Commun. Math. Phys. ■, ■–■ (2002)
Digital Object Identifier (DOI) 10.1007/s00220-002-0696-1

Communications in
**Mathematical
Physics**

Billiards With Pockets: A Separation Principle and Bound for the Number of Orbit Types

Charles I. Delman, Gregory Galperin

Mathematics Department, Eastern Illinois University, Charleston, IL 61920, USA.
E-mail: cfcid@eiu.edu; cfgg@eiu.edu

Received: 28 December 2001 / Accepted: 9 April 2002
Published online: ■ ■ ■ –

Abstract: We introduce and prove a Separation Principle, similar in form to the familiar Uncertainty Principle of quantum mechanics, which separates the position and direction of any two phase points on distinct unfoldings of (non-parallel) trajectories on a polygonal billiard table with pockets. Applying this principle, we demonstrate that the number of orbit types (that is, classes of trajectories, up to parallelism) on a polygonal billiard table with area A and pockets of area a is strictly bounded above by $\frac{\pi^2}{2} \cdot \frac{A}{a}$. More generally, the same bound applies to any compact polyhedral surface with pockets at its vertices. If the boundary is empty (so that billiard trajectories are just geodesics), the bound is reduced by a factor of two to $\frac{\pi^2}{4} \cdot \frac{A}{a}$. We believe the Separation Principle will also have fundamental applications to other problems in the theory of billiards and related dynamical systems.

Contents

1. Introduction

We consider billiard trajectories which “live” on a Euclidean polygonal table with a pocket at each vertex. (Later we will generalize our results to compact polyhedral surfaces.) By a pocket we mean a closed disk of small radius ε, centered at the vertex, which trajectories must not enter. A distinguishing feature of billiards with pockets is that all trajectories which do not terminate in a pocket are *periodic* [4].

In [2], Delman, Galperin, and Troubetzkoy proved that the number of orbit types (that is, classes of trajectories, up to parallelism) on a billiard table with pockets is finite. Furthermore, for a *rational* polygon – one in which all vertex angles are rational multiples of π – they proved in [3] that this number is proportional to $\frac{A}{a}$, where A is the area of the polygon or surface and $a = \pi \varepsilon^2$ is the area of a pocket.

Unfortunately, the above estimate, aside from being restricted to rational polygons, is unsatisfactory on several other counts. The constant of proportionality, c, is actually obtained as the product of π and three other constants: $c = \pi \cdot c_1 \cdot c_2 \cdot c_3$. The three constants c_1, c_2 and c_3 depend on the shape of the polygon, and two of them are unstable with respect to small perturbations of this shape. The first, c_1, bounds the ratio between the period of a trajectory and its length; it varies stably with the shape of the polygon. However, c_2 is the least common multiple of the denominators of the fractions expressing the angles, which varies wildly with small perturbations of the polygon (within the class of rational polygons) due to changes in the denominators. As for c_3, it is the coefficient in Masur's theorem [7] that the number of generalized diagonals of period less than or equal to T is bounded above by a multiple of T^2. It is also unstable with respect to the angles of the polygon. Moreover, the existence of c_3 was proven nonconstructively in [7], and the best estimates of it are huge (see Vorobets [8], Theorem 4.1).

In contrast, in the current paper we show that the number of orbit types on *any* polygon (rational or not) with pockets, indeed on any polygonal surface with pockets, is bounded proportionally – with a constant of proportionality *independent* of the polygon or surface – to the ratio of areas A/a, where A is, as before, the area of the polygon or surface and $a = \pi \varepsilon^2$ is the area of a pocket. Moreover, the constant of proportionality is just $\frac{\pi^2}{2}$, which is less than 5!

The earlier proof for rational polygons relies on the fact that every trajectory lies on a compact surface of area $2c_2 A$ (obtained by considering reflections of the polygon in all sides, with appropriate identifications; see [6] or the fine survey article [5]) which is invariant under the billard flow. (For example, the invariant surfaces for a square are tori consisting of four copies of the square, while those for an equilateral triangle are tori consisting of six copies.) Since a trajectory which misses all pockets must remain at a distance greater than ε from every vertex, it can be shown that a strip of width greater than 2ε around such a trajectory is embedded in this invariant surface, which limits the length of the trajectory to less than $\frac{2c_2 A}{2\varepsilon} = \frac{c_2 A}{\varepsilon}$. The remainder of the proof consists of bounding the period of a trajectory as a function of its length, which brings in the constant c_1, and applying Masur's theorem, here using the fact that every periodic orbit type corresponds to a generalized diagonal. (See [3].) Thus, in essence, the proof for rational polygons is based on the fact that a trajectory in a rational polygon with pockets cannot be too long. For a non-rational polygon this is not true, since its invariant domains consist of Riemann surfaces which are not compact. The proof of the current, general result relies instead on the finite volume of phase space.

Our proof, which is self-contained (except for use of the fact that all trajectories are periodic) and elementary, makes strong use of a fundamental theorem, which we call the "Separation Principle" and regard as the main result of this paper. The Separation Principle, which formally resembles the familiar Uncertainty Principle of quantum mechanics, states that two phase points which lie either on unfoldings of non-parallel trajectories or on distinct unfoldings of the same trajectory cannot be close together in both position and direction: *if the distance between the points is small, the angle between their directions must be large, and vice-versa.* Precisely, it is not possible both that the distance is less than 2ε and the angle is less than $\frac{2\varepsilon}{L_{min}}$, where L_{min} is the minimum length of the two trajectories involved.

It follows that each trajectory may be surrounded by a regular neighborhood, or "tube", whose cross-section is a rectangle of width 2ε and height $\frac{2\varepsilon}{L}$, where L is the length of the trajectory, and the tubes around non-parallel trajectories will be disjoint. We thus obtain pairwise disjoint tubes of volume $4\varepsilon^2$ around any representative collection of non-parallel trajectories in the phase space of the system. Dividing the total volume of phase space, $2\pi A$, by the volume of a tube gives the advertized bound.

A refinement of this bound is obtained by considering the reverse of each trajectory, by which we mean the trajectory that traverses the same trace in the opposite direction. The traces in phase space of a trajectory and its reverse are disjoint if they don't coincide, in which case they are sufficiently separated that the tubes surrounding them are also disjoint. We will call a trajectory which coincides with its reverse auto-reversing; a trajectory is auto-reversing if and only if it reflects of some side of the polygon at right angles. Because of the existence of auto-reversing trajectories (see, for example, [1]) we cannot reduce the bound by a factor of two by introducing reverse trajectories to the calculation (thereby obtaining two tubes for each class of parallel trajectories). Instead, if k denotes the number of auto-reversing trajectories and l denotes the number of remaining trajectories, up to parallelism, then $k + 2l < \frac{\pi^2}{2} \cdot \frac{A}{a}$.

More generally, we may consider billiards on any compact polyhedral surface with pockets at its vertices. The Separation Principle and all other results carry over to this context without significant modification. In particular, if the boundary of the surface is empty (so that billiard trajectories are just geodesics), then the number of orbit types is strictly bounded above by $\frac{\pi^2}{4} \cdot \frac{A}{a}$, because there are no auto-reversing trajectories on such a surface. For clarity of exposition, we introduce and prove all results in the familiar setting of polygons, leaving it to the reader to observe that nothing particular to this setting is required in the proofs.

2. Definitions and Results

2.1. Bound on the Number of Orbit Types (Theorem 2). Let Q be a polygon. A *billiard trajectory* on Q is a path which is geodesic on the interior of Q and, at points of ∂Q, satisfies the "billiard law" that each angle of incidence equals the corresponding angle of reflection. We imagine, of course, that each trajectory is the path of a particle bouncing off the sides of Q as it travels (at a constant speed, whose specific value we ignore). The image of the trajectory is called its *trace*. In this article, all polygons will be Euclidean. (See Fig. 1.)

Let Q_ε denote the billiard table obtained from Q by removing a pocket (that is, the intersection of Q with a closed disk) of radius ε centered at each vertex. The radius ε is assumed to be sufficiently small so that each pocket is disjoint from all other pockets and from all sides of Q except the two which meet at its center; if ε meets these conditions, we will say that Q admits pockets of radius ε. A billiard trajectory terminates on Q_ε if it enters a pocket. Whenever we refer to a trajectory on Q_ε, we shall mean a non-terminating trajectory.

Notation. Throughout this paper, A denotes the area of Q (or a more general polyhedral surface, where that is the context), and $a = \pi\varepsilon^2$, the area of the disks whose intersections with Q form the pockets of Q_ε.

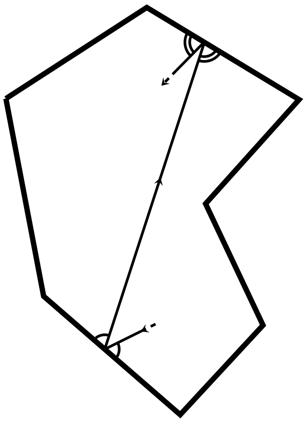


Fig. 1. Billiard trajectory

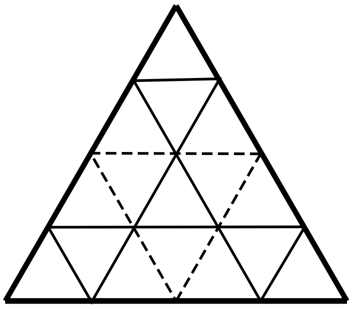


Fig. 2. A pair of equivalent trajectories

Recall that, as shown in [4], all trajectories on $\mathcal{Q}_\varepsilon$ are periodic. The *period* of a trajectory is the least number of reflections required before the billiard particle returns to a previously occupied position with the same direction.

