Mathematische Annalen (2019) 373:1151–1175
https://doi.org/10.1007/s00208-018-1781-x

Mathematische Annalen

# Markov spectra for modular billiards

Nickolas Andersen[1] · William Duke[1]

Received: 13 March 2018 / Published online: 27 November 2018


**Abstract**
The connection between Markov's theory of minima of indefinite binary quadratic forms and hyperbolic geodesics is well-known. We introduce some new analogues of the Markov spectrum defined in terms of modular billiards and consider the problem of characterizing that part of the spectrum below the lowest limit point.

## 1 Introduction

The abstract triangle group usually denoted by $\Delta(2, 3, \infty)$ is generated by $A, B, C$ subject to the relations $A^2 = B^2 = C^2 = (AB)^2 = (AC)^3 = 1$. The extended modular group $\Gamma = \mathrm{PGL}(2, \mathbb{Z})$ gives a faithful representation of this triangle group when we make the identifications:

$$A = \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \pm \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \pm \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{1.1}$$

The usual modular group $\mathrm{PSL}(2, \mathbb{Z})$ is the subgroup of index 2 consisting of all matrices in $\Gamma$ with determinant one.

Let $\mathbb{H}$ be the upper half-plane with its hyperbolic metric given by $ds = \frac{|dz|}{y}$. It is well known that $M = \pm \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ with $\det M = 1$ acts as an orientation preserving isometry of $\mathbb{H}$ through

In memory of Harvey Cohn (1923–2014)

Communicated by Kannan Soundararajan.

Supported by NSF Grant DMS 1701638.

✉ Nickolas Andersen
nandersen@math.ucla.edu

William Duke
wdduke@ucla.edu

[1] UCLA Mathematics Department, Box 951555, Los Angeles, CA 90095-1555, USA

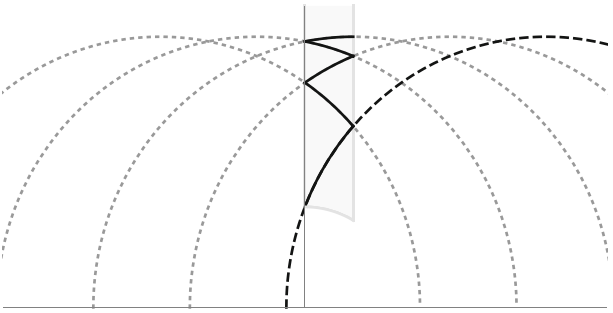


**Fig. 1** The modular triangle $\mathcal{T}$ and a modular billiard

$$z \mapsto M(z) = \frac{az+b}{cz+d}, \tag{1.2}$$

while when $\det M = -1$ it acts through $z \mapsto M(\overline{z})$ as an orientation reversing isometry. The generators $A, B, C$ give reflections across the unit circle, the $y$-axis, and the line $x = \frac{1}{2}$, respectively; $\Gamma$ acts as a reflection group. A convenient fundamental domain for $\Gamma$ is the solid hyperbolic triangle

$$\mathcal{T} = \left\{ z \in \mathbb{H}; 0 \le \operatorname{Re} z \le \tfrac{1}{2}, |z| \ge 1 \right\},$$

whose sides are fixed by the generating reflections and which is the shaded region depicted in Fig. 1.

Let $S$ be an oriented geodesic in $\mathbb{H}$. Thus $S$ is given either by a directed vertical half-line or a directed semi-circle that is orthogonal to $\mathbb{R}$ and is uniquely determined by ordering its endpoints, say $\alpha, \beta$, which are distinct elements of $\mathbb{R} \cup \{\infty\}$. More generally, for $z_1, z_2 \in \mathbb{H} \cup \mathbb{R} \cup \{\infty\}$ let $\langle z_1, z_2 \rangle$ denote the geodesic segment connecting $z_1$ to $z_2$. Hence we may write $S = \langle \alpha, \beta \rangle$ with $\alpha, \beta \in \mathbb{R} \cup \{\infty\}$.

The set of all geodesics splits into orbits $\Gamma S$ under the action of $\Gamma$, where $S$ is any geodesic in the orbit. Let $\mathcal{B}$ denote the set of distinct directed geodesic segments in $\mathcal{T}$ of an orbit $\Gamma S$. We will refer to $\mathcal{B}$ as the trajectory of a modular billiard, but usually call it simply a *modular billiard*. We will say that $\mathcal{B}$ is induced by $S$ for any $S$ in the orbit. Note that $\mathcal{B}$ can be thought of as the path of a point acting like a billiard ball bouncing off the sides of $\mathcal{T}$, with well-defined bounces from the corners of $\mathcal{T}$, which are at

$$z = i \quad \text{and} \quad z = \rho = \tfrac{1}{2} + \tfrac{\sqrt{-3}}{2}.$$

Unlike trajectories of the geodesic flow on $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$, which hit the vertical boundaries $\operatorname{Re}(z) = \pm 1/2$ and have to jump under $z \mapsto z \pm 1$, geodesics on the reflective triangle $\mathcal{T}$ really are billiards.

Suppose that $\mathcal{B}$ is induced by $\langle \alpha, \beta \rangle$. Define its *reversal* $\mathcal{B}^*$ to be the billiard induced by $\langle \beta, \alpha \rangle$. We say the billiard $\mathcal{B}$ is *non-orientable* if $\mathcal{B} = \mathcal{B}^*$, *orientable* otherwise. If $\mathcal{B}$ contains a vertical segment we say it is *improper*, otherwise *proper*. If the total hyperbolic length of the segments in $\mathcal{B}$ is finite, we call the billiard *periodic*. Clearly

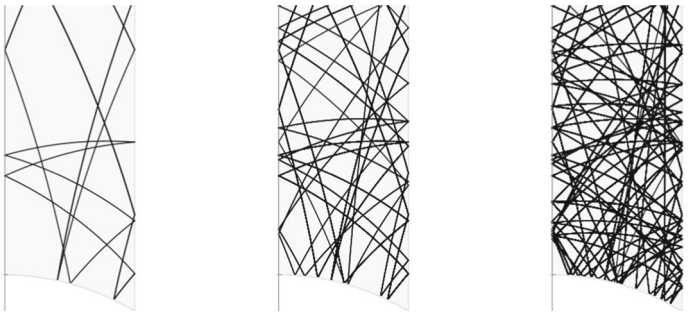


**Fig. 2** Increasing segments of a billiard exhibiting generic behavior

a periodic billiard is proper. The billiard illustrated in Fig. 1 is non-orientable and periodic.

Trajectories of the geodesic flow on $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$ corresponding to improper billiards are called *cuspidal* in the literature, while those corresponding to periodic billiards are *closed*. We remark that the reciprocal geodesic trajectories on $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$ studied by Sarnak in [23] give rise to some, but not all, non-orientable billiards.

The simplest modular billiard, which we will denote $\mathcal{C}_0$, is that induced by the imaginary axis $\langle 0, \infty\rangle$. It covers the segment connecting $i$ to infinity. The billiard induced by $\langle \frac{1}{2}, \infty\rangle$, denoted $\mathcal{C}_{\frac{1}{2}}$, covers the rest of the boundary of $\mathcal{T}$. Both $\mathcal{C}_0$ and $\mathcal{C}_{\frac{1}{2}}$ are improper and non-orientable.

In a prescient article of 1924, Artin [2] observed that properties of continued fractions imply that a generic modular billiard is dense in $\mathcal{T}$ (Fig. 2). On the other hand, the behavior of a non-generic billiard is subtle and can be quite interesting arithmetically. For instance, a modular billiard $\mathcal{B}$ has a maximal height, possibly infinite, defined to be the supremum of imaginary parts of points on $\mathcal{B}$. Let $\lambda_\infty(\mathcal{B})$ be *twice* this maximal height. Consider the set

$$\mathcal{M}_\infty = \{\lambda_\infty(\mathcal{B}); \mathcal{B} \quad \text{is a modular billiard}\}.$$

This is the *Markov spectrum*, which is usually defined (equivalently) in terms of the minima of indefinite binary quadratic forms. The Markov numbers are those positive integers $p$ for which there are $q, r \in \mathbb{Z}^+$ such that

$$p^2 + q^2 + r^2 = 3pqr.$$

These may be ordered into an infinite increasing sequence whose $n$th term is denoted by $p_n$:

$$\{1, 2, 5, 13, 29, 34, \ldots, p_n, \ldots\}.$$

The following result is a consequence of the fundamental work of Markov [21,22]:

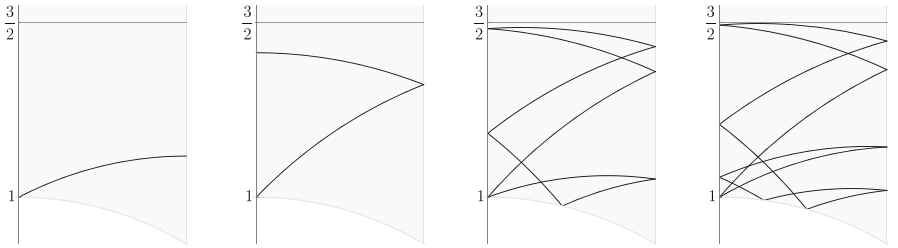



**Fig. 3** Billiards associated to the points $\sqrt{5}, \sqrt{8}, \frac{\sqrt{221}}{5}, \frac{\sqrt{1517}}{13} \in \mathcal{M}_\infty$

