Rational billiards and flat structures

Howard Masur[*] and Serge Tabachnikov[†]

Contents

[*]University of Illinois at Chicago

[†]University of Arkansas, Fayetteville

Acknowledgements

Both authors were supported by the National Science Foundation grants. The second author is grateful to Max-Planck-Institut in Bonn for its hospitality.

1 Billiards, polygonal billiards, rational billiards

1.1 General billiards

The topics of this survey are plane polygonal billiards, rational billiards, and their generalization to quadratic differentials or flat structures. However, to put the expositon into a perspective, we start with some generalities on the billiard dynamical systems (see, e.g., [10], [30], [48]).

A billiard table Q is a Riemannian manifold with a piece-wise smooth boundary ∂Q. The billiard dynamical system in Q is generated by the free motion of a mass-point (thought of as the billiard ball) subject to the elastic reflection in the boundary. This means that the point moves along a geodesic line in Q with a constant (say, unit) speed until it hits ∂Q. At a smooth boundary point the billiard ball instantaneously reflects so that the tangential component of its velocity remains the same while the normal component changes sign. In dimension two, the case of interest in this paper, the reflection is described by the familiar law of geometrical optics: *the angle of incidence equals the angle of reflection.* If the billiard ball hits a non-smooth boundary point (a corner) its further motion is not defined (with a few exceptions mentioned below).

The time-t billiard map acts on tangent vectors to Q, namely, on those tangent vectors whose trajectories undergo finitely many reflections in the boundary and avoid corners on the time interval $[0, t]$.

Thus the billiard flow is the geodesic flow on a manifold with boundary; this flow is discontinuous at the moments when the billiard ball reflects in the boundary. Denote by σ the billiard reflection at a smooth boundary point x; then σ is the linear reflection of the tangent space T_xQ in the hyperplane $T_x(\partial Q)$. It is convenient to identify the tangent vectors (x, v) and $(x, \sigma(v))$ for every smooth boundary point x and every tangent vector $v \in T_xQ$. These identifications being made, the billiard flow becomes continuous at the smooth boundary points; however the discontinuities, corresponding to the corners of ∂Q, persist.

One may consider the billiard flow as a Hamiltonian or as a Lagrangian dynamical system. This point of view will not play role in the present paper, so we mention it only in passing. The cotangent bundle T^*Q carries the canonical symplectic structure "$dp \wedge dq$"; using the identification of the tangent and cotangent spaces induced by the Riemannian metric, one obtains a symplectic structure on TQ. The billiard flow preserves this symplectic structure. The dual, Lagrangian viewpoint is that the billiard trajectories extremize the distance between their points. For example, a closed three-link billiard trajectory in Q has an extremal length among the geodesic triangles whose vertices lie on the boundary ∂Q.

Assume that Q is geodesically convex and compact. In the spirit of geometrical optics one may consider *the billiard transformation* acting on the oriented geodesics in Q (thought of as rays of light): the geodesic ray that contains a segment of the trajectory of the billiard ball, oriented in the direction of its motion, is sent to the ray that contains the next segment of this trajectory after the reflection in the boundary. The billiard transformation preserves the canonical symplectic structure of the space of oriented geodesics obtained from the symplectic structure in the cotangent bundle by symplectic reduction.

The described construction reduces the dimension of the phase space by one: the unit tangent bundle is replaced by its quotient space of geodesic rays, and the billiard flow – by the billiard map. Taking the first intersection point of a geodesic ray with the boundary, one also describes the phase space of the billiard map as the space of the unit tangent vectors $(x, v) \in TQ$ whose foot point x is a smooth boundary point and the vector v has the inward direction.

This paper concerns the dynamics of billiards. The billiard flow has an invariant measure (indiced by the invariant symplectic form): this measure in the unit tangent bundle is the product of the measure on Q induced by the Riemannian volume form and the Lebesgue measure on the unit tangent spheres. One can prove that for every billiard table the billiard flow is defined for all times on a subset of full measure in the unit tangent bundle of Q.

1.2 Polygonal billiards

Informally speaking, the theory of mathematical billiards can be partitioned into three areas: convex billiards with smooth boundaries, billiards in polygons (and polyhedra) and dispersing and semi-dispersing billiards (similarly to differential geometry in which the cases of positive, zero and negative curvature are significantly different). These areas differ by the types of results and the methods of study: in the former a prominent role is played by the KAM theory and the theory of area preserving twist maps; the latter concerns hyperbolic dynamics and has much in common with the study of the geodesic flow on negatively curved manifolds. The recent progress in the study of polygonal billiards is mostly due to applications of the theory of flat structures on surfaces (a.k.a. quadratic differentials) and the study of the action of the Lie group $\mathbf{SL}(2, \mathbf{R})$ on the space of quadratic differentials which is the main topic of this paper.

From now on we will be considering plane polygonal billiards; this means that Q is a compact polygon in the plane (in general, not necessarily convex or simply connected). We describe the invariant measure μ for the billiard transformation induced by the above mentioned invariant measure of the billiard flow.

The phase space of T consists of the inward unit tangent vectors (x, v) whose foot point x is an interior point of a side of Q. Let t be the length parameter along the perimeter of Q and let θ be the angle made by v with the respective side. Then the invariant measure is given by the next formula:

$$\mu = \sin\theta \; d\theta \; dt.$$

One often needs to consider parallel beams of billiard trajectories in polygons. For example, every even-periodic trajectory includes into a strip consisting of a one-parameter family of parallel periodic trajectories of the same period and length. An odd-periodic trajectory also includes into a parallel strip, this time consisting of trajectories whose period and length is twice as great. Note that the billiard reflection in a side of Q transforms a parallel beam to another family of parallel trajectories, and that the width of the beam remains the same.

Fig. 1

Part of the interest in polygonal billiards comes from the fact that they are closely related to problems of mechanics. We discuss one such simple example below.

1.3 Examples: a pair of ellastic point-masses on a segment and a triple of point-masses on a circle

Consider two points with masses m_1 and m_2 on the unit segment $[0,1]$. The points may elastically collide and reflect from the end points of the segment ("walls"). The configuration space of the system is given by the inequalities $0 \le x_1 \le x_2 \le 1$ where x_1 and x_2 are the coordinates of the points. Rescale the variables: $\bar{x}_i = \sqrt{m_i} x_i, \; i = 1,2$. The configuration space is now a right triangle in the $(\bar{x}_1 \bar{x}_2)$-plane with the acute angle $\tan^{-1}(\sqrt{m_1/m_2})$.

Consider a collision of the points. Let v_1, v_2 be the velocities before, and u_1, u_2 – after the collision. The conservation of momentum and energy laws read:

$$m_1 u_1 + m_2 u_2 = m_1 v_1 + m_2 v_2,$$

$$m_1 u_1^2/2 + m_2 u_2^2/2 = m_1 v_1^2/2 + m_2 v_2^2/2.$$

In the rescaled coordinates the velocities are rescaled by the same factors, therefore

$$\sqrt{m_1} \bar{u}_1 + \sqrt{m_2} \bar{u}_2 = \sqrt{m_1} \bar{v}_1 + \sqrt{m_2} \bar{v}_2,$$

$$\bar{u}_1^2 + \bar{u}_2^2 = \bar{v}_1^2 + \bar{v}_2^2.$$

The latter equation says that the magnitude of the vector $(\bar{u}_1, \bar{u}_2)$ does not change after the collision; the former one says that the scalar product of this vector with the vector $(\sqrt{m_1}, \sqrt{m_2})$ is preserved as well. The vector $(\sqrt{m_1}, \sqrt{m_2})$ is tangent to the side of the configuration triangle given by the equation $\bar{x}_1/\sqrt{m_1} = \bar{x}_2/\sqrt{m_2}$.

Thus the configuration trajectory reflects in this side according to the billiard reflection law. Likewise one considers collisions with the walls $x = 0$ and $x = 1$: they correspond to the billiard reflections in the other two sides of the configuration triangle. One concludes that the dynamical system of two elastic particles on a segment is isomorphic to the billiard in a right triangle whose shape depends on the ratio of the masses of the particles.

Similar arguments show that the system of three elastic point-masses on the circle with a fixed center of mass is isomorphic to the billiard in an acute triangle (we learned this example from E. Gutkin). Let the masses be m_1, m_2, m_3; then the angles α_i of the triangle are given by

$$\tan \alpha_i = m_i \sqrt{\frac{m_1 + m_2 + m_3}{m_1 m_2 m_3}}, \quad i = 1, 2, 3.$$

In the limit $m_3 \to \infty$ one obtains the previous example of two point-masses on a segment.

Many other mechanical models reduce to billiards, in particular, the model of gas in a closed vessel as a collection of elastic balls in a compact domain. Such models are of great importance in statistical physics.

1.4 Unfolding billiard trajectories, rational polygons

Instead of reflecting a billiard trajectory in a side of the billiard polygon Q one may reflect Q in this side and unfold the trajectory to a straight line. This process is iterated at each reflection: each succesive copy of the billiard polygon is obtained from the previous one by the reflection in the side, met by the straightened trajectory. This unfolding method has many applications in the study of polygonal billiards.

Fig. 2

Let $A(Q)$ be the group of motions of the plane generated by the reflections in the sides of Q. Denote the reflection in the side s by σ_s. Notice that for every two sides s and t of Q one has:

$$\sigma_{\sigma_s(t)} = \sigma_s \sigma_t \sigma_s.$$

It follows that every copy of Q involved in the unfolding is the image of Q under an element of the group $A(Q)$. The product of an even number of elements of this group preserves orientation while an odd number reverses it.

To keep track of the directions of billiard trajectories in Q consider the group $G(Q)$ that consists of the linear parts of the motions from $A(Q)$. This subgroup of the orthogonal group is generated by the reflections in the lines through the origin which are parallel to the sides of the polygon Q. The group $G(Q)$ acts on the unit circle. When a billiard trajectory in Q reflects in a side s its direction is changed by the action of the element of $G(Q)$ which is the projection of σ_s to $G(Q)$.

Definition 1.1 *A billiard polygon Q is called rational if the group $G(Q)$ is finite and irrational otherwise.*

If Q is a rational polygon then the group $G(Q)$ is the dihedral group of symmetries of a regular polygon. A necessary condition for Q to be rational is that all its angles are rational multiples of π. It is also sufficient if the boundary is connected; that is, the polygon is simply connected. The set of plane n-gons can be considered as a subset in $\mathbf{R}^{2n}$ (each vertex has two degrees of freedom). We give the space of n-gons the subspace topology. Rational polygons are dense in the space of polygons.

A given billiard trajectory in a rational billiard table will have only finitely many different directions; this finite collection of directions plays the role of an integral of motion. More precisely, let p be the composition of the projection of the unit tangent bundle $Q \times \mathbf{S^1}$ on $\mathbf{S^1}$ and the projection of $\mathbf{S^1}$ to the quotient space $\mathbf{S^1}/G(Q)$. Then the function p is constant along every billiard trajectory in a rational polygon.

We illustrate the unfolding procedure in the simplest example of a rational polygon, the square.

1.5 Example: billiard in the unit square

Unfolding a trajectory one obtains a line in the plane which is tiled by the unit squares, the images of the original square Q under the action of the group $A(Q)$. Two lines in the plane correspond to the same billiard trajectory in Q if they differ by a translation through a vector from the lattice $2\mathbf{Z} + 2\mathbf{Z}$.

Consider the fundamental domain of the group $2\mathbf{Z} + 2\mathbf{Z}$ which is the square made of four copies of Q. Identify the opposite sides to obtain a flat torus. A billiard trajectory becomes a geodesic line on this torus; every geodesic line has a constant slope λ. The unit tangent bundle of the torus is represented as the union of the tori, ennumerated by the slopes λ; the geodesic flow on each torus is a constant flow.

The dynamics on an individual invariant torus depends on whether λ is rational or irrational: in the former case the geodesic flow is periodic, and in the latter it is ergodic, and in fact, uniquely ergodic. In particular, a billiard trajectory with a rational slope is periodic, while the one with an irrational slope is dense in the square.

One can also analyse periodic trajectories. The unfolding of such a trajectory is a segment in the plane whose end-points differ by a vector from $2\mathbf{Z}+2\mathbf{Z}$. Every parallel trajectory is also periodic with the same period and length.

Assume that an unfolded periodic trajectory goes from the origin to the lattice point $(2p, 2q)$. If p and q are coprime this is a prime periodic trajectory, and if p and q have a comon multiplier then the periodic trajectory is multiple. The length of the trajectory is $2\sqrt{p^2+q^2}$, and to a choice of p and q there correspond two orientations of the trajectory. Thus the number of (strips of) parallel trajectories of length less than L equals the number of pairs of integers, satisfying the inequality $p^2+q^2 < L^2/2$.

This is the number of lattice points inside the circle of radius $L/\sqrt{2}$, centered at the origin. In the first approximation, this number equals the area $\pi L^2/2$, and to take only prime periodic trajectories into account one divides this estimate by $\pi^2/6$. Thus one obtains a quadratic asymptotic estimate on the number of periodic trajectories of the length not exceeding a fixed number. We will see in a later section that this result holds for general rational polygons.

What works so well in the above example is the fact that the images of Q under the group $A(Q)$ tile the plane. A similar consideration applies to rectangles, equilateral triangles and right triangles with an acute angle $\pi/4$ or $\pi/6$. These polygons are called *integrable*, and the billiard flow reduces to a constant flow on a torus. Note that in all these cases one may define the extension of a billiard trajectory through a vertex of the billiard polygon.

Although integrable polygons are exceptional, some of the features of the billiard dynamics in the integrable case extend to general rational polygons. These results will be discussed in succeeding sections.

One class of rational billiards for which one can make such precise statements is the class of *almost integrable* billiards, intermediate between integrable and general rational polygons, studied by Gutkin ([21]). A polygon Q is called almost integrable if the group $A(Q)$ is a discrete subgroup of the group of motions of the plane. There are exactly four such groups generated by the reflections in the sides of the four integrable polygons. An almost integrable polygon is one that can be drawn on the corresponding lattice.

Fig. 3

Given an almost integrable polygon, the billiard flow decomposes into directional flows F_θ (just as in the case of a square – see next section for a detailed discussion). Choose a basis e_1, e_2 of the respective lattice. A direction is called rational if it is given by a vector $a_1 e_1 + a_2 e_2$ with $a_1/a_2 \in \mathbf{Q}$. Gutkin proved that, similarly to the square case, the following conditions are equivalent:
(i) θ is an irrational direction;
(ii) F_θ is minimal;
(iii) F_θ is ergodic;
(iv) F_θ is aperiodic, that is, $F_\theta^t \neq id$ for all $t \neq 0$.

1.6 Rational billiard determines a flat surface

The following construction of a flat surface from a rational billiard table plays a central role in the present study (see [14], [31], [32], [45] and [34]; the later paper was the first to relate billiards in polygons and quadratic differentials).

Let Q be a rational polygon. The group $G(Q)$ is the dihedral group D_N generated by the reflections in the lines through the origin that meet at angles π/N where N is a positive integer. This group has $2N$ elements, and the orbit of a generic point $\theta \neq k\pi/N$ on the unit circle consists of $2N$ points. Let the angles of Q be $\pi m_i/n_i$ where m_i and n_i are coprime integers. Then N is the least common multiple of the denominators n_i.

Consider the unit tangent bundle $Q \times \mathbf{S^1}$, the phase space of the billiard flow, and let M_θ be the subset of points whose projection to $\mathbf{S^1}$ belongs to the orbit of θ under D_N. Then, M_θ is an invariant surface of the billiard flow in Q. This invariant surface is a level surface of the above mentioned function p, "the integral of motion".

Assume that $\theta \neq k\pi/N$ and ennumerate the angles in the D_N-orbit of θ on the unit circle counterclockwise: $\theta = \theta_1, \theta_2, ..., \theta_{2N}$. The surface M_θ is obtained from $2N$ copies of Q, namely, $Q \times \theta_i \subset Q \times \mathbf{S^1}$, $i = 1, ..., 2N$ by gluing their sides according to the action of D_N.

Consider $2N$ disjoint and parallel copies of Q in the plane. Call them $Q_1, ..., Q_{2N}$ and orient the even ones clockwise and the odd ones counterclockwise. Choose an index $i = 1, ..., 2N$ and a side s of Q_i; reflect the direction θ_i in this side. The resulting direction is θ_j for some $j = 1, ..., 2N$. Glue the side s of Q_i to the identical side of

Q_j. After these gluings are made for all values of i and all choices of the side s of Q_i, the sides of all the polygons Q_i are pasted pairwise, and the gluings agrees with the orientation. The result is an oriented compact surface that depends only on the polygon Q, but not on the choice of θ, and we denote it by M. The billiard flows on M in different directions are obtained one from another by rotation.

For example, if Q is a square then $N = 2$ and the result is a torus made of four identical squares. If Q is a right triangle with an acute angle equal to $\pi/8$ then the surface M_θ is obtained from a regular octagon, the result of gluing 16 copies of the triangle, by pairwise gluing its opposite sides; this surface has genus 2.

Fig. 4

The genus of M is given in the next lemma.

Lemma 1.2 *Let the angles of a billiard k-gon be $\pi m_i/n_i,\ i = 1, ..., k$ where m_i and n_i are coprime, and N be the least common multiple of n_i's. Then*

$$\text{genus } M = 1 + \frac{N}{2}\Big(k - 2 - \sum \frac{1}{n_i}\Big).$$

Proof. We need to analyse how the gluings are made around a vertex of Q. Consider the i-th vertex V with the angle $\pi m_i/n_i$. Let G_i be the group of linear transformations of the plane generated by the reflections in the sides of Q, adjacent to V. Then G_i consists of $2n_i$ elements.

According to the construction of M the number of copies of the polygons Q_j that are glued together at V equals the cardinality of the orbit of the test angle θ under the group G_i, that is, equals $2n_i$. Originally we had $2N$ copies of of the polygon Q, and therefore, $2N$ copies of the vertex V; after the gluings we have N/n_i copies of this vertex on the surface M.

It follows that the total number of vertices in M is $N \sum 1/n_i$. The total number of edges is Nk, and the number of faces is $2N$. Therefore the Euler characteristic of M equals

$$N \sum \frac{1}{n_i} - Nk + 2N = 2 - 2g$$

where g is the genus, and the result follows. $\square$

The billiard flow on M is obtained from the constant flows in the directions θ_i in the polygons Q_i. The result is a (unit) vector field on M with singularities at the vertices. The above proof shows that the i-th vertex of M is the result of gluing $2n_i$ copies of the angle $\pi m_i/n_i$ which sums up to an angle of $2\pi m_i$. One may realize such a singularity geometrically as follows. Take m_i copies of a Euclidean upper half plane H_j and m_i copies of a lower half plane L_j; $j = 1, ..., m_i$. Then glue the positive real axis of H_j to the positive real axis of L_j and glue the negative real axis of L_j to the negative real axis of H_{j+1} $(m_i + 1 = 1)$. The result is a singularity with a total angle of $2\pi m_i$. We will call it a cone angle $2\pi m_i$ singularity. From this description one can also see that for any direction θ, there are $2m_i$ separatrices emanating from the singularity in direction θ.

It is easy to describe the set of polygons for which all the angles are 2π, that is, $m_i = 1$; then the singularities of M are removable. The sum of interior angles of a k-gon is $\pi(k-2)$. Thus if $m_i = 1$ for all i then

$$\frac{1}{n_1} + ... + \frac{1}{n_k} = k - 2.$$

This equation has only four solutions with $n_i \geq 2$, considered up to permutations:

$$\Big(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\Big), \ \Big(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\Big), \ \Big(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\Big), \ \Big(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\Big).$$

These solutions correspond to the already mentioned integrable polygons, and in each case the invariant surface M is a torus with a constant flow.

To summarize the construction, given a rational billiard polygon Q, one constructs a compact surface M whose genus is given by the above lemma. This surface inherits a flat metric from Q with a finite number of cone-type singularities, corresponding to the vertices of Q, with cone angles multiples of 2π. The billiard flow on M is a constant flow in a fixed direction with singularities at the cone points.

1.7 Minimality of the billiard flow in rational polygons

The next result on the minimality of the billiard flow in a rational polygon is much easier than the stronger results on ergodicity described later in the paper. We discuss it because it serves a model for these harder theorems.

Definition 1.3 *A flow is called minimal if any of its orbits is dense in the phase space.*

Denote the billiard flow on M in the direction θ by F_θ.

