a triangle of area at least  $1/2$ , which cannot be contained in the triangle  $\Delta_{c_0, c_r}$  of strictly smaller area  $c_r - c_0$ . These  $r + 1$  points must then be equally spaced along the line they determine. Hence  $\{\tau_j\}$  is the image of a linear progression under a fractional linear transformation, and

$$(t_1, t_2, \dots, t_r) = (c_1 - c_0, c_2 - c_1, \dots, c_j - c_{j-1})$$

is in the image of a smooth function from  $\mathrm{SL}_2(\mathbb{R})/\mathrm{ASL}_1(\mathbb{R})$  to  $\mathbb{R}^r$ . The fact that the coset space  $\mathrm{SL}_2(\mathbb{R})/\mathrm{ASL}_1(\mathbb{R})$  has dimension  $3 - 1 = 2$  then explains why the intersection of the simplex (3.74) with the support of  $\Phi$  has

dimension at most 2. We claimed further that this intersection contains a positive proportion of the distribution, that it has dimension exactly 2 once  $r > 1$ , and that any two of the  $t_j$  map it injectively to  $\mathbb{R}^2$ . The first claim amounts to the easy fact that  $c_r - c_0 < 1/2$  on a subset of positive measure in  $E$ . The remaining claims follow quickly from an explicit parameterization of the  $t_j$ ; for instance, if  $c_j = (Aj + B)/(Cj + D)$  for some  $\begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$  then  $t_j = 1/(Cj + D)(C(j - 1) + D)$ .

Note that the fact that all of the  $N(|J_{i+1}|, |J_{i+2}|, \dots, |J_{i+r}|)$  are within  $o(1)$  of a subset of  $\mathbb{R}^r$  of dimension at most 5 requires only the elementary methods of this section; it is only to obtain their asymptotic distribution that we need the ergodic-theory results of §2.

As in the case of  $r = 1$ , very similar results hold if we fix  $\theta \in (0, 1)$  and cut  $\mathbb{R}/\mathbb{Z}$  at  $\{\sqrt{n}\}$  ( $\theta^2 N < n < N$ ). Again the same analysis applies, with the isosceles trapezoid (3.72) in place of the triangle  $\Delta_0$ . Since the largest triangles accommodated by this trapezoid have area  $(1 - \theta)/2$ , the asymptotic gap distribution now has 2-dimensional support in the simplex  $\sum_{j=1}^r t_j < 1/(2 - 2\theta)$ .

**Gaps in  $\sqrt{un + \nu} \bmod 1$ .** We obtain similar results if we fix a real  $\nu$  and a rational  $u > 0$  and ask for the distribution of gaps, or joint distribution of  $r$  consecutive gaps, in  $\{\sqrt{un + \nu}\}$  ( $0 < un + \nu < N$  or  $\theta^2 N < un + \nu < N$ ). We find again that these distributions, normalized by multiplying each gap by  $N$ , have asymptotic limits as  $N \rightarrow \infty$ . The limit distributions do not depend on  $\nu$  but do in general depend on  $u$ . To construct these distributions we must replace lattice translates by translates of more complicated periodic sets in  $\mathbb{R}^2$ . Still, the distributions for arbitrary  $u$  share the qualitative properties of the distributions described above for  $u = 1$ . For instance, the asymptotic density function for single normalized gaps is constant near  $t = 0$  and has a  $(C + o(1))/t^3$  tail as  $t \rightarrow \infty$ ; and the asymptotic joint distribution of  $r > 1$  consecutive normalized gaps is supported on a union of manifolds of dimension at most 5 in  $[0, \infty)^r$ , of which those that intersect small neighborhoods of the origin has dimension 2.

The argument generalizes our analysis of the case  $(u, \nu) = (1, 0)$ . In that case, we were led to a sequence of curves  $\{\Lambda_{s^2}(x): 0 < x < 1\}$  in the space  $\mathrm{ASL}_2(\mathbb{R})/\mathrm{ASL}_2(\mathbb{Z})$  of lattice translates. When  $u$  is an arbitrary positive rational, the same method leads us to curves in the space  $\mathrm{ASL}_2(\mathbb{R})/\Gamma$  for some congruence subgroup  $\Gamma$  of  $\mathrm{ASL}_2(\mathbb{Z})$ . This space is a finite cover of  $\mathrm{ASL}_2(\mathbb{R})/\mathrm{ASL}_2(\mathbb{Z})$  that parameterizes lattice translates with additional torsion structure. Happily we find that our ergodic result (Theorem 2.2) applies in this more general setting: our curves again approach uniform dis-

tribution in  $\mathrm{ASL}_2(\mathbb{R})/\Gamma$ . We thus establish the asymptotic gap distributions for  $\{\sqrt{un + \nu}\}$ .