The bi-infinite sequence of sides which a trajectory hits, up to translation and inversion, will be called its *code*. The code of a periodic trajectory is obviously periodic. Two trajectories will be considered *equivalent* if they have the same code. A pair of equivalent trajectories is shown in Fig. 2. An equivalence class of trajectories is called an *orbit type*.

It is very important that we may restrict our attention to trajectories of even period. To see both the justification and reason for doing so, consider a trajectory with odd period. Nearby equivalent trajectories switch sides after one period, rather than returning to their original position; therefore, these trajectories have period double that of the original, as illustrated by the example in Fig. 2. Moreover, the transverse orientation of a trajectory with even period is preserved. Preservation of transverse orientation is crucial to our arguments. A trajectory will be called *generic* if its period is even. It can be easily shown, using the method of unfolding described in the next section, that equivalent generic trajectories have the same period.

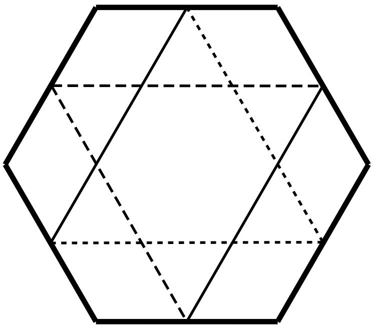


Fig. 3. Parallel legs on non-equivalent trajectories

Remark. Replacing a trajectory with a sufficiently close equivalent trajectory is possible because our pockets are closed.

Remark. Equivalent trajectories with distinct traces are parallel in the sense that the interior segments, or "legs", on these trajectories corresponding to the same part of the code are parallel. Note that legs from two non-equivalent trajectories or different legs of the same trajectory may be parallel in direction (see Fig. 3), but only equivalent trajectories admit a one-to-one correspondence pairing parallel legs with endpoints on the same sides of the polygon.

The *reverse* of a trajectory is the trajectory which traverses the same trace in the opposite direction. Note that a trajectory and its reverse are equivalent. We call a trajectory *auto-reversing* if it coincides with its reverse (that is, the particle returns at some time to its initial position with the opposite direction). It is easy to see that a trajectory is auto-reversing if and only if it reflects off some side of the polygon at right angles.

With these definitions, we are now prepared to state our bound, which we catalogue as the second theorem of the paper, as its proof depends on the Separation Principle to be introduced in the next section:

Theorem 2. *Given a compact Euclidean polygon $\mathcal{Q}$, let k be the number of equivalence classes of auto-reversing trajectories on $\mathcal{Q}_\varepsilon$, and let l be the number of remaining equivalence classes of trajectories. Then*

$$k + 2l < \frac{2\pi A}{4\varepsilon^2} = \frac{\pi^2}{2} \cdot \frac{A}{a}.$$

In particular, the total number of orbit types on $\mathcal{Q}_\varepsilon$ is strictly less than $\frac{\pi^2}{2} \cdot \frac{A}{a}$.

2.2. Separation Principle (Theorem 1). Some additional concepts and terminology are required to formulate the Separation Principle. Very important is the standard technique of *unfolding* a trajectory, in which an initial point of the trajectory (by which we mean both a position and a direction) is chosen in a polygon $\mathcal{Q}_0$ congruent to $\mathcal{Q}$, and the trajectory is represented as a straight line in the plane (called its unfolding) by reflecting in the sides hit by the billiard particle to obtain a succession of polygons $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, ..., as well as $\mathcal{Q}_{-1}$, $\mathcal{Q}_{-2}$, $\mathcal{Q}_{-3}$, ..., as the trajectory is followed in the reverse direction.

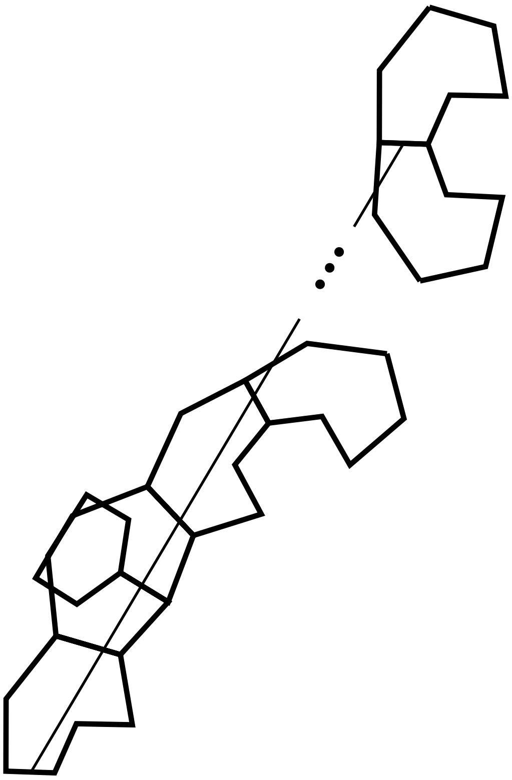


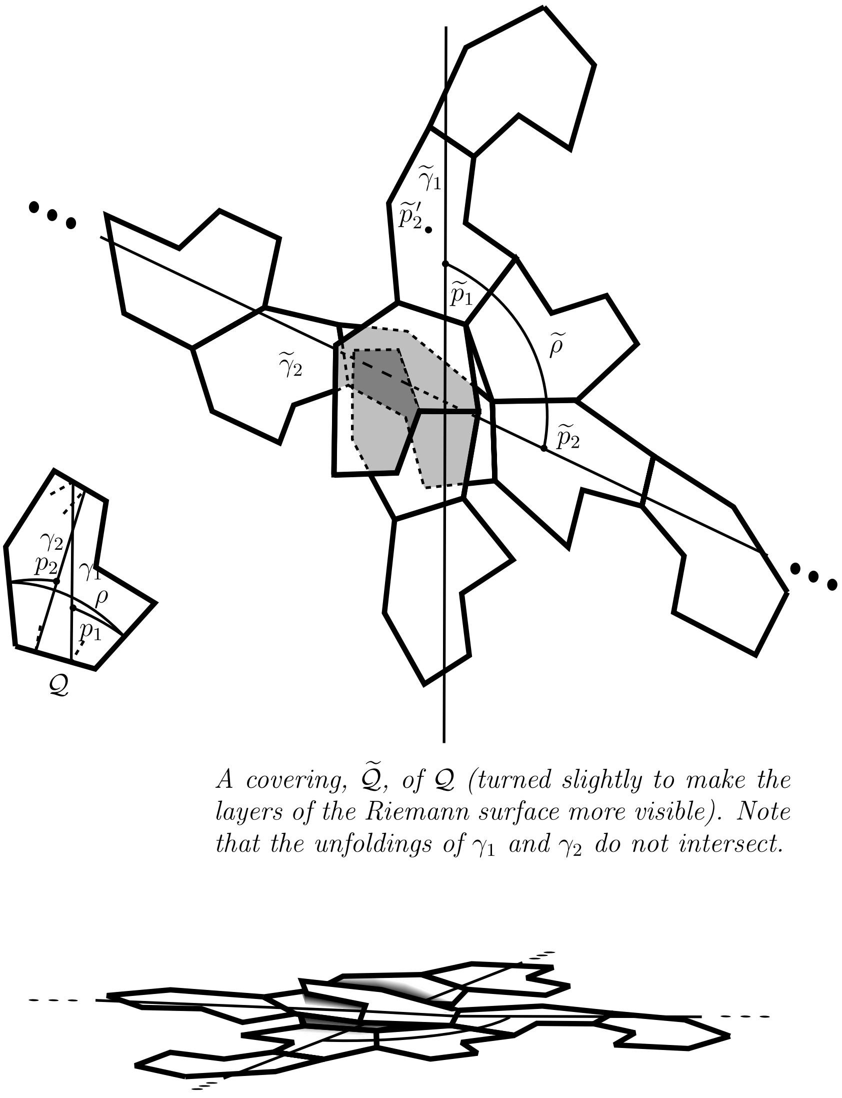
Fig. 4. The unfolding of a trajectory, with its corridor, showing one period

The collection of polygons $\{\mathcal{Q}_k\}_{k\in\mathbb{Z}}$ is called the *corridor* of the unfolding. The corridor of a ray or segment contained in an unfolding, or of any union of such rays or segments, is the minimal collection of polygons which contains it.

Note that the polygons in a corridor may overlap. These overlapping polygons should be viewed as lying in different copies of the plane. *Precisely, the corridor should be thought of as the Riemann surface, branched at some of its vertices, obtained from the disjoint union of its polygons by identifying their common edges.* (See Fig. 4.)

We will need to consider the following situation: two points, each on a trajectory in $\mathcal{Q}$ (possibly the same trajectory) are joined by a piecewise smooth path (which we assume to obey the billiard law at the boundary of $\mathcal{Q}$), and we wish to unfold the path and trajectories simultaneously. The corridors of the path and the two trajectories combine to form a connected Riemann surface, as in Fig. 5. This construction leads us naturally to define the following generalization of a corridor.

Definition 1. *A **covering** of $\mathcal{Q}$ is a connected Riemann surface obtained by identifying the common edges of a collection of polygons, each of which is obtained from a copy of $\mathcal{Q}$ by a sequence of reflections in edges. Every point (resp., set of points) in a covering corresponds to a unique point (resp., set of points) in the polygon $\mathcal{Q}$; we will say that it **covers** this point (resp., set of points).*



A covering, $\widetilde{\mathcal{Q}}$, of $\mathcal{Q}$ (turned slightly to make the layers of the Riemann surface more visible). Note that the unfoldings of γ_1 and γ_2 do not intersect.

Side view of $\widetilde{\mathcal{Q}}$.