Definition 1.4 *A saddle connection in direction θ is an orbit of F_θ that goes from a singularity to a singularity (possibly, the same one) and has no interior singularities.*

A saddle connection is also called a generalized diagonal since it corresponds to a billiard orbit in direction θ that goes from a vertex to another vertex.

We may decompose a saddle connection β as a sum of its horizontal part h_β and its vertical part v_β.

Definition 1.5 *A metric cylinder in direction θ is the isometric image of a straight Euclidean cylinder in $\mathbf{R}^3$ consisting of closed geodesics in direction θ. A metric cylinder is maximal if it cannot be enlarged.*

We already noted that a periodic orbit in direction θ for the billiard flow included into a strip of parallel periodic orbits thus giving rise to a metric cylinder in direction θ. The following lemmas and theorem are standard (see [47]), and the proofs do not use the group structure coming from billiards. Thus they will hold verbatim for the general situation of flat structures (that will be formally defined shortly).

Lemma 1.6 *If M has singularities, then the boundary of a maximal metric cylinder consists of a finite number of saddle connections in direction θ.*

Proof. If a metric cylinder fills the surface, then M is the flat torus which has no singularities. Thus we can assume the maximal cylinder has a boundary. The obstruction to enlarging a maximal cylinder are saddle connections on the boundary. $\square$

Lemma 1.7 *Suppose α^+ is a trajectory, infinite in the positive direction. Let β be an interval perpendicular to α^+ with initial point P_0 on α^+ as one of its endpoints. Then α^+ returns to β after P_0.*

Proof. Since there are a finite number of singularities, there are a finite number of trajectories starting at points of β that hit a singularity before crossing β again. By shortening β to a subinterval β' with one endpoint P_0 and some other endpoint Q_0, we can assume that no trajectory leaving β' hits a singularity before returning to β. Now flow the interval β' in the positive direction. The interval sweeps out rectangles of increasing area. Since the area of the surface is finite, the interval β' must return and overlap β. If α^+ itself does, we are done. Otherwise the trajectory leaving Q_0 returns to β' and some trajectory leaving a point $Q_1 \in \beta'$ returns to P_0. We now consider the interval $\beta'' \subset \beta'$ with endpoints P_0 and Q_1 and apply the previous analysis to it. Flowing in the forward direction it must return to β and now α^+ itself must return to β. □

Theorem 1.8 *For all but countably many directions θ the flow F_θ is minimal on the surface M.*

Proof. Since there are only finitely many singularities and countably many homotopy classes of arcs joining the singularities, it is clear that there are only countably many saddle connections (the same statement follows from the fact that the group $A(Q)$ is countable.) We claim that the flow F_θ is minimal if there is no saddle connection in direction θ.

By Lemma 1.6 there cannot be a metric cylinder for then there would be a saddle connection. Suppose there was an infinite trajectory l_θ in direction θ which is not dense. Let $A \neq M$ be the set of ω limit points of l_θ. Then A is invariant under the flow F_θ. Since $A \neq M$ one can choose a trajectory γ on ∂A; let P_0 be its initial point. We will show that γ is a saddle connection. If not then γ is infinite in at least one of the two directions. We show that there is an open neighborhood of P_0 contained in A, a contradiction to P_0 being a boundary point.

Let β be a perpendicular arc with P_0 an endpoint. It is enough to show that there exists a segment $[P_0, Q] \subset \beta$ which is contained in A. For then doing this on both sides we would have our open neighborhood. Now Lemma 1.7 implies that γ hits β again at some P_1. If the interval $[P_0, P_1] \subset A$ we are done. Suppose not. Then there exists $Q_1 \in [P_0, P_1]$ which is not in A. Since A is closed, there is a largest open subinterval $I_1 \subset [P_0, P_1]$ containing Q_1 which is in the complement of A. Let P_2 BE the endpoint of I_1 closest to P_0. Then $P_2 \in A$ and the trajectory through P_2 must be a saddle connection. For if it were infinite in either direction, it would intersect I_1. Since A is invariant under the flow, this contradicts that I_1 misses A. □

Projecting back to the rational billiard polygon Q we see that for all but countably many directions every billiard trajectory is dense in Q. Approximating a general billiard by a rational one we obtain the next result on topological transitivity of polygonal billiards ([31]).

Definition 1.9 *A flow is called topologically transitive if it has a dense orbit.*

Theorem 1.10 *For every $k \geq 3$ there exists a dense G_δ subset in the space of simply connected k-gons that consists of polygons with topologically transitive billiard flows.*

Proof. Identify the phase space of the billiard flow in each k-gon with $\mathbf{D^2} \times \mathbf{S^1}$, and assume that this identification depends continuously on the polygon. Let B_i be a countable basis for the topology of $\mathbf{D^2} \times \mathbf{S^1}$. Denote by X_n the set of k-gons Q such that for each open set U in the phase space, there exists a billiard trajectory starting in U that visits all the images of the sets $B_1, ..., B_n$ in the phase space of the billiard flow in Q. Each set X_n is open and their intersection is a G_δ set.

Let us show that this intersection is dense. Let Y_m be the set of rational k-gons with the angles $\pi m_i/n_i$, m_i and n_i coprime, such that the least common multiple of n_i's is at least m. For every polygon $Q \in Y_m$ the invariant surface M_θ is $1/m$-dense in the phase space. Therefore for every n there exists m such that for every $Q \in Y_m$ the surface M_θ intesects all the images of the sets $B_1, ..., B_n$ in the phase space of the billiard flow in Q. Since M_θ has a dense trajectory for all but countably many θ the space Y_m lies in X_n. Finally, Y_m is dense in the space of k-gons for every m, and so X_n is dense for every n as well. It follows that $\cap X_n$ is dense by Baire's theorem.

Let Q be a billiard polygon in $\cap X_n$. We claim that there is a dense billiard trajectory in the phase space of Q. Let U be a compact domain in the phase space $\mathbf{D^2} \times \mathbf{S^1}$. Suppose inductively we have chosen a compact neighborhood $U_{n-1} \subset U$. Since $Q \in \cap X_n$, one can find a billiard trajectory starting in U_{n-1} in Q that visits $B_1, ..., B_n$. By continuity there is a neighborhood $U_n \subset U_{n-1}$ such that each trajectory in U_n visits $B_1, \ldots, B_n$ within time T_n. Choose a phase point $v \in \cap U_n$. Then the trajectory of the vector v is dense in the phase space of Q. □

1.8 Rational billiards and interval exchange maps

In the case of rational billiards the billiard flow on the invariant surface can be reduced to a one-dimensional transformation; this reduction is a particular case of the reduction of the billiard flow to the billiard transformation.

Definition 1.11 *Let $(I_1, ..., I_n)$ be a partition of the interval $[0, 1)$ into nonintersecting semiclosed intervals, enumerated from left to right, and let σ be a permutation of n element. The respective interval exchange transformation $T : [0, 1) \to [0, 1)$ is a transformation whose restriction to every I_i is a parallel translation and such that the intervals $T(I_1), ..., T(I_n)$ follow from left to right in the order $\sigma(1), ..., \sigma(n)$.*

Clearly, an interval exchange transformation preserves the Lebesgue measure on the unit interval. For example, the exchange of two intervals $[0, a)$ and $[a, 1)$ is identified with the rotation of the circle $\mathbf{R}/\mathbf{Z}$ through $1 - a$.

The reduction of the billiard flow in a fixed direction θ to an interval exchange goes as follows. Recall that the invarinat surface M was constructed by pasting pairs of equal and parallel sides of $2N$ copies of the billiard polygon Q. Choose one side from each such pair and call these segments $s_1, ..., s_m$. Arrange these segments along a line and let I be their union. Give each segment the measure determined by the orthogonal projection on the direction, perpendicular to θ. Since "the width of a beam" is an invariant transversal measure of a constant flow, the billiard transformation induces a piecewise isometry T of the segment I. If the segments are appropriately oriented then T is orientation preserving. It remains to rescale I to the unit interval and to define T at its points of discontinuity so that it is continuous on the left. The result is an interval exchange transformation.

The reduction of the billiard flow in a fixed direction to an interval exchange transformation is by no means unique. For example, assume that the flow F_θ is topologically transitive. Choose an interval I in M, transverse to the flow, and give it the measure determined by the orthogonal projection on the direction, perpendicular to θ. Then the first return map to I along the trajectoires of F_θ is orientation and mesure preserving, that is, an interval exchange transformation. Note that since F_θ is topologically transitive, the first return map is defined no matter how small I is. The points of discontinuity of the first return map correspond to the orbits of the flow through the singular points of the flow F_θ.

1.9 Flat metrics and quadratic differentials

As we have seen a rational billiard defines a closed surface formed by gluing Euclidean polygons isometrically along their edges. The vertices of the polygon correspond to points with cone angle singularities of the metric. In this section we generalize this notion to what we will call flat structures with parallel line fields, or just flat structures for short, so that the set of rational billiards of a given genus is a subset of the space of flat structures. The main reason for this generalization is that the group $SL(2, \mathbf{R})$ acts on the space of flat structures while not preserving the space of rational billiards. The study of this action allows to make conclusions about flat structures in general, which then give us results about rational billiards in particular.

Let M be a compact C^∞ surface and Σ a finite set of points in M. On $M \setminus \Sigma$ we require coordinate charts $v = (x, y)$ such that the transition functions on the overlaps are of the form

$$v \to v + c \quad \text{or} \quad v \to -v + c.$$

That is, the transition functions are translations and reflections in the origin followed by translations. Since these preserve the Euclidean metric this allows one to define a locally Euclidean metric on M. In addition these transition functions preserve families of parallel lines in the plane so that for each direction θ, there is a well-defined foliation F_θ of the surface consisting of lines in direction θ.

Note that if the transition functions are all translations, then the line field defines a vector field and in each direction θ we have a flow F_θ. Now the same topological theorem Theorem 1.8 that holds for rational billiards hold in the more general case of a flat structure.

In a neighborhood of a point in Σ there are polar coordinates (r, θ) such that the metric can be written as

$$ds^2 = dr^2 + (crd\theta)^2$$

where c is a half integer. We say that the metric has cone type singularity with cone angle $2\pi c$. The curvature κ at a singular point is defined by the formula

$$\kappa = 2\pi - 2\pi c.$$

We may concretely describe the metric in a neighborhood of a singular point by gluing together half planes as described above. Notice that our previous discussion shows that a rational polygon determines such a flat structure with transition functions all of the first type (parallel translations) and c an integer.

Another way to define the same structure would be to begin by requiring that M have Gaussian curvature zero, or equivalently, it is locally isometric to $\mathbf{R}^2$ with the Euclidean metric, away from a finite set of points Σ and cone type singularities near points of Σ, where c is an arbitrary real. This in itself is not enough, for it does not rule out the possibility of rotations in the transition functions, which would not allow for parallel line fields. One way to guarantee the correct structure in this language is to consider parallel translation with respect to the connection determined by the metric. The global obstruction to parallel translation being well defined is the holonomy group which is a subgroup of $O(2)$. If the holonomy group is either trivial or $\{I, -I\}$ then parallel translation is well-defined and M possesses a parallel line field. The existence of such a line field implies that the cone angles are multiples of π. Pick one such line field and call it the vertical line field. In the neighborhood of each point of $M \setminus \Sigma$ we can construct a chart which is an orientation preserving isometry and takes the vertical line field to the vertical line field in $\mathbf{R}^2$. The change of coordinate functions between these charts then have the desired form.

The same structure can also be defined in terms of complex analysis. Recall that a Riemann surface or a complex analytic structure on a surface M consists of an atlas of charts (U_α, z_α) where $\{U_\alpha\}$ is a covering of M by open sets, $z_\alpha : U_\alpha \to \mathbf{C}$ is a homeomorphism and if $U_\alpha \cap U_\beta \neq \emptyset$, then $z_\alpha \circ z_\beta^{-1} : z_\beta(U_\alpha \cap U_\beta) \to \mathbf{C}$ is complex analytic. The maps z_α are called uniformizing parameters.

A meromorphic quadratic differential ϕ assigns to each uniformizing parameter z_α a meromorphic function $\phi_{z_\alpha}(z_\alpha)$ with the property that

$$\phi_{z_\beta}(z_\beta)(\frac{dz_\beta}{dz_\alpha})^2 = \phi_{z_\alpha}(z_\alpha)$$

in $U_\alpha \cap U_\beta$. Then the quadratic differential $\phi(z)dz^2$ is invariantly defined on M.

Although the value of ϕ is not well-defined, the set of zeroes and poles of ϕ and their orders are. It is a classic result in Riemann surface theory that ϕ has a finite number of zeroes with orders k_i and poles of order l_i satisfying

$$\sum k_i - \sum l_i = 4g - 4.$$

In this paper we will assume that if there are poles then they are simple. This implies that the norm or area of ϕ defined by

$$||\phi|| = \int_M |\phi(z)||dz|^2$$

is finite. Notice that the area element $|\phi(z)||dz^2|$ is well-defined independently of choice of coordinates. We may equally well consider ϕ to be a holomorphic quadratic differential on the surface punctured at the poles.

A quadratic differential determines a metric with the length element $|\phi(z)|^{1/2}|dz|$. If the quadratic differential has a simple pole then the metric is not complete on the punctured surface.

The vertical trajectories of ϕ are the arcs along which

$$\phi(z)dz^2 < 0$$

and the horizontal trajectories the arcs along which

$$\phi(z)dz^2 > 0.$$

We now indicate why a flat structure (with a parallel line field) and quadratic differential define the same object. First suppose ϕ is a quadratic differential on M

and $p \in M$ is a point which is neither a pole nor zero. We may choose a uniformizing parameter ζ in a neighborhood of p with p corresponding to $\zeta = 0$. We may then choose a branch of $\phi^{1/2}(\zeta)$ near $\zeta = 0$ and define a new uniformizing parameter w by

$$w = w(\zeta) = \int_0^\zeta \phi^{1/2}(\tau) d\tau.$$

Then, in the w coordinates, the quadratic differential is given by $\phi_w(w) \equiv 1$. Such coordinates are called natural. If w and w' define natural coordinates in overlapping neighborhoods, then

$$w' = \pm w + c$$

and so one has a flat structure. The Riemannian metric is the flat metric $|dw|$.

At a zero of ϕ of order k (at a simple pole take $k = -1$) the quadratic differential can be written as

$$\left(\frac{k+2}{2}\right)^2 \zeta^k d\zeta^2$$

for a choice of coordinate ζ. Then

$$w(\zeta) = \zeta^{\frac{(k+2)}{2}},$$

where the uniformizing parameter w is defined as above. The full angle $0 \leq \arg \zeta \leq 2\pi$ is subdivided into $k+2$ equal sectors

$$\frac{2\pi}{k+2} j \leq \arg \zeta \leq \frac{2\pi}{k+2}(j+1), \quad j = 0, ..., k+1.$$

Every sector is mapped by $w(\zeta)$ onto the upper or lower halfplane. Thus a zero of ϕ of order k corresponds to a singular point of a flat structure with the cone angle $\pi(k+2)$. The pre-images of the horizontal lines are the horizontal trajectories of the quadratic differential ϕ, and the trajectory structure has the form of a $k+2$ *pronged* singularity.

Fig. 5

Conversely, a flat structure with a parallel line field defines a quadratic differential. Let $v = (x, y)$ and $v' = (x', y')$ be coordinates in overlapping charts so that

$$v' = \pm v + c.$$

Setting $z = x + iy$ and $z' = x' + iy'$, the transition functions are complex analytic, so define a Riemann surface structure on $M \setminus \Sigma$, where Σ is the set of singularities. We may then define a quadratic differential ϕ on this Riemann surface by assigning the constant function 1 to the parameter z. The πc cone angle singularity corresponds to a zero of order $c - 2$.

There is an additional description of these structures via the theory of measured foliations, developed by Thurston. This description will not play any role here and we will not dwell on it.

We will go back and forth between the terminology of quadratic differentials and flat structures with parallel line fields.

We finish the section with an example taken from [59]: the Riemann surface corresponding to the billiard in a regular n-gon with n odd is conformally equivalent to the Fermat curve $x^n + y^n = 1$, and the respective quadratic differential is dx^2/y^4; see [3] for a similar description of Riemannian surfaces corresponding to rational triangles.

2 Teichmuller space, strata of quadratic differentials and $SL(2, \mathbf{R})$ action

2.1 Teichmuller space and mapping class group

A general reference for Teichmuller spaces, and compactifications is the paper of Bers [6]. Let M be a surface of genus g with n punctures. We assume $3g - 3 + n \geq 0$. Teichmuller space $T_{g,n}$ is the space of equivalence classes of complex structures X on M where $X_1 \sim X_2$ if there is a biholomorphic map from X_1 to X_2 which is isotopic to the identity on M. One may define the Teichmuller distance function $d_T(\cdot, \cdot)$ on T_g by

$$d_T(X_1, X_2) = 1/2 \inf \log K(f),$$

where the infimum is taken over all quasiconformal maps f isotopic to the identity on M and $K(f)$ is the maximal dilatation of f as measured by the complex structures X_1, X_2. The mapping class group $Mod(g, n)$ is the group $Diff^+(M)/Diff_0(M)$, the group of orientation preserving diffeomorphisms modulo those isotopic to the identity. It acts on $T_{g,n}$ by pull-back; given a complex structure X defined by coordinate charts (U_α, z_α) and $f \in Mod(g, n)$, we find a new complex structure $f \cdot X$ defined by the atlas of coordinate charts $(f(U_\alpha), z_\alpha \circ f^{-1})$ The quotient space $R_{g,n} = T_{g,n}/Mod(g, n)$

is the moduli space of Riemann surfaces.

2.2 Compactifications

The moduli space $R_{g,n}$ is well-known not to be compact. It is possible to deform a Riemann surface by pinching along one or more disjoint simple closed curves, by letting the hyperbolic length of the curves go to 0. The resulting surface then has nodes or punctures and may not be connected. For example if a closed surface of genus g is pinched along a single closed curve γ and γ does not disconnect M, the resulting surface has genus $g-1$ with 2 punctures, while if γ is dividing, the resulting surface has two components each of which has one puncture and the sum of their genera is g. We can compactify $R_{g,n}$ by adjoining the moduli spaces of surfaces obtained in this fashion.

Denote the compactification by $\bar{R}_{g,n}$. The topology has the following property. Let X_0 be any surface in the compactification and $X_n \to X_0$. Remove any neighborhood U of the punctures of X_0. Then for large enough n, there is a conformal embedding of $X_0 \setminus U \to X_n$.

The compactification by these moduli spaces is well-behaved with respect to quadratic differentials. For suppose ϕ_n are unit norm quadratic differentials on X_n which converge to $X_0 \in \bar{R}_{g,n}$. Then there is a subsequence of ϕ_n which converges uniformly on compact sets of X_0 via the conformal embedding to an integrable quadratic differential ϕ_0. However it may be the case that $\phi_0 \equiv 0$ on one or more components of X_0. This issue will be discussed further in the section on periodic orbits.

2.3 Strata of quadratic differentials

Fix a surface M of genus g. Let $\sigma = (k_1, \ldots, k_j, \epsilon)$ be an $j+1$ tuple where k_i are integers which are either positive or -1 satisfying $\sum k_i = 4g-4$ and $\epsilon = \pm$. Consider the set of quadratic differentials with zeroes of order k_i, and simple poles if $k_i = -1$, the sign $+$ if the quadratic differential is the square of a holomorphic 1 form, and $-$ otherwise. These spaces can be empty (see [43] for necessary and sufficient conditions) and need not be connected – [53] and [35]. We say that two such quadratic differentials are equivalent if there is a homeomorphism of M isotopic to the identity taking singular points to singular points of the same order and which at other points has the same local form as the change-of-coordinate functions. The resulting space is denoted by $Q(\sigma)$ and one can prove that this is a manifold. In the case that all $k_i = 1$ (and $\epsilon = -1$) $Q(\sigma)$ is called the principle stratum. For fixed g, the union of the $Q(\sigma)$ where all $k_i > 0$, is the space of all quadratic differentials Q_g over T_g. It is a well-known part of Teichmuller theory that T_g is a complex manifold and Q_g is the cotangent bundle, although we will not make use of this structure in this paper. The strata $Q(\sigma)$ are not closed subsets of Q_g, unless there is a single zero $(n = 1)$. This is because a sequence in $Q(\sigma)$ may collapse a pair of lower order zeroes into a higher order one.