**Theorem 1** *For any fixed $\kappa < 3$ there are only finitely many modular billiards $\mathcal{B}$ with $\lambda_\infty(\mathcal{B}) < \kappa$. The points in $\mathcal{M}_\infty$ less than 3 are given by the sequence*

$$\left\{\sqrt{5}, \sqrt{8}, \frac{\sqrt{221}}{5}, \frac{\sqrt{1517}}{13}, \frac{\sqrt{7565}}{29}, \ldots, \frac{\sqrt{9p_n^2-4}}{p_n}, \ldots.\right\},$$

*which is monotone increasing to the limit $3 \in \mathcal{M}_\infty$.*

It is also known that each of the points $< 3$ in $\mathcal{M}_\infty$ is actually attained by a non-orientable periodic billiard (see Theorem 75 of [11]) and it was conjectured by Frobenius [14], but is still open, that the multiplicity of each of these points is one, meaning that the associated billiard is unique. The part of the Markov spectrum that is $> 3$ is less understood but has been the subject of much research (see [9,20]). It is not hard to show that any open interval around 3 contains uncountably many points of $\mathcal{M}_\infty$ and that $\mathcal{M}_\infty$ is closed, but there are few completely definitive results known. Building on pioneering work of Hall [15,16], Freiman [13] obtained one such result. He showed that $[\mu, \infty) \subset \mathcal{M}_\infty$, where

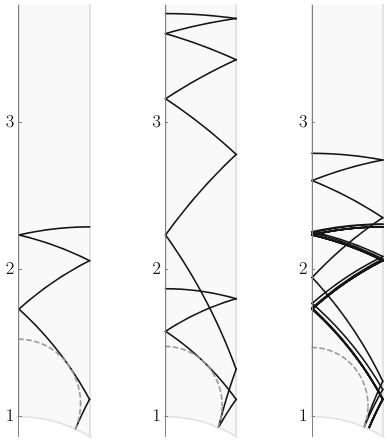
$$\mu = \frac{2221564096 + 283748\sqrt{462}}{491993569} = 4.527829566\ldots,$$

with no smaller value of $\mu$ being possible.

We remark that the seminal work of Harvey Cohn, beginning with his 1955 paper [7], greatly enhanced our understanding of the connections between the Markov spectrum, hyperbolic geometry and combinatorial group theory. Work of Cohn and several others revealed a completely unexpected relation between the Markov spectrum and the length spectrum of *simple* closed geodesics on the modular torus (see e.g. [8,17–19] and the references therein). This work has had a lasting impact on the study of simple closed geodesics on Riemann surfaces. It has also led to a better understanding of the Markov spectrum itself. As can be seen in Fig. 3, the modular billiards induced by the Markov geodesics are *not* simple in general, but if they are unfolded in the modular torus they become simple.

The value $\lambda_\infty(\mathcal{B})^{-1}$ may be thought of as a measure of how close the billiard $\mathcal{B}$ gets to the corner of $\mathcal{T}$ at the cusp $i\infty$. It is natural to ask how close a modular billiard must get to each of the other corners $i$ and $\rho$ of $\mathcal{T}$. By the distance of a billiard from

**Fig. 4** Billiards $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$




a point $z \in \mathcal{T}$, denoted by $\delta_z(\mathcal{B})$, we mean the infimum of the hyperbolic distance between points on the billiard and $z$. Let

$$\lambda_z(\mathcal{B}) = (\sinh \delta_z(\mathcal{B}))^{-1}. \tag{1.3}$$

A natural analogue of the Markov spectrum is

$$\mathcal{M}_z = \{\lambda_z(\mathcal{B}); \mathcal{B} \quad \text{is a modular billiard}\} \tag{1.4}$$

for a fixed $z \in \mathcal{T}$.

In this paper we will give results about $\mathcal{M}_\rho$ and $\mathcal{M}_i$ that correspond to Markov's for $\mathcal{M}_\infty$. The result for $z = \rho$ is quite easy to prove.

**Theorem 2** *The smallest value in $\mathcal{M}_\rho$ is $\sqrt{3}$, which is attained by $\mathcal{C}_0$. The value $\sqrt{3}$ is a limit point of $\mathcal{M}_\rho$.*

The result for $z = i$ is deeper and most of this paper is devoted to its proof.

**Theorem 3** *The three smallest values in $\mathcal{M}_i$ are*

$$\left\{\tfrac{1}{2}\sqrt{21}, \quad \tfrac{2}{3}\sqrt{14}, \quad \tfrac{1}{3}(3+\sqrt{21})\right\} = \{2.29129\ldots,\ 2.49444\ldots,\ 2.52753\ldots\}.$$

*These three values are attained, respectively, by unique billiards $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$, each proper and non-orientable. Here $\mathcal{C}_1$ and $\mathcal{C}_2$ are periodic billiards, while $\mathcal{C}_3$ is not periodic. The value $\frac{1}{3}(3+\sqrt{21})$ is a limit point of $\mathcal{M}_i$.*

Explicitly, $\mathcal{C}_1$ is induced by the geodesic $\langle \frac{1}{2}(1-\sqrt{21}), \frac{1}{2}(1+\sqrt{21})\rangle$, $\mathcal{C}_2$ is induced by the geodesic $\langle \frac{1}{2}(2-\sqrt{14}), \frac{1}{2}(2+\sqrt{14})\rangle$ and $\mathcal{C}_3$ is induced by the geodesic $\langle \frac{1}{2}(3-\sqrt{21}), \frac{1}{2}(5+\sqrt{21})\rangle$ (Fig. 4). We remark that $\mathcal{C}_3$ is an example of a non-periodic billiard that is not dense in $\mathcal{T}$, which is a reflection of the fact that $\frac{1}{2}(3-\sqrt{21})$ and $\frac{1}{2}(5+\sqrt{21})$ are not Galois conjugates. One can see from the doubly-infinite sequence $K_3$ which

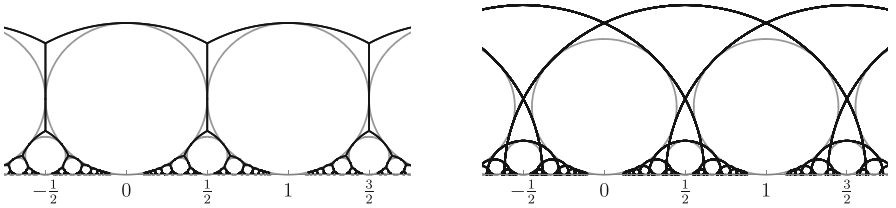



**Fig. 5** Ford circles

we use to represent $\mathcal{C}_3$ in (7.6) below that $\mathcal{C}_3$ is essentially a concatenation of two periodic billiards with a small perturbation in between.

In both cases the rest of the spectrum invites investigation. It is also of interest to consider the Markov spectrum $\mathcal{M}_z$ for other points in $\mathcal{T}$, in particular CM points. In addition to distances from a fixed point, there are other geometric quantities associated to non-generic modular billiards whose sets of values define Markov-type spectra. The purpose of this paper is to initiate a study of these generalizations by concentrating on the simplest and most natural examples and giving the analogues of Markov's results for them.

In the next section we give a geometric interpretation of Theorems 1–3 in terms of the packing of discs in tessellations formed by geodesic segments and prove the first statement of Theorem 2. In Sect. 3 we recall the connection between modular billiards and real indefinite binary quadratic forms and then in Sect. 4 give a formula for the hyperbolic distance between a billiard and a point. This formula is written in terms of the minimum of an indefinite quaternary quadratic form and is used to complete the proof of Theorem 2. In Sect. 5 we introduce reduced forms and express $\lambda_i(\mathcal{B})$ in terms of them. Then we give in Sect. 6 the correspondence between proper modular billiards and doubly-infinite sequences of positive integers that connects billiards to simple continued fractions. This connection is exploited in Sects. 7–9 to complete the proof of Theorem 3.

## 2 Packing discs in hyperbolic tessellations

Elementary geometric considerations provide some useful insight into Theorems 1–3 and serve to establish "trivial" bounds for $\lambda_z(\mathcal{B})$. The problem of finding points of $\mathcal{M}_z$ is equivalent to the problem of fitting geodesics in $\mathbb{H}$ between discs of varying radii around the images under $\Gamma$ of $z$.