Fig. 5. Two trajectories, γ_1 and γ_2, joined by a path, ρ, in $\mathcal{Q}$ and their unfoldings in a covering, $\widetilde{\mathcal{Q}}$, of $\mathcal{Q}$

We will sometimes wish to consider the set of points in the plane lying under some geometric object on this Riemann surface. Usually, it will be clear from the context when we are doing so, obviating the need for additional terminology; if there is any ambiguity, we will refer to this set as the *projection* of the object. (See Fig. 6.)

Notation. A covering of $\mathcal{Q}$ will be denoted by $\widetilde{\mathcal{Q}}$. Points in $\widetilde{\mathcal{Q}}$ which cover $p \in \mathcal{Q}$ will be denoted by $\widetilde{p}$, $\widetilde{p}\,'$, $\widetilde{p}\,''$, etc. The projection of $\tilde{p}$ onto the plane will be denoted by $\bar{p}$.

Notation. The segment in the plane with endpoints $\bar{p}_1$ and $\bar{p}_2$ will be denoted $\bar{p}_1\bar{p}_2$.

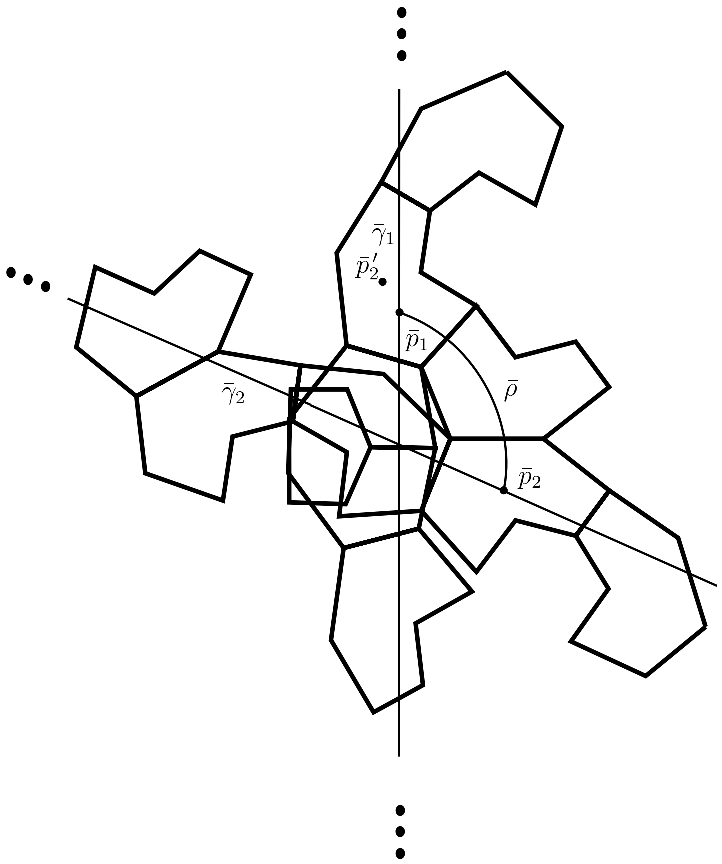


Fig. 6. The projection of a covering onto the plane

Definition 2. *The* ***distance in*** $\mathcal{Q}$ *between two points* p_1 *and* p_2 *is the Euclidean length of the shortest path in* $\mathcal{Q}$ *joining* p_1 *to* p_2*. More generally, the distance between two points* $\widetilde{p}_1$ *and* $\widetilde{p}_2$ *in a covering* $\widetilde{\mathcal{Q}}$ *of* $\mathcal{Q}$ *is the length of the shortest path in the Riemann surface* $\widetilde{\mathcal{Q}}$ *(meaning that it may pass from one polygon to another only by passing through a shared edge) joining* $\widetilde{p}_1$ *to* $\widetilde{p}_2$*.* (See Fig. 7.*)*

Notation. We denote the distance in $\widetilde{\mathcal{Q}}$ between $\widetilde{p}_1$ and $\widetilde{p}_2$ by $|\widetilde{p}_1\widetilde{p}_2|$.

Notation. Let $V = \{v_i\}_{i=1}^n$ be the set of vertices of $\mathcal{Q}$. If $\widetilde{\mathcal{Q}}$ is a covering of $\mathcal{Q}$, denote the set of vertices of $\widetilde{\mathcal{Q}}$ by $\widetilde{V}$.

It is easy to see that the shortest path on a covering between two points $\widetilde{p}_1$ and $\widetilde{p}_2$ is the union of segments with endpoints in the set $\widetilde{V} \cup \{\widetilde{p}_1, \widetilde{p}_2\}$. Also observe that projection from a covering onto the plane preserves path length.

Definition 3. *The* ***length*** *of a periodic trajectory is the (Euclidean) distance traveled by the billiard particle over the course of one period.*

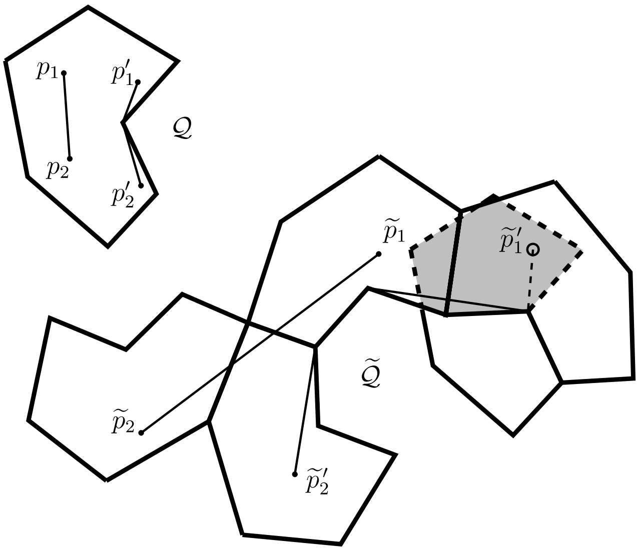


Fig. 7. Paths realizing the shortest distance between two points

The *phase space* of the billiard system on $\mathcal{Q}$, denoted $\mathfrak{P}(\mathcal{Q})$, is the quotient space of the unit tangent bundle over $\mathcal{Q} \setminus V$ obtained by identifying the points with the same position on a side of $\mathcal{Q}$ whose directions are reflections of each other in that side. (In other words, we identify position-direction pairs describing states which may be held simultaneously by a billiard particle.) We represent a *phase point* $\mathfrak{p} \in \mathfrak{P}(\mathcal{Q})$ by an ordered pair (p, φ), where $p \in \mathcal{Q}$ is the position of $\mathfrak{p}$ and $\varphi \in S^1 = [0, 2\pi]_{/0\sim 2\pi}$ is its direction in a chosen system of polar coordinates. Thus, the phase space may be thought of as the cylinder $(\mathcal{Q} \setminus V) \times [0, 2\pi]$ with appropriate identifications. With this geometric structure, it is a singular Euclidean 3-manifold whose singular points comprise a submanifold of dimension one (the points with position on a side of $\mathcal{Q} \setminus V$ and direction parallel to that side); hence, its volume is well-defined and is, clearly, $2\pi A$.

The phase space of unfoldings on a covering $\widetilde{\mathcal{Q}}$, which we will denote by $\mathfrak{P}(\widetilde{\mathcal{Q}})$, is by construction the trivial bundle $(\widetilde{\mathcal{Q}} \setminus \widetilde{V}) \times S^1$. We represent a phase point $\widetilde{\mathfrak{p}} \in \mathfrak{P}(\widetilde{\mathcal{Q}})$ by an ordered pair $(\widetilde{p}, \widetilde{\varphi})$, where $\widetilde{p} \in \widetilde{\mathcal{Q}}$ is its position and $\widetilde{\varphi} \in S^1$ is its direction in a chosen system of polar coordinates. Two unfoldings in a common covering $\widetilde{\mathcal{Q}}$ will be considered distinct if their corresponding trajectories in $\mathfrak{P}(\widetilde{\mathcal{Q}})$ have distinct traces. In particular, unfoldings which are reverses of each other are distinct.

Notation. If $\widetilde{\mathfrak{p}}_1 = (\widetilde{p}_1, \widetilde{\varphi}_1)$ and $\widetilde{\mathfrak{p}}_2 = (\widetilde{p}_2, \widetilde{\varphi}_2)$ are phase points written in a common system of coordinates, then we write $\Delta\widetilde{p} = |\widetilde{p}_1\widetilde{p}_2|$ and $\Delta\widetilde{\varphi} = |\widetilde{\varphi}_2 - \widetilde{\varphi}_1|$ (with the convention, of course, that $-\pi \leq \widetilde{\varphi}_2 - \widetilde{\varphi}_1 \leq \pi$).

We are now prepared to state the Separation Principle:

Theorem 1 (Separation Principle). *Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be distinct unfoldings in a common covering of generic trajectories γ_1 and γ_2, respectively, in $\mathcal{Q}_\varepsilon$. Suppose that γ_1 and γ_2 either have identical traces or are not equivalent. Let $\widetilde{\mathfrak{p}}_1 = (\widetilde{p}_1, \widetilde{\varphi}_1)$ and $\widetilde{\mathfrak{p}}_2 = (\widetilde{p}_2, \widetilde{\varphi}_2)$ be phase points of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, respectively, written in a common system of coordinates. Let L be the minimum of the lengths of γ_1 and γ_2. Then either*

$$\Delta\widetilde{p} > 2\varepsilon \quad \text{or} \quad \Delta\widetilde{\varphi} > \frac{2\varepsilon}{L}.$$

Remark. The well known Uncertainty Principle of quantum mechanics states that the position and velocity of a single particle cannot both be known with arbitrary precision. In our case, there are two particles whose position and velocity (direction) are in question, not one, and there is no uncertainty of measurement involved. Rather, the Separation Principle states that the two particles cannot be arbitrarily close in both respects. Nonetheless, we feel the formal analogy between the two principles is compelling.