The group $Mod(g)$ acts on each $Q(\sigma)$. The quotient is denoted by $QD(\sigma)$ and will play a crucial role in this paper. Since $Mod(g)$ does not act freely, the space $QD(\sigma)$ has the structure of an orbifold.

2.4 $SL(2, \mathbf{R})$ action on quadratic differentials

The group $SL(2, \mathbf{R})$ acts on quadratic differentials by linear transformations of local coordinates, preserving each stratum. If $\{u_\alpha\}$ is an atlas of natural coordinates defining a quadratic differential ϕ, and $A \in SL(2, \mathbf{R})$ then $\{Au_\alpha\}$ defines a new family of natural coordinates for a quadratic differential $A\phi$. We require that A takes singularites of ϕ to singularities of $A\phi$ of the same order.

The following one parameter subgroups of $SL(2, \mathbf{R})$ are of special interest:

$$g_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}, \quad r_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \quad h_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

referred to as the geodesic or Teichmuller, circular and horocyclic flows, respectively.

The effect of the flow $\phi \to g_t\phi$ is to stretch along the horizontal trajectories of ϕ by a factor of e^t and contract along the vertical trajectories by e^t. The flow contracts the transverse measure of the horizontal foliation and expands the transverse measure of the vertical foliation. Associated to g_t is the map f_t of the underlying Riemann surfaces $X \to X_t$, called the Teichmuller map. It is the famous theorem of Teichmuller that given any homeomorphism $f : X \to Y$ of Riemann surfaces of finite type, there is a unique Teichmuller map in the homotopy class which minimizes the maximal dilation.

The action of r_θ on a quadratic differential ϕ is the same as multiplying ϕ by $e^{2i\theta}$. This multiplication defines an action of the circle $\mathbf{R}/\pi\mathbf{R}$ on the set of quadratic differentials. Note that the action of r_θ leaves the flat metric invariant, but changes the vertical line field.

The action on saddle connections is as follows. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

be a matrix in $PSL(2, \mathbf{R})$. Let β be a saddle connection, and let h_1 and v_1 be its horizontal and vertical components with respect to ϕ. Then the horizontal and vertical components h_2 and v_2 with respect to $A\phi$ are given by matrix multiplication; namely

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} h_2 \\ v_2 \end{bmatrix}.$$

The orbit of a quadratic differential under $PSL(2, \mathbf{R})$ is called a Teichmuller disc and has been considered by numerous authors – see [36], for example. In a later section we will examine certain examples of Teichmuller discs that arise in the so-called Veech billiards and their generalizations.

The action of $SL(2, \mathbf{R})$ commutes with the action of $Mod(g)$ and thus descends to an action on $QD(\sigma)$. It is this action we will study in some detail.

It is possible to define an $SL(2, \mathbf{R})$ invariant measure μ_0 on each $QD(\sigma)$ which is absolutely continuous with respect to the orbifold structure. We show how to do this for the principle stratum $QD(\sigma)$. The measure is defined locally. A saddle connection of ϕ_0 persists under small perturbation of ϕ_0. There are a finite number of such saddle connections $\beta_1, \ldots, \beta_n$ whose horizontal and vertical components $(h_{\beta_1}, v_{\beta_1}, \ldots, h_{\beta_n}, v_{\beta_n})$ serve as local coordinates for $Q(\sigma)$ in a neighborhood of ϕ_0. Specifically, pass to a double cover $\pi : \tilde{M} \to M$, ramified over the zeroes of ϕ_0. There is an involution τ of $\tilde{M}$ which interchanges the sheets. We choose the β_i such that their lifts to $\tilde{M}$ form a basis for the first homology of $\tilde{M}$, odd with respect to the involution τ. The measure μ_0 is then Lebesgue measure on $\mathbf{R}^{2n}$ pulled-back to $Q(\sigma)$ via these local coordinates. One sees that μ_0 does not depend on choice of basis β_i, so actually defines a measure. It is easily seen to descend to $QD(\sigma)$ and to be $SL(2, \mathbf{R})$ invariant. The measure was shown to be finite and ergodic on the principle stratum in [37] and then on each component of general stratum in [52], [53]. An interesting question is to classify all ergodic $SL(2, \mathbf{R})$ invariant measures on $QD(\sigma)$.

3 Ergodicity

3.1 Veech nonuniquely ergodic example

Definition 3.1 *A foliation F is uniquely ergodic if it is minimal and the transverse measure is unique up to scalar multiplication.*

Unique ergodicity is equivalent to the following condition. Let G be any foliation with transverse measure, transverse to F. For any point x_0, the transverse foliation

G defines an arc length $l(t)$ on the leaf of F through x_0 in either direction. The pair (F,G) defines a measure μ on M. For every continuous function $f(x)$ on M and every point x_0 we have

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T f(l(t))dt = \int_M f(x)d\mu(x).$$

The classical example of a uniquely ergodic foliation is the irrational flow on the flat torus. In [51] Veech constructed examples of minimal and not uniquely ergodic dynamical systems that are skew products over irrational rotations of the circle. Take two copies of the unit circle and mark off a segment J of length $2\pi\alpha$ in the counterclockwise direction on each with one endpoint at 0. Now take θ irrational and consider the following dynamical system. Start with a point p, say in the first circle. Rotate counterclockwise by $2\pi\theta$ until the first time the orbit lands in J; then switch to the corresponding point in the second circle, rotate by $2\pi\theta$ until the first time the point lands in J; switch back to the first circle and so forth. Veech showed that for any irrational α there are an uncountable set of θ such that this system is minimal and not uniquely ergodic.

We give geometric description of Veech's examples that was worked out by John Smillie jointly with the first author (unpublished). In order to do that we first give a geometric criterion of how to find nonergodic foliations. This criterion was given in [42]. To set the notation let $E_0, E_1, \ldots$ be a sequence of subsets in M so that $E_n \cap E_{n+1} = \emptyset$ and each E_n is bounded by a set of saddle connections all in the same direction θ_n. Assume the directions θ_n converge to θ_∞. Let F_{θ_∞} be the foliation in direction θ_∞. Rotate the coordinate system so that θ_∞ is the vertical direction. Define a system of partitions $P_n = [A_n, B_n]$ of M as follows. If n is even, then $A_n = E_n$ and $B_n = E_n^c$ (complement of E_n), and if n is odd, then $A_n = E_n^c$ and $B_n = E_n$. Let h_n be the sum of the horizontal components of the saddle connections separating A_n and B_n. Let μ be the measure defined by the flat structure.

Theorem 3.2 *Suppose that*

- *(i)* $\lim_{n\to\infty} h_n = 0$
- *(ii) for some* c, c', $0 < c \le \mu(A_n) \le c' < 1$
- *(iii)* $\sum_{n=1}^{\infty} \mu(A_n \Delta A_{n+1}) < \infty$

Then the vertical foliation F_{θ_∞} *is nonergodic.*

Proof. Let

$$A_\infty = \liminf A_n = \{x : \exists N \text{ such that for } \ n \geq N, x \in A_n\}.$$

$$B_\infty = \liminf B_n = \{x : \exists N \text{ such that for } \ n \geq N, x \in B_n\}.$$

We first show that A_∞ and B_∞ satisfy

- (1) $\mu(M \setminus (A_\infty \cup B_\infty)) = 0$.
- (2) $A_\infty \cap B_\infty = \emptyset$
- (3) $\mu(A_\infty \Delta A_n) \to 0$ as $n \to \infty$.
- (4) $0 < \mu(A_\infty) < 1$.

By (iii) and the Borel-Cantelli Lemma the set of x' which are in infinitely many $A_n \Delta A_{n+1}$ has μ measure 0. ¿From this we have (1). Statement (2) is immediate. To see (3) note that

$$A_\infty \Delta A_n \subset \cup_{i=n}^\infty A_i \Delta A_{i+1}$$

so that

$$\mu(A_\infty \Delta A_n) \leq \sum_{i=n}^\infty \mu(A_i \Delta A_{i+1}).$$

Hypothesis (iii) implies that the right hand side goes to 0 as n goes to infinity, proving (3). Statement (4) follows from (3) and (ii).

Returning to the proof of Theorem 3.2 , we can assume that the vertical foliation F_{θ_∞} is orientable: if not, we may replace M by an orientable double cover and replace A_∞ and B_∞ by their lifts. Now let f_t be the flow along vertical leaves.

We claim that for any t the set A_∞ is μ a.e. invariant; that is,

$$\mu(f_t(A_\infty) \Delta A_\infty) = 0.$$

Suppose on the contrary, that for some t_0, we have $\mu(f_{t_0}(A_\infty) \Delta A_\infty) = \delta > 0$. By (3) and (i) we may choose n large enough so that

$$\mu(A_\infty \Delta A_n) < \delta/8 \ \text{ and } \ e^{t_0} h_n < \delta/8.$$

Since f_{t_0} is μ preserving, the first inequality above and the assumption give

$$\mu(f_{t_0}(A_n) \Delta A_n) \geq \delta - 2\delta/8 = 3\delta/4.$$

Thus at time t_0, $3\delta/8$ of the measure of A_n flows to its complement. However at most $e^{t_0}h_n < \delta/8$ of measure can cross the boundary of A_n, a contradiction, proving the claim. We would like to conclude that A_∞ is flow invariant, although by the claim it is only a.e. invariant for each time. The theorem is a consequence of the following general lemma.

Lemma 3.3 *Let f_t be a flow on a space X preserving a probability measure μ. Suppose there is a set A such that for every t, $\mu(f_t(A)\Delta A) = 0$. Then there is a set A' invariant under f_t with $\mu(A\Delta A') = 0$.*

Proof. Let λ be Lebesgue measure on $\mathbf{R}$. Let

$$A' = \{x : f_t(x) \in A \text{ for } \lambda \text{ } a.e.t\}.$$

It is clear that A' is f_t invariant. We will show that $\mu(A\Delta A') = 0$. Let

$$C_0 = \{(x,t) : x \in A\} \text{ and } C_1 = \{(x,t) : f_t(x) \in A\}.$$

For every t we have

$$\mu(\{x : (x,t) \in C_0\Delta C_1\}) = \mu(f_t(A)\Delta A) = 0.$$

This implies that $(\mu \times \lambda)(C_0\Delta C_1) = 0$. By Fubini there is a set X' of full μ measure so that for all $x \in X'$,

$$\lambda(\{t : (x,t) \in C_0\Delta C_1\}) = 0.$$

If $x \in X' \cap A$ then $(x,t) \in C_0$ so that $\lambda(\{t : (x,t) \notin C_1\}) = 0$. Therefore the set $\{t : f_t(x) \in A\}$ has full λ measure so $x \in A'$. If $x \in X' \setminus A$ then the set $\{t : (x,t) \notin C_1\}$ has full λ measure so that $x \notin A'$. Thus $A\Delta A'$ is contained in the complement of X' so $\mu(A\Delta A') = 0$. This finishes the proof of the Lemma and therefore the Theorem. $\square$

Now we wish to return to the Veech examples which will be described in terms of quadratic differentials. We start with a square torus T; that is, the unit square with lower left vertex at $(0,0)$ with opposite sides identified. Equivalently, it is $\mathbf{R}^2$ modulo the integer lattice. Fix a point $(x_0, y_0) \in R^2$. For integers m, n let w be a segment in $\mathbf{R}^2$ joining $(0,0)$ to $(x_0 + m, y_0 + n)$, projected to T. Take two copies of T, each slit along w, and identify the positive side of w on one copy to the negative side in another. It is easy to see that this results in a quadratic differential ϕ on a surface M of genus 2. Since the total angle around each of the points $(0,0)$ and (x_0+m, y_0+n) is 4π, they correspond to zeroes of order 2 of ϕ. The surface of genus 2 is partitioned into 2 sheets M_w^+ and M_w^- separated from each other by the union of the two slits.

Theorem 3.4 ([51]) *Suppose x_0, y_0 are not rationally related; that is, for any pair of rationals, r_1, r_2, the equation $r_1x_0 + r_2y_0 = 0$ implies $r_1 = r_2 = 0$. Then there exists uncountably many directions θ such that the flow on M in direction θ is minimal and not ergodic.*

In the special case that $x_0 = 0$ and $y_0 = \alpha$, so that the slit is along the vertical axis, the resulting surface M has two metric cylinders in the vertical direction, one for each sheet. Choose the core curve of each metric cylinder. The first return map to this pair of curves for the flow in direction θ is precisely the dynamical system studied by Veech.

This same dynamical system can be described in terms of rational billiards by taking a rectangle of length 2 and width 1, with an interior slit of length α from the midpoint of a horizontal side.

Fig. 6

Proof. Suppose w' is another segment on T with the same endpoints as w. Suppose w and w' intersect an odd number of times in their interior so they divide each other into an even number of pieces. This is equivalent to saying that w and w' are homologous mod 2 on T. The sheet interchange measured by $c = (M_w^+ \cap M_{w'}^-) \cup (M_w^- \cap M_{w'}^+)$ is a union of an even number of parallelograms with sides on w and w' (here w and w' are thought of as vectors). Thus the area of c is at most $|w \times w'|$.

Fig. 7

Now fix a sequence of positive numbers ρ_j with $\sum \rho_j < \infty$. We will build an infinite directed tree with each vertex leading to 2 further vertices. At level j there will be 2^{j-1} vertices. Each vertex will correspond to a pair of integers (p, q) which will determind a slit with endpoints $(0, 0)$ and $(x_0 + p, y_0 + q)$.

For any pair (p, q) form the quotient $(p+x_0)/(q+y_0)$, the slope of the corresponding slit. Let δ_j be the minimum distance between any two slopes as (p, q) varies over the vertices at level j. Now inductively, suppose we have determined the tree up to level j. For each vertex (p, q) at level j we will choose two vertices $(p', q'_)$ and $(p'', q,'')$ at level $j + 1$, and call these the two *children* of (p, q). To find such children begin by choosing an even integer d such that

$$\frac{\rho_j}{(q + y_0)(q + y_0 + d)} < \delta_j/4.$$

Then consider the inequality

$$d|(p + x_0)n - (q + y_0)m| < \rho_j$$

Since x_0 and y_0 are not rationally related, neither are $p_0 + n$ and $q + y_0$, and so there are infinitely many coprime solutions (m, n). Choose any two of them and set

$$p' = p + dm \text{ and } q' = q + dn,$$

calling the two resulting two pairs the children. A direct calculation shows that

$$\delta_j' = |\frac{p + x_0}{q + y_0} - \frac{p' + x_0}{q' + y_0}| < \delta_j/4.$$

That is to say, the distance between the slope of a slit and either child is bounded by $\delta_j/4$. Then the distance between the slopes of the two children of (p, q) is at most $\delta_j/2$, which implies that

$$\delta_{j+1} < \delta_j/2.$$

Let w be the segment corresponding to the parent and w' the segment of a child. Since d is even, w' is homologous to w mod 2. In fact $w' - w$ represents d times the primitive class of (m, n). As we have seen the area of the sheet interchange c is bounded by $|w' \times w|$, and an easy calculation shows that this is bounded by

$$|(p + x_0)n - (q + y_0)m]| < \rho_j.$$

Thus, in the situation of Theorem 3.2, we have constructed a sequence of partitions satisfying (iii). Now for any geodesic in the tree, the sequence of ratios $(p + x_0)/(q + y_0)$ is a Cauchy sequence hence converging to some θ_∞. From this, condition (i) of Theorem 3.2 follows. Condition (ii) holds since the partition divides the surface into two pieces of equal areas. Thus by Theorem 3.2 the limiting foliation F_{θ_∞} is not ergodic.

We now show that, in fact, there are uncountably many limiting directions. Since there are only countably many directions that are not minimal, we then can conclude that there are uncountably many limiting directions which are minimal. Since each pair (p, q) has two children there are clearly uncountably many sequences. We therefore need to show that any two limits are distinct. Suppose $r_j = (p_j + x_0)/(q_j + y_0) \to \theta$ and a distinct sequence $r_j' = (p_j' + x_0)/(q_j' + y_0) \to \theta'$. Suppose the sequences differ for the first time at stage j, so that the slopes r_j and r_j' satisfy

$$|r_j - r_j'| \geq \delta_j. \tag{4.5}$$

Then by (4.3) and (4.4) we have

$$|\theta - \frac{p_j + x_0}{q_j + y_0}| \le \sum_{k\ge j} \delta'_k < \sum_{k\ge j} \frac{\delta_k}{4} < \frac{\delta_j}{2},$$

and similarly $|\theta' - r'_j| < \delta_j/2$. Combined with (4.5), we have $\theta \neq \theta'$. □

The above theorem gives uncountably many minimal nonergodic directions. By requiring α to satisfy a Diophantine condition, it is even possible to show that there is a set of such examples of positive Hausdorff dimension. The main result of the next section however says that the set of nonergodic directions must have Lebesgue measure 0.

3.2 Ergodicity in almost every direction

Theorem 3.5 ([34]) *For any quadratic differential ϕ (rational billiard) the set of $\theta \in [0, 2\pi]$ such that $r_\theta\phi$ has minimal but non-uniquely ergodic vertical foliation has Lebesgue measure* 0.

The proof of this theorem uses the $SL(2, \mathbf{R})$ action on quadratic differentials and some combinatorial constructions related to quadratic differentials. Let QD_ϵ be the set of quadratic differentials all of whose saddle connections have length at least ϵ. Recall R_g is the moduli space of Riemann surfaces of genus g.

Proposition 3.6 *QD_ϵ is compact in QD.*

Proof. QD_ϵ is clearly closed in QD since if $\phi_n \to \phi$ and ϕ has a geodesic segment of length less than ϵ so does ϕ_n for n large. It is therefore enough to show that QD_ϵ lies over a compact set in R_g. If not, there is a sequence $\phi_n \in QD_\epsilon$ of quadratic differentials lying on Riemann surfaces X_n going to infinity in R_g. Passing to subsequences we may assume that X_n converges to a Riemann surface X_∞ with nodes or punctures acquired by pinching along a set of disjoint simple closed curves $\alpha_1, \ldots, \alpha_p$. By passing to a further subsequence we can assume that ϕ_n converges uniformly on compact sets to an integrable quadratic differential ϕ_∞ on X_∞.

Thus ϕ_∞ has at most simple poles at the nodes. There is therefore a curve α_j homotopic to a puncture with ϕ_∞ length at most $\epsilon/2$. For large n the uniform convergence implies that α_j has ϕ_n length at most $\epsilon/2$, a contradiction. □

Definition 3.7 *A quadratic differential ϕ is called divergent if the Riemann surfaces on which $g_t\phi$ lie, eventually leave every compact set of R_g as $t \to \infty$.*

Theorem 3.8 *Suppose the vertical foliation F_ϕ of ϕ is minimal but not uniquely ergodic. Then ϕ is divergent.*

Proof. Suppose the theorem is not true so that for some sequence $t_n \to \infty$ the Riemann surfaces $X_n = X_{t_n}$ converge in R_g to some X_0. By passing to a further subsequence we may assume that $\phi_n = g_{t_n}(\phi)$ converges to some ϕ_0 on X_0. Let f_{t_n} be the corresponding Teichmuller map. Denote by Σ and Σ_0 the sets of zeroes of ϕ and ϕ_0, rspectively.

The normalized transverse measures on the topological foliation F_ϕ form a finite dimensional convex set for which the extreme points ν_i are mutually singular ergodic measures – see [58]. Let ν_0 be the transverse measure of F_ϕ. We have $\nu_0 = \sum_{i=1}^m c_i\nu_i$, for constants c_i. Pick a segment I of a horizontal trajectory of ϕ so that $\nu_i(I) \neq \nu_j(I)$ for $i \neq j$. Let μ the transverse measure to the horizontal foliation defined by ϕ. This means that the area element $|\phi(z)dz^2| = \nu_0 \times \mu$. Let $E_i \subset X$ be the disjoint sets of generic points of the measure $\nu_i \times \mu$ for the interval I. We have $(\nu_i \times \mu)(E_i) = 1$. Each E_i is a union of leaves of F_ϕ.

