Annals of Mathematics, 124 (1986), 293–311

Ergodicity of Billiard Flows and Quadratic Differentials

By STEVEN KERCKHOFF, HOWARD MASUR and JOHN SMILLIE[1]

Consider the following simple mechanical system. Two objects with masses m_1 and m_2 are constrained to move along a straight, frictionless track. The objects collide elastically with each other and with barriers at either end of the track. Certain quantities of physical interest are defined as time averages along the trajectories of this system. These quantities are difficult to compute directly. If such a system is *ergodic* then, with probability one, these time averages are equal to integrals over phase space which are easy to calculate.

The question of the ergodicity of this system has been considered by a number of mathematicians. It is raised, for example, by Sinai ([S], p. 85). We will show in this paper that for a dense set of pairs (m_1, m_2) this system is ergodic. These are the first such examples for which ergodicity has been established.

The motion of two masses on an interval is equivalent to the motion of a single particle on a right triangular region of the plane where the particle obeys the laws of motion of a billiard ball. That is to say that the particle moves with constant velocity in the interior of the table and reflects off the boundary of the table so that the angle of incidence is equal to the angle of reflection. Such billiard flows are closely related to geodesic flows.

Let Q be a planar polygon. One can define a geodesic flow f_t on the unit tangent bundle $U(Q)$ so that orbits of this flow project to billiard ball paths on Q. The polygon Q is said to be *rational* if all of the angles of Q are rational multiples of π. When Q is rational the tangent vectors to a given orbit are parallel to a finite set of unit vectors. The orbits with initial direction θ lie in an invariant surface M_θ which consists of a finite number of copies of Q, one for each potential direction of an orbit with initial direction θ (cf. [F-K]). The dynamical analysis of f_t breaks up into an analysis of the flows $f_t|M_\theta$ as θ varies.

[1] This work was supported by the National Science Foundation, the Mathematical Sciences Research Institute and the Institute for Advanced Study.

A flow is *uniquely ergodic* if there is precisely one invariant probability measure. We will prove:

THEOREM 1. *For almost every θ the flow $f_t|M_\theta$ is uniquely ergodic.*

A flow is *ergodic* with respect to a probability measure if every invariant set has measure zero or one. It is easy to see that if a flow has only one invariant measure then the flow must be ergodic with respect to that measure. The surfaces M_θ have natural measures coming from Lebesgue measure on Q. These measures are invariant. As a corollary to the theorem we have: For almost every θ the flow $f_t|M_\theta$ is ergodic with respect to the natural measure on M_θ.

Theorem 1 has consequences for billiard tables which do not have rational angles. The set of all polygons with a given number of sides forms an open subset of a finite dimensional vector space. We would like to thank A. Katok and M. Boshernitzan for independently pointing out to us the following corollary of Theorem 1.

COROLLARY 1. *There is a dense G_δ in the space of polygons consisting of polygons for which the billiard flow, f_t, is ergodic.*

Theorem 1 follows from a result that we prove about Riemann surfaces and quadratic differentials. A quadratic differential q determines a vertical foliation defined by $\operatorname{Re} q^{1/2}\, dz = 0$. This foliation admits a transverse invariant measure. If it admits precisely one such measure up to scalar multiplication we say that it is uniquely ergodic.

THEOREM 2. *Given a compact Riemann surface M and a holomorphic quadratic differential q, then for almost all θ the vertical foliation of $e^{i\theta}q$ is uniquely ergodic.*

The results of [Z-K] and [B-K-M] show that for a typical direction θ the flow is minimal, i.e. all orbits are dense. It is a notorious fact however that minimality does not imply unique ergodicity for quadratic differentials or, equivalently, for interval exchange transformations. If Q is a rectangle or, more generally, if reflections through the sides of Q generate a tesselation of the plane, then theorem 1 is a consequence of Weyl's analysis of toral flows, as is pointed out in [F-K]. If the affine group generated by reflections in the sides of Q acts discretely on the plane, then Theorem 1 follows from results in [B] and [G]. In these special cases minimality does imply unique ergodicity.

Masur ([M]) proved that the set N of non-uniquely ergodic quadratic differentials has measure zero with respect to smooth measures on the space of quadratic differentials. (Veech in [V1] proved a related result for interval exchanges.) Theorem 2 shows that the intersection of N with each circle $\{e^{i\theta}q: 0 \leq \theta < 2\pi\}$ has measure zero. The Veech-Masur results follow from our result by application of the Fubini theorem. The Fubini theorem and our results also give the following new result:

COROLLARY 2. *On a given compact Riemann surface almost every holomorphic* 1-*form has a uniquely ergodic vertical foliation.*

Theorem 1 follows from Theorem 2. This reduction is discussed in Section 1. Theorem 2 is a consequence of Theorem 3 which is proved in Section 2 and Theorem 4 which is proved in Section 4. Corollary 1 is a consequence of the more explicit Proposition 4 which is proved in Section 5.

1. Preliminaries on quadratic differentials and billiard tables

We adopt a geometric approach to quadratic differentials. We define a quadratic differential on a surface to be a structure determined by an atlas of charts mapping open subsets of M to $\mathbf{R}^2$. We require that the change-of-coordinate functions be restrictions of maps of the form $v \to (v + c)$ or $v \to (-v + c)$. We also allow a finite number of singular points at which the charts are n-fold branched coverings of $\mathbf{R}^2/(\pm 1)$. Two quadratic differentials are isomorphic if there is a homeomorphism between them which takes singular points to singular points and which at other points takes the same local form as the change-of-coordinate functions.