Remark. In informal terms, what the Separation Principle says is that, in phase space, non-equivalent trajectories on a table with pockets must be somewhat spaced apart from each other, and each individual trajectory must be spaced out so it does not pass by itself too closely; moreover, the spacing increases with the size of the pockets.

2.3. Generalization to Polyhedral Surfaces. Let us now consider the more general situation of billiards on a Euclidean polyhedral surface, that is, a surface which is the union of Euclidean polygonal faces with certain pairs of edges identified. A billiard trajectory on a polyhedral surface is a trajectory which is (locally) geodesic on the interior and obeys the billiard law at the boundary; as in the polygonal case, if a trajectory hits a vertex (either in the interior or on the boundary) it terminates. Since the geometry on each face is that of the Euclidean plane, it is easy to see that when a trajectory passes through an interior edge, its angle of refraction equals its angle of incidence. Given a polyhedral surface $\mathcal{S}$, let $\mathcal{S}_\varepsilon$ denote the surface obtained by removing from each face of $\mathcal{S}$ a pocket of radius ε centered at each vertex. If S is compact, the result of [4] applies: every non-terminating trajectory on $\mathcal{S}_\varepsilon$ is periodic.

All of the concepts previously discussed extend naturally to this more general setting with only minor modifications. Indeed, the only significant changes are as follows: In the process of unfolding, we begin with a copy of the face containing the initial point, and if the trajectory passes to an adjacent face, we attach a copy of the new face (in the plane), identifying the common edge crossed by the trajectory. If the trajectory comes to an edge on the boundary of the surface, we reflect the polygon containing the corresponding point of the unfolding along the corresponding edge. More generally, a covering of a polyhedral surface $\mathcal{S}$ is defined as a connected Riemann surface obtained by identifying the common edges of a collection of polygons, each a copy of a face of $\mathcal{S}$, such that any one of these polygons is obtained from any other by a sequence of the operations just described. For billiards on a surface in which a pair of faces shares more than edge, the code of a periodic trajectory must indicate on which face the trajectory travels between two consecutive edges when ambiguity would otherwise arise. An example of such a surface is a doubled polygon, that is a polyhedron with two faces. In this case, since the faces alternate, a single assignment of either "+" or "−" to the code, indicating on which of the two faces the trajectory lies when leaving the first edge in the code, is sufficient. (The sign would, of course, switch with each cyclic permutation.) Note that trajectories whose codes have the same sequence of edges but different signs are *not* equivalent.

The *phase space* of the billiard system on $\mathcal{S}$, denoted $\mathfrak{P}(\mathcal{S})$, is the quotient space of the unit tangent bundle over $\mathcal{S} \setminus V$ obtained by identifying the points with the same position on an edge of the boundary of $\mathcal{S}$ whose directions are reflections of each other in that edge; in particular, if $\partial\mathcal{S} = \emptyset$, the phase space is just the unit tangent bundle over $\mathcal{S} \setminus V$. If S is compact with area A, then $\mathfrak{P}(\mathcal{S})$ has finite volume $2\pi A$.

The separation principle remains true in this broader context, from which our bound on the number of trajectories again follows:

Theorem 1 (Separation Principle, general statement). *Let S be any compact Euclidean polyhedral surface. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be distinct unfoldings in a common covering of generic trajectories γ_1 and γ_2, respectively, in $\mathcal{S}_\varepsilon$. Suppose that γ_1 and γ_2 either have identical traces or are not equivalent. Let $\widetilde{\mathfrak{p}}_1 = (\widetilde{p}_1, \widetilde{\varphi}_1)$ and $\widetilde{\mathfrak{p}}_2 = (\widetilde{p}_2, \widetilde{\varphi}_2)$ be phase points of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, respectively, written in a common system of coordinates. Let L be the minimum of the lengths of γ_1 and γ_2. Then either*

$$\Delta\widetilde{p} > 2\varepsilon \quad or \quad \Delta\widetilde{\varphi} > \frac{2\varepsilon}{L}.$$

Theorem 2 (General statement). *Given a compact Euclidean polyhedral surface $\mathcal{S}$, let k be the number of equivalence classes of auto-reversing trajectories on $\mathcal{S}_\varepsilon$, and let l be the number of remaining equivalence classes of trajectories. Then*

$$k + 2l < \frac{\pi^2}{2} \cdot \frac{A}{a},$$

where A is the area of $\mathcal{S}$ and $a = \pi\varepsilon^2$.

In particular, the total number of orbit types on $\mathcal{S}_\varepsilon$ is strictly less than $\frac{\pi^2}{2} \cdot \frac{A}{a}$.

Since a trajectory with no reflections at the boundary cannot coincide with its reverse, we obtain in the case $\partial S = 0$ a bound on the number of trajectories having the smaller coefficient of proportionality $\frac{\pi^2}{4}$.

Corollary. *For any compact Euclidean polyhedral surface $\mathcal{S}$ such that $\partial S = \emptyset$, the number of orbit types on $\mathcal{S}_\varepsilon$ is **strictly less** than*

$$\frac{\pi^2}{4} \cdot \frac{A}{a}.$$

Remark. It might be tempting to try to use the smaller bound for a surface without boundary to improve the bound in the general case; however, this fails. If one doubles a surface to eliminate the boundary, the area doubles but the number of trajectories does not: each trajectory on the original surface lifts to two trajectories on the doubled surface (which are reflections of each other in the obvious inversion), but exactly in the case of an auto-reversing trajectory, these lifts are equivalent. Thus the bound cannot be improved by this approach, but, rather, exactly the same result is obtained.

3. Propositions and Preliminary Lemmas

For completeness, we include the proof of the proposition that two equivalent generic trajectories have the same period, thus fully clarifying the nature of equivalence, and also that a trajectory is auto-reversing if and only if it has a right angle reflection. We then continue with some elementary lemmas which play a vital rule in proving the Separation Principle and Theorem 2. As all of these results are intuitively believable, the reader may wish to read only the statements and move on to the proof of the Separation Principle before coming back, if desired, for the technical details.

Proposition 1. *Equivalent generic trajectories have the same period.*

Proof. Let γ_1 and γ_2 be equivalent. Without loss of generality, we may choose initial points $\mathfrak{p}_1$ and $\mathfrak{p}_2$ for γ_1 and γ_2, respectively, which lie on a common side e of $\mathcal{Q}$ and have the same direction. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be unfoldings of γ_1 and γ_2 having initial points $\widetilde{\mathfrak{p}}_1$ and $\widetilde{\mathfrak{p}}_2$ covering $\mathfrak{p}_1$ and $\mathfrak{p}_2$, respectively, with positions $\widetilde{p}_1$ and $\widetilde{p}_2$ lying on a common edge $\widetilde{e}$. Since γ_1 and γ_2 are equivalent, the unfoldings $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ clearly share the same corridor. Furthermore, since the diameter of $\mathcal{Q}$ is finite, in order for $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ to share a common corridor their traces must be parallel.

After one period, starting from initial phase point $\widetilde{\mathfrak{p}}_1$, a billiard particle traveling on $\widetilde{\gamma}_1$ has position and direction represented by a phase point $\widetilde{\mathfrak{p}}_1'$ which again covers $\mathfrak{p}_1$. In particular, the edge $\widetilde{e}'$ containing $\widetilde{p}_1'$ must cover e, and it must be both parallel to $\widetilde{e}$ and oriented in the same direction if the corresponding endpoints are taken in the same order. (Note: At this point we require the fact that transverse orientation is preserved after a period.) Since the traces of $\widetilde{\gamma}_2$ and $\widetilde{\gamma}_1$ are parallel, after traveling the same distance (that is, the length of γ_1) from $\widetilde{p}_2$ in direction $\widetilde{\varphi}_2$, a particle on $\widetilde{\gamma}_2$ will be characterized by a phase point with position on $\widetilde{e}'$ which covers $\mathfrak{p}_2$. Thus, γ_2 has completed a whole number of periods; therefore, the period of γ_2 is less than or equal to that of γ_1. Reversing the roles of γ_1 and γ_2, we see that their periods are the same. □

Proposition 2. *Let γ be a trajectory, and let $-\gamma$ be the reverse of γ. If γ has no right angle reflection, then the traces of γ and $-\gamma$ in phase space are disjoint. If γ does have a right angle reflection, then the traces of γ and $-\gamma$ coincide.*

Proof. If γ does have a right angle reflection, it is clear that γ and $-\gamma$ coincide. Conversely, the traces of γ and $-\gamma$ intersect in phase space if and only if there are times $t_1 < t_2$ such that $\gamma(t_1)$ and $\gamma(t_2)$ have the same position and opposite directions. Following γ backwards from t_2 and forwards from t_1, we observe that for any real value Δt, $\gamma(t_1 + \Delta t)$ and $\gamma(t_2 - \Delta t)$ continue to have the same position and opposite directions; in particular, γ attains two opposite directions at time $t_{ave} = \frac{t_1+t_2}{2}$. This is only possible if γ reflects from an edge at right angles at time t_{ave}. □

The first six of the lemmas which follow concern the distances between points in a covering. The seventh places a lower bound on the length of a trajectory in $\mathcal{Q}_\varepsilon$, an important consideration in the formulation and proof of the Separation Principle.