Consider the case of the original Markov spectrum $\mathcal{M}_\infty$. A Ford circle is the horocycle around the reduced rational number $p/q$ with radius $\frac{1}{2q^2}$. Together, these circles form the $\Gamma$-orbit of the horocycle $\mathrm{Im}(z) = 1$. The set of all Ford circles form a packing of the tessellation $\Gamma\langle i, \rho\rangle$. See the left hand side of Fig. 5. It is obvious that every geodesic $S$ must intersect infinitely many Ford circles. This gives that $\lambda_\infty(\mathcal{B}) \geq 2$, or $\mathcal{M}_\infty \subset [2, \infty)$. Ford [12] proved that if we reduce the radii of the Ford circles to any

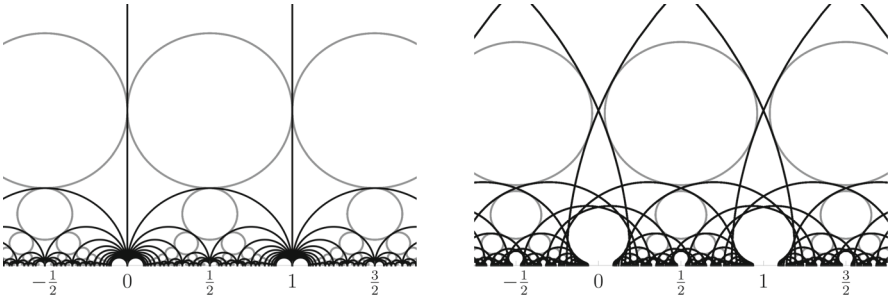



**Fig. 6** Illustrating Theorem 2 by disks around images of $\rho$

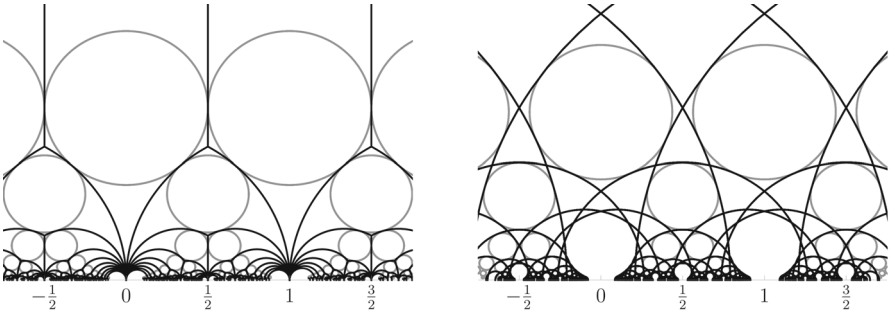



**Fig. 7** Disks around images of $i$

$$r \geq r_0 = \frac{1}{\sqrt{5}q^2}$$

it still forces intersection but if $r < r_0$ there are geodesics that intersect no circle. This is equivalent to shifting the cuspidal horocycle to $\mathrm{Im}(z) = 3/2$. See the right hand side of Fig. 5.

***Proof of first statement of Theorem 2*** The first statement of Theorem 2 may be proven this way. To show that $\lambda_\rho(\mathcal{B}) \geq \sqrt{3}$ first observe that the hyperbolic circles of radius $\frac{1}{2}\log 3$ around the points $\Gamma\rho$ form tangent sequences that approach a dense subset of $\mathbb{R}$. It is straightforward to show that they are tangent to the Farey triangulation $\Gamma\langle 0, \infty\rangle$. Any geodesic must intersect these circles if their radius is made any larger since its endpoints will be separated by a sequence of circles. See the left hand side of Fig. 6.

As we will prove below and is illustrated in the right hand side of Fig. 6, if the radii are reduced by any positive amount there are infinitely many inequivalent geodesics that intersect no circle.

Turning to Theorem 3, we can pack $\Gamma\langle i, \infty\rangle$ by geodesic circles centered at the points $\Gamma i$ of radius $\log(\frac{1+\sqrt{5}}{2}) = 0.481212\ldots$. This implies that $\lambda_i(\mathcal{B}) \geq 2$, as is illustrated in the left hand side of Fig. 7. This is weaker than the consequence of Theorem 3 that $\lambda_i(\mathcal{B}) \geq 2.29129\ldots$. as the right hand side of Fig. 7 illustrates.

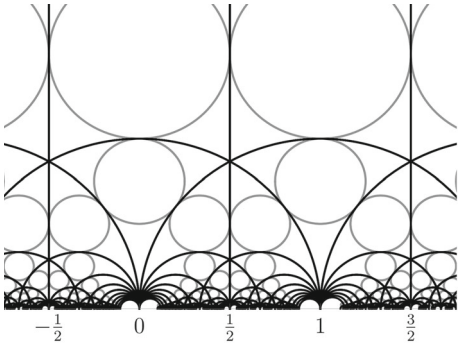



**Fig. 8** Disks around images of $\sqrt{-2}$

This point of view sheds light on why the Markov-type result for the distance problem is easier for $z = \rho$. The corresponding tessellation in this case comprises complete geodesics, while in the other two cases only geodesic segments.

For example, we easily get the first statement of the following result using the tessellation $\Gamma\langle\frac{1}{2}, \infty\rangle = \Gamma\langle i, \rho\rangle \cup \Gamma\langle\rho, \infty\rangle$, whose associated billiard is $\mathcal{C}_{\frac{1}{2}}$. Figure 8 illustrates the packing of this tessellation by disks around images of $\sqrt{-2}$ of radius $\frac{\log 2}{2}$.

**Theorem 4** *The smallest value in $\mathcal{M}_{\sqrt{-2}}$ is $\sqrt{8}$, which is attained by $\mathcal{C}_{\frac{1}{2}}$. The value $\sqrt{8}$ is a limit point of $\mathcal{M}_{\sqrt{-2}}$.*

## 3 Binary quadratic forms and billiards

To go beyond this basic geometric method we need a usable formula for the distance between a billiard and a point. Binary quadratic forms provide the key. In this section we establish their relation to modular billiards of the various kinds.

For $a, b, c \in \mathbb{R}$ with $d = \text{disc}(Q) = b^2 - 4ac \neq 0$ let

$$Q(x, y) = ax^2 + bxy + cy^2,$$

which is a non-singular real binary quadratic form. Sometimes we will write $Q = (a, b, c)$. Now $M = \pm \left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right) \in \Gamma = \text{PGL}(2, \mathbb{Z})$ acts on $Q$ by

$$(Q|M)(x, y) \stackrel{\text{def}}{=} (\det M) Q(a'x + b'y, c'x + d'y). \tag{3.1}$$

Clearly $Q|(M_1 M_2) = (Q|M_1)|M_2$ for $M_1, M_2 \in \Gamma$. We say that two such forms $Q_1$ and $Q_2$ are *equivalent* if there is an $M \in \Gamma$ such that

$$(Q_1|M)(x, y) = Q_2(x, y).$$

If $M \in \mathrm{PSL}(2,\mathbb{Z})$ then we say that $Q_1$ and $Q_2$ are *properly equivalent.* The class of forms that are equivalent to $Q$, but not necessarily properly equivalent to $Q$, will be denoted by $[Q]$. The discriminant $\mathrm{disc}(Q)$ is an invariant of $[Q]$.

If $Q = (0, b, c)$ with $b > 0$ let $\alpha_Q = -\frac{c}{b}$ and $\beta_Q = \infty$, while if $b < 0$ with let $\beta_Q = -\frac{c}{b}$ and $\alpha_Q = \infty$. Otherwise the roots of $Q(z, 1) = 0$ are given by

$$\alpha_Q = \frac{-b + \sqrt{d}}{2a} \quad \text{and} \quad \beta_Q = \frac{-b - \sqrt{d}}{2a}. \tag{3.2}$$

In all cases $\alpha_Q$ will be called the *first* root and $\beta_Q$ the *second* root of $Q$. One checks that $d$, $\alpha_Q$ and $\beta_Q$ uniquely determine $Q$. Furthermore, using the generators $A$, $B$, $C$ from (1.1), it follows that for each $j = 1, 2$ and for any $M = \pm \left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right) \in \Gamma$

$$\alpha_{Q|M} = M^{-1}(\alpha_Q) \quad \text{and} \quad \beta_{Q|M} = M^{-1}(\beta_Q),$$

with $M(z)$ given in the definition around (1.2) extended to all of $\mathbb{C}$. It is important that the action of $\Gamma$ on quadratic forms defined in (3.1) preserves the first and second roots.

The proofs of the following two results are straightforward.

**Proposition 1** (i) *The map $Q \mapsto \alpha_Q$ determines a bijection between classes $[Q]$ with a fixed negative discriminant that are represented by positive definite $Q$ and points of $\mathcal{T}$.*

(ii) *The map $Q \mapsto \langle \alpha_Q, \beta_Q \rangle$ determines a bijection $[Q] \leftrightarrow \mathcal{B}$ between the classes $[Q]$ with a fixed positive discriminant and the set of all modular billiards, where the associated billiard $\mathcal{B}$ is induced by $\langle \alpha_Q, \beta_Q \rangle$.*

Say $Q$ *represents zero* if $Q(x, y) = 0$ for some $x, y \in \mathbb{Z}$ not both zero, that $Q$ is *reciprocal* if $Q$ is equivalent to $-Q$ and that $Q = (a, b, c)$ is *primitive integral* if $a, b, c \in \mathbb{Z}$ with $\gcd(a, b, c) = 1$. We have the following characterization of improper, non-orientable and periodic billiards in terms of quadratic forms.

**Proposition 2** *Under the bijection $[Q] \leftrightarrow \mathcal{B}$ of (ii) of Proposition 1, for any $Q \in [Q]$*

(a) *$Q$ represents zero if and only if $\mathcal{B}$ is improper,*
(b) *$Q$ is reciprocal if and only if $\mathcal{B}$ is non-orientable,*
(c) *For some non-zero real $\kappa$ the form $\kappa Q$ does not represent zero and is primitive integral if and only if $\mathcal{B}$ is periodic.*

The study of periodic billiards is the same as that of (primitive) integral binary quadratic forms. It follows from Proposition 2 (c) that each periodic modular billiard $\mathcal{B}$ may be assigned a unique positive integer given by

$$d = \mathrm{disc}(\mathcal{B}) \stackrel{\mathrm{def}}{=} \kappa^2 \mathrm{disc}(Q)$$

for any $Q \in [Q]$. There are only finitely many periodic billiards with a given discriminant $d$ and each has the same length.

A formula for the length of $\mathcal{B}$ is determined by finding that solution $(t, u)$ with $t, u \in \mathbb{Z}^+$ of $t^2 - du^2 = \pm 4$ for which $\varepsilon = \frac{1}{2}(t + u\sqrt{d})$ is minimal. Then the length is $2 \log \varepsilon$. If $t^2 - du^2 = -4$ has a solution then proper equivalence is the same as equivalence. Otherwise each ordinary class consists of two proper classes.