Let  $u = u_1/u_2$  with  $u_1, u_2$  coprime positive integers. As in the Introduction, fix  $t > 0$ , and for  $N \gg 0$  consider an interval  $I = [x, x + t/N] \subset \mathbb{R}/\mathbb{Z}$  with  $x$  chosen uniformly at random from  $[0, 1]$ . Then

$$\begin{aligned} & \{\sqrt{un + \nu}\} \in I, \quad \text{some } n \text{ with } 0 \leq un + \nu \leq N \\ \iff & \sqrt{un + \nu} \in I + a, \quad \text{some } a \leq \sqrt{N}, \\ \iff & un + \nu \in (I + a)^2 \approx a^2 - x^2 + 2(a + x)I. \end{aligned}$$

Again we can replace  $(I + a)^2$  by  $a^2 - x^2 + 2(a + x)I$  without changing the asymptotic gap distribution. Now  $un + \nu \in a^2 - x^2 + 2(a + x)I$  if and only if

$$\frac{2(a + x)I - x^2 - \nu}{u} \ni n - \frac{a^2}{u} = n - \frac{u_2}{u_1}a^2.$$

Comparing this with our earlier computation for  $(u, \nu) = (1, 0)$ , we find three differences: first, the triangle  $T$  of (1.3) must be translated by  $(0, -\nu)$ ; second, it must be scaled by  $u^{-1}$  in the vertical direction; third, the lattice  $\mathbb{Z}^2$  must be replaced by

$$\{(a, b) \in \mathbb{R}^2 : a \in \mathbb{Z}, b \in \mathbb{Z} + u^{-1}a^2\} = \omega_{1/u}\mathbb{Z}^2, \quad (3.75)$$

where  $\{\omega_v : v \in \mathbb{R}\}$  is the one-parameter family of transformations of  $\mathbb{R}^2$  defined by

$$\omega_v(x, y) := (x, y + v^2x). \quad (3.76)$$

(Note that  $\omega_v\mathbb{Z}^2$  depends only on the class of  $v \bmod 1$ .)

The first two changes are minor: the translation still yields an asymptotically random unimodular lattice translate, and the scaling divides its volume by  $u$ , which corresponds to the fact that  $\{n : 0 < un + \nu < N\}$  contains  $u^{-1}N + O(1)$  integers rather than  $N + O(1)$ . But equation (3.75) represents a more profound change:  $\omega_{1/u}\mathbb{Z}^2$  is not a lattice on  $\mathbb{R}^2$  unless  $u_1 = 1$  or  $u_1 = 2$ . Still,  $\omega_{1/u}\mathbb{Z}^2$  is a disjoint union of at most  $u_1$  translates of the same lattice, and its stabilizer in  $\mathrm{ASL}_2(\mathbb{R})$ , call it  $\Gamma(u)$ , contains a congruence subgroup of  $\mathrm{ASL}_2(\mathbb{Z})$ , namely the  $3 \times 3$  matrices in  $\mathrm{ASL}_2(\mathbb{Z})$  congruent to the identity mod  $u_1$ . Hence  $\Gamma(u)$  has finite covolume in  $\mathrm{ASL}_2(\mathbb{R})$ , and the moduli space

$$E(u) := \mathrm{ASL}_2(\mathbb{R})/\Gamma(u)$$

of the images of  $\omega_{1/u}\mathbb{Z}^2$  under  $\mathrm{ASL}_2(\mathbb{R})$  inherits a probability measure from Haar measure on  $\mathrm{ASL}_2(\mathbb{R})$ .

We can now describe the asymptotic gap distribution of  $\{\{\sqrt{un + \nu}\} : 0 < un + \nu < N\}$ .

**Theorem 3.17** Fix a positive  $u \in \mathbb{Q}$  and any  $\nu \in \mathbb{R}$ . Let  $p_u(t)$  be the probability that a given triangle  $S_t$  of area  $t$  meets a random image of  $\omega_{1/u}\mathbb{Z}^2$  under  $\text{ASL}_2(\mathbb{R})$ . Then  $F_u(t) := -p''(t)$  is the asymptotic gap distribution for  $\{\sqrt{un + \nu}\}$ . That is, for any interval  $[t_0, t_1] \subset [0, \infty)$ , the proportion of gaps in  $\{\{\sqrt{un + \nu}\} : 0 < un + \nu < N\}$  whose length is contained in  $[t_0u/N, t_1u/N]$  approaches  $\int_{t_0}^{t_1} F_u(t) dt$  as  $N \rightarrow \infty$ .