Fix a complex structure on $\mathbf{R}^2$. Let M be a surface of genus g. Each quadratic differential on M gives rise to a complex structure on M minus the set of singularities of M. This complex structure extends to all of M. The space of all isomorphism classes of complex structures on M is the moduli space R_g of Riemann surfaces of genus g. If we consider complex structures with respect to the equivalence relation of being isomorphic by means of a homeomorphism homotopic to the identity then the resulting space of equivalence classes is the Teichmüller space T_g. Similarly we can consider the space of quadratic differentials with respect to the equivalence relation of being isomorphic by means of a homeomorphism homotopic to the identity. We denote the resulting space by Q_g which can be identified with the cotangent bundle of T_g and given the bundle topology. The mapping class group Mod(g) acts on T_g and on Q_g. The quotient space of T_g by Mod(g) is just R_g. Let Q_g^0 denote the space of quadratic differentials with area one. Let $QD = QD(g)$ denote the quotient of Q_g^0 by Mod(g).

A quadratic differential defines a metric on M. Outside of the singular set this metric is a Riemannian metric with zero curvature. We say that a metric has a cone type singularity at a point p if there is a neighborhood of p which is isometric to a neighborhood of the origin in $\mathbf{R}^2$ with the metric $ds^2 = dr^2 + (cr\,d\theta)^2$. In this case we say that the cone angle at p is $2\pi c$. The metric defined on M by the quadratic differential has cone type singularities at the singularities of the quadratic differential. Each cone angle is a multiple of π. We can speak of parallel translation with respect to the connection determined by this metric. The global obstruction to parallel translation being well-defined is the holonomy group which is a subgroup of $O(2)$. If a metric of zero curvature arises from a quadratic differential, then the holonomy group is either the trivial group or $\{I, -I\}$. One says that the quadratic differential is orientable in the first case and nonorientable in the second case. Note that the underlying surface is always orientable. Conversely if a surface of zero curvature has cone type singularities and the holonomy group is one of these two groups, then the metric comes from a quadratic differential.

A geodesic is a curve with the property that the tangent lines at any two points of the curve are parallel. We will have occasion to consider geodesics which pass through singularities. Such curves can also be thought of as unions of geodesic segments with singularities as endpoints.

A quadratic differential defines a pair of foliations on M. The horizontal foliation is the foliation induced on M by the foliation of $\mathbf{R}^2$ by horizontal lines. The vertical foliation is induced by the foliation of $\mathbf{R}^2$ by vertical lines. Both of these foliations have transverse measures determined by the quadratic differential. If the quadratic differential is orientable then we can define horizontal and vertical flows.

If a geodesic path is contained in a coordinate chart then its image in $\mathbf{R}^2$ is a line segment. We define the horizontal and vertical components of the curve to be the horizontal and vertical components of its image. If a geodesic path or curve is not contained in a single chart then we define its horizontal and vertical components by dividing it into short pieces and summing the components of the pieces. Two isotopic geodesic curves have the same horizontal and vertical components.

Let $\{\phi_i\}$ be an atlas and let $a \in \mathrm{SL}(2, \mathbf{R})$. We can define a new atlas $\{a\phi_i\}$. In this way we define an action of $\mathrm{SL}(2, \mathbf{R})$ on the set of quadratic differentials. Note that if $a = -I$ then the two atlases give isomorphic structures. Thus this action descends to a well-defined action of $\mathrm{PSL}(2, \mathbf{R}) = \mathrm{SL}(2, \mathbf{R})/\{I, -I\}$. The notion of parallel translation remains unchanged when the coordinate charts are changed as above. In particular the set of geodesic curves remains unchanged. The components of a geodesic do change however.

The following one parameter subgroups of PSL(2, $\mathbf{R}$) will be important in what follows.

$$g_t = \begin{bmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{bmatrix},$$

$$r_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix},$$

$$h_s = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}.$$

We refer to g as the geodesic or Teichmüller flow. This can be thought of as the geodesic flow on the moduli space. The action of r_θ on a quadratic differential q is the same as multiplying q by $e^{2i\theta}$. Since $r_{\theta+\pi} = -r_\theta$, multiplication by r_θ defines an action of the circle $\mathbf{R}/\pi\mathbf{Z}$ on the set of quadratic differentials. Note that the action of r_θ leaves the "flat metric" invariant. Any two quadratic differentials which determine the same metric and orientation on M differ by the action of this subgroup. We refer to h as the horocycle flow.

The groups g and r act continuously on QD. Since PSL(2, $\mathbf{R}$) is generated by these subgroups it acts continuously on QD.

The dynamics of these flows have been considered in [M] and [V2]. The orbit of a quadratic differential under PSL(2, $\mathbf{R}$) is called a Teichmüller disc and has been considered by many authors (cf. [K]).

Action of PSL(2, $\mathbf{R}$) on geodesic segments: Let A be a matrix representing an element of PSL(2, $\mathbf{R}$). Let β be a geodesic segment. Let h_1 and v_1 be the horizontal and vertical components of β with respect to q. Let h_2 and v_2 be the coordinates of β with resect to Aq. Then:

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

Note that in this equation the matrix and two vectors are well defined modulo multiplication by -1.

Reduction of billiard problem to quadratic differential problem. We will consider only billiard tables which are subsets of the plane though this discussion can be carried out for more general "tables" (cf. [G]). Let Q be a polygonal region of the plane. We do not require that Q be convex or that the boundary of Q be connected. Let V be the set of vertices of Q. Let $O(2)$ denote the group of linear isometries of the plane. For each side, s_i, of Q let $\rho_i \in O(2)$ be the reflection through a line parallel to s_i and passing through the origin. We construct the phase space for the billiard flow as follows. Let P be the product space $(Q - V) \times S^1$ with the following identifications. For $x \in s_i$ we identify (x, v) with $(x, \rho(v))$. After we make these identifications, the flow f_t is continuous although the continuation of an orbit through a vertex is not defined.

Definition. Q is a rational polygon if the subgroup of $O(2)$ generated by reflections in the sides of Q is finite.

Note that if the boundary of Q is connected then Q is rational if and only if every vertex angle is a rational multiple of 2π. If the boundary is not connected then the condition on the angles is necessary but not sufficient for the rationality of Q. For the remainder of the section assume that Q is rational and let $\Gamma \subset O(2)$ be the group generated by the reflections in the sides.