Classically one says that a primitive integral $Q$ is improperly equivalent to itself when $Q|M = -Q$ where $M \in \Gamma$ with $\det M = -1$ since then the change of variables $(x, y) \mapsto (x, y)M^t$ preserves $Q$. It follows from Theorem 90 of [10] that this holds if and only if some $(a, b, c) \in [Q]$ has $a|b$. These forms are called *ambiguous* and one may say that the billiard $\mathcal{B}$ associated to $[Q]$ is ambiguous. By Proposition 2 (b) we have that an ambiguous $\mathcal{B}$ is non-orientable. If $t^2 - du^2 = -4$ has a solution then every non-orientable billiard is ambiguous. Otherwise it is possible for a billiard to be non-orientable without being ambiguous. This happens when an associated $Q$ is properly equivalent to $-Q$. Markov's billiards are examples. If $d$ is fundamental, the number of non-orientable billiards is $2^{\omega(d)-1}$, where $\omega(d)$ is the number of distinct prime factors dividing $d$.

## 4 Quaternary quadratic forms

A consequence of Proposition 2, one that is crucial for us, is a formula for

$$\sinh \delta_z(\mathcal{B}) = (\lambda_z(\mathcal{B}))^{-1}$$

from (1.3). This involves finding the minimum of a certain indefinite quaternary quadratic form. For a fixed $z \in \mathcal{T}$ let $Q'(x, y) = a'x^2 + b'xy + c'y^2$ with $a' > 0$ and $d' = b'^2 - 4a'c' < 0$ represent $z$. Also, let $Q(x, y) = ax^2 + bxy + cy^2$ with $d = b^2 - 4ac > 0$ represent the modular billiard $\mathcal{B}$.

**Proposition 3** *Notation as above,*

$$\sinh \delta_z(\mathcal{B}) = (d|d'|)^{-\frac{1}{2}} \inf_{[Q]} |2c'a + 2a'c - b'b|.$$

***Proof*** A standard exercise in hyperbolic geometry shows that the hyperbolic distance $\delta(z, S)$ from $z$ to the geodesic $S = \langle \alpha_Q, \beta_Q \rangle$ satisfies

$$\sinh \delta(z, S) = (d|d'|)^{-\frac{1}{2}} |2c'a + 2a'c - b'b|. \tag{4.1}$$

See e.g. [3, p. 162]. The result follows. □

We are now able to justify the second statements of Theorems 2 and 4. For fixed $\ell \in \mathbb{Z}^+$ let $\mathcal{A}_\ell$ be the billiard associated to the quadratic form

$$Q(x, y) = x^2 - \ell xy - y^2.$$

The case when $\ell = 5$ is illustrated in Fig. 1, which is typical in that of those geodesics in the orbit intersecting $\mathcal{T}$, the one corresponding to $Q$ is the one that gives the closest approach to $\rho$. By Proposition 3 we have

$$\sinh \delta_\rho(\mathcal{A}_\ell) = \frac{\ell}{\sqrt{3\ell^2+12}}.$$

Therefore $\lambda_\rho(\mathcal{A}_\ell) = \frac{\sqrt{3\ell^2+12}}{\ell}$, which decreases to the limit $\sqrt{3}$ as $\ell \to \infty$. This completes the proof of Theorem 2. The second statement in Theorem 4 follows in like manner.

Note that we may rewrite (4.1) using (1.3) as

$$\lambda_z(\mathcal{B})^{-1} = (d|d'|)^{-\frac{1}{2}} \inf_{x_1x_4-x_2x_3=\pm 1} |Q''(x_1, x_2, x_3, x_4)|, \tag{4.2}$$

where for a *fixed* choice of $Q(x, y) = ax^2 + bxy + cy^2$ representing $\mathcal{B}$ we have

$$\begin{aligned} Q''(x_1, x_2, x_3, x_4) = {} & 2a'cx_1^2 + 2a'ax_2^2 + 2c'cx_3^2 + 2c'ax_4^2 \\ & + 2b'ax_2x_4 + 2b'cx_1x_3 - 2c'bx_3x_4 \\ & - 2a'bx_1x_2 - b'bx_1x_4 - b'bx_2x_3. \end{aligned}$$

Here $Q''$ is an indefinite quaternary quadratic form of signature $(2, 2)$. Observe that we must minimize $Q''(x_1, x_2, x_3, x_4)$ subject to

$$x_1x_4 - x_2x_3 = \pm 1 \tag{4.3}$$

whereas it is more usual to only require that $(x_1, x_2, x_3, x_4) \neq (0, 0, 0, 0)$.

The formula for $\lambda_\infty(\mathcal{B})$ from the Markov spectrum corresponding to (4.2) is simply

$$\lambda_\infty(\mathcal{B})^{-1} = d^{-\frac{1}{2}} \inf_{(x_1,x_2)\neq(0,0)} |Q(x_1, x_2)|. \tag{4.4}$$

In this sense the problem of finding $\lambda_i(\mathcal{B})$ is more difficult than that of finding $\lambda_\infty(\mathcal{B})$. The study of the minima of certain indefinite quaternary forms subject to (4.3) goes back at least to a 1913 paper of Schur [24], which was an inspiration for this paper and deserves to be better known.

At this point we may obtain a good lower bound for $\lambda_i(\mathcal{B})$ when $\mathcal{B}$ is improper. It is easy to check that an improper billiard is determined by some $\langle \alpha, \infty \rangle$, where $0 \leq \alpha \leq \frac{1}{2}$, or equivalently by the form $Q = y(x - \alpha y)$ for this $\alpha$.

**Proposition 4** *For $\mathcal{B}$ an improper billiard we have that*

$$\lambda_i(\mathcal{B}) \geq 3$$

*and this is attained by the billiard determined by $\langle \frac{1}{3}, \infty \rangle$.*

***Proof*** Let $Q(x, y) = y(x - \alpha y)$. Then $\lambda_i(\mathcal{B})^{-1} \leq \min(\alpha, 1 - 2\alpha)$, which is found by applying Proposition 3 in the form (4.2) to

$$\tfrac{1}{2}Q''(x_1, x_2, x_3, x_4) = -\alpha(x_1^2 + x_3^2) - x_3x_4 - x_1x_2$$

and taking $(x_1, x_2, x_3, x_4) = (1, 0, 0, 1), (1, -1, 1, 0)$. Thus $\alpha = 1/3$ gives the minimum. □

## 5 Proper billiards and reduced forms

We say that a form $Q = (a, b, c)$ with discriminant $d > 0$ that does not represent zero is *reduced* if

$$-1 < \beta_Q < 0 \quad \text{and} \quad \alpha_Q > 1,$$

where $\alpha_Q$ and $\beta_Q$ were defined in (3.2). A classical argument given in the proof of Theorem 76 in [10] may be adapted to prove that for any proper billiard $\mathcal{B}$ the corresponding class $[Q]$ (as in Proposition 2) contains such reduced forms.

Given a form $Q = (a, b, c)$, let

$$Q^*(x, y) = -Q(-y, x). \tag{5.1}$$

If $Q$ is reduced then $Q^*$ is a reduced form that is properly equivalent to $-Q$.

Note that the geodesics associated to the reduced forms do not necessarily account for all of the geodesic segments comprising a modular billiard. This fact is illustrated in Fig. 1, where the single geodesic associated to a reduced form is shown in black. In the proof of Markov's Theorem 1 the maximal height of a billiard $\mathcal{B}$ will be approached by the heights of geodesics associated to reduced forms. Thus by (4.4) it follows that we have the simple formula

$$\lambda_\infty(\mathcal{B})^{-1} = d^{-\frac{1}{2}} \inf_{Q \in [Q]} \inf_{\text{reduced}} |a|. \tag{5.2}$$

To obtain an analogous formula for $\lambda_i(\mathcal{B})^{-1}$ we must consider some transforms of reduced forms. This motivates the following definition. For any for $Q = (a, b, c)$ with discriminant $d > 0$ define

$$\nu(Q) = d^{-\frac{1}{2}} \min(|a + c|, |2a + b + c|, |2c + b + a|). \tag{5.3}$$

**Proposition 5** *For a proper $\mathcal{B}$ we have*

$$\lambda_i(\mathcal{B})^{-1} = \inf_{Q \in [Q] \text{ reduced}} \nu(Q),$$

*where the class $[Q]$ corresponds to $\mathcal{B}$.*

***Proof*** To prove this we will show that for any geodesic in $\mathbb{H}$ that intersects $\mathcal{T}$, we can find a form $Q = (a, b, c)$ such that either $Q$ or $-Q$ is reduced and that the geodesic corresponding to $Q(x, y)$, $Q(x + y, y)$, or $-Q(x, x - y)$ is as close or closer to $z = i$. The result then follows by Proposition 3 applied with $Q' = (1, 0, 1)$ since the terms in (5.3) correspond exactly to these three cases.

Note that we may restrict our attention to geodesics $S$ that either (i) cross both vertical sides of the boundary of $\mathcal{T}$ or (ii) cross the right vertical side and the circular arc of this boundary. This is because the reflection across the $y$-axis of a geodesic that crosses the left hand vertical side and the circular arc will cross both vertical sides and will also have the same distance from $z = i$.