**Proof.** By our analysis above, it is enough to show that as  $N \rightarrow \infty$  the relevant image of  $[0, 1]$  in  $E(u)$  becomes uniformly distributed. The case  $u = 1$  of this is Theorem 1.2. But the proof of that result, culminating in Theorem 2.2, applies for arbitrary rational  $u > 0$ : the image of  $[0, 1]$  is still  $A_s \cdot \sigma$  for some nonlinear horocycle section  $\sigma$  of finite period  $u_1$ . The remainder of the argument proceeds as before, and we conclude that  $F_u$  gives the asymptotic gap distribution of  $\{\sqrt{un + \nu}\}$  for each  $\nu$ . ■

In general,  $F_u$  does not coincide with the distribution  $F = F_1$  given by the explicit formulas of Theorem 3.14; but it shares the same qualitative behavior. For instance, any three points of  $\omega_u\mathbb{Z}^2$  are either collinear or span a triangle of area at least  $1/2u_1$  or  $1/u_1$  according as  $u_1$  is odd or even. Generalizing our above arguments for  $u = 1$ , we deduce from this that  $F_u(t)$  is constant for  $t$  between 0 and  $1/2u_1$  or  $1/u_1$  respectively. Computing this constant takes more work. For  $u = 1$ , the stabilizer  $\Gamma(1) = \text{ASL}_2(\mathbb{Z})$  of  $\omega_1\mathbb{Z}^2 = \mathbb{Z}^2$  acts transitively on pairs of points that differ by a primitive vector. This made it easy to describe unimodular lattice translates that meet two sides of a triangle but not its interior. For general  $u$ , the stabilizer  $\Gamma(u)$  still acts transitively on  $\omega_{1/u}\mathbb{Z}^2$ ; indeed  $\Gamma(u)$  contains the abelian group freely generated by  $(m, n) \mapsto (m, n+1)$  and  $(m, n) \mapsto (m+1, n+(2m+1)/u)$ , which acts on  $\omega_{1/u}\mathbb{Z}^2$  simply transitively. But there is in general more than one  $\Gamma(u)$  orbit of pairs of points in  $\omega_{1/u}\mathbb{Z}^2$  that are “primitive” in the sense that the line segment joining them contains no further points of  $\omega_{1/u}\mathbb{Z}^2$ . The value of  $F_u(t)$  for small  $t$  is a sum over these orbits. We have not computed this sum in general, but we can show that  $F_u(0) < 1$  for all  $u$ , that  $F_u(0) \rightarrow 1$  as  $u_1 \rightarrow \infty$  (uniformly in  $u_2$ ), and that if  $u_1 = p$  or  $2p$  for some odd prime  $p$  then

$$F_u(0) = \frac{1}{\zeta(2)} \left[ 1 + \frac{p-1}{p} \sum_{i=2}^{p-1} \left( \frac{1}{i^2} - \frac{1}{p^2} \right) \right].$$

As we did for  $u = 1$ , we also obtain for every  $\theta \in (0, 1)$  a function  $F_u^{(\theta)}$  giving the asymptotic distribution of gaps in  $\{\{\sqrt{un + \nu}\} : \theta^2 N < un + \nu < N\}$  by truncating triangles to trapezoids, and an asymptotic joint distribution of  $r$

consecutive gaps each of whose support's components has dimension at most  $5 = \dim E(u)$ , with all components near the origin having dimension 2.

### 3.6 Open questions

We conclude this section with a few open questions suggested by our results.

**Rates of convergence.** We have obtained explicit formulas for the function  $F(x)$  that describes the asymptotic normalized distribution of gaps as  $N \rightarrow \infty$ , but no estimates on how quickly the error approaches zero as  $N$  increases. Numerical data such as exhibited in Figure 1 suggests that  $F(x)$  approximates the actual gap distribution quite well for  $x > 1/2$ , but underestimates the distribution considerably for small  $x$  while overestimating it slightly for  $x \in [\epsilon, 1/2]$ . Can this behavior be explained?