In the rational case we can construct an "integral of motion". Note that Γ is a dihedral group. Choose an interval J contained in S^1 which is a fundamental domain for the action of Γ. We can also think of J as the quotient of S^1 by the action of Γ. We can define a function from $Q \times S^1$ to J by composing the projection onto S^1 with the quotient map from S^1 to J. This composition is compatible with the identifications defining P and therefore induces a function π from P to J. Also π is constant on orbits of f_t.

Let θ be an interior point of the interval J. There is a surface $\pi^{-1}(\theta)$ with punctures corresponding to vertices of Q. Denote this surface by M_θ. Since each such surface is invariant it makes sense to analyze the flow restricted to any such surface. Each surface inherits from Q a flat metric with cone type singularities. The flow $f_t|M_\theta$ is generated by a vector field on M_θ which is parallel with respect to the flat structure on M_θ. With respect to their flat metrics the surfaces M_θ are all isometric. We will describe a natural identification of any two of them. Let θ and λ be points in the interior of the interval J. If (x, v) is in M_θ then we identify it with (x, v'), where v' corresponds to λ under the quotient identification and v' belongs to the same fundamental domain as v under the action of Γ on S^1. With this identification M_θ, when thought of as a Euclidean manifold, is independent of θ. We will denote it by M. We can choose a direction on M and thus a quadratic differential q so that each of the flows $f_t|M_\theta$ is equivalent to one of the flows generated by some $r_\theta q$ for θ in J. We see from this that Theorem 1 follows from Theorem 2.

Section 2

Definition. A quadratic differential q is *divergent* if $g_t(q)$ eventually leaves all compact sets as t increases to infinity. A quadratic differential is *recurrent* if it is not divergent. Note that q is recurrent if and only if the ω-limit set of q is non-empty.

The proof of Theorem 2 divides into two cases, depending on whether $r_\theta(q)$ is recurrent or divergent. The divergent case is treated in Section 4. In this section we consider the case of recurrent $r_\theta(q)$ and prove the following:

THEOREM 3. *The set of θ for which $r_\theta(q)$ is recurrent and not uniquely ergodic has measure zero.*

The following formulas will be used in the proof of this theorem.

$$g_t h_{s \cdot \exp(-t)} g_{-t} = h_s, \tag{1}$$

$$\lim_{t \to \infty} g_t r_{s \cdot \exp(-t)} g_{-t} = h_s. \tag{2}$$

Formula (1) expresses the fact that the horocycle orbits are uniformly expanded by the Teichmüller flow. Formula (2) shows that the horocycle flow is the limit of circle actions. Both formulas follow from a straightforward calculation in $\mathrm{PSL}(2, \mathbf{R})$.

Proof of theorem. According to [M] there is a subset B of quadratic differential space such that if q is not uniquely ergodic then every ω-limit point of the Teichmüller orbit of q is contained in B. For our purposes it suffices to note two properties of B. First, B is a closed set and second, B consists of quadratic differentials each of which contains a closed vertical geodesic (which may contain singularities). The second property implies that the intersection of B with any horocycle orbit is countable. To see this let q be a quadratic differential. Let α be a closed geodesic with components h and v. The horizontal component of α with respect to $h_s(q)$ is $h + sv$. Also α is vertical when $h + sv = 0$. This occurs for at most one value of s. If β is isotopic to α then it will have the same components and be vertical for the same s. The number of isotopy classes of geodesics is countable; thus the number of values s for which $h_s(q)$ can have a vertical geodesic is countable.

Assume that the theorem is false. Let U be a positive measure set of θ so that for θ in U, $r_\theta(q)$ is not uniquely ergodic and the orbit of $r_\theta(q)$ under the Teichmüller flow, g_t, has non-trivial ω-limit set. Partition quadratic differential space into a countable collection of compact sets. For each of these sets K consider the set of directions θ for which the ω-limit set of $r_\theta(q)$ intersects K. Now U is the union of this countable collection of sets; thus at least one of these sets must have positive measure. Call this set V and the corresponding compact set K.

Let B_ε be the closed ε neighborhood of B.

LEMMA. *For every δ there is an ε such that $\mu\{s \in [-1, 1]: h_s(k) \in B_\varepsilon\} < \delta$ for each k in K.*

Proof. Let $f_n(k) = \mu\{s \in [-1, 1]: h_s(k) \in B_{1/n}\}$. The lemma is equivalent to the assertion that the supremum of f_n goes to 0 as n increases. This is a consequence of the following two properties of a sequence of functions defined on a compact set: f_n is a non-increasing sequence of functions with pointwise limit 0. Each function f_n is upper semicontinuous. We will verify these two conditions but leave the proof of their sufficiency to the reader.

We begin by showing that f_n is an upper semicontinuous function. That is to say that, for every r, the set $\{k: f_n(k) < r\}$ is open. Fix a k for which $f_n(k)$ is less than r. Let $S = [-1, 1] - \{s \in [-1, 1]: h_s(k) \in B_{1/n}\}$. By construction, the measure of S is greater than $2 - r$. Let C be a compact subset of S with measure $2 - r$. Let B' be the subset of pairs (k', s) in $K \times [-1, 1]$ for which $h_s(k')$ is not in $B_{1/n}$. Using the compactness of C and the fact that B' is relatively open we can find a neighborhood U of k such that $C \times U$ is contained in B'. It follows that for $k \in U$, $f_n(k) < r$. This shows semicontinuity.

Fix k and let $T_n = \{s \in [-1, 1]: h_s(k) \in B_{1/n}\}$, $f_n = \mu(T_n)$. Now T_n is a decreasing sequence of sets so that $f_n(k)$ is non-increasing and $\lim f_n(k)$ is the measure of $\bigcap T_n$. Since B is closed the intersection of the sets T_n is just the set of $s \in [-1, 1]$ such that $h_s(k) \in B$. The set of such s is countable and therefore has measure zero. This completes the proof of the lemma. We return to the proof of the theorem.