In case (i) we may assume that the apex of the geodesic $S$ lies on or to the right of the $y$-axis; if not, the reflection across the $y$-axis of $S$ will have that property and be the same distance from $i$. Let $\widetilde{\alpha}, \widetilde{\beta}$ be the roots of a form $\widetilde{Q}$ associated to $S$. Then we have

$$\widetilde{\alpha} > \tfrac{1}{2}, \quad \widetilde{\beta} < 0, \quad \widetilde{\alpha} - \widetilde{\beta} \geq 2, \quad \text{and} \quad \widetilde{\alpha} + \widetilde{\beta} \geq 0.$$

Thus there is a unique integer $n \geq 0$ such that the form $Q(x, y) = \widetilde{Q}(x - ny, y)$, obtained by shifting $S$ to the right $n$ units, is reduced. If $n = 0$ or 1 then we are done because either $\widetilde{Q}$ is reduced already or $\widetilde{Q}(x, y) = Q(x + y, y)$. Suppose that $n \geq 2$. Then $\widetilde{\beta} < -2$ and $\widetilde{\alpha} > 2$ which implies that the roots $\alpha, \beta$ of $Q$ satisfy

$$-1 < \beta < 0 \quad \text{and} \quad \alpha > 4.$$

If $\langle \alpha, \beta \rangle$ intersects the $y$-axis above $i$, then a simple geometric argument shows that $\langle \alpha, \beta \rangle$ is closer than $S$ to the point $z = i$. If not, then the geodesic $\langle \alpha - 1, \beta - 1 \rangle$ associated to $Q(x + y, y)$ crosses the $y$-axis above $i$ (since $(\alpha - 1)(\beta - 1) \leq -2$); hence either $\langle \alpha - 1, \beta - 1 \rangle$ or $\langle \alpha, \beta \rangle$ is closer than $S$ to the point $z = i$.

In case (ii) we may assume that either $Q = \widetilde{Q}$ is reduced or that

$$0 < \widetilde{\beta} < \tfrac{1}{2} \quad \text{and} \quad \widetilde{\alpha} > 1,$$

in which case $Q(x, y) = -\widetilde{Q}(x, x - y)$ is reduced. Since this is equivalent to the identity

$$\widetilde{Q}(x, y) = -\, Q(x, x - y),$$

we are done. □

## 6 Billiards and sequences

Simple continued fractions are crucial in Markov's proof of Theorem 1 and in our proof of Theorem 3. We denote one by

$$[k_1, k_2, k_3, k_4, \dots] = k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cfrac{1}{k_4 + \dots}}},$$

whose finite version ending in $\frac{1}{k_n}$ is written $[k_1, k_2, \dots, k_n]$. Here we will present the beginnings of Markov's method in a form that we adapt in the next section to prove Theorem 3. The method relates chains of reduced quadratic forms to doubly-infinite sequences of positive integers. We give a somewhat novel treatment of this correspondence based on equivalence rather than proper equivalence.

For a reduced $Q$ define the doubly infinite sequence $K_Q$ by expanding

$$\alpha_Q = [k_1, k_2, k_3, \dots] \quad \text{and} \quad -\beta_Q = [0, k_0, k_{-1}, k_{-2}, \dots] \tag{6.1}$$

into simple continued fractions and setting

$$K_Q = (\dots, k_{-1}, k_0, k_1, k_2, \dots).$$

We shall refer to $k_1$ as the first entry of $K$.

Say two doubly infinite sequences of positive integers $K = (k_n)$ and $L = (\ell_n)$ are *equivalent* if there is a $j \in \mathbb{Z}$ such that $k_n = \ell_{n+j}$ for all $n \in \mathbb{Z}$ and *properly equivalent* if there is a $j \in 2\mathbb{Z}$ such that $k_n = \ell_{n+j}$ for all $n \in \mathbb{Z}$. If $K = (k_n)$ define the *reversal* of $K$ to be $K^* = (k_{1-n})$.

**Proposition 6** *The map $Q \mapsto K_Q$ determines a bijection between classes of forms with a fixed positive discriminant that do not represent zero and equivalence classes of sequences of positive integers. It also determines a bijection between proper equivalence classes of forms with a fixed positive discriminant that do not represent zero and proper equivalence classes of sequences of positive integers. Furthermore $K_Q^* = K_{Q^*}$.*

***Proof*** For each $n \in \mathbb{Z}$ let

$$r_n = [k_n, k_{n+1}, \dots] \quad \text{and} \quad s_n = [0, k_{n-1}, k_{n-2}, \dots]. \tag{6.2}$$

Thus $r_1 = \alpha_Q$ and $s_1 = -\beta_Q$ and also

$$r_{n-1} = \frac{1}{r_n} + k_{n-1} \quad \text{and} \quad s_{n-1} = \frac{1}{s_n} - k_{n-1}. \tag{6.3}$$

Define $a_n, b_n > 0$ for each $n \in \mathbb{Z}$ by

$$\frac{a_n}{\sqrt{d}} = \frac{1}{r_{n-1} + s_{n-1}} \quad \text{and} \quad \frac{b_n}{\sqrt{d}} = \frac{r_n - s_n}{r_n + s_n}. \tag{6.4}$$

It can be seen that using (6.3) that $d = \operatorname{disc}(Q) = b_n^2 + 4a_n a_{n+1}$ for all $n \in \mathbb{Z}$. Let

$$Q_n = (a_{n+1}, -b_n, -a_n).$$

A calculation shows that

$$\alpha_{Q_n} = r_n \quad \text{and} \quad -\beta_{Q_n} = s_n.$$

It follows that each $Q_n$ is reduced and equivalent to $Q$. We claim that every such form occurs as a $Q_n$. Further, each $Q_{2n+1}$ is properly equivalent to $Q$ and every reduced form that is properly equivalent to $Q$ is one of the $Q_{2n+1}$. Again, these statements follow from variations on the arguments given in Chaper VII of [10]. That the claimed bijections are well-defined and injective follows. Clearly every sequence $K$ arises from some reduced form so the maps are also surjective.

Turning to the last statement, recall that $Q^*$ was defined in (5.1) and observe that

$$\beta_{Q^*} = -\tfrac{1}{\alpha_Q} \quad \text{and} \quad \alpha_{Q^*} = -\tfrac{1}{\beta_Q}.$$

Let $r_n^*$ and $s_n^*$ correspond to $Q^*$ as in (6.2). By (6.1) we have

$$r_1^* = -\tfrac{1}{\beta_Q} = [k_0, k_{-1}, k_{-2}, k_{-3}, \dots] \quad \text{and} \quad s_1^* = \tfrac{1}{\alpha_Q} = [0, k_1, k_2, k_3 \dots],$$

giving the result.

This completes the proof of Proposition 6. □

Say that a sequence $K = (k_n)$ is *periodic* if there is an $N \in \mathbb{Z}^+$ so that $k_{n+N} = k_n$ for all $n \in \mathbb{Z}$ and *palindromic* if $K^*$ is equivalent to $K$. Combining Propositions 2 and 6 we derive the following correspondence.

**Theorem 5** *There is a bijection between proper modular billiards and equivalence classes of doubly infinite sequences of positive integers. Under this correspondence a billiard is periodic precisely when the sequence is periodic and non-orientable precisely when the corresponding sequence is palindromic.*

From the first formula of (6.4) and (5.2) we have

$$\lambda_\infty(\mathcal{B}) = \sup_{n\in\mathbb{Z}}(r_n + s_n). \tag{6.5}$$

This is the starting point of the proof of Markov's Theorem 1. The main difficulty is in understanding which $K$ cannot have any small values of $r_n + s_n$. The first observation is that if any $k_m > 2$ we must have $\lambda_\infty(\mathcal{B}) > 3$. The complete result requires an ingenious analysis of continued fractions all of whose partial quotients are either 1 or 2. A treatment of Markov's method and a proof of Theorem 1 based on it can be found in Dickson's book [11]. Other useful references are [1,4–6].

## 7 Sequences and the spectrum

To return to the proof of Theorem 3, recall from (1.4) that $\mathcal{M}_i$ is defined in terms of $\lambda_i(\mathcal{B})$, which was given in (1.3). We now find a formula for $\lambda_i(\mathcal{B})$ that is analogous to (6.5) when $\mathcal{B}$ is a proper modular billiard. For our problem we are led to estimate certain quantities involving pairs of successive values of $r_n$ and $s_n$ from (6.2), rather than simply $r_n + s_n$.