In our analysis there are two sources of error: the linear approximation 3.13, which lets us approximate gap lengths  $L_N(t)$  by  $L(\Lambda_{s^2}(t))$ ; and the ergodic theory argument for Theorem 1.2, which lets us approximate  $\Lambda_{s^2}(t)$  by a random lattice translate. Estimating the former error should be elementary if not pleasant. But the ergodic theory that proves Theorem 1.2 yields no error estimates at all. Can such error estimates be obtained at least for the convergence of the specific family of curves  $\Lambda_{s^2}(x)$  to the uniform distribution? Which of the two error sources accounts for the effects observed numerically for  $x < 1/2$ ?

Naturally we ask the same questions for the gap distributions that arise in each the tractable generalizations we presented earlier, even in the cases where we have not obtained an elementary closed form for the asymptotic distribution.

**Behavior of, and explicit formulas for, the asymptotic answers to our generalized gap distribution problems.** Can  $F^{(\theta)}(t)$  be expressed as an elementary function of  $t, \theta$ , and if so what is this function? Same question for the function  $F_u^{(\theta)}$  for rational  $u > 0$  with  $2/u \notin \mathbb{Z}$ . Can the joint distributions of  $r$  consecutive gaps be obtained explicitly for any  $r > 1$ ? If not, can one at least locate the phase transitions and/or supports of these joint distributions on  $[0, \infty)^r$ ?

**Dependence on  $u$ ; gaps in  $\{\sqrt{un}\}$  for irrational  $u$ ?** Numerical experiments suggest that, for fixed irrational  $u > 0$ , the gaps in  $\{\sqrt{un}\}$  approach the exponential distribution of gaps between random numbers in  $\mathbb{R}/\mathbb{Z}$ . Suppose  $u_i$  ( $i = 1, 2, 3, \dots$ ) are rational numbers approaching  $u$ . Does  $F_{u_i}(t) \rightarrow e^{-t}$  for each  $t \geq 0$ ? (We know this only for  $u = 0$ .) If that is true, might it be used to prove that the gaps in  $\{\sqrt{un}\}$  approach exponential distribution, at least for some irrational values of  $u$ ? More gen-

erally one may ask this question for  $F_{u_i}^{(\theta)}$  and for the joint distributions of  $r$  consecutive gaps.

## References

- [Bo] M. D. Boshernitzan. Uniform distribution and Hardy fields. *J. Anal. Math.* **62**(1994), 225–240.
- [El] N. D. Elkies. Rational points near curves and small nonzero  $|x^3 - y^2|$  via lattice reduction. In *Algorithmic Number Theory : 4th International Symposium, ANTS-IV*, volume 1838 of *Lecture Notes in Computer Science*, pages 33–63. Springer, 2000.
- [EsM] A. Eskin and C. McMullen. Mixing, counting and equidistribution in Lie groups. *Duke Math. J.* **71**(1993), 181–209.
- [Fe] W. Feller. *An Introduction to Probability Theory and its Applications. Vol. I.* John Wiley & Sons, Inc., 1968.
- [HW] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, 1979.
- [Kn] A. W. Knap. *Representation Theory of Semisimple Groups.* Princeton University Press, 1986.
- [KN] L. Kuipers and H. Niederreiter. *Uniform distribution of sequences.* Wiley, 1974.
- [Ph] Robert R. Phelps. *Lectures on Choquet's Theorem.* D. Van Nostrand Co., Inc., 1966.
- [Rat] M. Ratner. On Raghunathan's measure conjecture. *Annals of Math.* **134**(1991), 545–607.
- [RS] Z. Rudnick and P. Sarnak. The pair correlation function of fractional parts of polynomials. *Comm. Math. Phys.* **194**(1998), 61–70.
- [Sa] R. Salem. *Algebraic Numbers and Fourier Analysis.* Wadsworth, 1983.
- [Sos] V. T. Sós. On the distribution mod 1 of the sequence  $n\alpha$ . *Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Math.* **1**(1958), 127–134.

- [Sw] S. Świerczkowski. On successive settings of an arc on the circumference of a circle. *Fund. Math.* **46**(1958), 187–189.
- [Wa] Peter Walters. *An Introduction to Ergodic Theory*, volume 79 of *Graduate Texts in Mathematics*. Springer-Verlag, 1982.
- [We] H. Weyl. Über die Gleichverteilung von Zahlen mod Eins. *Math. Ann.* **77**(1916), 313–352.
- [Zim] R. Zimmer. *Ergodic Theory and Semisimple Groups*. Birkhäuser, 1984.

MATHEMATICS DEPARTMENT  
HARVARD UNIVERSITY  
1 OXFORD ST  
CAMBRIDGE, MA 02138-2901