For each i let $A_i \subset X_0$ denote the set of accumulation points of $f_{t_n}(x) \in X_n$ for $x \in E_i$. Now let $U \subset X_0$ be any open set. We claim that there is a j such that

$$U \cap A_j \neq \emptyset.$$

To prove the claim, choose an open W so that $\bar{W} \subset U$. Since $X_n \to X_0$, for each n, W may be considered as a subset of X_n with area bounded below by $\delta > 0$. Consider $W_n = f_{t_n}^{-1}(W) \subset X$. Since f_{t_n} preserves area, $(\nu_0 \times \mu)(W_n) > \delta$. Then for each n, there is a $j = j(n)$ such that

$$(\nu_j \times \mu)(E_j \cap W_n) = (\nu_j \times \mu)(W_n) \geq \frac{\delta}{mc_j}.$$

Choose a j so that this inequality holds for infinitely many n. This implies that there exists some $x \in E_j \cap W_n$ for infinitely many n. Since $f_{t_n}(x) \in W$, $f_{t_n}(x)$ has an accumulation point in $\bar{W}$, proving the claim.

Partition X_0 into a finite number of rectangles in the natural coordinates of ϕ_0 whose sides are horizontal and vertical segments. Adjacent rectangles meet along a common horizontal or vertical segment of a side. Each point of Σ_0 is required to be a vertex of a rectangle. For each pair of adjacent rectangles R_1, R_2 choose open sets $U_i \subset R_i$ with the property that if $y_i \in U_i$ there is a vertical segment l_i with one endpoint y_i such that l_1, l_2 are two sides of a coordinate rectangle contained in $R_1 \cup R_2$. For any two points y_1, y_2 in the same coordinate rectangle there are vertical segments l_i with endpoints y_i which are the vertical sides of a coordinate rectangle. Now by the claim, by passing to further subsequences of t_n, for each open set U_i we may choose $y_i \in U_i$ and $x_i \in \cup_{j=1}^m E_j$ such that $y_i = \lim f_{t_n}(x_i)$.

Fig. 8

Now suppose U_1, U_2 are such open sets contained in either adjacent or the same rectangle with corresponding y_i and x_i. We claim that x_1 and x_2 must belong to the same E_j. Assume otherwise. By renumbering assume $x_i \in E_i, i = 1, 2$. Let l_i be the vertical segments with endpoints y_i and which are vertical segments of a coordinate rectangle. They have the same length. Each l_i is the limit of vertical segments $l_{i,n}$ of equal length of $g_{t_n}\phi$ with one endpoint $f_{t_n}(x_i)$. Each $l_{i,n}$ is the image under f_{t_n} of a vertical segment $L_{i,n} \subset E_i$ passing through x_i; $L_{i,n}$ have equal length which go to ∞ with n. Therefore the number of intersections of $L_{i,n}$ with I goes to ∞ and since $L_{i,n} \subset E_i$,

$$\lim_{n\to\infty} \frac{\mathrm{card}(l_{i,n} \cap f_{t_n}(I))}{\mathrm{card}(l_{2,n} \cap f_{t_n}(I))} = \lim_{n\to\infty} \frac{\mathrm{card}(L_{1,n} \cap I)}{\mathrm{card}(L_{2,n} \cap I)} = \frac{\nu_1(I)}{\nu_2(I)} \neq 1.$$

But since $l_{i,n}$ has limit l_i and every horizontal segment of ϕ_0 that intersects l_1 intersects l_2 and vice versa,

$$\frac{\mathrm{card}(l_{1,n} \cap f_{t_n}(I))}{\mathrm{card}(l_{2,n} \cap f_{t_n}(I))}$$

can be made arbitrarily close to 1 by taking n large enough. This is a contradiction, proving the claim.

Any two rectangles R_1, R_n can be connected by a chain of rectangles $R_1, \ldots, R_n$ where R_i and R_{i+1} are adjacent along a side. This fact and the last claim imply that there is a single index j such that $x_i \in E_j$ for all x_i.

Choose some index $i \neq j$ and $x \in E_i \setminus \Sigma$. Let $y \in X_0$ be an accumulation point of $f_{t_n}(x)$. By passing to further subsequences of t_n we can assume that there is a single rectangle R such that $f_{t_n}(x) \in R$ and $\lim_{n\to\infty} f_{t_n}(x) = y$. Let $U_k \subset R$ be one of the open sets found previously. We may find vertical segments $l \subset R$ with one endpoint y and $l' \subset R$ with endpoint $y_k = \lim f_{t_n}(x_k)$ such that l, l' are two vertical sides of a coordinate rectangle contained in R. We may find vertical segments l_n and $l_{n'}$ of $g_{t_n}\phi$ of equal length, $l_n \to l$ and $l'_n \to l'$, l_n has endpoint $f_{t_n}(x)$ and l'_n has endpoint $f_{t_n}(x_k)$. As before

$$\frac{\operatorname{card}(l_n \cap f_{t_n}(I))}{\operatorname{card}(l'_n \cap f_{t_n}(I))}$$

can be made arbitrarily close to 1 by taking n large enough, and yet since x and x_k are points of E_i and E_j respectively, the ratio has limit

$$\frac{\nu_i(I)}{\nu_j(I)} \neq 1,$$

and we have our contradiction. $\square$

The proof of Theorem 3.5 follows from the above theorem and the next

Theorem 3.9 *The set of $\theta \in [0, 2\pi]$ such that $r_\theta\phi$ is divergent, has Lebesgue measure zero.*

3.3 A combinatorial construction

In order to prove this theorem we will need to discuss certain a combinatorial construction. We fix a stratum $QD(\sigma)$.

Definition 3.10 *Two saddle connections are* disjoint *if their only common points are vertices.*

Denote by $p_0 = p_0(\sigma)$ the maximum number of disjoint saddle connections. A set of disjoint saddle connections Γ is called a *system*. By $|\Gamma|$ we mean the maximum length of any member of Γ. By a *complex* K we mean a subset of M consisting of disjoint saddle connections and triangles, such that a triangle is in K if and only if its sides are in K. An ϵ complex is a complex all of whose saddle connections have length at most ϵ. We denote the (topological) boundary of K by ∂K. Then ∂K is a system of saddle connections.

Note that if K is a complex, then $K \setminus \partial K$ (viewed as a subset of M) is a (possibly empty) domain. We say that the *complexity* of a complex is j if it has j saddle connections, including those on the boundary. The main construction is as follows.

Proposition 3.11 *If K is a complex of complexity i and σ is a saddle connection crossing ∂K or disjoint from K, then there is a complex K' of complexity $j > i$ with $K \subset K'$ and $|\partial K'| \leq 3|\partial K| + |\sigma|$.*

Proof. If σ is disjoint from K, we may just add σ to K. Thus we may assume that σ crosses ∂K at a point Q on a segment ω with endpoints A, B. In following σ from Q into the exterior of K to its other endpoint P, suppose σ hits P before again crossing ∂K. We may form a polygon with no interior singularities whose sides are ω, a possibly broken geodesic joining A to P and a possibly broken geodesic joining B to P. The edges and diagonals of this polygon have lengths bounded by $|\sigma| + |\omega|$. Since ω is a boundary edge of K not all of the edges and diagonals of the polygon can be in K. We may therefore add an edge that is not in K.

Fig. 9

Thus assume that before next hitting P, σ again crosses ∂K at Q_1, an interior point of a segment ω_1. Let the endpoints of ω_1 be denoted C, D. Let $\hat{\sigma}$ be the subsegment of σ joining Q to Q_1. Now move $\hat{\sigma}$ parallel to itself with one endpoint moving along ω toward A , the other along ω_1 toward D. This can be done until for a first time a segment σ_1 parallel to $\hat{\sigma}$ either contains a singularity P_0 in its interior or is an endpoint of ω or an endpoint of ω_1, say A. We have $|\sigma_1| \leq |\sigma| + |\omega| + |\omega_1|$. In the first case we build a polygon as before, with σ_1 replacing σ. In the second case also replace σ with σ_1 and also replace ω with ω_1 and P with A. Then we may use σ_1, A, and ω_1 to form the desired polygon. □

Definition 3.12 *Let $\epsilon < C$. A saddle connection γ is (ϵ, C) isolated if it has length less than ϵ and every saddle connection that crosses γ in its interior has length greater than C.*

Since any two (ϵ, C) isolated curves are disjoint, the number of such curves is bounded above by p_0.

Definition 3.13 *For a quadratic differential ϕ let $n_\epsilon(\phi)$ be the maximum number of simplices in a connected ϵ complex.*

Definition 3.14 *For a stratum $QD(\sigma)$ let $N(\epsilon, C)$ be the set of quadratic differentials in the stratum which possess (ϵ, C) isolated saddle connections.*

Proposition 3.15 *Let μ denote Lebesgue measure on the circle. Suppose there is a set S of angles θ of positive measure so that $r_\theta \phi$ is divergent for $\theta \in S$. Then there are*

- *(1) a sequence of times* $T_i \to \infty$
- *(2) a sequence of sets* $S_i \subset S$ *and a number* $\delta > 0$ *such that* $\mu(S_i) \geq \delta$
- *(3) a sequence* $\epsilon_i \to 0$
- *(4) a positive constant* C *such that* $g_{T_i} r_\theta \phi \in N(\epsilon_i, C)$ *for* $\theta \in S_i$.

The idea behind this Proposition is that since the Riemann surface of $g_t r_\theta \phi$ eventually leaves every compact set, there will be times T_i after which there will always be a segment of length less than ϵ_i in the corresponding metrics. For a fixed C, however, the sets $N(\epsilon_i, C)$ do not form a neighborhood basis of infinity so it may not be the case that for sufficiently large T_i that there will be always (ϵ_i, C) isolated segments. Nevertheless by passing to slightly smaller sets than S_i we will be able to make the statement for isolated segments.

Proof. Choose a sequence $\epsilon_i \to 0$. Since for each ϵ_i the set QD_{ϵ_i} is compact, for each divergent ray $r_\theta(q)$ there are times T_i such that $g_t r_\theta(q) \in QD \setminus QD_{\epsilon_i}$ for all $t \geq T_i$. Then T_i can be chosen so that this relationship holds for all $\theta \in S'$ where $S' \subset S$ and $\mu(S') > \mu(S)/2$.

Consider now triples of sequences (ϵ_i, T_i, S_i) such that $\epsilon_i \to 0$, $T_i \to \infty$, $\mu(S_i)$ bounded away from 0 and $g_{T_i} r_\theta \phi \in QD \setminus QD_{\epsilon_i}$ for $\theta \in S_i$. The set of triples is nonempty for we can take $S_i = S'$ for all i. Choose a sequence of triples that maximizes

$$\min_i \min_{\theta \in S_i} n_\epsilon(g_{T_i} r_\theta \phi).$$

For the corresponding sequence, for each i and $\theta \in S_i$ let $C_{i,\theta}$ be the length of the shortest saddle connection that crosses a boundary segment of the ϵ_i complex of $g_{T_i} r_\theta \phi$. For each i, let m_i be a number so that

$$\mu(\{\theta \in S_i : C_{i,\theta} \geq m_i\}) = \mu(S_i)/2.$$

We claim that the numbers m_i are bounded below. Suppose not. Choose a subsequence converging to 0 and for that subsequence let

$$S_i' = \{\theta \in S_i : C_{i,\theta} < m_i\},$$

so $\mu(S_i') = \mu(S_i)/2$ is bounded below, and let

$$\epsilon_i' = 3\epsilon_i + m_i \to 0.$$

Fix $\theta \in S_i'$. Since there is an ϵ_i complex and a saddle connection of length at most m_i crossing it, by Proposition 3.11 we can find an ϵ_i' complex on $g_{T_i} r_\theta \phi$ with more simplices, contradicting the maximality of our sequence of triples. Thus the set of m_i is bounded below, by some $C > 0$. If we replace S_i with the set of θ such that $C_{i,\theta} \geq m_i \geq C$, every saddle connection crossing the ϵ_i complex has length at least C, proving the proposition. $\square$

Let α be a geodesic segment. For some θ_α , the segment α is vertical with respect to $r_{\theta_\alpha}\phi$. Rotate the coordinate system so that $\theta_\alpha = 0$. Denote by $l(t,\theta)$ the length of α with respect to the structure $g_t r_\theta \phi$, so that $|\alpha|_\phi = l(0,0)$. In these coordinates

$$l(t,\theta) = l(0,0)(e^t \sin^2\theta + e^{-t}\cos^2\theta)^{1/2}.$$

Fix $t \geq 0$ and choose $\epsilon > 0$. Define two intervals $I_\alpha \subset J_\alpha \subset (-\pi/2, \pi/2)$ with respect to the rotated coordinates:

$$I_\alpha = \{\theta : |\sin\theta| \leq \epsilon/(ve^{t/2})\},$$

$$J_\alpha = \{\theta : |\sin\theta| \leq C/(2ve^{t/2})\}.$$

Lemma 3.16 *For $\theta \notin I_\alpha$ one has: $l(t,\theta) \geq \epsilon$. There are constants ϵ_1, T, K independent of α such that for $\epsilon < \epsilon_1$ and $t > T$ we have $\mu(I_\alpha)/\mu(J_\alpha) < K\epsilon$. In addition if $l(t,\theta) < \epsilon$ for some θ, then $l(t,\theta) < C$ for all $\theta \in J_\alpha$.*

Proof. The first statement is immediate from the definition. If $l(t,\theta) < \epsilon$ for some θ then, since $l(t,\theta)$ attains its minimum at $\theta = 0$,

$$ve^{-t/2}|\cos\theta| \leq ve^{-t/2} \leq l(t,\theta) < \epsilon.$$

For $\theta \in J_\alpha$, we then have

$$l(t,\theta) \leq (C^2/4 + \epsilon^2 e^{-t})^{1/2} \leq C/2 + \epsilon.$$

If we choose $\epsilon_1 = C/2$ and $\epsilon < \epsilon_1$ then $l(t,\theta) < C$. This proves the last statement.

Let I_α' and J_α' be the images of I_α and J_α under the sine function. If $C/(2ve^{t/2}) < 1/2$ then J_α' is contained in the interval $(-1/2, 1/2)$ and $\mu(I_\alpha')/\mu(J_\alpha') = 2\epsilon/C$. Since the arcsine function restricted to the interval $(-1/2, 1/2)$ is Lipschitz and has a Lipschitz inverse, it changes the lengths of intervals by a bounded amount. Thus we can choose a constant K depending on C such that $\mu(I_\alpha)/\mu(J_\alpha) < K\epsilon$ as claimed.

To ensure that $C/(2ve^{t/2}) < 1/2$ choose T such that $C/(2me^{T/2}) < 1/2$ where m is the length of the shortest saddle connection on ϕ. Since $v \geq m$ and $t > T$, the inequality holds and the proof of the lemma is finished. $\square$

The last technical statement we need is as follows.

Proposition 3.17 *There are constants T, ϵ_1, K' such that for $t > T$ and $\epsilon < \epsilon_1$ one has:*

$$\mu(\{\theta : g_t r_\theta \phi \in N(\epsilon, C)\}) < K'\epsilon.$$

Proof. Fix T and $\epsilon < \epsilon_1$ as in the preceding lemma. Consider all saddle connections α which are (ϵ, C) isolated with respect to $g_t r_\theta \phi$ for some θ. For each such α let $\hat{I}_\alpha$ be the smallest interval of $\theta's$ for which α is (ϵ, C) isolated. Then $\hat{I}_\alpha \subset I_\alpha \subset J_\alpha$.

Construct a new open interval $\hat{J}_\alpha$ as follows: the left-hand endpoint of $\hat{J}_\alpha$ will be halfway between the lefthand endpoints of $\hat{I}_\alpha$ and J_α, and the righthand endpoint of $\hat{J}_\alpha$ will be halfway between the righthand endpoints of $\hat{I}_\alpha$ and J_α. The point of this construction is the fact that if $\hat{J}_\alpha$ and $\hat{J}_\beta$ intersect, then either J_α intersects $\hat{I}_\beta$ or J_β intersects $\hat{I}_\alpha$. Assume the first. We can find $\theta \in J_\alpha \cap \hat{I}_\beta$ such that β is (ϵ, C) isolated with respect to $g_t r_\theta q$. Now, with respect to the metric of $g_t r_\theta \phi$, the saddle connection α has length less than C, and β cannot cross any geodesic segment of length less than C. Thus α and β are disjoint. The maximum number of disjoint segments is p_0. Therefore no θ can lie in more than p_0 segments $\hat{J}_\alpha$. The sum of the lengths of the $\hat{J}_\alpha$ is at most $2\pi p_0$. The sum of the lengths of the J_α is at most $4\pi p_0$. The sum of the lengths of the intervals I_α is at most $4K\pi\epsilon p_0$, and the intervals I_α cover the set of θ for which $g_t r_\theta \phi \in N(\epsilon, C)$. Thus we can take $K' = 4K\pi p_0$. $\square$

Proof of Theorem 3.9. Assume the theorem is false. Then choose constants $\epsilon_i \to 0$, times $T_i \to \infty$, a constant C, and sets S_i so that Proposition 3.15 holds. In particular, $\mu(S_i) \geq \delta > 0$. Let K', T be as in Proposition 3.17; then, for ϵ_i small enough and $T_i > T$ large enough, we have: $\mu(S_i) < K'\epsilon_i$. However, for ϵ_i small enough, one has: $K'\epsilon_i < \delta$, a contradiction. $\square$

3.4 Further ergodicity results

The measure 0 result in Theorem 3.5 was improved in [41] to give a statement about Hausdorff dimension.

Theorem 3.18 *For any quadratic differential ϕ the set of $\theta \in [0, 2\pi]$ such that $r_\theta \phi$ has minimal but non-uniquely ergodic vertical foliation has Hausdorff dimension at most* $1/2$.

On the other hand positive Hausdorff dimension turns out to be typical as was shown in [42]. Recall the stratum $QD(\sigma)$ and $SL(2, \mathbf{R})$ ergodic invariant measure μ_0 on each component of $QD(\sigma)$. Assume that $\sigma \neq (\emptyset; +), (1, -1; -), (-1, -1, -1, -1; -)$, that is, the quadratic differentials do not define a flat torus, once punctured torus, or four times punctured sphere.

Theorem 3.19 *For each component C of $QD(\sigma)$ there is a $\delta = \delta(C) > 0$ such that for μ_0 almost all $\phi \in C$, the set of $\theta \in [0, 2\pi]$ such that $r_\theta \phi$ has minimal but non-uniquely ergodic vertical foliation has Hausdorff dimension δ.*

3.5 Ergodicity of general polygonal billiards

Now we apply Theorem 3.5 to polygonal billiards. As before, the collection of plane polygonal regions with n vertices and a given combinatorial type is identified with a subset of $\mathbf{R}^{2n}$ and is given a subspace topology. Recall that $G(Q)$ denotes the group generated by reflections in the sides of a polygon Q.

The next result is in the spirit of Theorem 1.10 but is stronger.

Theorem 3.20 ([34]) *Let X be a closed subset of the space of billiard tables with the property that for any number N the set of rational tables $Q \in X$ with* $\mathrm{card}(G(Q)) \geq N$ *is dense. Then ergodic tables in X form a dense G_δ subset.*

Proof. We can assume that $X \subset \mathbf{R}^{2n}$ is compact. Each $Q \in X$ is a polygonal region in the plane. We can assume that the area of Q is 1. Let PX be the bundle whose base space is X and the fiber P_Q over $Q \in X$ is the phase space of the corresponding billiard table Q. Note that $PX \subset X \times \mathbf{R}^2 \times \mathbf{S^1}$. Let μ_Q the product of area measure on Q with unit Lebesgue measure on $\mathbf{S^1}$, and let ϕ_t be the billiard flow on PX.

Choose a sequence of continuous functions $f_1, f_2 \ldots$ on PX which, when restricted to each P_Q, are dense in $L^2(P_Q)$. We make the further assumption that if v is an outward and v' is the corresponding inward vector on the boundary of Q, then $f_i(v) = f_i(v')$. Let $E(i, n, T)$ be the set of $Q \in X$ for which

$$\int_{z \in P_Q} (\frac{1}{T} \int_0^T f_i(\phi_t(z)) dt - \int_{P_Q} f_i d\mu)^2 d\mu < 1/n.$$

Let

$$E(i,n) = \cup_{T=1}^{\infty} E(i,n,T),$$

and let

$$E = \cap_{i=1}^{\infty} \cap_{n=1}^{\infty} E(i,n).$$

The set of $Q \in X$ for which ϕ_t restricted to P_Q is ergodic is precisely E – see [44]. We will prove that

- (1) The sets $E(i,n,T)$ are open, and
- (2) For a given i and n, there is an N such that $E(i,n)$ contains all rational tables Q for which $\mathrm{card}(G(Q)) > N$.