Choose ε so that the fraction of points in each horocycle orbit segment which lies in B_ε is less than $1/2$. Let K' be the union of sets $h_s(K)$ for $s \in [-1, 1]$. Let K'' be the union of closed unit balls with centers in K'. K'' is compact. Let V_n be the set of θ's in V with the property that for $t > n$ if $g_t(q) \in K''$ then $g_t(q) \in B_{\varepsilon/2}$. The union of the sets V_n is V; thus there is some set V_n with positive measure. Choose one such set and call it V'. Let θ' be a point of density in V'. By changing coordinates we may assume that $\theta' = 0$.

The limit in formula (2) converges uniformly for s in any compact set. Choose n large enough so that for $t > N$ and $k \in K$ the distance between the left hand side of (2) and the right hand side is less than $\varepsilon/2$ when applied to any $k \in K$ and for $s \in [-1, 1]$. Consider a $t > N$ for which $g_t(q) \in K$. Let $\theta \in [-e^{-t}, e^{-t}]$. Let $s = e^t \cdot \theta$. Applying the left and right sides of (2) to $g_t(q)$ shows that the distance between $g_t r_\theta(q)$ and $h_s g_t(q)$ is less than $\varepsilon/2$. Claim: $g_t r_\theta(q) \in K''$. By assumption $g_t(q) \in K$. Since $s \in [-1, 1]$ $h_s g_t(q)$ is in K'. The distance between $h_s g_t(q)$ and $g_t r_\theta(q)$ is less than one so that $g_t r_\theta(q)$ is in K''. For $\theta \in V'$, $g_t r_\theta(q)$ is in $B_{\varepsilon/2}$ when it is in K''. If $g_t r_\theta q$ is in $B_{\varepsilon/2}$ then $h_s g_t(q)$ is in B_ε. This can be the case for at most half of the s's in $[-1, 1]$. This shows that the density of V' in the interval $[-e^{-t}, e^{-t}]$ is less than $1/2$. This argument holds for a sequence of arbitrarily large values of t. This contradicts the assumption that θ' was a point of density for V'. □

Section 3

In this section we will investigate the geometry of the flat metric induced by a quadratic differential. We will relate the divergence of $r_\theta(q)$ to the pinching of curves on M with respect to the flat metric $g_t r_\theta(q)$. Our results are summarized in Proposition 2. This proposition will be used in the proof of Theorem 4 in the next section.

Definition. Let $QD = QD(k, \chi)$ be the space of quadratic differentials with area one on a surface M of Euler characteristic χ with k singularities.

Convention. A geodesic segment will be a geodesic interval with singularities as endpoints which contains no singularities in its interior.

PROPOSITION 1. *The set QD_ε consisting of quadratic differentials with no geodesic segment of length less than ε is compact.*

Proof. QD_ε is clearly closed in QD, for if $q_n \to q_0$ and q_0 has a trajectory β of length less than ε then the length of β is less than ε for n large enough. It remains to be shown that QD_ε lies over a compact set in R_g. If not there is a sequence q_n in QD_ε lying on Riemann surfaces X_n going to infinity in R_g. Passing to subsequences we may assume X_n converges to a Riemann surface X_∞ with nodes acquired by pinching along a set of disjoint curves $\alpha_1, \ldots, \alpha_p$ (see [Be]). Again passing to subsequences, the q_n can be assumed to converge uniformly on compact sets to an integrable q_∞ on X_∞. This means q_∞ has at most simple poles at the punctures. For any δ there is a curve homotopic to punctures with q_∞ length less than δ which means there are curves homotopic to α_i with q_n length less than δ for large n. This contradiction completes the proof. □

This proposition shows that if $r_\theta(q)$ is divergent then, as t increases to infinity, the flat metrics corresponding to $g_t r_\theta(q)$ possess arbitrarily short curves. In the remainder of this section we will show that we can choose short curves which are not crossed by other short curves.

Definition. A subcomplex is a triangulation of a subset of M so that the vertices are singularities, the edges are geodesic segments and the faces are triangles which do not contain singularities. We do not require that the vertices of a triangle be distinct.

We assume that if three sides of the complex bound a triangular region containing no singularities then that region is included in the complex.

LEMMA. *The maximum number of segments in a subcomplex is* $3(k - \chi)$. *The maximum number of triangles is* $2(k - \chi)$.

Proof. By adding a finite collection of segments to a subcomplex we may assume that the complementary regions are disks. These regions can be further subdivided into triangles. Thus we may assume that we have a decomposition of M into triangular regions with k vertices. Let e be the number of edges and let t be the number of triangles. Euler's formula gives $k - e + t = \chi$. Each edge is contained in 2 triangles and each triangle has 3 edges; thus $3t = 2e$. These two equations imply that $t = 2(k - \chi)$ and $e = 3(k - \chi)$.

Definition. An ε-subcomplex is a subcomplex in which each side has length less than or equal to ε.

Definition. A boundary edge is an edge which bounds less than two triangles.

LEMMA. *If* ε *is less than* $((k - \chi)3^{1/2}/2)^{-1/2}$ *then an* ε*-subcomplex has a boundary edge.*

Proof. A triangle with sides of length less than ε has area less than $\varepsilon^2 3^{1/2}/4$. The number of triangles is less than $2(k - \chi)$. If ε is smaller than the number above then the area of the ε-subcomplex is less than one. If the ε-subcomplex is not a triangulation of the entire surface then it must have some boundary edge.

Definition. A subset X of M is *convex* if every path (not necessarily joining singularities) lying in X, which is homotopic relative to endpoints to a geodesic path, is homotopic to a geodesic path lying in X.