Let $K$ be a doubly-infinite sequence of positive integers. The quantities we need are the following:

$$\mu_n'(K) = \left| \frac{1}{r_n + s_n} - \frac{1}{r_{n-1} + s_{n-1}} \right| \tag{7.1}$$

$$\mu_n''(K) = \left| \frac{2 - r_n + s_n}{r_n + s_n} - \frac{1}{r_{n-1} + s_{n-1}} \right| \tag{7.2}$$

$$\mu_n'''(K) = \left| \frac{2}{r_{n-1} + s_{n-1}} + \frac{r_n - s_n - 1}{r_n + s_n} \right|. \tag{7.3}$$

Also set

$$\mu'(K) = \inf_{n \in \mathbb{Z}} \mu_n'(K), \quad \mu''(K) = \inf_{n \in \mathbb{Z}} \mu_n''(K), \quad \mu'''(K) = \inf_{n \in \mathbb{Z}} \mu_n'''(K) \tag{7.4}$$

and

$$\mu(K) = \min\left(\mu'(K), \mu''(K), \mu'''(K), \mu'(K^*), \mu''(K^*), \mu'''(K^*)\right). \tag{7.5}$$

**Proposition 7** *Let $K$ correspond to a proper $\mathcal{B}$. Then we have*

$$\lambda_i(\mathcal{B})^{-1} = \mu(K).$$

***Proof*** By Proposition 6 and (6.4)

$$\mu_n'(K) = d^{-\frac{1}{2}} |a_{n+1} - a_n|, \quad \mu_n''(K) = d^{-\frac{1}{2}} |2a_{n+1} - b_n - a_n|,$$
$$\mu_n'''(K) = d^{-\frac{1}{2}} |2a_n + b_n - a_{n+1}|.$$

The result follows from Proposition 5 since every reduced form of the class $[Q]$ is found among the $Q_n$. □

In the following proposition we show that

$$\left\{ \tfrac{1}{2}\sqrt{21}, \quad \tfrac{2}{3}\sqrt{14}, \quad \tfrac{1}{3}(3 + \sqrt{21}) \right\} \in \mathcal{M}_i,$$

and that each value is attained. For $j = 1, 2, 3$ and some fixed choice of the first entry in each, define doubly-infinite sequences

$$K_1 = (\overline{1, 3}), \quad K_2 = (\overline{1, 2, 1, 6}), \quad K_3 = (\overline{3, 1}, 4, \overline{1, 3}). \tag{7.6}$$

Here an overlined subsequence adjacent to a parenthesis indicates that one must concatenate the subsequence infinitely many times in the direction of the parenthesis.

**Proposition 8** *Let $\mathcal{C}_j$ be the modular billiard associated to $K_j$ for $j = 1, 2, 3$. Then*

$$\lambda_i(\mathcal{C}_1) = \tfrac{1}{2}\sqrt{21}, \quad \lambda_i(\mathcal{C}_2) = \tfrac{2}{3}\sqrt{14}, \quad \lambda_i(\mathcal{C}_3) = \tfrac{1}{3}(3 + \sqrt{21}). \tag{7.7}$$

*Each value $\lambda_i(\mathcal{C}_j)$ is attained. All three billiards are proper and non-orientable; $\mathcal{C}_1, \mathcal{C}_2$ are periodic, while $\mathcal{C}_3$ is not periodic.*

***Proof*** That (7.7) holds and that each $\lambda_i(\mathcal{C}_j)$ is attained is a straightforward application of Proposition 7. Clearly $K_1, K_2$ are periodic and palindromic and $K_3$ is not periodic but is palindromic. Thus the final statement follows from Theorem 5. □

## 8 Exceptional sequences

In this and the next section we complete the proof of Theorem 3. By Proposition 4 we may assume that the billiard $\mathcal{B}$ is proper. The main result of this section is the following proposition, which (together with Proposition 7) shows that the only proper billiards that stay farther away from $z = i$ than $\mathcal{C}_3$ are $\mathcal{C}_2$ and $\mathcal{C}_1$. The method is completely elementary and amounts to finding inequalities determined by continued fractions.

**Proposition 9** *Unless $K$ is equivalent to $K_j$ for $j = 1, 2, 3$ we have that*

$$\mu(K) < \tfrac{1}{4}\left(\sqrt{21} - 3\right) = 0.395644\ldots.$$

We will say that any $K$ with $\mu(K) \geq 0.395644\ldots$ is *exceptional.* The proof of Proposition 9 consists of a series of results that successively eliminate configurations of subsequences in a $K$ that force it to not be exceptional. Since $\mu(K) = \mu(K^*)$, it is clearly permissible to only prove it for either $K$ or $K^*$. Hence we will often only provide estimates for one of them and might not mention when reversals must also be considered in order to cover all cases.

**Proposition 10** *An exceptional $K$ must have the form*

$$K = (\ldots, 1, m_1, 1, m_2, 1, \ldots)$$

*where $m_j \geq 2$. If any $m_j = 2$ then $K$ is equivalent to $K_2 = (\overline{1, 2, 1, 6})$.*

Proposition 10 will be proven in the four lemmas that follow.

**Lemma 1** *Given any $K$, if $K$ contains any subsequence of the form*

$$(1, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (m, m') \tag{8.1}$$

*for $3 \leq m \leq m'$ then $\mu'(K) \leq \frac{1}{3}$ and therefore $K$ is not exceptional.*

***Proof*** If $K$ contains the subsequence $(m, m')$ for any $m, m' \in \mathbb{Z}^+$ there is an $n \in \mathbb{Z}$ so that

$$r_{n-1} = [m, m', \ldots], \quad r_n = [m', \ldots], \quad s_{n-1} = [0, \ldots], \quad s_n = [0, m, \ldots].$$

Thus $[m, m', 1] < r_{n-1} < [m, m']$, $0 < s_{n-1} < 1$, $[m'] < r_n < [m', 1]$, $[0, m, 1] < s_n < [0, m]$ and so it follows that

$$\frac{m}{mm' + m + 1} < \frac{1}{r_n + s_n} < \frac{m + 1}{mm' + m' + 1}$$

and

$$\frac{m'}{mm' + m' + 1} < \frac{1}{r_{n-1} + s_{n-1}} < \frac{m' + 1}{mm' + m + 1}.$$

The result now follows from (7.1) and (7.4) since for the pairs in (8.1) it can be easily verified that

$$\left| \frac{m - m' - 1}{mm' + m + 1} \right|, \left| \frac{m' - m - 1}{mm' + m' + 1} \right| \leq \frac{1}{3}.$$

□

A useful consequence of Lemma 1 is that (1, 1) cannot occur as a subsequence in an exceptional $K$. We will in several places use this fact without further mention.

**Lemma 2** *If $K$ contains the sequence $(\ell', 2, \ell)$ where $\ell \neq 1$ or $\ell' \neq 1$ then $K$ is not exceptional.*

***Proof*** By Lemma 1 we may assume that either (i) $\ell, \ell' \geq 7$ or (ii) that $\ell' = 1$ and $\ell \geq 7$. In case (i) there is an $n \in \mathbb{Z}$ so that

$$r_{n-1} = [\ell', 2, \ell, \dots], \quad r_n = [2, \ell, \dots], \quad s_{n-1} = [0, \dots], \quad s_n = [0, \ell', \dots],$$

hence

$$\begin{aligned} r_{n-1} &> [7, 2, 7] = \tfrac{112}{15}, \quad 0 < s_{n-1} < 1, \\ 2 < r_n &< [2, 7]7 = \tfrac{15}{7}, \quad 0 < s_n < [0, 7] = \tfrac{1}{7}. \end{aligned}$$

It follows that

$$0 < \frac{1}{r_{n-1} + s_{n-1}} < \tfrac{15}{112} \quad \text{and} \quad -\tfrac{1}{16} < \frac{2 - r_n + s_n}{r_n + s_n} < \tfrac{1}{14}$$

so that by (7.2) we have that $\mu_n''(K) < \frac{11}{56} = 0.196429\dots$.

In case (ii) either (a) $K$ contains $(m', 1, m, 1, 2, \ell)$ where $m, m' \geq 2$ or (b) $K$ contains $(m, 2, 1, 2, \ell)$ where $m \geq 7$.

In case (a) there is an $n \in \mathbb{Z}$ so that

$$\begin{aligned} r_{n-1} &= [m, 1, 2, \ell, \dots], \quad r_n = [1, 2, \ell, \dots], \\ s_{n-1} &= [0, 1, m', \dots], \quad s_n = [0, m, 1, m', \dots]. \end{aligned}$$

Thus

$$m+\tfrac{2}{3} < r_{n-1} < m+\tfrac{15}{22}, \quad \tfrac{2}{3} < s_{n-1} < 1, \quad \tfrac{22}{15} < r_n < \tfrac{3}{2}, \quad \tfrac{1}{m+1} < s_n < \tfrac{1}{m+\frac{2}{3}}.$$

Hence

$$\frac{(m+3)(3m+2)}{3(m+1)(3m+4)} < \frac{2-r_n+s_n}{r_n+s_n} < \frac{(m+1)(24m+61)}{(3m+2)(22m+37)}$$

and

$$\frac{1}{m+\frac{37}{22}} < \frac{1}{r_{n-1}+s_{n-1}} < \frac{1}{m+\frac{4}{3}}.$$

It now follows easily that $\mu_n''(K) < \frac{4}{11} = 0.363636\ldots$.

The same kind of computation shows that in case (b) we have

$$\tfrac{8}{3} < r_{n-1} < \tfrac{59}{22}, \quad 0 < s_{n-1} < \tfrac{1}{7}, \quad \tfrac{22}{15} < r_n < \tfrac{3}{2}, \quad \tfrac{7}{15} < s_n < \tfrac{1}{2}$$

and so $\mu_n''(K) < \frac{157}{870} = 0.18046\ldots$.

By (7.4) and (7.5) the result follows. □

Lemmas 1 and 2 prove the first statement of Proposition 10.