Lemma 1. *Let $\widetilde{\mathcal{Q}}$ be a covering of $\mathcal{Q}$. Let $\widetilde{p} \in \widetilde{\mathcal{Q}}$ be a point on the trace of an unfolding of a trajectory in $\mathcal{Q}_\varepsilon$, and let $\widetilde{v} \in \widetilde{V}$. Then $|\widetilde{p}\,\widetilde{v}| > \varepsilon$.*

Proof. Let $\widetilde{\rho}$ be the shortest path from $\widetilde{p}$ to $\widetilde{v}$. Without loss of generality, assume $\widetilde{v}$ is the only vertex on this path. (If not, then $\widetilde{v}$ is further from $\widetilde{p}$ than any other vertex on the path, in particular, the one closest to $\widetilde{p}$.)

First suppose that segment $\widetilde{p}\,\widetilde{v}$ intersects an edge of $\widetilde{\mathcal{Q}}$; let $\widetilde{e}$ be the edge closest to $\widetilde{v}$. Then $\widetilde{v}$ is a vertex of a polygon containing $\widetilde{e}$; hence, the pocket centered at $\widetilde{v}$ is disjoint from $\widetilde{e}$. It follows that $|\widetilde{p}\,\widetilde{v}| > \varepsilon$. (See Fig. 8a.) On the other hand, if no edge intersects $\widetilde{p}\,\widetilde{v}$ then this segment lies in a single polygon. Since $\widetilde{p}$ lies outside the pocket centered at $\widetilde{v}$, it again follows that $|\widetilde{p}\,\widetilde{v}| > \varepsilon$. (See Fig. 8b.) □

An immediate and useful consequence of Lemma 1 is that if $\widetilde{p}_1$ and $\widetilde{p}_2$ lie on traces of unfoldings of trajectories in $\mathcal{Q}_\varepsilon$ and the shortest path between them contains a vertex, then $|\widetilde{p}_1\widetilde{p}_2| > 2\varepsilon$.

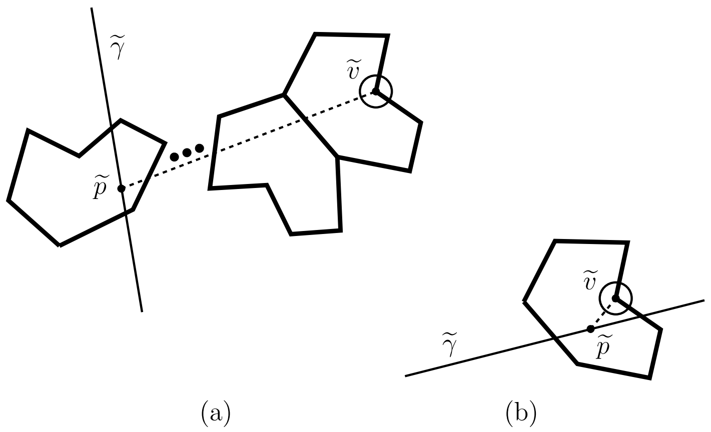


Fig. 8. $|\widetilde{p}\,\widetilde{v}| > \varepsilon$

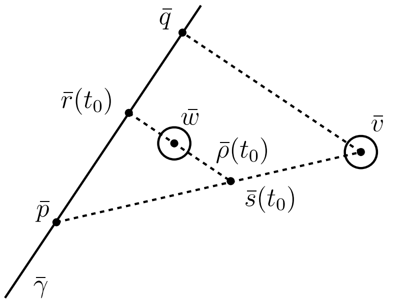


Fig. 9. The distance from $\bar{v}$ to $\bar{\gamma}$ is greater than ε

Lemma 2. *Let $\widetilde{Q}$ be a covering of Q. Let $\widetilde{\gamma}$ be an unfolding on $\widetilde{Q}$ of a trajectory on Q_ε, and let $\widetilde{v}$ be a vertex of $\widetilde{Q}$. If the trace of $\widetilde{\gamma}$ contains a point $\widetilde{p}$ such that the shortest path from $\widetilde{p}$ to $\widetilde{v}$ is a segment, then the (orthogonal) distance from $\bar{v}$, the projection of $\widetilde{v}$, to $\bar{\gamma}$, the projection of the trace of $\widetilde{\gamma}$, is greater than ε.*

Proof. Let $\bar{q}$ be the foot of the perpendicular from $\bar{v}$ to $\bar{\gamma}$, and let $\widetilde{q}$ be the point on $\widetilde{\gamma}$ lying above $\bar{q}$. For $t \in [0, 1]$, let $\widetilde{r}(t)$ be the point on $\widetilde{\gamma}$ at distance $t|\widetilde{p}\,\widetilde{q}\,|$ from $\widetilde{p}$ in the direction of $\widetilde{q}$, and let $\widetilde{s}(t)$ be the point on $\widetilde{p}\,\widetilde{v}$ at distance $t|\widetilde{p}\,\widetilde{v}|$ from $\widetilde{p}$ in the direction of $\widetilde{v}$. Let $\widetilde{\rho}(t)$ be the shortest path in $\widetilde{Q}$ from $\widetilde{r}(t)$ to $\widetilde{s}(t)$, and let t_0 be the smallest value of t for which $\widetilde{\rho}(t)$ contains a vertex. (The value t_0 exists since the set of values $t \in [0, 1]$ for which $\widetilde{\rho}(t)$ contains a vertex is clearly non-empty, since $\widetilde{\rho}(1)$ contains $\widetilde{s}(1) = \widetilde{v}$, and closed.)

Let $\widetilde{w}$ be a vertex of $\widetilde{\rho}(t_0)$. Its projection, $\bar{w}$, lies inside or on the triangle with vertices $\bar{p}, \bar{q}$ and $\bar{v}$. Furthermore, $\bar{r}(t_0)\bar{w} \parallel \bar{q}\bar{v}$. It follows from Lemma 1 that $|\bar{q}\bar{v}| \geq |\bar{r}(t_0)\bar{w}| > \varepsilon$. (See Fig. 9.) □

Lemma 3. *Let $\widetilde{Q}$ be a covering of Q. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be unfoldings on $\widetilde{Q}$ of trajectories on Q_ε such that $\bar{\gamma}_1 \parallel \bar{\gamma}_2$ but $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not share the same corridor. If the traces of $\widetilde{\gamma}_1$*

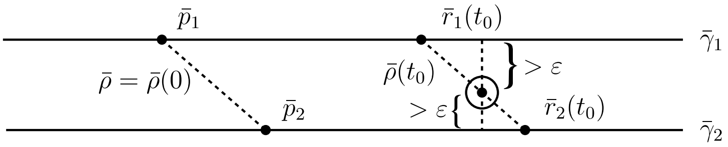


Fig. 10. The distance between $\bar{\gamma}_1$ and $\bar{\gamma}_2$ is greater than 2ε

and $\widetilde{\gamma}_2$ contain points $\widetilde{p}_1$ and $\widetilde{p}_2$, respectively, such that the shortest path from $\widetilde{p}_1$ to $\widetilde{p}_2$ is a segment, then the distance between $\bar{\gamma}_1$ and $\bar{\gamma}_2$ is greater than 2ε.

Proof. By reversing the direction of one of the trajectories, if necessary, we may assume without loss of generality that the projections of points forward from $\widetilde{p}_1$ and $\widetilde{p}_2$ on their respective trajectories lie on the same side of line $\overleftrightarrow{\bar{p}_1\bar{p}_2}$. For $t \in [0, \infty)$, let $\widetilde{r}_1(t)$ and $\widetilde{r}_2(t)$ be the points on $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ at distance t from $\widetilde{p}_1$ and $\widetilde{p}_2$, respectively, in the forward direction. Let $\widetilde{\rho}(t)$ be the shortest path in $\widetilde{Q}$ from $\widetilde{r}_1(t)$ to $\widetilde{r}_2(t)$. For some value of t, $\widetilde{\rho}(t)$ contains a vertex, else $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ would share the same corridor. (Every edge which intersected $\widetilde{\gamma}_1$ would also intersect $\widetilde{\gamma}_2$, and vice-versa.)

Let t_0 be the smallest such value. (It is possible, of coure, that $t_0 = 0$.) Then $\widetilde{\rho}(t_0)$ is a segment containing a vertex. The projection of this vertex lies in the "strip" bounded by $\bar{\gamma}_1$ and $\bar{\gamma}_2$, and by Lemma 2, its distance from each of them is greater than ε. The conclusion follows. (See Fig. 10.) □

Observe that even if the projections onto the plane of two unfoldings are not parallel, the unfoldings themselves, which lie on a Riemann surface, may not intersect. The following lemma shows that if this situation occurs for unfoldings which lift trajectories on Q_ε, then points lying on them must be more than 2ε apart.

Lemma 4. *Let $\widetilde{Q}$ be a covering of Q. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be unfoldings on $\widetilde{Q}$ of trajectories on Q_ε such that $\bar{\gamma}_1 \not\parallel \bar{\gamma}_2$ but $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not intersect. Then $|\widetilde{p}_1\widetilde{p}_2| > 2\varepsilon$ for all points $\widetilde{p}_1$ and $\widetilde{p}_2$ on $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, respectively.*

Proof. Given two points, $\widetilde{p}_1$ and $\widetilde{p}_2$, let $\widetilde{\rho}$ denote the shortest path between them in $\widetilde{Q}$, and let $\bar{\rho}$ denote its projection onto the plane. Let $\bar{q}$ be the point at which lines $\bar{\gamma}_1$ and $\bar{\gamma}_2$ intersect, and let $\widetilde{q}_1$ and $\widetilde{q}_2$ be the lifts of $\bar{q}$ to the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, respectively.