Assuming these statements, the theorem is proved as follows. The first stetement implies that the sets $E(i,n)$ are open. The second one implies that the $E(i,n)$ are dense. Then E is a G_δ, and it follows from the Baire category theorem that E is dense in X.

Statement (1) is a consequence of the next result.

Lemma 3.21 *Let $T > 0$ be fixed, and let f be a continuous function on PX respecting the boundary identifications. Then*

$$\int_{z \in P_Q} \left(\frac{1}{T}\int_0^T f(\phi_t(z))dt - \int_{P_Q} f d\mu_Q\right)^2 d\mu_Q$$

depends continuously on Q.

Proof. For $Q \in X$ let $a(Q) = \int_{P_y} f d\mu_Q$. Clearly, $a(Q)$ depends continuously on Q.

Replace the function f by the function $\tilde{f}$ defined as follows: for $z \in P_Q$ let $\tilde{f}(z) = f(z) - a(Q)$. Then the proof of the lemma reduces to the proof of the continuity of the following function:

$$c(Q) = \int_{z \in P_Q} \left[\frac{1}{T}\int_0^T \tilde{f}(\phi_t(z))dt\right]^2 d\mu_Q.$$

The difficulty is that, due to discontinuities of the billiard flow at the corners, the billiard orbits of close velocity vectors may diverge. Roughly speaking, this difficulty is overcome by deleting the set of velocity vectors whose trajectories hit a corner; this set has zero measure, and the time T trajectory depends continuously on the velocity vector in its complement.

More specifically, for $z \in P_Q \subset PX$ introduce an auxiliary function $l(z)$ as the infimum of the distance from the time T trajectory of vector z to the set of vertices of the polygon Q. Clearly $l(z)$ depends continuously on z, and the zero level set of $l(z)$, that is, the set of velocity vectors whose time T trajectory hits a vertex, has zero measure.

Let $\epsilon > 0$ be fixed, and let $M = \sup \tilde{f}$. For a neighborhood N_1 of Q in the space of polygons denote by Q_1 the intersection of all polygons from N_1 (note that Q_1 is not a polygon). We can choose N_1 so small that the area of Q_1 is at least $1 - \epsilon/6M^2$.

Let $C_\delta \subset P_Q$ be the set of velocity vectors z for which $l(z) \geq \delta$. Choose δ small enough so that the measure of C_δ is at least $1 - \epsilon/6M^2$. Let $D_\delta \subset C_\delta$ consist of pairs $(x, v) \in \mathbf{R}^2 \times \mathbf{S^1}$ with $x \in Q_1$. Note that the measure of D_δ is at least $1 - \epsilon/3M^2$. Since $Q_1 \times \mathbf{S^1}$ is contained in P_R for all $R \in N_1$, we can identify $N_1 \times D_\delta$ with a subset of PX in a natural way. Let $\bar{l}(R)$ be the infimum of $l(z)$ for $z \in (P_R \cap (N_1 \times D_\delta))$. Now $\bar{l}$ is continuous and $\bar{l}(Q) = \delta$. We can find a neighborhood $N_2 \subset N_1$ so that for all $R \in N_2$, $\bar{l}(R) > 0$. Let $d(R)$ denote

$$d(R) = \int_{z \in (P_R \cap (N_1 \times D_\delta))} [\frac{1}{T} \int_0^T \tilde{f}(\phi_t(z))dt]^2 d\mu_R.$$

Then

$$|d(R) - c(R)| \leq M^2 \mu(P_R \setminus (P_R \cap (N_1 \times D_\delta))) \leq \epsilon/3$$

for each $R \in N_2$. For $R \in N_2$ the function $f(\phi_t(z))$ depends continuously on $t \in [0, T]$ and $z \in D_\delta$. Thus we can find a neighborhood N_3 of Q in which d varies by less than $\epsilon/3$. Then, for $R \in N_3$,

$$|c(Q) - c(R)| \leq |c(Q) - d(Q)| + |d(Q) - d(R)| + |d(R) - c(R)| \leq \epsilon.$$

This completes the proof of continuity of c and the proof of the Lemma. $\square$

To finish the proof of the Theorem one needs to establish statement (2); this statement follows from the next lemma.

Lemma 3.22 *Fix $n > 0$ and let f be a continuous function on PX. Choose $\delta > 0$ so that if $|\theta_1 - \theta_2| < \delta$ then $|f(\theta_1) - f(\theta_2)| < 1/2n$. Let $N \geq 2/\delta$, and let Q be a rational polygon with $|G(Q)| \geq N$. Then for sufficiently large T one has:*

$$[\int_{z \in P_Q} [\frac{1}{T} \int_0^T f(\phi_t(z))dt - \int_{P_x} f d\mu_Q]^2 d\mu_Q]^{1/2} < 1/n.$$

Proof. Since Q is fixed we will drop the subscripts from P, G and μ. For $\theta \in \mathbf{S^1}$ let $u(\theta) = 1/|G| \int_{z \in M_\theta} f(z) dA$ where dA is the area measure on M_θ; here, as before, M_θ is the invariant surface of the billird flow in direction θ. For $z \in M_\theta$ let $u'(z) = u(\theta)$. For $z \in P$ let $v_T(z) = 1/T \int_0^T f(\phi_t(z)) dt$. The quantity which appears in the lemma is the norm in the space $L^2(P)$ of the function $v_T - \int f d\mu$. We claim that

$$\lim_{T \to \infty} ||v_T - u'|| = 0.$$

To prove the claim, notice that the surfaces M_θ are parametrized by $\theta \in \mathbf{S^1/G}$. We evaluate the norm by integrating first with respect to M_θ and then with respect to θ:

$$||v_T - u'|| = [|G| \int_{\theta \in \mathbf{S^1/G}} \frac{1}{|G|} \int_{z \in M_\theta} (v_T(z) - u'(z))^2 dA d\theta]^{1/2}.$$

Let

$$w_T(\theta) = [\frac{1}{|G|} \int_{z \in M_\theta} (v_T(z) - u'(z))^2 dA]^{1/2}.$$

Then

$$||v_T - u'|| = [|G| \int_{\theta \in \mathbf{S^1/G}} w_T(\theta)^2 d\theta]^{1/2}.$$

For a given θ the ergodicity of ϕ restricted to M_θ implies that $\lim_{T \to \infty} w_T(\theta) = 0$. Theorem 3.5 implies ergodicity for almost all θ. Since the functions w_T are bounded and converge pointwise almost everywhere to 0, they converge to 0 in norm. This completes the proof of the claim.

The second claim is that

$$||u' - \int f d\mu|| \leq 1/2n.$$

Now $||u' - \int f d\mu||$ is equal to the norm of $u - \int f d\mu$ in $L^2(\mathbf{S^1})$. Let θ_1 and θ_2 be points in the circle at which u assumes its minimum and maximum values m and M, respectively. Note that u is constant on the orbits of G. The distance between neighboring points in a G orbit is less than $2/|G| < 2/N < \delta$. By replacing θ_2 by $g(\theta_2)$, where $g \in G$, we may assume that $|\theta_1 - \theta_2| < \delta$. It follows from the continuity assumption on f that since $|\theta_1 - \theta_2| < \delta$, then $|u(\theta_1) - u(\theta_2)| < 1/2n$. Since u is defined by averaging f, $\int_P f = \int_{S^1} u$. Thus $m \leq \int f \leq M$; hence $|u(\theta) - \int f| \leq 1/2n$ and $||u - \int f|| \leq 1/2n$. This completes the proof of the claim.

We now complete the proof of the lemma. Choose T sufficiently large so that $||v_T - u'|| \leq 1/2n$. Then

$$||v_T - \int f|| \leq ||v_T - u'|| + ||u' - \int f|| \leq 1/2n + 1/2n = 1/n,$$

and we are done. $\square$

3.6 Construction of polygons with ergodic billiard flow

Ya. Vorobets gave a constructive description of a topologically massive set of polygons with ergodic billiard flow – see [56]. We describe his result without proof.

Definition 3.23 *Let $\phi(N)$ be a positive function of a positive integer N whose limit is zero as $N \to \infty$. Let Q be a plane k-gon with angles $\alpha_1, ..., \alpha_k$ between the adjacent sides. We say that Q admits approximation by rational polygons at the rate $\phi(N)$ if for every $n > 0$ there is $N > n$ and positive integers $n_1, ..., n_k$, each coprime with N, such that $|\alpha_i - \pi n_i/N| < \phi(N)$ for all i.*

The next theorem gives an explicit estimate of how well a polygon should be approximated by rational ones to guarantee ergodicity of the billiard flow.

Theorem 3.24 *Let Q be a polygon that admits approximation by rational polygons at the rate*

$$\phi(N) = \left(2^{2^{2^{N}}}\right)^{-1}.$$

Then the billiard flow in Q is ergodic.

The paper [56] also contain constructive versions of other results on polygonal billiards and flat surfaces. In particular it gives a different proof of the quadratic upper bound growth rate for the number saddle connections, a result we will discuss in more detail in the next section.

4 Periodic orbits

4.1 Periodic directions are dense

Suppose β is a closed geodesic of a quadratic differential ϕ that does not pass through a zero. As we have observed, β can be moved parallel to itself and sweeps out a cylinder of homotopic geodesics; we called such a cylinder a *metric cylinder*. The boundary of the metric cylinder is a union of parallel saddle connections. In the special case of the flat torus, there are closed geodesics in a dense set of directions, and, of course, for each such direction the closed geodesics fill the surface. For surfaces of higher genus, the boundary of the metric cylinder contains saddle connections. We have already seen that for a rational billiard, there are periodic trajectories that leave a side orthogonally. The main objective of this section is to prove the following theorem,

Theorem 4.1 ([38]) *For any quadratic differential ϕ_0 there is a dense set of directions $\theta \in \mathbf{S^1}$ such that ϕ_0 has a metric cylinder in the direction θ.*

Corollary 4.2 *For any rational billiard table there is a dense set of directions with a periodic orbit in that direction.*

Theorem 4.1 was strengthened in the following result of Boshernitzan, Galperin, Kruger and Troubetzkoy – see [9].

Theorem 4.3 *For any quadratic differential ϕ_0 there is a dense set of vectors in the tangent space to the surface such that the orbit determined by that vector is closed.*

The theorem is proved in [9] for rational billiards. The proof holds verbatim for the general case of a quadratic differential.

Recall that in the compactification of R_g we allow pinching Riemann surfaces along $p \leq 3g-3$ simple closed curves to produce Riemann surface X_0 with nodes or punctures. Also recall that if ϕ_n is a sequence of unit norm quadratic differentials on X_n, we can pass to a subsequence and assume that ϕ_n converges uniformly on compact sets of X_0 to a finite norm quadratic differential ϕ_∞ on X_0. We single out the a set of limiting quadratic differentials.

Definition 4.4 *A limiting quadratic differential ϕ_∞ on X_0 is called exceptional if*

(a) *some component of X_0 is a torus and ϕ_∞ defines the flat metric on the torus or*

(b) *some component of X_0 is a punctured sphere and ϕ_∞ has no zeroes or*

(c) *$\phi_\infty \equiv 0$ on every component of X_0.*

We note that in case (b), since ϕ_∞ has no zeroes, there must be exactly four punctures at which ϕ_∞ has a simple pole, and at any other puncture, ϕ_∞ is regular. The importance of this definition is illustrated by the following proposition. We are indebted to Yair Minsky for suggesting the proof. A different proof appeared in [38] based on work of [37].

Proposition 4.5 *There exists $M > 0$ with the property that if $X_n \to X_0$ in $\bar{R}_g$ and ϕ_n are quadratic differentials on X_n which converge uniformly on compact sets of X_0 to an exceptional ϕ_∞, then for large enough n, ϕ_n has a metric cylinder of length at most M.*

Proof. In cases (a) and (b) there is a dense set of directions such that ϕ_∞ has a metric cylinder in that direction. Choose one such metric cylinder and a neighborhood of the punctures such that the cylinder contains a closed curve missing that neighborhood. The uniform convergence of ϕ_n to ϕ_∞ on the complement of this neighborhood implies that for large n, ϕ_n has a closed regular geodesic with some uniformly bounded length.

Thus we can assume we are in case (c). For any $\epsilon > 0$ and for each loop C surrounding a puncture of X_0 we have $|C|_{\phi_n} < \epsilon$ for large n. The loops C divide X_n into p annuli B_i homotopic to the pinching curves and $\phi_n \to \phi_\infty$ uniformly on $X_0 \setminus (\cup B_i)$. Since $\phi_\infty \equiv 0$, the ϕ_n area of $X_0 \setminus (\cup B_i)$ goes to 0 and so the ϕ_n area of $\cup B_i$ goes to 1. For large n, for at least one annulus B, the ϕ_n area of B is at least $1/(3g-3)$. This annulus has two boundary components C.

Fix such a large n. If there is no metric cylinder in the homotopy class of C then we may let α be the unique ϕ_n geodesic homotopic to the loop C. Then α passes through singularities, at some of which the angle is in excess of π. We will arrive at a contradiction. First assume that α is embedded. For each r let $N(r)$ be the r neighborhood of α; the set $N(r)$ is convex. For small values of r, $N(r)$ is an annulus containing α in its interior (here is where we use that α is embedded). For all r it is a domain with nonpositive Euler characteristic. Let $\alpha_r = \partial N(r)$. For all except an isolated set of r, the curve α_r is smooth.

The Gauss-Bonnet formula says that

$$\int_{N(r)} K + \int_{\alpha_r} \kappa ds = 2\pi\chi(N(r)) \leq 0,$$

where K is the Gaussian curvature and κ is the geodesic curvature. Now $K = 0$ except at the interior points x of $N(r)$, where the curvature is negative and concentrated so that the contribution to the first term is $-m\pi$, for m a positive integer. The number m may jump a bounded number of times as more zeroes are included in $N(r)$ and so is maximized by some M. Thus

$$\int_{\alpha_r} \kappa ds \leq m\pi \leq M\pi.$$

with equality in the first inequality as long as $N(r)$ remains an annulus. Let $A(r) = ||N(r)||$, the area of $N(r)$, and let $L(r)$ denote the length of α_r. Then

$$A'(r) = L(r) \quad \text{and} \quad L'(r) = \int_{\alpha_r} \kappa ds \leq m\pi \leq M\pi,$$

again with equality in the first inequality as long as $N(r)$ remains an annulus. In particular this gives $L(r) \geq \pi r$, as long as $N(r)$ remains an annulus. Now the fact that $A''(r) \leq M\pi$ together with $A(0) = 0$ and $A'(0) \leq \epsilon$, gives

$$A(r) \leq M\pi r^2 + \epsilon r.$$

Choose ϵ small enough so that

$$A(\epsilon) \leq \frac{1}{3g-3}.$$

We show that $B \subset N(\epsilon)$ and that this leads to the desired contradiction. Let $r_0 = \epsilon/\pi$. The first step is to notice that each component C of the boundary of B must intersect $N(r_0)$. If $N(r_0)$ is an annulus this follows from the fact that $|\alpha_{r_0}| \geq r_0\pi = \epsilon$ and $N(r_0)$ is convex, so any curve, namely C, of length smaller than ϵ cannot be homotopic to and exterior to $N(r_0)$. If $\chi(N(r_0)) < 0$, then C must intersect $N(r_0)$ since C is homotopic to α. Since C has length at most ϵ the fact that C intersects $N(r_0)$ implies that C must be contained in $N(\epsilon(1/\pi + 1/2)) \subset N(\epsilon)$. Then both boundary components C of B are contained in $N(\epsilon)$, so we must have $B \subset N(\epsilon)$. But now

$$||B|| \leq A(\epsilon) \leq \frac{1}{(3g-3)},$$

which is a contradiction.

In case α is not embedded, so that α is not in the interior of $N(r)$, we lift α to the annular cover so that it is embedded. We have the same lower bound growth estimates for the length of curves in a neighborhood of the lifted α as well as the upper bound for the area of the neighborhood. This neighborhood projects to $N(r)$. The rest of the proof then goes through. □

The following Lemma says that the Teichmuller flow extends continuously to the compactification.

Lemma 4.6 *If ϕ_0 is not identically 0 on every component of X_0, then for any t, $g_t(\phi_n) \to g_t(\phi_0)$.*

Proof. The proof is immediate by viewing the flow in the natural coordinates of quadratic differentials. $\square$

For a given genus g, a stratum $QD(\sigma)$ is not a closed subset of Q_g, unless σ consists of a single zero. This is because a sequence in $QD(\sigma)$ may collapse a pair of lower order zeroes into a higher order zero. In addition, the degeneration of Riemann surfaces allows limits of quadratic differentials on which the topology of the surface has changed. Because a degenerated surface need not be connected, in the following definition we allow disconnected surfaces and corresponding strata.

Definition 4.7 *A stratum $QD(\sigma')$ is a degeneration of a stratum $QD(\sigma)$ if $QD(\sigma') \neq QD(\sigma)$ and there exists a sequence $\phi_n \in QD(\sigma)$ on Riemann surfaces $X_n \in \bar{R}_g$ such that $X_n \to X_0$ in $\bar{R}_g$ and ϕ_n converges uniformly on compact sets to some $\phi_\infty \in QD(\sigma')$ where ϕ_∞ is not identically zero.*

For any quadratic differential ϕ which is not identically zero, denote by $l(\phi)$ the length of the shortest saddle connection of ϕ. Notice that $l(\cdot)$ is not continuous under degeneration within a fixed genus. This is because zeroes are collapsed. Thus if $\phi_n \in QD(\sigma)$ is a sequence converging to $\phi_\infty \in QD(\sigma')$ then $l(\phi_n) \to 0$ while $l(\phi_\infty) > 0$.

Now suppose a nonzero quadratic differential ϕ_0 on a connected surface is given. For any closed interval $I \subset [0, 2\pi]$ let

$$E_I(\phi_0) = \{\phi : \exists \theta_n \to \theta \in I, \exists t_n \to \infty, \exists \psi_n : r_{\psi_n} g_{t_n} r_{\theta_n} \phi_0 \to \phi\}.$$

That is to say, E_I is the set of all ω limit points ϕ in the $SL(2, \mathbf{R})$ orbit of ϕ_0 where the limiting initial rotation angle lies in I. Here the limiting ϕ may belong to any stratum. Note that θ_n need not lie in I. Now Theorem 4.1 will follow from the next proposition.

Proposition 4.8 *For any ϕ_0, $E_I(\phi_0)$ contains an exceptional ϕ.*

Proof. Since I is closed, $E_I = E_I(\phi_0)$ is compact. Assume E_I does not contain any exceptional ϕ; we will arrive at a contradiction. Let

$$E_I^T = \{\phi \in E_I : \forall \phi' \in E_I, QD(\sigma'(\phi')) \text{ is not a degeneration of } QD(\sigma(\phi))\}.$$

That is, E_I^T is the set of $\phi \in E_I$ which belong to a stratum that cannot be further degenerated within E_I. Let

$$l = \inf_{\phi \in E_I^T} l(\phi).$$

If $l = 0$ choose a sequence $\phi_n \in E_I^T$ with $\lim l(\phi_n) \to 0$. By passing to a subsequence we may assume that $\phi_n \to \phi_\infty \in E_I$. Since $l = 0$ then either the Riemann surface of ϕ_n has degenerated or some set of zeroes of ϕ_n have been collapsed to a higher order zero of ϕ_∞. In either case ϕ_∞ belongs to a stratum which is a degeneration of the stratum containing ϕ_n, a contradiction to the definition of E_I^T.