Let B be a vertex in an ε-complex. The segments that contain B are cyclically ordered. Let AB and BC be adjacent edges. We say that ABC is an external angle if the segments AB and BC are not contained in a triangle ABC which is contained in the ε-complex.

LEMMA. *An* ε*-subcomplex is convex if and only if the measure of each external angle is greater than* π.

Proof. Denote the ε-complex by X. The second condition is the criterion for local convexity. It is clear that if X is convex then it is locally convex. To prove the converse choose a path in X homotopic rel endpoints to a geodesic segment not lying in X. Find the shortest path in this relative homotopy class that does lie in X. This path is a piecewise geodesic but not a geodesic. Where this path crosses a singularity we can measure the angle of bending on the right side of the path or on the left side of the path. The sum of these two angles will be the cone angle at the singularity. Since this path can be shortened, the measure of some bending angle ABC at some vertex B is less than π. The path can be shortened by moving it into the interior of the angle ABC. Since this shortened path is not contained in X there must be some sector of the angle ABC not contained in X.

LEMMA. *If a boundary edge of a connected ε-subcomplex is crossed by a geodesic segment of length less than C, where $C > \varepsilon$, then there is a connected $(C + \varepsilon)$-subcomplex with a larger number of simplices.*

Proof. Let X be the ε-subcomplex. If X is not convex then we can find some boundary angle ABC with measure less than π. If there is a segment AC so that the region bounded by the segments AB, BC and AC is a disk with no singularities then add the segment AC and the triangle ABC. The length of AC is less than 2ε. Now assume that this is not the case. Let M_r be the point on the segment AB such that the ratio of the length of BM_r to the length of AB is r. Let N_r be the point on the segment CB such that the ratio of the length of BN_r to length of CB is r. For small values of r we can construct a triangle M_rBN_r which contains no singularities. Let r_0 be the first r for which such a triangle cannot be constructed. The obstruction is a singularity P in the closure of the union of segments M_rN_r for $r < r_0$. We can construct a segment BP. The segment BP cannot intersect any segment in the ε-complex. Add BP to the ε-complex. The length of BP is less than ε.

Assume that X is convex. If a geodesic segment crosses the boundary of X then at least one endpoint is not contained in X. Thus there is a singularity not in X within distance C of X. Consider the singularity closest to X. Let this point be P. Let AB be the segment in the boundary of X which contains the point closest to P. The triangle ABP cannot contain any singularities not contained in X; otherwise these would be closer to X than P. The triangle cannot contain any point Q in X for if it did by convexity it would contain the triangle ABQ. This would contradict the assumption that the closest point to P lies on AB. We can add the triangle ABP to the original subcomplex. The longest side of ABP is less than $C + \varepsilon$. □

Definition. Let ε be less than C. A geodesic segment AB is (ε, C)-isolated if it has length less than ε and every geodesic that crosses AB in its interior has length greater than C.

Definition. $N(\varepsilon, C)$ is the set of quadratic differentials which possess (ε, C)-isolated curves.

We consider (ε, C)-isolated segments because there is a fixed bound to the number that can exist in any of the flat metrics. This follows from the fact that distinct (ε, C)-isolated segments are disjoint and the number of disjoint segments is less than or equal to $3(k - \chi)$. This bound plays a key role in the proof of Theorem 4 in the next section.

We are now in a position to establish the main result of this section.

PROPOSITION 2. *Suppose that there is a set S of positive measure so that $r_\theta(q)$ is divergent for $\theta \in S$. Then there are:*

1) *a sequence of times T_i increasing to infinity,*
2) *a sequence of subsets S_i contained in S such that $\mu(S_i) \geq \delta > 0$,*
3) *a sequence of $\varepsilon_i > 0$ converging to 0,*
4) *a positive constant C such that:*

$$g_{Ti} r_\theta(q) \in N(\varepsilon_i, C) \text{ for } \theta \in S_i.$$

The idea behind Proposition 2 is that if $r_\theta(q)$ is divergent then there will be times T_i after which there will always be a segment of length less than ε_i in the corresponding metrics. Unfortunately, for a fixed C, the sets $N(\varepsilon_i, C)$ do not form a neighborhood basis for infinity; so we cannot make the same statement for (ε_i, C)-isolated segments. However, by considering subsequences of times and subsets of S, we are able to make an analogous statement.

Definition. Let $U(\varepsilon)$ be the set of quadratic differentials with a geodesic segment of length less than ε.

Definition. Let $n_\varepsilon(q)$ be the maximum number of simplices in a connected ε-subcomplex.

Proof of Proposition 2. Choose a sequence $\varepsilon_i \to 0$. By Proposition 1, the sets $U(\varepsilon_i)$ have compact complements. We conclude that for each divergent ray $r_\theta(q)$ there are times T_i such that $g_t r_\theta(q) \in U(\varepsilon_i)$, for all $t > T_i$. The T_i can be chosen so that this relationship holds for all $\theta \in S'$ where $S' \subset S$ and $\mu(S') > \mu(S)/2$.

We have now constructed:

1) a sequence of times T_i increasing to infinity,
2) a set S' of θ's with positive measure,
3) a sequence of positive numbers ε_i converging to zero such that

$$g_{Ti} r_\theta(q) \in U(\varepsilon_i) \text{ for } \theta \in S_i'.$$

In order to prove Proposition 2 we will construct sets $S_i \subset S'$ and find a constant C such that:

$$g_{Ti} r_\theta(q) \in N(\varepsilon_i, C) \text{ for } \theta \in S_i.$$

Consider triples of sequences ε_i, T_i, and S_i with the property that ε_i goes to zero, T_i goes to infinity, the measure of S_i is bounded below and for $\theta \in S_i$ the surface $g_{T_i} r_\theta q$ is an element of $U(\varepsilon_i)$. We can construct one such triple by taking ε_i and T_i as above and taking $S_i = S$ for all i. For each such triple we can compute the minimum $n_\varepsilon(g_{T_i} r_\theta q)$. Choose a triple of sequences that maximizes this number. For each i and each θ in S_i let $C_{i,\theta}$ be the length of the shortest geodesic segment that crosses a boundary segment of the ε_i-subcomplex of $g_{T_i} r_\theta q$.