If $\widetilde{\rho}$ contains a vertex, then $|\widetilde{p}_1\widetilde{p}_2| > 2\varepsilon$ by the observation following Lemma 1. On the other hand, suppose $\widetilde{\rho}$ contains no vertex. For $t \in [0, 1]$, let $\widetilde{r}_1(t)$ be the point on the trace of $\widetilde{\gamma}_1$ between $\widetilde{p}_1$ and $\widetilde{q}_1$ and at a distance of $t|\widetilde{p}_1\widetilde{q}_1|$ from $\widetilde{p}_1$. (Hence $\widetilde{r}_1(0) = \widetilde{p}_1$ and $\widetilde{r}_1(1) = \widetilde{q}_1$.) Similarly define $\widetilde{r}_2(t)$ on the trace of $\widetilde{\gamma}_2$, and let $\widetilde{\rho}(t)$ be the shortest path from $\widetilde{r}_1(t)$ to $\widetilde{r}_2(t)$.

Since $\widetilde{q}_1 \neq \widetilde{q}_2$, $|\widetilde{q}_1\widetilde{q}_2| \neq 0$; hence, for some value $t < 1$, $\widetilde{\rho}(t)$ contains a vertex. Let t_0 be the smallest such value. Then $\widetilde{\rho}(t_0)$ is a segment whose length is clearly less than or equal to that of $\widetilde{\rho}$. (See Fig. 11.) It follows that $|\widetilde{p}_1\widetilde{p}_2| > 2\varepsilon$. □

Lemma 5. *Let $\widetilde{Q}$ be a covering of Q. Let $\widetilde{v}$ and $\widetilde{w}$ be any two (distinct) vertices of $\widetilde{Q}$. If Q admits pockets of radius ε, then $|\widetilde{v}\,\widetilde{w}| > 2\varepsilon$.*

Proof. Without loss of generality, assume that the shortest path from $\widetilde{v}$ to $\widetilde{w}$ contains no vertices in its interior. If this path intersects an edge of $\widetilde{Q}$, then the distance from $\widetilde{v}$ to

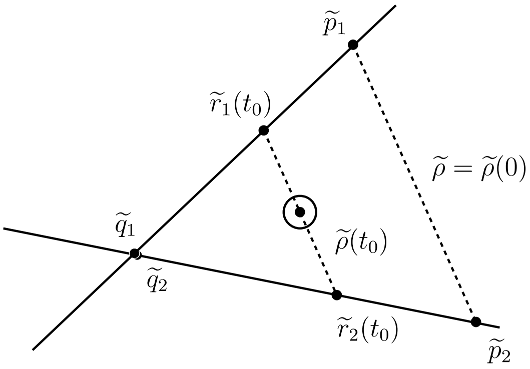


Fig. 11. Segment $\widetilde{\rho}(t_0)$

the point of intersection with the edge nearest $\widetilde{v}$ must be greater than ε, since this edge is in the same polygon as $\widetilde{v}$ and, hence, must be disjoint from the pocket centered at $\widetilde{v}$. Similarly, the distance from $\widetilde{w}$ to the point of intersection with the edge nearest $\widetilde{v}$ must be greater than ε. On the other hand, if segment $\widetilde{v}\,\widetilde{w}$ intersects no edge, then $\widetilde{v}$ and $\widetilde{w}$ lie in the same polygon, so the pockets centered at $\widetilde{v}$ and $\widetilde{w}$ must be disjoint. In either case, it is clear that $|\widetilde{v}\,\widetilde{w}| > 2\varepsilon$. □

Lemma 6. *Let $\widetilde{Q}$ be a covering of Q. Let $\widetilde{v}$ be a vertex of $\widetilde{Q}$, and let $\widetilde{p}$ be a point on any edge of $\widetilde{Q}$ not containing $\widetilde{v}$. If Q admits pockets of radius ε, then $|\widetilde{v}\,\widetilde{p}| > \varepsilon$.*

Proof. Similar to that of the preceding lemma. □

Lemma 7. *Let γ be any trajectory on Q_ε. Let L be the length of γ. Then $L > 2\varepsilon$.*

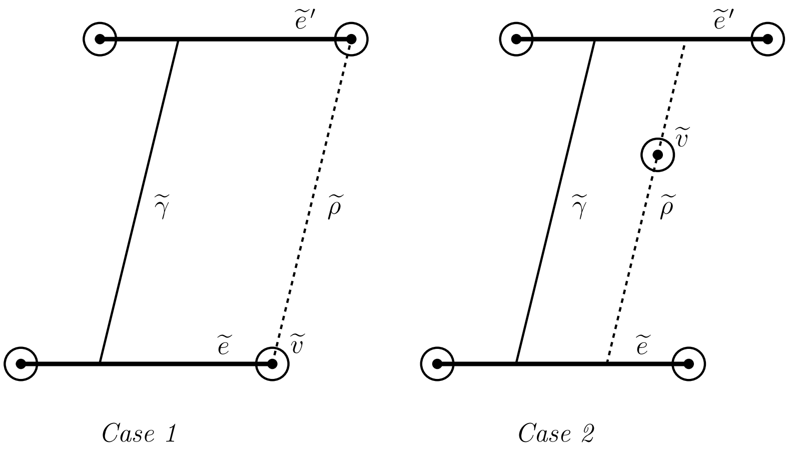
Proof. Let $\mathfrak{p}$ be a phase point of γ whose position lies on an edge e of Q. Let $\widetilde{\gamma}$ be an unfolding of γ, and let $\widetilde{\mathfrak{p}}$ and $\widetilde{\mathfrak{p}}'$ be phase points of $\widetilde{\gamma}$ which cover $\mathfrak{p}$ and whose positions are distance L apart. Then $\widetilde{p}$ and $\widetilde{p}'$ are corresponding points of parallel edges $\widetilde{e}$ and $\widetilde{e}'$, respectively, covering e. Let $\widetilde{v}$ be a vertex whose distance to segment $\widetilde{p}\,\widetilde{p}'$ is minimal, and let $\widetilde{\rho}$ be the segment parallel to $\widetilde{p}\,\widetilde{p}'$, and joining points of $\widetilde{e}$ and $\widetilde{e}'$, which passes through $\widetilde{v}$. (See Fig. 12.)

Case 1. Vertex $\widetilde{v}$ is an endpoint of edge $\widetilde{e}$ or $\widetilde{e}'$. Then $\widetilde{\rho}$ joins a pair of corresponding endpoints of $\widetilde{e}$ and $\widetilde{e}'$. It follows from Lemma 5 that $L > 2\varepsilon$.

Case 2. Vertex $\widetilde{v}$ does not lie on edge $\widetilde{e}$ or $\widetilde{e}'$. Then it follows from Lemma 6 that the segments on $\widetilde{\rho}$ joining $\widetilde{v}$ to $\widetilde{e}$ and $\widetilde{e}'$ have length greater than ε; hence, $L > 2\varepsilon$. □

4. Proof of the Separation Principle

The idea of the Separation Principle grew out of the proof of the first theorem of [2]. Although the proof given here is completely self-contained, the reader may find it instructive and motivational to consult the proof of the earlier theorem.



Case 1 *Case 2*

Fig. 12. The length of a trajectory on $\mathcal{Q}_\varepsilon$ is greater than 2ε

Proof of the Theorem 1 (Separation Principle). Let L_1 and L_2 be the lengths of γ_1 and γ_2, respectively, and assume without loss of generality that $L_1 \le L_2$, so $L = L_1$. The proof divides into two cases.

Case 1. $\Delta\widetilde{\varphi} = 0$ or $\Delta\widetilde{\varphi} = \pi$. Since $\pi > \frac{2\varepsilon}{L}$ by Lemma 7, we may assume that $\Delta\widetilde{\varphi} = 0$; we will prove that $\Delta\widetilde{p} > 2\varepsilon$. Since $\Delta\widetilde{\varphi} = 0$, the projections onto the plane of the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ are parallel.

Let $\widetilde{\rho}$ be the shortest path from $\widetilde{p}_1$ to $\widetilde{p}_2$. If $\widetilde{\rho}$ contains a vertex then, as previously observed, $\Delta\widetilde{p} > 2\varepsilon$, and we are done. On the other hand, suppose $\widetilde{\rho}$ is a segment containing no vertex. By Lemma 3, it suffices to prove that $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not share the same corridor, since clearly $|\widetilde{p}_1\widetilde{p}_2|$ is at least as great as the distance between $\bar{\gamma}_1$ and $\bar{\gamma}_2$. Thus the result in this case follows from:

Claim. The unfoldings $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not share the same corridor.

Suppose to the contrary that $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ share the same corridor. Without loss of generality, choose initial points $\widetilde{\mathfrak{p}}_1$ and $\widetilde{\mathfrak{p}}_2$, respectively, whose positions lie on a common edge $\widetilde{e}$.

The trajectories γ_1 and γ_2 are clearly equivalent. Equivalent trajectories satisfy the hypothesis of the theorem only if they have the same trace. Furthermore, by assumption γ_1 and γ_2 traverse their common trace in the same direction; that is, $\gamma_1 = \gamma_2 = \gamma$ (except for the choice of initial point). We may therefore consider the phase point of $\widetilde{\gamma}_1$ nearest $\widetilde{\mathfrak{p}}_1$ in the forward direction that covers $\mathfrak{p}_2$; denote this point by $\widetilde{\mathfrak{p}}_2'$. Because the two unfoldings are parallel, $\widetilde{p}_2'$ lies on an edge $\widetilde{e}'$ of $\widetilde{\mathcal{Q}}$ which is parallel to the edge $\widetilde{e}$ (and also covers e). (See Fig. 13.)