Suppose on the other hand that $l > 0$. Choose $t_n \to \infty$, $\theta_n \to \theta_1 \in I$ such that

$$g_{t_n} r_{\theta_n} \phi_0 \to \phi_1 \in E_I^T.$$

Then $l(\phi_1) \geq l$. Let γ be a saddle connection of ϕ_1 such that $l(\phi_1) = |\gamma|_{\phi_1}$. Let θ_γ be such that γ is vertical with respect to $r_{\theta_\gamma}\phi_1$. As $s \to \infty$ the length of γ with respect to the metric of $g_s r_{\theta_\gamma}\phi_1$ goes to 0. Thus we can choose s such that the length of γ with respect to the metric of $g_s r_{\theta_\gamma}\phi_1$ is less than $l/2$. We will have a contradiction if we can show that $g_s r_{\theta_\gamma}\phi_1 \in E_I^T$. Since the stratum containing ϕ_1 cannot be degenerated within E_I we only need to show that $g_s r_{\theta_\gamma}\phi_1 \in E_I$. Now by Lemma 4.6

$$g_s r_{\theta_\gamma}\phi_1 = \lim_{n\to\infty} g_s r_{\theta_\gamma} g_{t_n} r_{\theta_n}\phi_0.$$

We claim that there exist sequences $\theta'_n \to 0$, $t'_n \to \infty$, and $\psi_n \to \theta$ such that

$$g_s r_{\theta_\gamma} g_{t_n} = r_{\psi_n} g_{t'_n} r_{\theta'_n}.$$

Canonically identify G with the unit tangent vectors to the upper half-plane with Id identified with the vertical vector at $i = \sqrt{-1}$. Then $SL(2,\mathbf{R})$ acts by Mobius transformations. Furthermore ∞ is an attracting fixed point of the geodesic flow g_t. Therefore for each n, there is a unique rotation $r_{-\psi_n}$ such that $r_{-\psi_n} g_s r_\theta g_{t_n}$ takes the vertical vector at i to a vector based on the imaginary axis and Then for some some small $\theta_{n'}$ and large t'_n, $r_{-\theta'_n} g_{-t'_n}$ takes this vector to the vertical vector at i, proving the claim. Finally,

$$g_s r_{\theta_\gamma}\phi_1 = \lim_{n\to\infty} r_{\psi_n} g_{t'_n} r_{\theta_n+\theta'_n}\phi_0,$$

that is, $g_s r_{\theta_\gamma}\phi_1 \in E_I$, and we are done. □

Proof of Theorem 4.1. Choose a closed subinterval I' of I contained in the interior of I. Then Proposition 4.8 says that there must be some exceptional $\phi \in E_{I'}$. By Proposition 4.5, for a sequence $t_n \to \infty$ and $\theta_n \to \theta_0 \in I'$, $g_{t_n} r_{\theta_n}\phi_0$ has a metric cylinder in the homotopy class of some β_n and the length is uniformly bounded above by some M. Since $t_n \to \infty$ and g_{t_n} expands horizontal lengths, the horizontal length of β_n in the metric of $r_{\theta_n}\phi_0$ goes to 0.

Thus for some $\theta_n' \to 0$, the quadratic differential $r_{\theta_n'} r_{\theta_n} \phi_0$ has a metric cylinder in the class of β_n and since $\theta_n + \theta_n' \in I$, for large n, we are done. □

Proof of Theorem 4.3. Let $T_1(S) = \{(p, \theta)\}$ denote the unit tangent space to S; here $p \in S$ and $\theta \in \mathbf{S^1}$. Let $\epsilon > 0$. We begin by fixing a minimal uniquely ergodic direction $\theta \in \mathbf{S^1}$. Unique ergodicity implies that we can choose N large enough so that for all $x \in S$ the leaf through x is $\epsilon/2$-dense in S.

There are only a finite number of saddle connections of length less than or equal to $2N$ and so their tangent vectors lie in a finite set of directions. This means that we can choose δ sufficiently small so that if θ' satisfies $|\theta - \theta'| < \delta$, then any saddle connection in direction θ' has length at least $2N$. We may also choose $\delta < \epsilon/2N$. ¿From Theorem 4.1 we know that there is a point (x_0, θ_0) on a closed leaf L_{θ_0} in a metric cylinder such that $|\theta_0 - \theta| < \delta$. Now let L_θ be the leaf through x_0 in direction θ extended to length N in each direction, or to a zero. Now we claim that in at least one of the two possible directions, moving distance N, the endpoint of L_θ and the endpoint of L_{θ_0} are within $\epsilon/2$ of each other. To prove the claim, for each direction, form the simply connected domain bounded by the segments and a geodesic joining their endpoints. If both simply connected domains contained a zero in their interior, there would be a saddle connection γ of length at most $2N$ in direction θ' satisfying $|\theta' - \theta| < \delta$, contradicting the choice of δ. Thus in one direction the simply connected domain is actually a metric triangle with no interior zeroes. The choice of $\delta < \epsilon/2N$ says now that the endpoints are within $\epsilon/2$ of each other, proving the claim.

Since the segment of L_θ is $\epsilon/2$ dense in S, the leaf L_{θ_0} is ϵ dense. This can be done for every uniquely ergodic direction θ. The set of vectors determining uniquely ergodic directions is dense in the whole phase space $T_1 S$. Since ϵ is arbitrary this completes the proof. □

4.2 Counting saddle connections and maximal cylinders

In this section we discuss some results on the asymptotics of the number of saddle connections and periodic orbits. No proofs will be given. As noted in the section on billiards in the unit square, the number of (parallel families of) periodic orbits of length at most L grows asymtotically as $\pi L^2/2\zeta(2)$. For a general rational billiard or flat structure, asymptotics are not known. However we do have quadratic upper and lower bounds. To be specific, given a flat structure ϕ and $L > 0$, let $N_1(L)$ be the number of saddle connections of length at most L and $N_2(L)$ the number of maximal metric cylinders of length at most L. The following theorem was proved in [39] and

[40].

Theorem 4.9 *There exists positive constants c_1, c_2 such that*

$$c_2 L^2 \leq N_2(L) \leq N_1(L) \leq c_1 L^2.$$

Different proofs of the upper estimate are given in [56] and [13]. It is possible to prove precise asymptotics for generic flat structures.

Theorem 4.10 ([13]) *There exists constants $c = c(\sigma)$ and $s = s(\sigma)$ such that on each component of $QD(\sigma)$, for μ_0 a.e. $\phi \in QD(\sigma)$ one has: $N_1(L) \sim cL^2$ and $N_2(L) \sim sL^2$.*

In the next section we discuss some circumstances under which these precise asymptotics are known to hold.

5 Veech groups

5.1 Definition and examples of Veech groups

In this section we discuss the groups of affine transformations associated to flat structures. These groups give rise to subgroups in $PSL(2, \mathbf{R})$. If these groups are lattices in $PSL(2, \mathbf{R})$, then the flat structure has particularly nice properties.

Let ϕ be a quadratic differential on the Riemann surface X. Let Σ be a subset of X which contains the zeroes of ϕ (but may be larger). Let $X' = X \setminus \Sigma$. Following Veech we define the affine group $Aff^+(\phi)$ to be the group of orientation preserving self homeomorphisms of X that map X' to itself and are affine with respect to the natural coordinates of ϕ. This means that for each point $p \in X'$ we choose coordinates z near p so that $\phi = dz^2$ and coordinates w near $f(p)$ so that $\phi = dw^2$, and in these coordinates, the map f is an affine homeomorphism. We call the derivative (i.e., linear part) of that homeomorphism the derivative of f. It is well-defined independently of p, z and w up to a factor of $\pm I$. Let $a(f) = \pm A$ be the derivative. Then by an observation of Veech [54] we have

$$|f^*\phi| = det(A)|\phi|$$

and

$$\int |\phi| = \int |f^*(\phi)|,$$

which implies that the determinant is 1 and so A is an element of $G = PSL(2, \mathbf{R})$. We define the Veech group $\Gamma(\phi)$ to be the image of $Aff^+(\phi)$ under the derivative map.

The case of most interest is when Γ is a lattice in G. Recall that a lattice Γ in G is a discrete subgroup such that G/Γ has finite volume; a lattice is nonuniform if the quotient is not compact (see, e.g., [5]).

We present several examples. The first is essentially due to Thurston [49]. Recall that a (right) Dehn twist about a simple closed curve α on a surface is a homeomorphism of the surface which fixes α and twists any curve β crossing α once to the right about α.

Take the unit square with lower left vertex at the origin and identify opposite vertical sides. Mark off 4 points on both the top and bottom of the resulting cylinder C, dividing the top and bottom into equally spaced intervals and so that the points on top lie directly above the points on the bottom. Arrange the intervals on top into two pairs and glue each pair isometrically as indicated in the figure. Do the same for the bottom.

Fig. 10

The Euclidean metric on the rectangle extends to a quadratic differential $\phi = dz^2$ on the glued surface X, which has genus 2. The 4 marked points on the top identify to a single zero of order 2; similarly for the bottom. Let Σ consist of these two points. We define an affine map $f_1 : X \to X$ which in the coordinates (x, y) is given by

$$(x, y) \to (x + y \ mod(1), y).$$

The map is easily seen to be well-defined on the identification space, fixes Σ and the top and bottom of C, and acts as a right Dehn twist on the core curve of C. Its derivative is

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

To find a second affine map f_2 notice that the gluings force all trajectories in the vertical direction to be closed as well. Indeed, an upward vertical trajectory hits the top horizontal side of the square at some point S and continues downwards from the point T of the same horizontal side which is identified with S. The surface X decomposes into vertical metric cylinders C_1, C_2 of circumference 2 and height 1/4 each composed of 2 vertical strips glued together; C_1 is formed from the rectangles $0 \leq x \leq 1/4$ and $1/2 \leq x \leq 3/4$ glued along tops and bottoms and similarly for C_2.

(Note that here circumference is the vertical dimension of the cylinders while height is the horizontal one). This allows us to introduce coordinates ($0 \leq x' \leq 1/4, 0 \leq y' \leq 2$) in C_1 by $x' = x, y' = y$ for $0 \leq x \leq 1/4$ and $x' = 3/4 - x, y' = 2 - y$ for $1/2 \leq x \leq 3/4$ and similarly for C_2. The map f_2 is a right Dehn twist in each C_i. It is defined by

$$(x', y') \to (x', y' + 8x' mod(2)).$$

It again fixesd Σ and its derivative is

$$A_2 = \begin{pmatrix} 1 & 0 \\ 8 & 1 \end{pmatrix}.$$

The matrices A_1, A_2 generate a finite index subgroup of $PSL(2, Z)$.

Gutkin and Judge [24] showed that the Veech group is commensurable with $SL(2, Z)$ if and only if the surface can be tiled by a single Euclidean parallelogram. Recall that two subgroups Γ_1, Γ_2 of G are commensurable if there is some $g \in G$ such that $g\Gamma_1 g^{-1} \cap \Gamma_2$ has finite index in both $g\Gamma_1 g^{-1}$ and Γ_2.

The next set of examples are due to Veech – see [54], [55]. Part of the motivation was to construct an example of a non-arithmetic group, that is, a group, not commensurable to $PSL(2, Z)$. These examples are the flat structures ϕ_n associated with billiards in right triangles with angles $\pi/n, n \geq 5$. Veech proved the following theorem. The set Σ consists of the zeroes of ϕ_n.

Recall that the (p, q, r) triangle group is the index 2 subgroup of the group generated by reflections in the sides of a triangle with angles $\pi/p, \pi/q, \pi/r$ in the hyperbolic plane; the triangle group is a subgroup of G, the group of isometries of the hyperbolic plane. If an angle is 0, then the vertex of the triangle is at infinity (see, e.g., [5]).

Theorem 5.1 *For n odd, $\Gamma(\phi_n)$ is a $(2, n, \infty)$ triangle group. For even $n = 2m$, $\Gamma(\phi_n)$ is the (m, ∞, ∞) triangle group. In either case the Riemann surface X is the nonsingular surface associated to $y^2 + x^n = 1$.*

One can check that the invariant surfaces for this triangle and the isosceles triangle with equal angles of π/n are translation isomorphic so that they determine the same Veech group.

We will not give a proof of the theorem but illustrate the main idea by an example. It is convenient to discuss $\Gamma(\phi_8)$ as this invariant surface was already descirbed in figure 4. The invariant surface is the regular octagon P with opposite sides identified. The surface X has genus 2 and the quadratic differential ϕ_8 has a single zero of order

4 corresponding to the identification of all of vertices of P. The counterclockwise rotation f_1 of order 8 about the center of P is obviously affine with respect to ϕ_8 and is holomorphic with respect to X. It fixes the zero. Its derivative is

$$A_1 = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}.$$

To find a second affine map, rotate the octagon so that two of the opposite sides are vertical. Then we claim that every horizontal trajectory is closed and that the surface decomposes into two metric cylinders. The first cylinder C_1 consists of horizontal segments joining the opposite vertical sides. Its boundary consists of the two horizontal segments joining opposite vertices. Next notice that because of the identifications made, each horizontal segment leaving one of the nonvertical sides is also closed; it first hits the other nonvertical side before closing. The family of these closed trajectories determine the second metric cylinder C_2. The boundary includes the first boundary together with an extra closed saddle connection which is the top and bottom horizontal segments of the octagon. If we normalize so that the sides of P are of length 1, then C_1 has circumference $\cot \pi/8 = 1 + \sqrt{2}$ and height 1, while C_2 has circumference $1 + \cot \pi/8 = 2 + \sqrt{2}$ and height $\sqrt{2}/2$. Notice that the ratio of height to circumference for the first cylinder is twice that for the second one. The ratio of height to circumference is called the *modulus* of a metric cylinder.

Fig. 11

Then there exists an affine map which preserves each cylinder fixing the boundary pointwise and whose derivative is

$$A_2 = \begin{pmatrix} 1 & 2\cot \pi/8 \\ 0 & 1 \end{pmatrix}.$$

The map is the square of a right Dehn twist on C_1 and a right Dehn twist on C_2. Now A_1, A_2 generate a $(8, \infty, \infty)$ triangle group which is not commensurable with $PSL(2, Z)$ ([5]).

The following theorem combines results from [12],[57] and [59], and complements Theorem 5.1.

Theorem 5.2 *The Veech groups of the flat surfaces associated with the three series of triangles with angles*

$$(\pi/2n, \pi/n, (2n-3)\pi/2n), \quad ((2k-1)\pi/4k, (2k-1)\pi/4k, \pi/2k)$$

and

$$(k\pi/2k+1, k\pi/2k+1, \pi/2k+1), \quad n \geq 4, \ k \geq 2,$$

are, respectively,

$$(3, n, \infty), \ (2k, \infty, \infty) \ \text{ and } \ (2k+1, \infty, \infty)$$

triangle groups. In all cases the set Σ is the set of zeroes.

5.2 Veech dichotomy

We now discuss some properties of $\Gamma(\phi)$. These results are all due to Veech – [54], [55].

Lemma 5.3 *Let $f \in Aff^+(\phi)$ such that $A = a(f) \neq \pm Id$, and let $\zeta \in \mathbf{R}^2 \setminus \{0\}$ such that $A\zeta = \pm\zeta$. Then every leaf in directon corresponding to ζ is either a saddle connection or closed so that the surface decomposes into cylinders $C_1, \ldots, C_p$ in direction ζ. Moreover there exists a positive integer k such that for each i the map f^k preserves C_i, fixing the boundary, and is a power of a Dehn twist in C_i.*

Proof. Clearly f fixes the foliation in direction ζ. For some k, that can be assumed even, f^k maps each separatrix to itself. Since $A\zeta = \pm\zeta$, f^k must be the identity on each separatrix. Each separatirix is either a saddle connection or dense in an open set. In the latter case, since f^k is real analytic, f^k would be the identity in the open set, and so $A^k = \pm Id$. However A is an unipotent, therefore $A^k \neq \pm Id$. Thus every separatrix is a saddle connection which means that the surface decomposes into cylinders such that f^k is the identity on the boundary of each cylinder. As it is linear on each cylinder, it must be a power of a Dehn twist. □

Veech also shows that the moduli of the cylinders are rationally related.

Lemma 5.4 *If $a(f) \neq \pm Id$ then f is not isotopic to the identity.*

Proof. Since f is affine in the coordinates determined by ϕ, it is a Teichmuller map determined by ϕ. Teichmuller maps are extremal quasiconformal maps within their homotopy class. This means that f cannot be homotopic to the identity, since the identity is conformal. □

Corollary 5.5 *$Aff^+(\phi)$ is a subgroup of the mapping class group of the Riemann surface.*

Lemma 5.6 *$\Gamma(\phi)$ is a discrete subgroup of G.*

Proof. Suppose $f_n \in Aff^+(\phi)$ is such that $a(f_n) \to \pm Id$. By passing to a subsequence and using the Arzela-Ascoli theorem we can assume that f_n converges uniformly to a diffeomorphism f. For large m, $f_m \circ f_{m+1}^{-1}$ is isotopic to the identity, which by Lemma 5.4 gives $a(f_{m+1}) = a(f_m)$ and so $a(f_m) = \pm Id$ for large m. $\square$

We now consider consequences of the assumption that $\Gamma(\phi)$ is a lattice. Begin again by canonically identifying G with the unit tangent vectors to the upper half-plane with Id identified with the vertical vector at $i = \sqrt{-1}$. Then $SL(2, \mathbf{R})$ acts on itself by Mobius transformations. Recall that $SL(2, \mathbf{R})$ also acts on quadratic differentials. If we associate the identity in G to ϕ, then we canonically identify the $SL(2, \mathbf{R})$ orbit of ϕ with $SL(2, \mathbf{R})$.

Theorem 5.7 *(Veech dichotomy). Suppose $\Gamma(\phi)$ is a lattice. For any direction either the foliation in that direction is minimal or its every leaf is closed or a saddle connection. In the latter case the surface decomposes into cylinders of closed leaves and there is $f \in Aff^+(\phi)$ with $a(f) \neq I$ which preserves these cylindrers and acts as Dehn twists on them. Furthermore Γ is a nonuniform lattice.*

Proof. If the foliation is not minimal there is a saddle connection in that direction. By rotating, we can assume the direction is vertical. Then $g_t\phi$ has a saddle connection which approaches 0 in length. This means that the ω limit set under the flow g_t of Id in G/Γ is empty. Since Γ is a lattice this can only happen if the point at infinity is a fixed point for some parabolic element of Γ. In particular, the lattice is not uniform. Such a parabolic element fixes the vertical vector $\zeta = [0, 1]$. It remains to apply Lemma 5.3. $\square$

A different proof of the above theorem can be found in [57].

Theorem 5.7 says that, just as in the case of the flat torus, there is a dichotomy: either the flow is minimal or every orbit is closed. Since the triangle groups are lattices, the Veech dichotomy applies, in view of Theorem 5.1, to billiards in the isosceles triangles with equal angle π/n; it applies to regular polygons as well.

In fact, the analogy with the flat torus extends to ergodicity as well.

Theorem 5.8 *Suppose $\Gamma(\phi)$ is a lattice. If the foliation in a direction is minimal, then it is uniquely ergodic.*

Proof. Again we can assume the direction is vertical. If the foliation is not uniquely ergodic then by Theorem 3.8 $g_t\phi$ eventually leaves every compact set of moduli space. In particular, it cannot have an ω limit point in G/Γ. Again this implies that Γ has a parabolic element fixing infinity. By Lemma 5.3 the leaves in the vertical direction are all closed, which contradicts the minimality. □

5.3 Asymptotics

Let ϕ be a quadratic differential. Recall that we defined $N_2(\phi, T)$ as the number of homotopy classes of closed geodesics of length less than T or equivalently, the number of maximal cylinders whose circumference has length at most T. Recall in Theorem 4.9 there are upper and lower quadratic bounds for the number of maximal cylinders. Veech (op. cit.) has shown the following theorem.

Theorem 5.9 *Suppose $\Gamma(\phi)$ is a lattice. Then there exists a constant $c = c(\phi)$ such that $N_1(T) \sim cT^2$.*

The proof is using Eisenstein series. The constant c can be computed explicitly in the case of isosceles triangles T_n with angles $(\pi/n, \pi/n, (n-2)\pi/n)$. In that case

$$c = \frac{n(n^2-1)}{48\pi(n-2)|T_n|},$$

where $|T_n|$ is the area of the triangle.

Gutkin and Judge [24] have given another proof of the same theorem using the mixing of the horocyclic flow. The constant c is given in their work in purely topological terms.

5.4 Covers

Constructing a flat surface M from a rational billiard polygon P it may happen that some of the cone angles are equal to 2π or equivalently, the points are not zeroes of the quadratic differential. For example, if P is a square then M is a flat torus with no singular points. From the billiard view point this means that one can continuously define the extention of the billiard trajectory through the vertex of P; this is possible if and only if the angle is of the form π/n. Adding such a point to Σ, the set of points that are required to be preserved by the affine diffeomorphisms may change the Veech group. It is particularly important to keep track of removable singular points in the study of covers of flat structures.