**Lemma 3** *Suppose that $K$ contains the subsequence $(1,2,1,m)$. If $m \neq 6$ then $K$ is not exceptional.*

***Proof*** Suppose that $1 < m < 5$. By Lemmas 1 and 2 we may assume that for some $n \in \mathbb{Z}$

$$r_{n-1} = [2,1,m,1,m',\ldots], \quad s_{n-1} = [0,1,m'',\ldots],$$
$$r_n = [1,m,1,m',\ldots] \quad \text{and} \quad s_n = [0,2,1,m'',\ldots],$$

where $m', m'' \geq 2$. It follows that

$$[2,1,m,1,2] < r_{n-1} < [2,1,m,1] \quad \text{and} \quad \tfrac{2}{3} < s_{n-1} < 1,$$

while

$$[1,m,1] < r_n < [1,m,1,2] \quad \text{and} \quad [0,2,1] < s_n < [0,2,1,2].$$

Hence

$$\frac{m+2}{4m+7} < \frac{1}{r_{n-1}+s_{n-1}} < \frac{9m+15}{33m+46} \quad \text{and} \quad \frac{8(3m+2)}{33m+46} < \frac{1}{r_n+s_n} < \frac{3(m+1)}{4m+7}.$$

A calculation now shows that for $m \leq 4$

$$\left|\frac{1}{r_n + s_n} - \frac{1}{r_{n-1} + s_{n-1}}\right| < \frac{2m+1}{4m+7} < \frac{9}{23} = 0.391304\ldots$$

and the statement of the Lemma in this case follows by (7.1), (7.4) and (7.5).

Now assume that $m \geq 7$. For this we will apply (7.3) to the reversed sequence $(m, 1, 2, 1)$. By the above we may assume that

$$r_{n-1} = [m, 1, 2, 1, m', 1, \ldots], s_{n-1} = [0, 1, m'', \ldots],$$
$$r_n = [1, 2, 1, m', \ldots], s_n = [0, m, 1, m'', \ldots]$$

where $m' \geq 5$ and $m'' \geq 2$. Then a calculation using

$$r_{n-1} > [m, 1, 2, 1, 5], \quad [1, 2, 1] < r_n < [1, 2, 1, 5],$$
$$s_{n-1} > [0, 1, 2], \quad [0, m, 1] < s_n < [0, m, 1, 2]$$

gives

$$0 < \frac{2}{r_{n-1} + s_{n-1}} < \frac{2}{m + \frac{97}{69}} \quad \text{and} \quad \frac{3m-7}{9m+6} < \frac{r_n - s_n - 1}{r_n + s_n} < \frac{3(6m-11)}{17(4m+7)}.$$

Thus we have

$$\left|\frac{2}{r_{n-1} + s_{n-1}} + \frac{r_n - s_n - 1}{r_n + s_n}\right| < \frac{3\left(414m^2 + 2951m + 4407\right)}{17(4m+7)(69m+97)} \leq \frac{2721}{6902}$$
$$= 0.394234\ldots$$

for $m \geq 7$.

Suppose now that $K$ contains $(1, 2, 1, 5)$. By the above we may assume that $K$ contains $(1, 5, 1, 2, 1, 5, 1)$ or $(1, 6, 1, 2, 1, 5, 1)$.

If $K$ contains $(1, 5, 1, 2, 1, 5, 1)$ we have for some $n$ that

$$[1, 2, 1, 5, 1] < r_{n-1} < [1, 2, 1, 5, 1, 2], \quad [0, 5, 1] < s_{n-1} < [0, 5, 1, 2]$$

and

$$[2, 1, 5, 1, 2] < r_n < [2, 1, 5, 1], \quad [0, 1, 5, 1, 2] < s_n < [0, 1, 5, 1].$$

This gives

$$0.65473\ldots = \tfrac{969}{1480} < \frac{1}{r_{n-1} + s_{n-1}} < \tfrac{60}{91} = 0.659341\ldots \quad \text{and}$$
$$0.269231\ldots = \tfrac{7}{26} < \frac{1}{r_n + s_n} < \tfrac{10}{37} = 0.27027\ldots$$

and hence

$$\mu'_n(K) < \tfrac{71}{182} = 0.39011\ldots.$$

Similarly, if $K$ contains $(1, 5, 1, 2, 1, 6, 1)$ we have for some $n$ the same inequalities for $s_{n-1}$ and $s_n$ while

$$[1, 2, 1, 6, 1] < r_{n-1} < [1, 2, 1, 6, 1, 2] \quad \text{and} \quad [2, 1, 6, 1, 2] < r_n < [2, 1, 6, 1].$$

This gives

$$0.655757\ldots = \tfrac{1122}{1711} < \frac{1}{r_{n-1} + s_{n-1}} < \tfrac{138}{209} = 0.669856\ldots \quad \text{and}$$
$$0.267943\ldots = \tfrac{56}{209} < \frac{1}{r_n + s_n} < \tfrac{460}{1711} = 0.268849\ldots.$$

This gives

$$\mu'_n(K) < \tfrac{82}{209} = 0.392344\ldots.$$

Hence by (7.4) and (7.5) we are done. □

If we now assume that $K$ contains $(1, 2, 1, 6)$ then in fact we may assume that $K$ contains $(1, 2, 1, 6, 1, m, 1)$ where $m \geq 2$.

**Lemma 4** *Suppose that $K$ contains $(1, 2, 1, 6, 1, m, 1)$ where $m > 2$. Then $K$ is not exceptional.*

***Proof*** We may assume that for some $n$

$$[6, 1, m, 1, 2] < r_{n-1}, \quad [0, 1, 2] < s_{n-1}$$

and

$$[1, m, 1] < r_n < [1, m, 1, 2], \quad [0, 6, 1, 2, 1] < s_n < [0, 6, 1, 2].$$

Thus we have that

$$0 < \frac{2}{r_{n-1} + s_{n-1}} < \frac{6(3m + 5)}{69m + 106}$$

and

$$\frac{-9m^2 + 45m + 34}{(m + 1)(69m + 106)} < \frac{r_n - s_n - 1}{r_n + s_n} < -\frac{(m + 1)(12m - 73)}{(3m + 2)(31m + 58)}.$$

Hence

$$\left|\frac{2}{r_{n-1}+s_{n-1}}+\frac{r_n-s_n-1}{r_n+s_n}\right| < \frac{846m^3+9975m^2+20671m+11218}{(3m+2)(31m+58)(69m+106)} \leq \frac{185848}{519893} = 0.357474\ldots.$$

for $m \geq 3$. □

This gives the second statement and thus completes the proof of Proposition 10. Suppose now that $K = (\ldots, 1, m_1, 1, m_2, 1, \ldots)$ is exceptional with $K \neq K_1$ and $K \neq K_2$. By Proposition 10 and (7.6) we have that $m_j \geq 3$ for all $j$ with at least one $m_j > 3$. Now for $K_3$ given in (7.6) we have that

$$\mu_n'''(K_3) = \left|\frac{2}{r_{n-1}+s_{n-1}}+\frac{r_n-s_n-1}{r_n+s_n}\right| = \frac{1}{4}\left(\sqrt{21}-3\right),$$

when

$$r_n = [1,3,1,3,1,\ldots] \quad \text{and} \quad s_n = [0,4,1,3,1,3,\ldots].$$

Since $\mu(K) \leq \mu_n'''(K)$ by (7.4) and (7.5), Proposition 9 is a consequence of the following lemma, which implies that

$$\mu_n'''(K) < \mu_n'''(K_3)$$

unless $K$ is equivalent to $K_3$.

**Lemma 5** *Suppose that*

$$r_n = [1, m_1, 1, m_2, 1, \ldots], \quad s_n = [0, m_0, 1, m_{-1}, 1, \ldots]$$

*and*

$$r_n' = [1, m_1', 1, m_2', 1, \ldots], \quad s_n' = [0, m_0', 1, m_{-1}', 1, \ldots]$$

*with $m_j' \geq m_j \geq 1$ for all $j \in \mathbb{Z}$. Then*

$$\left|\frac{2}{r_{n-1}'+s_{n-1}'}+\frac{r_n'-s_n'-1}{r_n'+s_n'}\right| \leq \left|\frac{2}{r_{n-1}+s_{n-1}}+\frac{r_n-s_n-1}{r_n+s_n}\right|,$$

*with equality if and only if $m_j' = m_j$ for all $j \in \mathbb{Z}$.*

***Proof*** Under our assumptions we have $r_{n-1} \le r'_{n-1}$ and $r_n \ge r'_n$. Now $s_{n-1} = \frac{1}{s_n} - m_0$ and $r_{n-1} = \frac{1}{r_n} + m_0$ and similarly for $r'$, $s'$. Thus

$$\frac{2}{r_{n-1} + s_{n-1}} + \frac{r_n - s_n - 1}{r_n + s_n} = \frac{2r_n s_n + r_n - s_n - 1}{r_n + s_n}.$$

The result now follows since the function

$$F(x, y) = \frac{2xy + x - y - 1}{x + y}$$

satisfies $0 < F(x_1, y_1) \le F(x_2, y_2)$ whenever $1 < x_1 \le x_2 < 2$ and $0 < y_1 \le y_2 < 1$, with equality if and only if $x_1 = x_2$ and $y_1 = y_2$. To see this, use that the gradient of $F(x, y)$ is given by

$$\nabla F(x, y) = \left( \tfrac{2y(y+1)+1}{(x+y)^2}, \tfrac{2(x-1)x+1}{(x+y)^2} \right).$$

□

## 9 Approximating $\mathcal{C}_3$

We conclude by justifying the final statement of Theorem 3. For $\ell \ge 1$, let

$$K_\ell = (\overline{3, 1}, 4, 1, 3, 1, \ldots, 3, 1, 4, \overline{1, 3}), \quad k_0 = k_{2\ell} = 4,$$

where the number of 1's between the 4's is given by $\ell$. Propositions 7 and 9 show that the billiard $\mathcal{B}_\ell$ associated to $K_\ell$ satisfies $\lambda_i(\mathcal{B}_\ell) > \frac{1}{3}(3 + \sqrt{21})$.