The trace of the unfolding $\widetilde{\gamma}_2$ intersects $\widetilde{e}'$ at a point $\widetilde{p}_3'$ whose distance from $\widetilde{p}_2'$ is $\Delta\widetilde{p}$, the same as the distance from $\widetilde{p}_1$ to $\widetilde{p}_2$. Let $\widetilde{p}_3$ be the translation by vector $\overrightarrow{\widetilde{p}_2'\widetilde{p}_2}$ of $\widetilde{p}_3'$. Then $\widetilde{p}_3$ is a point of edge $\widetilde{e}$ at distance $\Delta\widetilde{p}$ from $\widetilde{p}_2$ and on the opposite side of $\widetilde{p}_2$ from $\widetilde{p}_1$. Let $\widetilde{\gamma}_3$ denote the lift of γ parallel to $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ which passes through $\widetilde{p}_3$; the trace of $\widetilde{\gamma}_3$ intersects ray $\overrightarrow{\widetilde{p}_2'\widetilde{p}_3'}$ at a point $\widetilde{p}_4'$ whose distance from p_3' is $\Delta\widetilde{p}$.

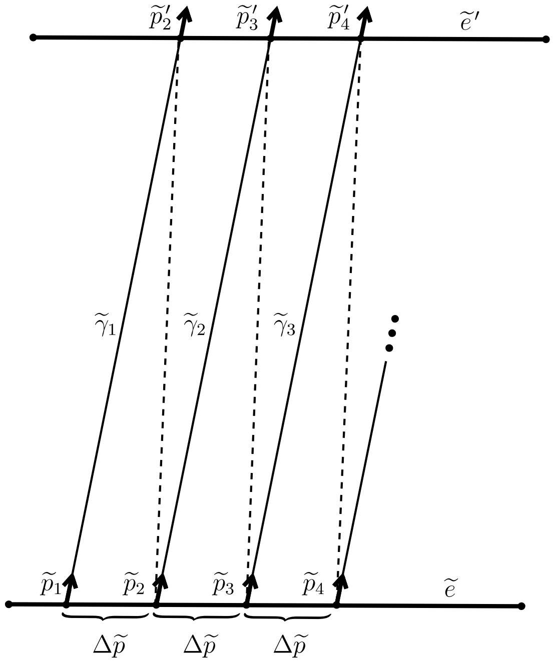


Fig. 13. Repeatedly laying off $\widetilde{p}_1\widetilde{p}_2$ on edge $\widetilde{e}$

Furthermore, since every point lying between the traces of $\widetilde{\gamma}_2$ and $\widetilde{\gamma}_3$ covers the same point of Q as its translate by the vector $\overrightarrow{\widetilde{p}_2\widetilde{p}_2'}$, which lies between the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, no vertex lies between $\widetilde{\gamma}_2$ and $\widetilde{\gamma}_3$. Thus $\widetilde{\gamma}_3$ also lies in the corridor containing $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, and $\widetilde{p}_4'$ lies on edge $\widetilde{e}'$. Let $\widetilde{p}_4$ be the corresponding point on edge $\widetilde{e}$. (See Fig. 13.)

Continuing in this fashion, we obtain an infinite sequence of points $\widetilde{p}_1, \widetilde{p}_2, \widetilde{p}_3, \widetilde{p}_4, \ldots$ on edge $\widetilde{e}$ spaced the fixed distance $\Delta\widetilde{p}$ apart, which contradicts the obvious fact that the length of edge $\widetilde{e}$ is finite. We conclude that $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not share the same corridor.

Case 2. $\Delta\varphi \neq 0, \pi$. By Lemma 4, if the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ do not intersect, then $\Delta p > 2\varepsilon$. Thus we have reduced to the case that the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ intersect. We will show in this case that $\Delta\varphi > \frac{2\varepsilon}{L}$.

Denote by $\widetilde{p}$ the point at which the two unfoldings intersect and, without loss of generality, assume $\widetilde{p}_1 = \widetilde{p}_2 = \widetilde{p}$. Recall that $L = L_1$, the length of γ_1, is the minimum of the lengths of γ_1 and γ_2. Let n be the period of γ_1, and let Q_0 be the polygon of $\widetilde{Q}$ containing $\widetilde{p}$. Consider, in the corridor of $\widetilde{\gamma}_1$, the polygons $Q_0, Q_1, Q_2, \ldots, Q_n$, and let $\widetilde{p}'$ be the point on $\widetilde{\gamma}_1$ at distance L in the forward direction from $\widetilde{p}$. Since n is the period of γ_1, Q_n is a parallel translate of Q_0 by distance L in the direction of $\widetilde{\gamma}_1$. Therefore, there is an unfolding $\widetilde{\gamma}_2'$ of γ_2 through $\widetilde{p}'$ whose projection onto the plane is parallel to $\bar{\gamma}_2$. (See Fig. 14.)

Simple trigonometry shows that the distance between the projections $\bar{\gamma}_2$ and $\bar{\gamma}_2'$ is $L \sin \Delta\varphi < L\Delta\varphi$. Furthermore, by the claim proven in Case 1, $\widetilde{\gamma}_2$ and $\widetilde{\gamma}_2'$, being lifts of the same trajectory, cannot share the same corridor. Finally, segment $\widetilde{p}\,\widetilde{p}'$ lies in $\widetilde{Q}$. Thus,

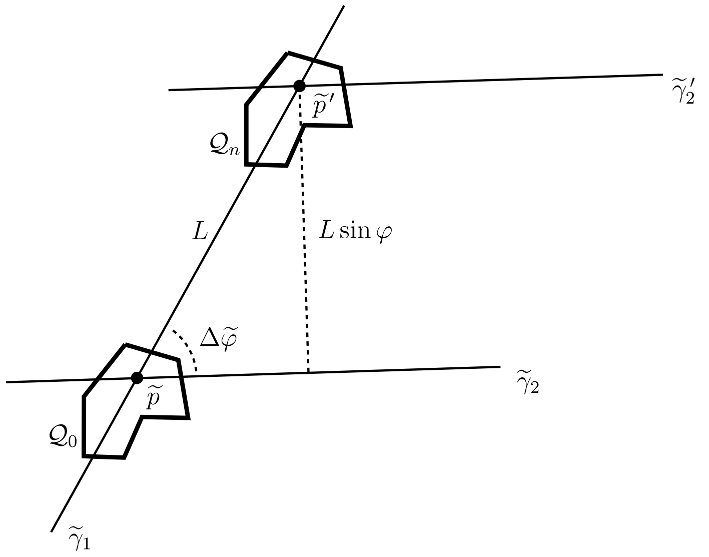


Fig. 14. Two parallel unfoldings of γ_2

Lemma 3 implies that the distance between $\bar{\gamma}_2$ and $\bar{\gamma}_2'$ is greater than 2ε. Combining this inequality with the one above, we obtain $L\Delta\varphi > 2\varepsilon$; hence, $\Delta\varphi > \frac{2\varepsilon}{L}$. □

Remark. Given a generic trajectory γ with length L, the argument of Case 2, along with the fact that $\sin\Delta\varphi < 1$, provides an alternative means of seeing that $L > 2\varepsilon$ (see Lemma 7) in the case that some trajectory (possibly γ itself) intersects γ transversely. However, there is no guarantee that such a trajectory exists.

5. Proof of Theorem 2

For each trajectory γ in $\mathcal{Q}$, we consider the corresponding trajectory (which we will also denote by γ, as the context will be clear) in $\mathfrak{P}(\mathcal{Q})$, the phase space of $\mathcal{Q}$, given by its position and direction. We think of the directional "axis" of the phase space as vertical and oriented so that the values of $[0, 2\pi]$ increase in the upward direction. Let L be the length of γ. (This length is independent of whether we regard γ as being in $\mathfrak{P}(\mathcal{Q})$ or $\mathcal{Q}$, since its image in $\mathfrak{P}(\mathcal{Q})$ is horizontal.) As always, we assume that γ is parametrized at unit speed.

From this point on, we fix ε and assume γ is a non-terminating generic trajectory on $\mathcal{Q}_\varepsilon$. We will, in a natural way, map a Euclidean solid torus T with core of length L and cross-section $[-\varepsilon, \varepsilon] \times [-\frac{\varepsilon}{L}, \frac{\varepsilon}{L}]$ into phase space so that its core maps onto γ and its image forms a neighborhood of γ. Clearly, the volume of T is $4\varepsilon^2$.

For each $t \in \mathbb{R}$, let $\tau(t)$ be the direction of γ at time t, let β be the upward pointing unit vector, and let $\nu(t)$ be the horizontal unit vector normal to γ at time t such that the local basis $(\beta, \nu(t), \tau(t))$ is positively oriented. Given a point $(t, x, y) \in T$ (where T is parametrized as $[0, L] \times [-\varepsilon, \varepsilon] \times [-\frac{\varepsilon}{L}, \frac{\varepsilon}{L}]$ with each point $(0, x, y)$ identified with the point (L, x, y)), let $r = \sqrt{x^2 + y^2}$, the distance of (t, x, y) from $(t, 0, 0)$, the origin of the cross-section at t. Define a map $f : T \to \mathfrak{P}(\mathcal{Q})$ by setting $f(t, x, y)$ to be the point at distance r from $\gamma(t)$ in the direction of $x\nu(t) + y\beta$.

We first prove that for each trajectory γ, the map f is an isometric embedding; hence, γ is surrounded by a tube in phase space of volume $4\varepsilon^2$. The map f is a local isometry by construction; therefore, it suffices to prove f is 1-1. The Separation Principal provides the essential tool for doing so, for it (along with Lemma 7, which shows that the height of T, $\frac{2\varepsilon}{L}$, is less than 2π) will show that the dimensions of T are sufficiently small that its image can have no self-intersections.