For each i, let m_i be the median value of $C_{i,\theta}$ on S_i. That is, m_i is some value so that the measure of points with greater or equal value is equal to the measure of points with smaller or equal value. We claim that $m_i \geq C$ for some $C > 0$. Suppose not. Then we could replace S_i by the set of θ's with values of $C_{i,\theta}$ less than or equal to m_i and replace ε_i by $\varepsilon_i + m_i$. We could then choose a subsequence so that m_i converged to zero. The new sequence of ε_i's still goes to zero and the measure of the new sets S_i is still bounded below. According to a previous lemma we could find ε-subcomplexes with more simplices. This contradicts the maximality of our triple of sequences. Thus the set of m_i is bounded below as claimed.

If we replace S_i by the set of θ with $C_{i,\theta} \geq m_i \geq C$ we ensure that every geodesic segment crossing the ε_i-complex has length at least C. Proposition 2 follows. □

Section 4

In this section we consider the case of divergent $r_\theta(q)$ and prove the following:

THEOREM 4. *The set of θ's such that $r_\theta(q)$ is divergent has measure zero.*

Proof. The argument is by contradiction. We assume that there is a set of θ's of positive measure for which $r_\theta(q)$ is divergent. With this assumption Proposition 2 applies. In this section we prove the following estimate:

PROPOSITION 3. *Fix $C > 0$. There are constants T, ε_1 and K such that, for $t > T$ and $\varepsilon < \varepsilon_1$, the measure of the set of θ such that $g_t r_\theta(q)$ is in $N(\varepsilon, C)$ is less than $K\varepsilon$.*

When $K\varepsilon$ is less than the lower bound ρ for the measures of the sets S_i given by Proposition 2, we have a contradiction. Thus no non-trivial set of divergent geodesics can exist. This completes the proof of Theorem 4, assuming Propositions 2 and 3. □

The remainder of this section is devoted to the proof of Proposition 3.

Behavior of the length function: Let α be a geodesic segment. Denote by $l(t, \theta)$ its length with respect to the structure $g_t r_\theta(q)$. The segment α is vertical in $r_\theta(q)$; i.e. $h = 0$ and $v = l(0, \theta)$, for some value $\theta(\alpha)$. Rotate the circular coordinate so that $\theta(\alpha) = 0$. In these coordinates

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

Fix a value of t and choose $\varepsilon > 0$. For each geodesic segment α, define two intervals $I_\alpha \subset J_\alpha \subset (-\pi/2, \pi/2)$ with respect to the rotated coordinates for α.

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

LEMMA. *There are constants ε_1, T and K, independent of α, such that for $\varepsilon < \varepsilon_1$ and $t > T$ we have $\mu(I_\alpha)/\mu(J_\alpha) < K\varepsilon$. If $l(t, \theta) < \varepsilon$ for some θ, then it is less than C on J_α and is greater than ε outside I_α.*

Proof. If $l(t, \theta) < \varepsilon$ for some θ, then

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

since $l(t, \theta)$ attains its minimum at 0. From (1) we have

$$l(t, \theta) < C/2 + \varepsilon \quad \text{for } \theta \in J_\alpha.$$

If we choose $\varepsilon_1 = C/2$ and $\varepsilon < \varepsilon_1$ then $l(t, \theta)$ is less than C on J_α as desired.

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 $\mu(I'_\alpha)/\mu(J'_\alpha) < 2\varepsilon/C$. Since the function arcsin 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 K depending on C such that $\mu(I_\alpha)/\mu(J_\alpha) < K\varepsilon$, 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 geodesic segment on q. Since $v \geq m$ and $t > T$, the inequality holds, which completes the proof of the lemma.

Proof of Proposition 3. Fix $t > T$ and $\varepsilon < \varepsilon_1$ so that the above lemma applies. Consider all geodesic segments α which are (ε, C)-isolated with respect to $g_t r_\theta(q)$ for some θ. For each α let $\hat{I}_\alpha$ be the smallest interval of θ's which contains the set of θ's for which α is (ε, C)-isolated. Then $\hat{I}_\alpha \subset I_\alpha \subset J_\alpha$, where I_α and J_α are defined as above.

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. The right-hand endpoint of $\hat{J}_\alpha$ will be halfway between the right-hand endpoints of $\hat{I}_\alpha$ and J_α. The point of this construction is the following fact. 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 that J_α intersects $\hat{I}_\beta$. We can find a θ in the intersection so that β is (ε, C)-isolated with respect to $g_t r_\theta(q)$.

Now with respect to $g_t r_\theta(q)$, α has length less than C and β cannot cross any geodesic segment of length less than C. Thus the interiors of α and β are disjoint. By a previous lemma the maximum number of disjoint geodesic segments is $G = 3(k - \chi)$.

We conclude that no θ can lie in more than G sets $\hat{J}_\alpha$. The sum of the lengths of the intervals $\hat{J}_\alpha$ is at most πG. The sum of the lengths of the intervals J_α is at most $2\pi G$. Thus by the above lemma, the sums of the lengths of the intervals I_α is less than $\varepsilon K 2\pi G$. The intervals I_α cover the set of θ's for which $g_t r_\theta(q)$ is in $N(\varepsilon, C)$; so Proposition 3 follows. □

Section 5

The collection of polygonal regions in $\mathbf{R}^2$ with n vertices and a given combinatorial type can be identified with a subset of $\mathbf{R}^{2n}$. We define the space of billiard tables to be the set of all such regions topologized as a disjoint union of subsets of $\mathbf{R}^{2n}$. Now let X be an arbitrary closed set of the space of billiard tables. For a rational table $x \in X$, let $\Gamma_x \subset O(2)$ be the group generated by reflections in the sides.

