Introduction to rational billiards

Emil Ledberg
Introduction to rational billiards

Emil Ledberg

June 1, 2018
Abstract

In this paper we introduce and illustrate some properties of billiards, later we will focus on rational polygonal billiards and develop tools and prove a theorem regarding them. We will also see these tools are well suited to study other related problems. The last section will include some intuition and counterexamples regarding existence of billiards with strange properties.
1 Introduction

1.1 Historical background

In 1963 Yakov Sinai published a paper in which he proved that the Sinai billiard is chaotic and ergodic. In the Sinai billiard a particle bounces around in a square with a circle cut out from the center. The Sinai billiard was the first dynamical system which was proved chaotic and ergodic [4]. A system is chaotic if small changes in initial values have significantly different behaviour and ergodic means that the particle pass through all available space. Since his publication the subject of billiards have continued to be studied and extended upon.

![Figure 1: The Sinai billiard](image)

1.2 Overview

The study of billiards is not a single confined theory or subject. The subject is too vast to be covered all at once and there are many different approaches to billiards and methods to study them. In his lectures Billiard table as a mathematician’s playground A.B. Katok says the following: "In a more serious sense, the expression “playground” should be understood as “testing grounds”: various questions, conjectures, methods of solution, etc. in the theory of dynamical systems are “tested” on various types of billiard problems." [3]. There are many known results concerning rational billiards. They are a subgroup of billiards which are reasonably easy to work with. As soon as we don’t restrict ourselves we come across simple questions whose answer is not known, e.g. "Do all triangular billiards have periodic orbits?". It have been proven all triangles with degrees less than 100° have periodic trajectories [5].

1.3 Definitions

**Definition 1.1.** A two dimensional billiard is a domain bounded by a piecewise smooth closed curve in $\mathbb{R}^2$ and a point-mass. The point moves with unit speed along straight lines and when it hits the boundary its tangential velocity remains the same and the normal velocity changes sign. In the case the ball hits the curve on a point where it is not smooth, the reflection is undefined and the motion terminates.

**Lemma 1.2.** When a billiard trajectory is reflected in the boundary the law of reflection holds i.e the angle of incidence is equal to the angle of reflection.
Proof. Let $\vec{v}_1, \vec{v}_2$ be the velocity vector of the particle before and after reflection, $\vec{u} = (u_1, u_2)$ and $\vec{w} = (w_1, w_2)$ be the tangential and normal components respectively such that $\vec{v}_1 = \vec{u} + \vec{w}$ and $\vec{v}_2 = \vec{u} - \vec{w}$. Let $\theta_1$ be the angle of incidence and $\theta_2$ be the angle of reflection. Then $\cos \theta_1 = \frac{||\vec{u}||}{||\vec{v}_1||} = \frac{||\vec{u}||}{||\vec{u}||} = \frac{||\vec{u}||}{||\vec{u}||}$ and because $\theta_1, \theta_2 \in [0, \frac{\pi}{2}]$ we can draw the conclusion $\theta_1 = \theta_2$. \hfill $\Box$

The study of billiards revolve around the general behaviour of this motion. This definition can be generalized to n-dimensional, non-euclidian space.

A rational polygonal billiard is a billiard whose boundary is a polygon whose angles are rational multiples of $\pi$.

1.4 Motivation

Billiards can be used to model mechanical systems but is also of interest to quantum physics, e.g the phase space of a one-dimensional mechanical system with two elastic particles can be represented as a two-dimensional billiard.

I came in contact with billiards when I was presented a counterexample to a question I had about non-dense, non-periodic geodesics on polyhedral surfaces. In fact, a billiard and its trajectories can be seen as geodesics on a flat polyhedron. To further empathize billiards as a "mathematicians playground" we will see that some of the tools we will develop in this paper are indeed very well suited to study geodesics on polyhedrons.

![Figure 2: A reflection in a straight line.](image1)

![Figure 3: An example of a billiard table on which there exists non-periodic trajectories that are not dense on the billiard. The angle of which the trajectory hits the boundary is $\frac{\pi}{4}$.](image2)
Figure 4: By elongating the L-shaped billiard table we can construct a polyhedron with geodesics with similar properties as the trajectories on the L-shaped billiard. Whenever we exit an L-shaped face we will spin around once and enter the opposite L-face at the same point. The geodesic on the L-shaped faces are going to be similar to that of the trajectory on the L-shaped billiard, non-periodic and non-dense.

1.5 Circular billiards

Figure 5: To the left a 4-periodic orbit in black and a 3-periodic orbit in red. To the right the closure of a non-periodic orbit in grey.

The circle billiard is perhaps the most simple. It has rotation symmetry and the trajectory always makes the same angle with the boundary at impact. If the angle at impact is $\alpha$ then the next impact point can be found by rotating the circle $2\alpha$. From this it is obvious that the trajectory is periodic if and only if $\alpha = \frac{p}{q}\pi$, $p, q \in \mathbb{Z}$ as the trajectory will return to a previous impact point after a while and the angle is the same.