Lemma 8. *The map f is an isometry.*

Proof. Suppose that $f(t_1, x_1, y_1) = f(t_2, x_2, y_2)$. Let r_1 be the (oriented) straight line path from $(t_1, 0, 0)$ to (t_1, x_1, y_1), and let r_2 be the (oriented) straight line path from (t_2, x_2, y_2) to $(t_2, 0, 0)$. Let $\mathfrak{r}_i$ denote $f \circ r_i$, for $i = 1, 2$. The concatenation of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ is a connected path from $\mathfrak{p}_1 = \gamma(t_1)$ to $\mathfrak{p}_2 = \gamma(t_2)$, since $f((t_1, x_1, y_1)) = f((t_2, x_2, y_2))$; denote this path by $\mathfrak{r}$, and its projection onto the positional coordinate by ρ. (We similarly denote the projections of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ onto the positional coordinate by ρ_1 and ρ_2, respectively; ρ is the union of ρ_1 and ρ_2.) (See Fig. 15.)

Let $\widetilde{\rho}$ be an unfolding of ρ; denote its endpoints by $\widetilde{p}_1$ and $\widetilde{p}_2$, where $\widetilde{p}_1$ covers p_1, the position of $\mathfrak{p}_1$, and $\widetilde{p}_2$ covers p_2, respectively. Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be unfoldings of γ with initial phase points $\widetilde{\mathfrak{p}}_1$ and $\widetilde{\mathfrak{p}}_2$, covering $\mathfrak{p}_1$ and $\mathfrak{p}_2$ and having positions $\widetilde{p}_1$ and $\widetilde{p}_2$, respectively. Each path ρ_i has length less than or equal to ε, so $\Delta\widetilde{p} \leq 2\varepsilon$. Moreover, $\Delta\widetilde{\varphi} = |y_2 - y_1| \leq \frac{2\varepsilon}{L}$. Thus, by the Separation Principle, as $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ are unfoldings of the same trajectory lying in a common covering, $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ must coincide.

Noting that $\widetilde{\rho}_1$ is a segment perpendicular to the trace of $\widetilde{\gamma}_1$ and $\widetilde{\rho}_2$ is a segment perpendicular to the trace of $\widetilde{\gamma}_2$, we deduce that the traces of $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ coincide only if $\widetilde{p}_1 = \widetilde{p}_2$ and $\Delta\widetilde{\varphi} = 0$. Thus $|y_1 - y_2| = 0$ and $|x_1 - x_2| = \Delta p = 0$. Finally, since $\mathfrak{p}_1 = \mathfrak{p}_2$ and γ traverses a single period between $t = 0$ and $t = L$, $t_1 = t_2$. We conclude that f is 1-1. □

Next, we show that the tubes around non-equivalent trajectories are disjoint. Once again, the Separation Principle lies at the crux of the proof.

Lemma 9. *Let γ_1 and γ_2 be non-equivalent trajectories on $\mathcal{Q}_\varepsilon$, with tubular neighborhoods given as the images of maps $f_1 : T_1 \to \mathfrak{P}(\mathcal{Q})$ and $f_2 : T_2 \to \mathfrak{P}(\mathcal{Q})$, respectively, as described above. Then the images of f_1 and f_2 are disjoint.*

Proof. Suppose, to the contrary, that there are points $(t_1, x_1, y_1) \in T_1$ and $(t_2, x_2, y_2) \in T_2$ such that $f_1((t_1, x_1, y_1)) = f_2((t_2, x_2, y_2))$. By a method similar to that used in the proof of the preceding lemma, we obtain unfoldings $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ (which must have distinct traces, since γ_1 and γ_2 are not equivalent) with initial phase points $\widetilde{\mathfrak{p}}_1$ and $\widetilde{\mathfrak{p}}_2$, respectively, such that $\Delta\widetilde{p} \leq 2\varepsilon$ and $\Delta\widetilde{\varphi} \leq \frac{2\varepsilon}{L}$, in contradiction to the Separation Principle. □

The fact that $\frac{2\varepsilon}{L} < \pi$ suggests that the tubes around reverse trajectories will also be disjoint. Indeed this is true, and the key ingredient of the proof has already been incorporated into the Separation Principle. (See the proof of the Separation Principle, Case 1, which makes direct use of the fact that $\frac{2\varepsilon}{L} < \pi$ at the outset.)

Lemma 10. *Let γ be a trajectory on $\mathcal{Q}_\varepsilon$ which is not auto-reversing. Let $-\gamma$ denote its reverse. Then the tubes around γ and $-\gamma$ are disjoint.*

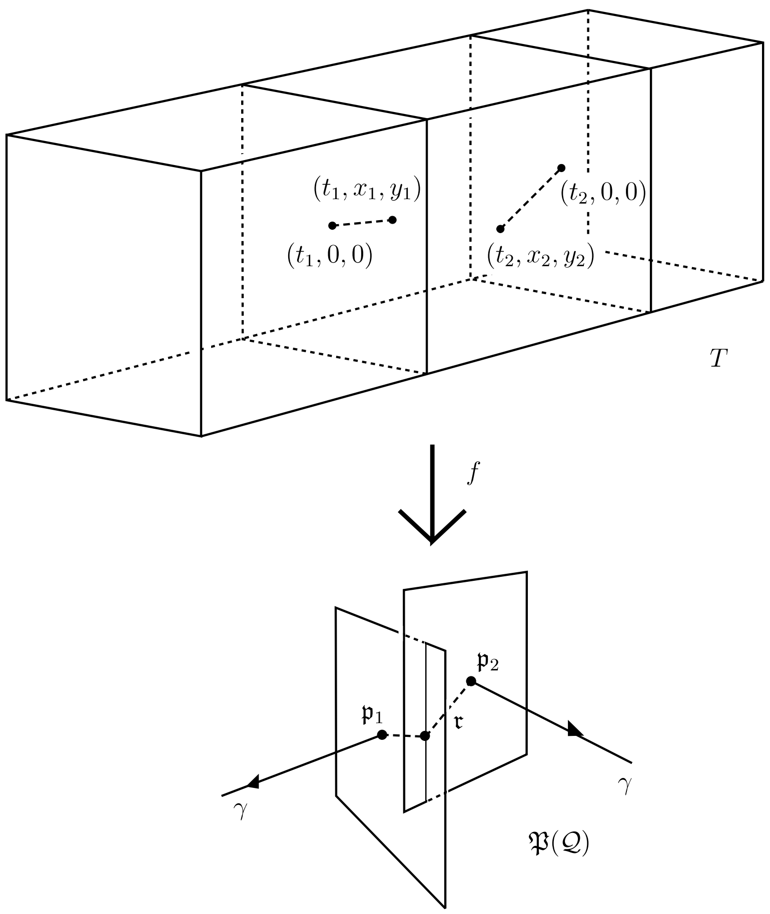


Fig. 15. The path $\mathfrak{r}$ from $\mathfrak{p}_1$ to $\mathfrak{p}_2$

Proof. Let $f^+ : T \to \mathfrak{P}(Q)$ and $f^- : T \to \mathfrak{P}(Q)$ be the maps into phase space whose images are the tubes around γ and $-\gamma$, respectively. (T is the solid torus obtained by identifying the ends of the rectangular solid $[0, L] \times [-\varepsilon, \varepsilon] \times [-\frac{\varepsilon}{L}, \frac{\varepsilon}{L}]$, where L is the common length of γ and $-\gamma$.) Suppose that $f^+(t_1, x_1, y_1) = f^-(t_2, x_2, y_2)$. Proceeding as in the proofs of Lemmas 8 and 9, we obtain unfoldings $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ with initial phase points $\widetilde{\mathfrak{p}}_1$ and $\widetilde{\mathfrak{p}}_2$, respectively, such that $\Delta\widetilde{p} \le 2\varepsilon$ and $\Delta\widetilde{\varphi} \le \frac{2\varepsilon}{L}$. Since γ and $-\gamma$ have the same trace, $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ must coincide (in both trace *and* direction), which implies that γ and $-\gamma$ coincide (in both trace *and* direction), in contradiction to the hypothesis that γ is not auto-reversing. □

The proof of Theorem 2. Applying Lemmas 8, 9 and 10 to a maximal collection of non-equivalent generic trajectories and their reverses, we obtain a pairwise disjoint family of tubes, each of volume $4\varepsilon^2$ and one for each trajectory, in the phase space of the billiard system. Since the volume of the phase space is $2\pi A$, the sum of the number of equivalence classes of auto-reversing trajectories and twice the number of equivalence classes of trajectories which are not auto-reversing can be no more than $\frac{2\pi A}{4\varepsilon^2} = \frac{\pi^2}{2} \cdot \frac{A}{a}$. □

Remark. Clearly the proof above requires no prior knowledge about the cardinality of the set of trajectories; thus, it is independent of the previous results of [2] and others regarding the cardinality of this set. (It is of course necessary to know a priori that all trajectories are periodic.)

References

1. Cipra, B., Hanson, R., Kolan, A.: Periodic trajectories in right triangle billiards. Preprint 1994
2. Delman, C., Galperin, G., Troubetskoy, S.: Finiteness of the set of orbit types for billiards in polygons with pockets. In preparation
3. Delman, C., Galperin, G., Troubetskoy, S.: A bound on the length and number of orbits for billiards in rational polygons with pockets. In preparation
4. Galperin, G., Krüger, T., Troubetskoy, S.: Local instability of orbits in polygonal and polyhedral billiards. Commun. Math. Phys. **169**, 463–473 (1995)
5. Gutkin, E.: Billiards in Polygons. Physica **19D**, 311–333 (1986)
6. Zemlyakov, A.N., Katok, A.B.: Topological transitivity of billiards in polygons. Mathematical Notes of the Academy of Sciences of the USSR **18**, 760–764 (1975)
7. Masur, H.: Lower bounds for the number of saddle connections and closed trajectories of a quadratic differential. Ergodic Theory and Dynamical Systems **10**(1), 151–176 (1990)
8. Vorobets, Y.B.: Ergodicity of billiards in polygons. Sbornik: Mathematics **188**(3), 389–434 (1997)

Communicated by P. Sarnak