Proposition 4. *Let X be a space of billiard tables as above with the property that for any number N the set of rational tables x in X with card$(\Gamma_x) \geq N$ is dense. Then the ergodic billiard tables in X form a dense G_δ.*

If we take X to be an appropriate space of right triangles then this proposition implies the existence of a dense G_δ of right triangular ergodic billiard tables as claimed in the introduction.

Proof of proposition. It suffices to prove the result for any compact subset of X. Without loss of generality we may assume that X is compact and contained in a single copy of $\mathbf{R}^{2n}$. Each $x \in X$ corresponds to a polygonal region Q_x of the plane. Assume that the area of each Q_x is one. Let PX be the bundle with base space X so that the fiber, P_x, over x is the phase space of the corresponding billiard table Q_x. Note that PX can be thought of as a subset of $X \times \mathbf{R}^2 \times S^1$. Let μ_x be the product of the area measure on Q_x with the unit Haar measure on S^1. Let φ_t denote the billiard flow on PX.

Choose a sequence of continuous functions $f_1, f_2 \ldots$ on PX which when restricted to P_x, for any $x \in X$, are dense in $L^2(P_x)$. We make the further assumption that each f_i respects the natural identifications on each P_x in the sense that if v is an outward pointing vector on the boundary of a polygon and v' is the corresponding inward pointing vector then $f(v) = f(v')$. Let $E(i, n, T)$

be the set of $x \in X$ for which

$$\int_{z \in P_x} \left[\frac{1}{T} \int_0^T f_i(\phi_t(z))\, dt - \int_{P_x} f_i \, d\mu_x \right]^2 d\mu_x < \frac{1}{n}.$$

Let $E(i, n) = \bigcup_{T=1}^{\infty} E(i, n, T)$, and let $E = \bigcap_{i=1}^{\infty} \bigcap_{n=1}^{\infty} E(i, n)$.

The set of $x \in X$ for which $\phi_t | P_x$ is ergodic is precisely E (cf. [O-U]). 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 x for which $\mathrm{card}(\Gamma_x) > N$.

The first statement implies that the sets $E(i, n)$ are open. The second statement implies that the sets $E(i, n)$ are dense in X. Since E is the countable intersection of sets $E(i, n)$, it is a G_δ and the Baire category theorem implies that E is dense in X. Statement 1) is a consequence of the following lemma. Statement 2) follows from Lemma 5.2.

Lemma 5.1. *Let $T > 0$ be fixed. Let f be a continuous function on PX respecting the identifications. Then*

$$\int_{z \in P_x} \left[\frac{1}{T} \int_0^T f(\phi_t(z))\, dt - \int_{P_x} f d\mu_x \right]^2 d\mu_x$$

depends continuously on x.

Proof. For $y \in X$ let $a(y)$ be the integral of f over P_y. We begin by proving that $a(y)$ is a continuous function of y. Let $x \in X$ and let $\varepsilon > 0$ be given. Let $N_1 \subset X$ be a neighborhood of x. For each $y \in N_1$, Q_y is a subset of the plane. Let Q_1 be the intersection of the sets Q_y where $y \in N_1$. Let $M = \sup|f|$. By choosing N_1 sufficiently small we may assume that the area of Q_1 is at least $1 - \varepsilon/3M$. Let $P_1 = Q_1 \times S^1$ and let μ_1 be the product of Lebesgue measure on Q_1 with unit Haar measure on S^1. We can view $N_1 \times P_1$ as a subset of PX.

Let $b(y) = \int_{p \in P_1} f(y, p)\, d\mu_1$. The continuity of b follows from the continuity of f. Furthermore b is close to a:

$$|a(y) - b(y)| \le \int_{P_y - yxP_1} f d\mu \le \mu(P_y - yxP_1) \cdot M \le \frac{\varepsilon}{3}.$$

Let $N_2 \subset N_1$ be a neighborhood of x consisting of points y for which $|b(x) - b(y)| < \varepsilon/3$. Then for $y \in N_2$ we have

$$|a(x) - a(y)| \le |a(x) - b(x)| + |b(x) - b(y)| + |b(y) - a(y)| \le \varepsilon.$$

This completes the proof of the continuity of the function a.

We replace the function f by the function $\bar{f}$ defined as follows: For $z \in P_y$, $\bar{f}(z) = f(z) - a(y)$. Then the proof of the lemma reduces to the proof of the continuity of the following function:

$$c(x) = \int_{z \in P_x} \left[\frac{1}{T} \int_0^T \bar{f}(\phi_t(z))\, dt \right]^2 d\mu_x.$$

We begin by introducing an auxiliary function. Let $\operatorname{vert}(y)$ denote the set of vertices of Q_y. Let π be the projection from P_y to Q_y. For $z \in P_y \subset PX$ define

$$l(z) = \inf_{\substack{0 \le t \le T \\ v \in \operatorname{vert}(y)}} d(\pi\phi_t(z), v).$$

If $\phi_t(z)$ is not defined for some t between 0 and T then we take $l(z)$ to be 0. Since $\pi(\phi_t(z))$ is a smooth function of t and z when $\phi_t(z)$ is defined, it is easily seen that l is a smooth function of z.

We will prove the continuity of the function c. Fix $x \in X$. Let $\varepsilon > 0$ be given. Choose a neighborhood N_1 of x such that the area of the set Q_1 is at least $1 - \varepsilon/10M$, where Q_1 and M are defined as before.