Definition 5.10 *Let $(X_i, \phi_i), i = 1, 2$ be flat structures and $\Sigma_i \subset X_i, i = 1, 2$ sets that contain the zeroes of X_i. Then (X_1, ϕ_1, Σ_1) covers (X_2, ϕ_2, Σ_2) if there exists a continuous map $f : X_1 \to X_2$, called a cover, that sends Σ_1 to Σ_2 and which is given, in local coordinates on the complement of these sets by parallel translation. The multiplicity of a cover is the number of preimages of a non-singular point in X_2 (independent of the point).*

In particular, a cover is a branched holomorphic cover of X_2 by X_1 branched over Σ_2.

Definition 5.11 *Two covers $f_1 : X_1 \to Y$ and $f_2 : X_2 \to Y$ are called isomorphic if there exists an isomorphism of the flat structures $g : X_1 \to X_2$ such that $f_1 = f_2 g$. Similarly one defines an isomorphism of covers $f_1 : Y \to X_1$ and $f_2 : Y \to X_2$.*

The following finitness property holds (see [57] for a proof).

Proposition 5.12 *A given flat structure admits a finite number of isomorphic classes of covers with a given multiplicity. Likewise, the number of isomorphic classes of covers realizable by a given flat structure is finite.*

The next construction provides examples of covers of flat structures associated with billiards in polygons. Let P_1 and P_2 be rational polygons, and assume that P_2 tiles P_1 by reflections. This means that P_1 is partitioned into a number of isometric copies of P_2, each two either disjoint, or having a common vertex or a common side, and if two of these polygons have a side in common then they are symmetric with respect to this side.

For example, a right triangle with angle π/n tiles by reflections a regular n-gon – see figure 4 for $n = 8$. Another example is given by the above mentioned Gutkin's almost integrable billiard polygons: if such a polygon is drawn on a, say, integer square lattice in the plane, then it is tiled by the unit square.

Let (X_i, ϕ_i) be the flat structures associated with the polygon P_i, $i = 1, 2$, and assume that P_2 tiles P_1. Let Σ_1 be the set of zeroes of ϕ_1 (cone points with angles greater than 2π.

Lemma 5.13 *There is a set Σ_2 containing the zeroes of ϕ_2 such that (X_1, ϕ_1, Σ_1) covers (X_2, ϕ_2, Σ_2).*

Proof. Denote by G_1 and G_2 the groups generated by the linear parts of the reflections in the sides of the polygons P_1 and P_2. Then $G_1 \subset G_2$. Each copy of P_2, involved in the tiling of P_1, is identified with P_2 by an isometry; this isometry is a composition of reflections in the sides of P_2. These identifications, combined, define a projection p from P_1 to P_2. Given a point $x \in P_1$, let $\alpha(x) \in G_2$ be the linear part of the isometry that takes P_2 to the tile (a copy of P_2) that contains x.

Consider the map $g : P_1 \times G_1 \to P_2 \times G_2$ that sends (x, β) to $(p(x), \beta\alpha(x))$. Notice that the surfaces X_i are quotient spaces of $P_i \times G_i$, $\; i = 1, 2$, and the map g determines a map $f : X_1 \to X_2$. This is the desired cover. The multiplicity equals n/m where n is the number of tiles in the tiling and $m = [G_2 : G_1]$.

Some singular points of the first structure may project to removable singular points of the second one. This happens only if P_2 has a vertex angle π/n. One then adds these points to Σ_2. $\square$

Note that one does not need to add removable singular points if none of the angles of P_2 is of the form π/k.

As an example to Lemma 5.13 consider again a regular octagon P_1 tiled by right triangles P_2 with angle $\pi/8$. The flat surface M_2 for the triangle is the regular octagon whose opposite sides are identified, while the flat surface M_1 for the regular octagon is a union of 8 octagons with some gluings of the sides. Thus M_1 covers M_2 with multiplicity 8.

The next result relates the Veech groups of flat structures one of which covers the other; it was obtained independently by Gutkin and Judge [23], [24] and by Vorobets [57]. Recall that two subgroups of G are commensurate if they share a common subgroup that has a finite index in each.

Theorem 5.14 *If (X_1, ϕ_1, Σ_1) covers (X_2, ϕ_2, Σ_2) then $\Gamma(\phi_1)$ and $\Gamma(\phi_2)$ are commensurate. In particular, one of the groups $\Gamma(\phi_1)$ and $\Gamma(\phi_2)$ is a lattice if and only if the other is.*

Proof. Consider the set of triples $S = \{(X, \phi, f)\}$ where (X, ϕ) is a flat structure and $f : X_1 \to X$ is a cover. Let $\bar{S}$ be the set of equivalence classes of such triples considered up to isomorphism of covers.

Let $g \in Aff^+(\phi_1)$ and let $A = a(g)$ be its derivative. Then fg is a cover of the flat structure $(X, A^{-1}(\phi))$ by (X_1, ϕ_1). This defines a right action of $Aff^+(\phi_1)$ on S, and this action descends to an action on $\bar{S}$. Consider the subgroup $Aff_0^+(\phi_1) \subset Aff^+(\phi_1)$ that consists of the affine transformations acting trivially on $\bar{S}$. By Proposition 5.12 $\bar{S}$ is finite, therefore this subgroup has finite index. Let $\Gamma_0 \subset \Gamma(\phi_1)$ consist of the derivatives of the elements of $Aff_0^+(\phi_1)$; then Γ_0 has finite index too.

Let $g \in Aff_0^+(\phi_1)$ and $A = a(g)$. Then, for every $(X, \phi, f) \in S$ the flat structures (X, ϕ) and $(X, A^{-1}\phi)$ are isomorphich. Therefore $A \in \Gamma(\phi)$. Since (X_1, ϕ_1) covers (X_2, ϕ_2), it follows that $\Gamma_0 \subset \Gamma(\phi_2)$. Thus $\Gamma_0 \subset \Gamma(\phi_1) \cap \Gamma(\phi_2)$ and has finite index in $\Gamma(\phi_1)$. To show that Γ_0 has finite index in $\Gamma(\phi_2)$ as well one applies a similar argument to the set of covers of (X_2, ϕ_2). $\square$

Corollary 5.15 *([28]). If a rational polygon P_2 has no angles of the form π/n and P_2 tiles P_1 by reflections then the respective Veech groups are commensurable.*

Hubert and Schmidt [28] have shown that the Veech group associated to the isosceles triangle with angles $2\pi/n, (n-2)\pi/2n, (n-2)\pi/2n$ is not a lattice even though the right triangle with angle π/n tiles it with one reflection. This is due to the appearance of removable singular points that we discussed above.

6 Miscellaneous results

6.1 Stable periodic trajectories

The material in this subsection is taken from the paper [18].

Definition 6.1 *A periodic billiard trajectory in a polygon Q is called stable if an arbitrary small perturbation of the polygon Q leads to a perturbation of this trajectory but not to its destruction.*

For example, a 2-periodic trajectory in a square, perpendicular to a side, is not stable: a deformation of the square to a quadrilateral without parallel sides destroys this trajectory. The Fagnano 3-periodic trajectory in an acute triangle, connecting the foot points of the altitudes, is stable, and the same holds for the 6-periodic trajectory which is the double of the Fagnano one – see figure 1 again.

What follows is a criterium for the stability of periodic trajectories. Label the consecutive sides of the polygon Q by $1, 2, ..., k$. Without loss of generality, assume that the periodic trajectory has an even number of links $2n$ (otherwise, double it). Then the trajectory is encoded by the sequence $i_1, ..., i_{2n}$ of the labels of the consecutively visited sides of Q.

Lemma 6.2 *The trajectory is stable if and only if the labels in the sequence $i_1, ..., i_{2n}$ can be partitioned into pairs of identical labels such that the label in each pair appears once at an odd position and once at an even one.*

Proof. Denote by α_i the angle, made by i-th side of Q with a fixed direction. Unfold the periodic trajectory to a straight line l. Then the $(2n)$-th copy of Q along this line, denoted by Q_{2n}, is parallel to Q and, moreover, the parallel translation t that takes Q to Q_{2n} preserves l. The composition of two reflections is a rotation through the angle, twice that between the axes of reflection. It follows that Q_{2n} is obtained from Q by a rotation through the angle

$$2(\alpha_{i_1} - \alpha_{i_2} + \alpha_{i_3} - ... + \alpha_{i_{2n-1}} + \alpha_{i_{2n}}),$$

and this angle is a multiple of 2π.

Thus for the trajectory to be stable it is necessary that the variation of this angle satisfies the equation

$$\delta\alpha_{i_1} - \delta\alpha_{i_2} + \delta\alpha_{i_3} - ... + \delta\alpha_{i_{2n-1}} + \delta\alpha_{i_{2n}} = 0$$

for every perturbation of the polygon. An arbitrary variation of Q gives rise to an arbitrary variation of the directions of its sides. Hence the above relation holds only if its terms cancel pairwise, as claimed.

Conversely, assume that the condition of the lemma on the coding of a periodic trajectory is satisfied. Using the same notation as before, one has: $t(l) = l$. Let Q' be a polygon, sufficiently close to Q, whose sides are labeled the same way as in Q. Reflect Q' consecutively in its sides according to the sequence $i_1, ..., i_{2n}$. Then Q'_{2n} is parallel to Q'. Let t' be the respective parallel translation; this translation is close to t. Choose a point $x \in l \cap Q \cap Q'$ and let l' be the line through x, invariant under t'. If Q' is sufficiently close to Q then l' is the unfolding of a billiard trajectory in Q' labeled $i_1, ..., i_{2n}$, and since $t'(l') = l'$ this trajectory is $(2n)$-periodic. $\Box$

The lemma implies the next property of irrational billiards.

Corollary 6.3 *If the angles of a billiard k-gon are maximally independent over the rationals, that is, the dimension of the linear space over* $\mathbf{Q}$, *generated by the angles, equals $k - 1$, then every periodic billiard trajectory in this polygon is stable.*

Proof. Consider a periodic billiard trajectory, labeled $i_1, ..., i_{2n}$. Then

$$\alpha_{i_1} - \alpha_{i_2} + \alpha_{i_3} - ... + \alpha_{i_{2n-1}} + \alpha_{i_{2n}} = \pi m, \quad m \in \mathbf{Z}.$$

Rewrite this as

$$(\alpha_{i_1} - \alpha_1) - (\alpha_{i_2} - \alpha_1) + (\alpha_{i_3} - \alpha_1) - ... + (\alpha_{i_{2n-1}} - \alpha_1) + (\alpha_{i_{2n}} - \alpha_1) = \pi m.$$

If the labels do not satisfy the condition of the above lemma, that is, do not cancel pairwise, one obtains a non-trivial relation over $\mathbf{Q}$ on the angles $\alpha_2 - \alpha_1, \alpha_3 - \alpha_1, ..., \alpha_k - \alpha_1$. It remains to notice that the linear space over $\mathbf{Q}$, generated by these angles, coincides with that, generated by the angles of the billiard polygon. $\Box$

6.2 Encoding billiard trajectories. Polygonal billiards have zero entropy

The encoding of billiard trajectories by the consecutively visited sides of the billiard polygon provides a link between billiard and symbolic dynamics. Let us consider this encoding in more detail. We follow the paper [17].

Assume that the billiard k-gon Q is simply connected (most of the results hold without this assumption), and its sides are labeled $1, ..., k$. The phase space of the billiard transformation T consists of the inward unit vectors whose foot points are on the boundary ∂Q and whose forward orbits never hit a vertex of Q. Given a phase point x, denote by $w(x)$ the infinite sequence of labels $1, ..., k$ that encodes the forward T-orbit of x. Given an infinite word w, denote by $X(w)$ the set of phase points x with $w(x) = w$. Call a subset $S \subset X(w)$ a strip if all vectors in S are parallel and their foot points constitute an interval on a side of Q. An open strip is defined analogously. The billiard transformation takes a strip to a strip, preserving its width.

Start with two simple observations. First, if $w(x) = w(y)$ then the vectors x and y are parallel. If not, the unfolded trajectories of x and y linearly diverge, so a vertex of a copy of Q will fall into the angle between them. First time this happens the reflections of the two trajectories occur at different sides of the polygon, thus $w(x) \neq w(y)$.

Fig. 12

Secondly, if x and y are parallel vectors in $X(w)$ then every parallel vector, whose foot point lies between those of x and y also belongs to $X(w)$. This is true because Q is simply connected. It follows that $X(w)$ is the maximal strip corresponding to the word w. The trajectories of its boundary points come arbitrarily close to vertices of Q.

The next theorem from [17] shows that the encoding is one-to-one on the set of non-periodic billiard trajectories.

Theorem 6.4 *If w is a $(2n)$-periodic word then each vector from $X(w)$ has a $2n$-periodic trajectory and $X(w)$ is an open strip. If w is aperiodic then the set $X(w)$ consists of at most one point.*

Proof. Start with the first claim. Consider $x \in X(w)$ and unfold its trajectory to the line l. One claims that the $(2n)$-th copy of Q along this trajectory denoted, as before, by Q_{2n}, is parallel to Q. If not, x and $T^{2n}(x)$ have their foot points on the same side of Q but they are not parallel. According to the remark, preceding the theorem, $w(x) \neq w(T^{2n}(x))$ which contradicts the periodicity of w.

Let t be the parallel translation that takes Q to Q_{2n}. One wants to show that $t(l) = l$. Suppose not. Then, for a sufficiently great integer k, the polygon $t^k(Q)$ does not intersect l. This is a contradiction because

this polygon is Q_{2nk}. Finally, the strip $X(w)$ is open because a periodic trajectory stays a bounded distance from the vertices of the polygon.

Now consider the second claim. Assume, to the contrary, that the set $X(w)$ is a maximal strip of non-zero width; call this strip S. Let $x \in S$ be the vector whose base point lies exactly in the middle of S.

The proof is especially simple if Q is a rational polygon, and we start with this case. Since Q is rational, the trajectory of x has a finite set of directions. It follows that this trajectory will visit some side of Q infinitely often with some fixed direction. Call these points $x_i = T^{n_i}(x)$ and let $S_i = T^{n_i}(S)$. All these strips have the same width, therefore they must intersect. Suppose S_i and S_j intersect. Then the maximal width strips $\bar{S}_i$ and $\bar{S}_j$, containing S_i and S_j, respectively, are also parallel and intersect. These maximal strips cannot coincide, otherwise w would be periodic. Therefore a part of the boundary of $\bar{S}_i$ lies inside $\bar{S}_j$ (or vice versa). Since the trajectories of the boundary points of the maximal strip $\bar{S}_i$ come arbitrarily close to vertices of Q, there are vertices inside $\bar{S}_i$, a contradiction.

Consider a general polygon Q. One cannot claim anymore that the trajectory of x visits some side of Q infinitely often with some fixed direction. This claim will be replaced by a weaker recurrence result.

Let Y be the forward limit set of x under the mapping T. Since the width of S is positive, the trajectories of the points from Y stay bounded distance from the vertices of Q. Therefore the restriction of T to Y is continuous. Note that Y is closed and bounded, hence compact.

One uses the strengthened version of the Poincaré recurrence theorem, due to H. Furstenberg [15]. This theorem implies that there exists a uniformly recurrent point $y \in Y$. This means the following: for every neighborhood U of y there is a constant C such that the return times $n_i > 0$, defined by $T^{n_i}(y) \in U$, satisfy the inequalities $n_{i+1} - n_i < C$.

Fix a sufficiently small $\epsilon > 0$ and consider the unfolded maximal strip S' of y together with its ϵ neighborhood S'_ϵ. Since S' is maximal, some vertices of copies of Q, unfolded along this strip, will fall into S'_ϵ. *Claim*: they fall with uniformly bounded gaps into each of the two components of $S'_\epsilon - S'$.

This claim follows from the uniform recurrency of y. More specifically, let z be the leftmost point of S'. Then there is $m > 0$ such that the foot point of $T^m(z)$ lies within $\epsilon/2$ of a vertex v of Q; let m be the first such number. Let U be a neighborhood of y such that for every $u \in U$ the m-th iterate $T^m(u)$ is $\epsilon/2$ close to $T^m(y)$; such a neighborhood exists since T is continuous at y. Let n_i be the return times of y to U. Then $T^{m+n_i}(y)$ is $\epsilon/2$ close to $T^m(y)$, and therefore $T^{m+n_i}(z)$ is $\epsilon/2$ close to $T^m(z)$. It follows that the foot points of $T^{m+n_i}(z)$ lie within ϵ from the vertex v. The claim is proved.

To finish the proof of the theorem recall that y is a forward limit of x: there is a sequence $n_i \to \infty$ such that $x_i = T^{n_i}(x) \to y$. The argument, given above in the case of a rational polygon Q, shows that no two vectors x_i and x_j are parallel. Consider the intersection of the unfolded maximal strip for x_i with $S'_\epsilon - S'$. Let α_i be the angle between these strips of width ϵ and, say, δ. Then $\alpha_i \to 0$ as $i \to \infty$. The intersection is a parallelogram which, by elementary geometry, contains a rectangle of width ϵ and length $l_i = (\delta - \epsilon \cos \alpha_i)/ \sin \alpha_i$. Since the gaps between the vertices in $S'_\epsilon - S'$ are uniformly bounded and $l_i \to \infty$ as $i \to \infty$, a vertex will eventually appear inside the unfolded maximal strip for x_i. This is a contradiction. $\square$

As a consequence, the closure of every non-periodic billiard orbit contains a vertex of the billiard polygon.

An important corollary of the theorem concerns the entropy of polygonal billiards.

Corollary 6.5 *The metric entropy of the billiard mapping with respect to any invariant measure is zero.*

Proof. Morally, the argument goes as follows: the "past" is uniquely determined by the "future", therefore the entropy vanishes.

More specifically, consider the set Σ of words in $1, ..., k$ that encode the billiard trajectories for time from $-\infty$ to ∞, and let S be the shift transformation of Σ. Then the encoding map conjugates the billiard transformation and S. By the theorem, the shift S has a one-sided generator, the partition into k parts according to the value of the zero symbol in a word. This means that for this partition η the σ-algebra of measurable sets is generated by $\vee_{i=0}^{\infty} S^{-i}(\eta)$. It is a standard result in ergodic theory that the metric entropy of a transformation with a one-sided generator vanishes (see, e.g., [10]). $\square$

The variational principle implies that the topological entropy vanishes as well.

The zero entropy result was proved in [7] and [46] for the canonical billiard measure, and in [29], [17] and [22] for topological entropy. The latter work concerns a broader class of transformations, the so-called, polygon exchanges.

The zero entropy result implies that a number of quantities, associated with a polygonal billiard, grow slower than exponentially: the number of different words of length n in Σ, the number of strips of n-periodic trajectories or periodic trajectories of length not greater than L, etc. Conjecturally, all these numbers have a polynomial growth.

6.3 Complexity of billiard trajectories in rational polygons

Consider an aperiodic billiard trajectory in a rational polygon in a direction θ. According to Theorem 1.4 its code w is an aperiodic sequence. Generically this sequence enjoys a weaker periodicity property, called quasiperiodicity, described in the next result.

Proposition 6.6 *For all but countably many directions θ each finite segment of w appears in w infinitely many times.*

Proof. Let $i_1, ..., i_n$ be a segment of w, and let x be a phase point on the trajectory under consideration whose first n successive reflections occur in the sides labeled $i_1, ..., i_n$. There is a neighborhood U of x such that for every phase point $y \in U$ the first n reflections occur in the same sides of the polygon. By Theorem 1.8, for all but countably many directions θ, the trajectory is dense on the invariant surface. It follows that x returns to U infinitely many times, and each time the segment $i_1, ..., i_n$ reappears in w. $\square$

A convenient way to measure the complexity of aperiodic trajectories and their codes is provided by the next definition.

Definition 6.7 *The complexity function $p(n)$ of an infinite sequence w is the number of distinct n-element segments of w.*

We start with the complexity of billiard trajectories in a square; the results are due to Hedlund and Morse [26]. We slightly modify the encoding using only two symbols, say, 0 and 1, to indicate that a trajectory reflects in a horizontal or a vertical side, respectively.

Theorem 6.8 *For every aperiodic trajectory one has: $p(n) = n + 1$.*

The sequences with complexity $p(n) = n + 1$ are called Sturmian.