**Proposition 11** *Let $\mathcal{B}_\ell$ be the billiard associated to the class of $K_\ell$ by Theorem 5. Then*

$$\lim_{\ell \to \infty} \lambda_i(\mathcal{B}_\ell) = \tfrac{1}{3}(3 + \sqrt{21}).$$

***Proof*** For a fixed $\ell$, define $r_n = r_n(\ell)$ and $s_n = s_n(\ell)$ as in (6.2) for the sequence $K_\ell$. Then

$$r_1 = [1, 3, 1, \ldots, 3, 1, 4, \overline{1, 3}] \quad \text{and} \quad s_1 = [0, 4, \overline{1, 3}] = \tfrac{1}{2}(5 - \sqrt{21}),$$

where the bar indicates a repeated sequence, and the number of 1's before the 4 in $r_1$ is given by $\ell$. One can easily compute an explicit formula for $r_1$ using the recurrence relation

$$r_1(\ell + 1) = 1 + \frac{1}{3 + \frac{1}{r_1(\ell)}}, \quad r_1(1) = \tfrac{1}{2}(7 - \sqrt{21}). \tag{9.1}$$

Let $\varepsilon = \frac{1}{2}(5+\sqrt{21})$ denote the fundamental unit in $\mathbb{Q}(\sqrt{21})$. Then by (7.3) and (6.3) we have the formula

$$\mu_1'''(K_\ell) = \frac{(-3+2\sqrt{21})\varepsilon^\ell - 3\varepsilon^{1-\ell}}{(11+\sqrt{21})\varepsilon^\ell - \frac{1}{2}\varepsilon^{-\ell}}.$$

We claim that $\mu(K_\ell) = \mu_1'''(K_\ell)$; the result then follows easily since $\lambda_i(\mathcal{B}_\ell) = \mu(K_\ell)^{-1}$.

It is straightforward to check that $\mu(K_\ell) = \mu'''(K_\ell)$ for small $\ell$, so we assume $\ell \geq 3$. To prove that $\mu(K_\ell) = \mu'''(K_\ell)$, we first show that $\mu_n'(K_\ell), \mu_n''(K_\ell) > \frac{-3+\sqrt{21}}{4} = .395\ldots$ for all $n$ and that $\mu_n'''(K_\ell) > \frac{-3+\sqrt{21}}{4}$ if $n \neq \pm 1, 2\ell \pm 1$. If $n$ is odd, then

$$\begin{aligned} \tfrac{1+\sqrt{2}}{2} = [\overline{1,4}] \leq r_n \leq [\overline{1,3}] = \tfrac{3+\sqrt{21}}{6}, \\ \tfrac{-1+\sqrt{2}}{2} = [0,\overline{4,1}] \leq s_n \leq [0,\overline{3,1}] = \tfrac{-3+\sqrt{21}}{6}. \end{aligned}$$

If $n$ is even and $k_n = 4$, then

$$\begin{aligned} \tfrac{5+\sqrt{21}}{2} = [4,1,\overline{3,1}] \leq r_n \leq [\overline{4,1}] = 2+2\sqrt{2}, \\ \tfrac{-3+\sqrt{21}}{2} = [0,\overline{1,3}] \leq s_n \leq [0,1,3,\overline{1,4}] = \tfrac{3-\sqrt{2}}{2}, \end{aligned}$$

while if $n$ is even and $k_n \neq 4$, then

$$\begin{aligned} \tfrac{3+\sqrt{21}}{2} = [\overline{3,1}] \leq r_n \leq [3,1,\overline{4,1}] = 1+2\sqrt{2}, \\ \tfrac{-3+\sqrt{21}}{2} = [0,\overline{1,3}] \leq s_n \leq [0,\overline{1,4}] = -2+2\sqrt{2}, \end{aligned}$$

It follows from (7.1), (7.2), and (6.3) that in each case, $\mu_n'(K_\ell), \mu_n''(K_\ell) > .399$. Similarly, $\mu_n'''(K_\ell) > .399$ if $n$ is even. If $n$ is odd and $k_{n+1}, k_{n-1} \neq 4$ then

$$r_n \geq [1,3,\overline{1,4}] = \tfrac{6+2\sqrt{2}}{7}, \quad s_n \geq [0,3,1,\overline{4,1}] = \tfrac{-1+2\sqrt{2}}{7},$$

from which it follows that $\mu_n'''(K_\ell) > .399$.

It remains to show that $\mu_{-1}(K_\ell) > \mu_1(K_\ell)$ since, by symmetry, we have $\mu_{2\ell+1}(K_\ell) > \mu_{2\ell-1}(K_\ell) = \mu_1(K_\ell)$. By (6.3) we have

$$\mu_{-1}'''(K_\ell) = \mu_1'''(K_\ell) + \frac{6(4r_1s_1 + s_1 - r_1)}{r_1 + s_1}.$$

Thus $\mu_{-1}(K_\ell) > \mu_1(K_\ell)$ if and only if $r_1 < \frac{s_1}{1-4s_1} = \frac{1}{6}(3+\sqrt{21})$. This inequality follows from the relation (9.1), which completes the proof. □

**Acknowledgements** The second author thanks Alex Kontorovich for some enlightening discussions on the topics of this paper. The authors thank the referee for several constructive comments that have improved the exposition of this paper.

## References


1. Aigner, M.: Markov's Theorem and 100 Years of the Uniqueness Conjecture. A Mathematical Journey from Irrational Numbers to Perfect Matchings. Springer, Cham (2013)
2. Artin, E.: Ein mechanisches System mit quasiergodischen Bahnen. Hamb. Math. Abh. **3**, 170–177 (1924)
3. Beardon, A.F.: The Geometry of Discrete Groups. Graduate Texts in Mathematics, vol. 91. Springer, New York (1983)
4. Berstel, J., Lauve, A., Reutenauer, C., Saliola, F.: Combinatorics on Words. Christoffel Words and Repetitions in Words. CRM Monograph Series, vol. 27. AMS, Providence (2009)
5. Bombieri, Enrico: Continued fractions and the Markoff tree. Expos. Math. **25**(3), 187–213 (2007)
6. Cassels, J.W.S.: An Introduction to Diophantine Approximation. Cambridge Tracts in Mathematics and Mathematical Physics, No. 45. Cambridge University Press, New York (1957)
7. Cohn, H.: Approach to Markoff's minimal forms through modular functions. Ann. Math. (2) **61**, 1–12 (1955)
8. Cohn, H.: Markoff geodesics in matrix theory. In: Number Theory with an Emphasis on the Markoff Spectrum (Provo, UT, 1991), Lecture Notes in Pure and Applied Mathematics, vol. 147, pp. 69–82. Dekker, New York (1993)
9. Cusick, T.W., Flahive, M.E.: The Markoff and Lagrange Spectra. Mathematical Surveys and Monographs, vol. 30. American Mathematical Society, Providence (1989)
10. Dickson, L.E.: Modern Elementary Theory of Numbers. University of Chicago Press, Chicago (1939)
11. Dickson, L.E.: Studies in the Theory of Numbers. Chicago University Press, Chicago (1930)
12. Ford, L.R.: A geometrical proof of a theorem of Hurwitz. Proc. Edinb. Math. Soc. **35**, 59–65 (1917)
13. Freiman, G.A.: Diofantovy priblizheniya i geometriya chisel (zadacha Markova). (Russian) [Diophantine approximations and the geometry of numbers (Markov's problem)] Kalinin. Gosudarstv. Univ., Kalinin (1975)
14. Frobenius, G.: Über die Markoffschen Zahlen, Preuss. Akad. Wiss. Sitzungberichte (1913) 458–487 (also in: G. Frobenius, Gesammelte Abhandlungen, Bd. 3, Springer, Berlin, Heidelberg, New York, 1968, pp. 598–627)
15. Hall Jr., M.: On the sum and product of continued fractions. Ann. Math. (2) **48**, 966–993 (1947)
16. Hall Jr., M.: The Markoff spectrum. Acta Arith. **18**, 387–399 (1971)
17. Haas, A., Series, C.: The Hurwitz constant and Diophantine approximation on Hecke groups. J. Lond. Math. Soc. (2) **34**(2), 219234 (1986)
18. Haas, A.: Diophantine approximation on hyperbolic Riemann surfaces. Acta Math. **156**(1–2), 3382 (1986)
19. Lehner, J., Sheingorn, M.: Simple closed geodesics on $H^+/\Gamma(3)$ arise from the Markov spectrum. Bull. Am. Math. Soc. (N.S.) **11**(2), 359362 (1984)
20. Malyshev, A.V.: Markov and Lagrange spectra [Survey of the literature]. Zap. Nauch. Sem. Lenin. Otd. Math. Inst. V.A. Steklova AN SSSR **67**, 5–38 (1977). (English translation in J. Soviet Math. **16** (1981) 767–788)
21. Markoff, A.: Sur les formes quadratiques binaires indéfinies. Math. Ann. **15**, 381–409 (1879)
22. Markoff, A.: Sur les formes quadratiques binaires indéfinies. Math. Ann. **17**, 379–399 (1880)
23. Sarnak, P.: Reciprocal geodesics. In: Analytic Number Theory, pp. 217–237, Clay Mathematics Proceedings, vol. 7, American Mathematical Society, Providence, RI (2007)
24. Schur, I.: Zur Theorie der indefiniten binären quadratischen Formen, Sitzungberichte der Preussischen Akademie Wiss. 212–231 (1913)