The gradient of l at $z \in P_x$ where $l(z) = 0$ is non-zero. So, by the implicit function theorem, the set of vectors z in P_x for which $l(z) = 0$ has codimension 1 in P_x. Thus it has measure zero. Let C_δ be the set of $z \in P_x$ for which $l(z) \ge \delta$. Choose δ small enough that the measure of C_δ is at least $1 - \varepsilon/10M$. Let D_δ consist of pairs $(p, v) \in \mathbf{R}^2 \times S^1$ such that $p \in Q_1$ and $(p, v) \in C_\delta$. Since $Q_1 \times S^1$ is contained in P_x for all $x \in N_1$, we can identify $N_1 \times D_\delta$ with a subset of PX in a natural way. Let $\bar{l}(y)$ be the infimum of $l(z)$ for $z \in (P_y \cap N_1 \times D_\delta)$. Now $\bar{l}$ is continuous and $\bar{l}(x) = \delta$. We can find a neighborhood $N_2 \subset N_1$ so that for $y \in N_2$, $\bar{l}(y)$ is positive. Let $d(y)$ denote the integral

$$d(y) = \int_{z \in y \times D_\delta} \left[\frac{1}{T} \int_0^T \bar{f}(\phi_t(z))\, dt \right]^2 d\mu_y.$$

Since $f(\phi_t(z))$ is a continuous function of t and z for $t \in [0, T]$ and $x \in N_2 \times D_\delta$, the integral varies continuously. We can find a neighborhood N_3 of x on which d varies by less than $\varepsilon/3$. The difference between $d(y)$ and $c(y)$ is less than $\varepsilon/3$. For $y \in N_3$,

$$|c(x) - c(y)| \le |c(x) - d(x)| + |d(x) - d(y)| + |d(y) - c(y)| \le \varepsilon.$$

This completes the proof of the continuity of c and the proof of the lemma.

Lemma 5.2. *Fix $n > 0$ and a continuous function f on PX. Choose $\delta > 0$ so that if the distance between two points θ_1 and θ_2 is less than δ then $|f(\theta_1) - f(\theta_2)| < 1/2n$. (Recall that PX is compact.) Let N be greater than $2/\delta$. Let Q_x*

be a rational polygon with $|\Gamma_x| \geq N$. *Then for T sufficiently large*

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

Proof. Since x is fixed we will drop the subscripts from P, Q, Γ and μ. For θ in S^1 let $u(\theta) = 1/|\Gamma|\int_{z\in M_\theta} f(z)\,dA$, where dA is the area measure on M_θ. For $z \in P$ let $u'(z) = u(\theta)$ where $z \in M_\theta$. For z in P let $v_T(z) = 1/T\int_0^T f(\phi_t(z))\,dt$. Let $\operatorname{int} f$ denote the integral of f with respect to μ. The quantity which appears in the lemma is the norm in the space $L^2(P)$ of the function $v_T - \operatorname{int} f$.

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

Remark. It is in the proof of this claim that essential use is made of the central result, Theorem 1.

The surfaces M_θ are parametrized by $\theta \in S^1/\Gamma$. We evaluate the norm by integrating first with respect to M_θ, then with respect to θ.

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

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

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

For a given θ the ergodicity of $\phi|M_\theta$ implies that $\lim_{T\to\infty} w_T(\theta) = 0$. Theorem 1 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.

Claim. $\|u' - \operatorname{int} f\| \leq 1/2n$.

The norm of the function $u' - \operatorname{int} f$ in $L^2(P)$ is equal to the norm of the function $u - \operatorname{int} f$ in $L^2(S^1)$. Let θ_1 be a point in the circle at which u assumes its maximum value M. Let θ_2 be a point at which u assumes its minimum value m. Note that u is constant on the orbits of Γ. The distance between neighboring points in a Γ orbit is less than $2/|\Gamma| < 2/N < \delta$. By replacing θ_2 by some $\gamma\theta_2$, where $\gamma \in \Gamma$, we may assume that the distance between θ_1 and θ_2 is less than δ. It follows from the continuity assumption on f that since θ_1 and θ_2 are closer than δ then $|u(\theta_1) - u(\theta_2)|$ is less than $1/2n$. Since u is defined by averaging f, the integral of f over P is equal to the integral of u over S^1. Thus $m \leq \operatorname{int} f \leq M$; hence $|u(\theta) - \operatorname{int} f| \leq 1/2n$ and $\|u - \operatorname{int} f\| \leq 1/2n$. This completes the proof of the claim.

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

$$\|v_T - \operatorname{int} f\| \leq \|v_T - u'\| + \|u' - \operatorname{int} f\| \leq 1/2n + 1/2 = 1/n.$$

Stanford University, Stanford, California
University of Illinois at Chicago, Illinois
Lehman College, the Bronx, New York and Graduate Center, New York, New York

References

[Be] L. Bers, Finite dimensional Teichmüller spaces and generalizations, Bull. AMS **5** (1981) 131–172.

[B-K-M] C. Boldrighini, M. Keane and F. Marchetti, Billiards in polygons, Ann. of Prob. **6**, No. 4 (1978), 532–540.

[B] M. Boshernitzan, A condition for minimal interval exchange maps to be uniquely ergodic, preprint.

[C-F-S] I. P. Cornfeld, S. V. Fomin and Ya. G. Sinai, *Ergodic Theory*, Springer-Verlag (1982).

[F-K] R. H. Fox and R. B. Kershner, Concerning the transitive properties of geodesics on a rational polyhedron, Duke Math. J. **2** (1936), 147–150.

[G] E. Gutkin, Billiards on almost integrable polyhedral surfaces, preprint.

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

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

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

[S] Ya. G. Sinai, *Introduction to Ergodic Theory*, Princeton University Press (1976).

[V1] W. A. Veech, Gauss measures for transformations on the space of interval exchange maps, Ann. of Math. **115** (1982), 201–242.

[V2] ______, Dynamical systems on analytic manifolds of quadratic differentials, III to be published in Ann. of Math.

[Z-K] A. N. Zemlyakov and A. B. Katok, Topological transitivity of billiards in polygons, Matem. Zametki **18**:2 (1975), 291–300 (English translation in Math. Notes **18**:2 (1976), 760–764).

(Received June 14, 1985)