Proof. Start with the following observation. Every aperiodic trajectory is dense in the respective invariant torus. Therefore $p(n)$ can be computed as the number of different initial segments of length n in the codes $w(x)$ where x ranges over the phase vectors having a fixed direction.

Unfold the billiard trajectory to a line l. Partition the square grid in the plane into "ladders", going in the south-east – north-west direction, as shown in figure 11. The n-th symbol in the code of the trajectory is 0 or 1, according as l meets a horizontal or a vertical segment of n-th ladder.

Fig. 13

Let (e_1, e_2) be the orthonormal frame. Consider the linear projection of the plane onto the diagonal $x + y = 0$ whose kernel is parallel to the line l. Factorize the diagonal by the translation through $e_1 - e_2$ and identify the quotient space with the unit circle $\mathbf{S}^1$. The vertices of the first ladder are the lattice points (a, b) with $a+b = 1$ or $a + b = 2$. Since l has an irrational slope the projections of the vertices of the first ladder partition the circle into two irrational arcs. Let T be the rotation of $\mathbf{S}^1$ through the length of an arc, that is, through the projection of e_1.

The number of different initial n-segments corresponding to the lines, parallel to l, equals the number of segments into which the projections of the vertices of the first n ladders partition $\mathbf{S}^1$. Each ladder is obtained from the first one by the translation through e_1. It follows that $p(n)$ equals the cardinality of the orbit $T^i(0),\ i = 0, ..., n$. Since T is an irrational rotation of the unit circle all points of this orbit are distinct, and $p(n) = n + 1$. □

There are generalizations of the above theorem to multi-dimensional cubes – see [2] and [4].

Next we consider the complexity of billiard trajectories in convex rational polygons. The following result is due to P. Hubert ([27]).

Let Q be a convex rational k-gon with the angles $\pi m_i/n_i,\ i = 1, ..., k$ (m_i and n_i are coprime), and let N be the least common multiple of n_i's. Let θ be a direction such that there are no generalized diagonals in this direction. Consider a billiard trajectory in direction θ which avoids the vertices. Let w be its coding in the alphabet $\{1, ..., k\}$, and denote by $p(n)$ the complexity of w.

Theorem 6.9 *For all sufficiently great n one has: $p(n) = n(k-2)N + 2N$.*

In particular, if Q is a square, one obtains: $p(n) = 4(n+1)$. This does not contradict the previous theorem because the encoding there was different and, actually, 4 times less precise: the pair of parallel sides were labeled identically.

Proof. Consider the invariant surface M of the billiard flow. Recall that M is the result of pairwise identifying the sides of $2N$ copies of Q. Thus M has the structure of a complex; denote by v, e and f the number of vertices, edges and faces. Note that $e = Nk$ and $f = 2N$. The edges can be labeled by pairs $(i, j),\ i = 1, ..., k,\ j = 1, ..., N$ so that the (i, j)-th edge corresponds to the i-th side of Q. Let $E \subset M$ be the union of the edges.

The billiard trajectory under consideration can be encoded in an obvious way in the alphabet $\{(i, j)\}$; denote its code by W. The projection $(i, j) \to i$ sends W to w (we will always denote by W words in the alphabet $\{(i, j)\}$ and by w their projections). In other words, the new encoding takes into account not only the side of Q in which a reflection occurs but also the angle of reflection that can take finitely many values.

Let $P(n)$ be the complexity of W.

Claim: $P(n) = n(k-2)N + 2N$ for all $n \geq 1$.

Arguing as in the proof of the preceding theorem, $P(n)$ equals the number of distinct initial segments of length n in the codes $W(x)$ where x ranges over points of E.

Next, given a word W_n of length n, consider the set $X(W_n)$ consisting of the points $x \in E$ such that the initial segment of $W(x)$ is W_n. Then, arguing as in the previous section, $X(W_n)$ is an interval on an edge from E.

Denote by T the first return map of the directional flow F_θ to E (that is, the billiard map). Consider the set of initial segments of length $(n+1)$ in $W(x),\ x \in X(W_n)$. The word W_n will have different successors if and only if $T^n(X(W_n))$ contains a vertex V of M. We say that W_n splits at V.

To find $P(n+1) - P(n)$ one needs to learn how many words of length n split at each vertex. Let V be a vertex with cone angle $2\pi c$. Then there are c incoming saddle connections. Trace each one back n steps, and let $W^1, ..., W^c$ be the respective n-length words. Then $T^n(X(W^\alpha))$ contains V for each $\alpha = 1, ..., c$.

The words W^α are all different since their $(n-1)$-st letters are distinct. The latter hold because in the polygon Q there is only one trajectory that starts at a given side in a given direction and goes straight to a given vertex.

Thus all the n-length words that split at V are $W^1, ..., W^c$. It follows that

$$P(n+1) - P(n) = \sum c(V),$$

sum over all vertices. To evaluate this sum note that the Euler characteristic of M equals the sum of indices of singular points of the flow F_θ; the index at vertex V equals $1 - c(V)$. Hence

$$\chi(M) = \sum (1 - c(V)) = v - \sum c(V).$$

On the other hand, $\chi(M) = v - e + f$, and it follows that $\sum c(V) = (k-2)N$. Therefore $P(n+1) - P(n) = (k-2)N$, and since $P(1) = kN$, the claim follows.

To finish the proof of the theorem it remains to show that $P(n) = p(n)$ for all sufficiently great n. Suppose not; then there exist arbitrary long words $W \neq W'$ that are the initial segments of the codes of phase points $x \in E$ such that $w = w'$.

The billiard trajectories, corresponding to W and W', start at the same side of Q but make different angles with it: otherwise, since $w = w'$, they would reflect at the same sides of Q and meet them at the same angles which would imply that $W = W'$. Note that the angle between the two trajectories is bounded below by a constant α depending on Q and θ.

Unfold the trajectories to straight lines. Since these lines linearly diverge, there is a constant m, depending on Q and α, such that after at most m reflections of Q along either of the unfolded trajectories a vertex of a copy of Q will fall into the angle between the lines. Therefore if the length of W and W' is greater than m the codes w and w' are distinct, a contradiction. $\Box$

If Q is a general k-gon let N be the least common denominator of its π-rational angles and s be the number of its distinct π-irrational angles. Then one has the next upper bound for the complexity $p_\theta(n)$ of the billiard trajectories in Q in a fixed initial direction θ – see [25].

Theorem 6.10 *For all n one has:*

$$p_\theta(n) \leq kNn(1 + \frac{n}{2})^s.$$

Using techniques, similar to the ones in the proof of Theorem 6.9, S. Troubetzkoy obtained in [50] a complexity lower bound for arbitrary polygonal billiards. Unlike the previous results, his theorem concerns billiard trajectories in all directions, so that $p(n)$ now means the number of distinct words of length n in the coding in the alphabet $\{1, ..., k\}$ of all billiard orbits in a k-gon.

Theorem 6.11 *For every polygon there exists a constant c such that $p(n) \geq cn^2$ for all $n \geq 0$.*

A similar estimate is proved in [50] for d-dimensional polyhedra with the exponent 2 replaced by d.

6.4 Periodic trajectories in some irrational billiards

The following elementary argument shows that every rational polygon has a periodic billiard trajectory of a special kind; it was found independently by A. Stepin ([18]) and by M. Boshernitzan ([8]).

Shoot the billiard ball in the direction, perpendicular to a side of the polygon Q. By Poincaré's recurrence theorem, for almost every initial position, the ball will return to the original side at an angle, arbitrarily close to $\pi/2$. Since the set of possible directions of the ball is discrete in a rational polygon, this angle will be equal to $\pi/2$. After the ball bounces off of the side it will backtrack the same trajectory, so the trajectory is periodic.

Surprisingly, a variation of this argument applies to some irrational polygons. This was observed by B. Cipra, R. Hanson and A. Kolan ([11]), and then generalized by E. Gutkin and S. Troubetzkoy ([25]). The next result is due to the former authors.

Theorem 6.12 *Given a right triangle with π-irrational acute angles, almost every (in the sense of measure) billiard trajectory, that starts at a side of the right angle in the direction, perpendicular to this side, returns to the same side in the same direction.*

Proof. The idea is to apply the construction which gave the invariant surfaces of the billiard flow in rational polygons. A significant difference with the rational case is that the invariant surface will not be compact.

Start with reflecting the triangle in the sides of the right angle to obtain a rhombus R. The study of the billiard inside the triangle reduces to that inside the rhombus. Let α be the acute angle of R.

Fig. 14

Consider the beam of horizontal trajectories, starting at the upper half of the vertical diagonal of the rhombus, and construct the invariant surface of the phase space, corresponding to this beam. This surface consists of rhombi, obtained from R by the action of the group $A(R)$, whose sides are pasted pairwise in an appropriate manner. Since α is π-irrational, this surface is not compact. The invariant surface is partially foliated by the parallel trajectories from the horizontal beam, and this foliation has an invariant transversal measure, "the width of a beam".

Each rhombus involved is obtained from the original one by a rotation through the angle $n\alpha$ where $n \in \mathbf{Z}$. Such a rhombus will be referred to as the n-th rhombus and denoted by R_n; in particular, $R = R_0$. A trajectory from the beam under consideration may leave the n-th rhombus through a side which has one of the two possible directions. Call such a side positive if the trajectory enters an $(n+1)$-st rhombus and negative if it enters an $(n-1)$-st one.

One wants to show that almost every trajectory returns to R_0 (where they hit the vertical diagonal in the perpendicular direction). We will prove that for every $\epsilon > 0$ the relative measure of the set of trajectories that do not return to R_0 is less than ϵ.

Let ϵ be given. Since α is π-irrational, there exists $n > 0$ such that the vertical projection of the positive side of R_n is less than ϵ. This implies that the relative measure of the set of trajectories that make it to $(n+1)$-st rhombi is less than ϵ.

The rest of trajectories are bound to stay in $R_0, R_1, ..., R_n$; call this set of trajectories S. The union of the rhombi from 0 through n is compact, therefore the Poincaré recurrence theorem applies to S. It follows that almost every trajectory in S is recurrent, that is, returns to R_0. $\square$

In fact, the above proof gives more: for every direction in a rhombus almost every billiard trajectory in this direction returns arbitrarily close to itself and in the same direction. A relevant definition is proposed in [25].

Definition 6.13 *Given a billiard polygon Q, a direction is called recurrent if almost every billiard trajectory in this direction returns arbitrarily close to itself and in the same direction. A polygon Q is called strongly recurrent if every direction is recurrent.*

In particular, every rational polygon is strongly recurrent. An argument, similar to the preceding proof, gives the next result from [25], applicable, in particular, to all parallelograms.

Theorem 6.14 *Let Q be a polygon whose sides have one of the two fixed directions. Then Q is strongly recurrent. In particular, the billiard orbits, perpendicular to each side of Q, are periodic with probability one.*

6.5 A non-periodic trajectory that is not dense in the configuration space

An example of such a trajectory was constructed by G. Galperin in [16].

Fig. 15

Consider a centrally symmetric hexagon (so that its opposite sides are parallel and congruent) such that the bisector of the angle $BAC = \alpha$ is perpendicular to the side AF and the the bisector of the angle $ABF = \beta$ is perpendicular to the side BC.

Since the bisector of α is perpendicular to BC, the reflection of AB in BC is the line BF. Thus the reflection of any vertical segment leading to BC is parallel to BF. Similarly, the reflection of vertical segments in CD are parallel to AC. After another reflection the segment becomes vertical again. Choose a line, perpendicular to the side AB; let X, Y, Z be the projections on this line of the points A, E, F, respectively. The set of vertical lines is identified with a horizontal segment XY, and the second iteration of the billiard transformation induces the exchange of the intervals XZ and ZY. An exchange of two intervals is equivalent to a circle rotation. For generic angles α and β each orbit of this circle rotation is dense. Therefore each vertical billiard trajectory is dense in the hexagon.

Extend the non-vertical sides of the hexagon to obtain a parallelogram. Then the two triangles, added to the hexagon, are never visited by any vertical trajectory, that starts inside the original hexagon.

Similar but more tedious considerations prove the following result by Galperin.

Theorem 6.15 *There exists an acute angle α_0 such that for almost every (in the sense of measure) $\alpha \in (0, \alpha_0)$ a right triangle with the acute angle α contains a non-periodic and not everywhere dense billiard trajectory.*

References

[1] P. Arnoux. Ergodicité Générique des Billards Polygonaux. Sém. Bourbaki, No 696, (1987-88).

[2] P. Arnoux, C. Mauduit, I. Shiokawa, J.-I. Tamura. Complexity of sequences defined by billiards in the cube. Bull. SMF, 122, F.1, (1994), 1-12.

[3] E. Aurell, C. Itzykson. Rational billiards and algebraic curves. J. Geom. and Phys., 5, (1988), 191-208.

[4] Yu. Baryshnikov. Complexity of trajectories in rectangular billiards. Comm. Math. Phys., 174, (1995), 43-56.

[5] A. Beardon. The geometry of discrete groups. Springer-Verlag, 1983.

[6] L. Bers. Finite dimensional Teichmuller spaces and generalizations, Bull. AMS, 5, (1981), 131-172.

[7] C. Boldrighini, M. Keane, F. Marchetti. Billiards in polygons. Ann. of Prob., 6, (1978), 532-540.

[8] M. Boshernitzan. Billiards and rational periodic directions in polygons. Amer. Math. Monthly, 99, (1992), 522-529.

[9] M. Boshernitzan, G. Galperin, T. Kruger, S. Troubetzkoy. Periodic billiard orbits are dense in rational polygons. Trans. AMS, 350 (1998), 3523-3535.

[10] I. Cornfeld, S. Fomin, Ya. Sinai. Ergodic theory. Springer-Verlag, 1982.

[11] B. Cipra, R. Hanson, A. Kolan. Periodic trajectories in right triangle billiards. Phys. Rev., E 52, (1995), 2066-2071.

[12] C. Earle, F. Gardiner. Teichmuller disks and Veech's $\mathcal{F}$-structures. Contemp. Math., 201, AMS, Providence, 1997, 165-189.

[13] A. Eskin, H.Masur. Pointwise asymptotic formulas on flat surfaces. Preprint

[14] R. Fox, R. Kershner. Geodesics on a rational polyhedron. Duke Math. J., 2, (1936), 147-150.

[15] H. Furstenberg. Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton Univ. Press, 1981.

[16] G. Galperin. Nonperiodic and not everywhere dense billiard trajectories in convex polygons and polyhedrons. Comm. Math. Phys., 91, (1983), 187-211.

[17] G. Galperin, T. Kruger, S. Troubetzkoy. Local instability of orbits in polygonal and polyhedral billiards. Comm. Math. Phys., 169, (1995), 463-473.

[18] G. Galperin, A. Stepin, Ya. Vorobets. Periodic billiard trajectories in polygons: generating mechanisms. Russ. Math. Surv., 47, No 3, (1992), 5-80.

[19] E. Gutkin. Billiards in polygons. Physica D, 19, (1986), 311-333.

[20] E. Gutkin. Billiards in polygons: survey of recent results. J. Stat. Phys., 83, (1996), 7-26.

[21] E. Gutkin. Billiards on almost integrable polyhedral surfaces. Ergod. Th. and Dyn. Syst., 4, (1984), 569-584.

[22] E. Gutkin, N. Haydn. Topological entropy of generalized polygon exchanges. Bull. AMS, 32, (1995), 50-57.

[23] E.Gutkin, C.Judge. The geometry and arithemtic of translation surfaces with applications to polygonal billiards. Math. Res. Lett., 3, (1996), 391-403.

[24] E.Gutkin, C.Judge. Affine mappings of translation surfaces: geometry and arithmetic. Preprint

[25] E. Gutkin, S. Troubetzkoy. Directional flows and strong recurrence for polygonal billiards. in Pitman Res. Notes Math. Ser., 362, 1996.

[26] G. Hedlund, M. Morse. Symbolic dynamics 2. Sturmian trajectories. Amer. J. of Math., 62, (1940), 1-42.

[27] P. Hubert. Complexité de suites définies par des billards rationnels. Bull. Soc. Math. France, 123, (1995), 257-270.

[28] P.Hubert, T.Schmidt, Veech groups and polygonal coverings. Preprint

[29] A. Katok. The growth rate for the number of singular and periodic orbits for a polygonal billiard. Comm. Math. Phys., 111, (1987), 151-160.

[30] A. Katok, B. Hasselblatt. Introduction to the modern theory of dynamical systems. Cambridge Univ. Press, 1995.

[31] A. Katok, A. Zemlyakov. Topological transitivity of billiards in polygons. Math. Notes, 18, (1975), 760-764.

[32] M. Kean. Coding problems in ergodic theory. Proc. Int. School of Math. Phys., Univ. Camerino, 1974.

[33] R. Kenyon, J. Smillie. Billiards in rational-angled triangles. Preprint

[34] S. Kerckhoff, H. Masur, J. Smillie. Ergodicity of Billiard Flows and Quadratic Differentials. Ann. of Math., 124, (1986), 293-311.

[35] M. Kontsevich, A. Zorich. Connected components of the moduli spaces of holomorphic differentials with prescribed singularities. Preprint

[36] I.Kra. On the Nielsen-Thurston-Bers type of some self-maps of Riemann surfaces. Acta. Math., 146, (1981), 231-270.

[37] H.Masur. Interval exchange transformations and measured foliations. Ann. of Math., 115, (1982), 169-200.

[38] H. Masur, Closed trajectories for quadratic differentiasl with an application to billiards. Duke Math. J., 53, (1986), 307-314.

[39] H. Masur. Lower bounds for the number of saddle connections and closed trajectories of a quadratic differential. In Holomorphic Functions and Moduli, Vol.1, D. Drasin ed., Springer-Verlag, 1988.

[40] H.Masur. The growth rate of trajectories of a quadratic differential. Erg. Th. Dyn. Syst., 10, (1990), 151-176.

[41] H. Masur. Hausdorff dimension of the set of nonergodic foliations of a quadratic differential. Duke Math. J., 66, (1992), 387-442.

[42] H. Masur, J.Smillie. Hausdorff dimension of sets of nonergodic foliations, Ann. of Math., 134, (1991), 455-543

[43] H. Masur, J. Smillie. Quadratic differential with prescribed singularities and pseudo-Anosov diffeomorphisms. Comm.. Math. Helv., 68, (1993), 289-307.

[44] J.C.Oxtoby, S.M.Ulam. Measure preserving homeomorphisms and metrical transitivity, Ann. of Math., 42, (1941), 874-920.

[45] R. Richens, M. Berry. Pseudointegrable systems in classical and quantum mechanics. Physica D, 2, (1981), 495-512.

[46] Ya. Sinai. Introduction to ergodic theory. Princeton Univ. Press, 1976.

[47] K. Strebel. Quadratic differentials. Springer-Verlag, 1984.

[48] S. Tabachnikov. Billiards. SMF Panoramas et Synthèses, 1995, No 1.

[49] W.P. Thurston. On geometry and dynamics of diffeomorphisms of surface. Bull Amer. Math. Soc., 18, (1988), 417-431.

[50] S. Troubetzkoy. Complexity lower bounds for polygonal billiards. Chaos, 8, (1998), 242-244.

[51] W.Veech. Strict ergodicity in zero dimensional dynamical systems and the Kronecker-Weyl theorem mod 2. Trans. AMS, 140, (1969), 1-34.

[52] W.Veech. Teichmuller geodesic flow. Ann. of Math., 124, (1986), 441-530.

[53] W. Veech. Moduli spaces of quadratic differentials. J. D'Analyse Math., 55, (1990), 117-171.

[54] W.Veech. Teichmuller curves in moduli space. Eisenstein series and and an application to triangular billiards. Invent. Math., 97, (1990), 117-171.

[55] W.Veech The billiard in a regular polygon, Geom.Funct.Anal. 2 (1992) 341-379

[56] Ya. Vorobets. Ergodicity of billiards in polygons. Mat. Sb., 188, (1997), No 3, 65-112.

[57] Ya. Vorobets. Plane structures and billiards in rational polygons: the Veech alternative. Russ. Math. Surv., 51, (1996), No 5, 3-42.

[58] P. Walters. An Introduction to Ergodic Theory. Springer-Verlag, 1982.

[59] C. Ward. Calculation of Fuchsian groups associated to billiards in a rational triangle. Ergod. Th. and Dyn. Syst., 18, (1998), 1019-1042.