**Lemma 1.3.** If $\alpha$ is not $\pi$-rational then the points of impact will be dense on the circle.

**Proof.** Let $p$ be our first point of impact. Let $d_1$ be the length of the arc between two points of impact. Let $x$ be the first point where the arclength between $x$ and $p$ is less than $d_1$. Let $y$ be the point with arclength $d_1$ from $x$ such that $p$ is on the arc $xy$. The arclength between $p$ and one of $x$ and $y$ is always going to be less than $d_1/2$, call this distance $d_2$. We can then apply the same argument with $d_2$ to find $d_3$ and so on, we note that $\lim_{n \to \infty} d_n = 0$. Any interval $I$ with length $l$ will contain a point of impact as we can find a small enough interval $d_n$ such that $d_n < l$. \hfill $\square$

**Lemma 1.4.** The closure of a non-periodic orbit in a circular billiard is an annulus with inner radius $r \cdot \cos \alpha$, where $r$ is the radius of the circle.
Proof. The trajectory can not be closer to the center than $r \cdot \cos \alpha$. Assume there is an open ball $B$ in the annulus with outer radius $r$ and inner radius $r \cdot \cos \alpha$ which is not in the closure of our open trajectory. Take two points $a, b \in B$ such that their clockwise tangent with the circle with radius $r \cdot \cos \alpha$ does not coincide. Project the segment $ab$ onto an arc on the outer circle by drawing the clockwise tangents to the smaller circle for each point on the segment $ab$. The intersection between this set of tangents will be two arcs, chose one of them. As seen in the previous proof there will be points that belong to the trajectory on this arc, thus there will be points on the segment $ab$ that belong to the trajectory which is a contradiction. The closure must be the entire annulus. 

1.6 Elliptic billiards

The behaviour of the trajectories in the ellipse can be categorized in three categories: focal, elliptic and hyperbolic.

If a trajectory travels through one of the focal points it will alternate travelling through the two focal points. Over time this motion will converge to the major axis. If a trajectory is traveling
clockwise with respect to the center then the clockwise arc length between two points of impact will always be shorter than half the circumference of the ellipse, two points of impact will always be on opposite sides of the major axis. This will make the points of impacts converge to the intersection between the ellipse and its major axis. The only periodical trajectory that travels through the focal points is the one that coincides with the major axis.

If the trajectory intersects the major axis between the focal points the trajectory will be bounded by a hyperbola with the same focal points, if the trajectory is non periodic the closure will be the entire region which is bounded by the ellipse and the hyperbola.

Similarly if the trajectory intersects the major axis outside of the segment between the focal points it will be bounded by a ellipse with the same focal points, again if it is non-periodic the closure will be the entire region between the two ellipses.

1.7 Rectangular billiards

The rectangular billiard is also a very simple one. We can break down this motion by analyzing its horizontal and vertical component individually. Because of the nature of reflections the horizontal component of the velocity will only be affected by reflection in the sides and the vertical in the top and bottom of the rectangle. When the particle hits the top or bottom the vertical velocity changes sign while the horizontal remains the same, if it hits either side the horizontal component will change sign and the vertical will be unaffected. Looking at periodic trajectories we see that a trajectory that have travelled 2b in the vertical direction will have the same vertical position and velocity as before. If we have travelled 2a in the horizontal direction we will have the same horizontal position and velocity.
Say that our starting point is \((x_1, y_1)\) and our initial velocity have the same direction as the slope \(y = kx\). After impact the velocity will be parallel to \(y = -kx\). To find periodic trajectories we want to know if the equation \(2b = 2ka \cdot t, t \in \mathbb{Q}\) have solutions. If we simplify we see that the equation have solutions iff \(\frac{a}{b}k\) is a rational. It is not hard to see that the trajectory is periodic iff \(\frac{a}{b}k\) is a rational number.

2 Triangular billiards

2.1 Periodic trajectories of acute triangles

A trajectory is periodic if there exist an \(h\) such that the particle have the same position and velocity at \(t\) and \(t + h\) for almost all \(t\). The period of a trajectory is the number of times it bounces before it returns to a previous visited point. In an acute triangle very easy to find periodic trajectories even if it is not rational. The periodic trajectory that connects the base points of the altitudes is called the Fagano billiard trajectory.

Lemma 2.1. The triangle connecting the base points of the altitudes in an acute triangle is a 3-periodic trajectory.

Proof. \(BCR\) and \(BAP\) are similar because they share \(\angle CBA\) and both have a right angle. Thus \(\frac{BP}{PA} = \frac{BR}{PC} \iff \frac{BC}{PA} = \frac{BR}{PC}\). This means \(BRP\) and \(BAC\) are all similar. Similarly we can conclude that the triangles \(ABC, PBR, PQC, AQR\) are similar. Thus \(\angle BAC = \angle RPB = \angle CPQ, \angle CBA = \angle PQC = \angle AQC, \angle ACB = \angle QRA = \angle BRP\). A particle following the red path will be reflected at \(Q, P, R\) as depicted in the picture and the triangle \(QRP\) is a billiard trajectory.

Later we will see that all trajectory have a family of trajectories travelling parallel to them. For now we will simply observe this and note that the black dashed "boundary" of the gray strip is a terminating trajectory in both directions.
2.2 Periodic trajectories in right angled triangles

For right triangles it the fagano trajectory terminates but we can still find its family of trajectories.

![Image of Right Triangle with Fagano Trajectory](image)

Figure 11: The degenerated Fagano trajectory in red and a blue 6-periodic trajectory. A gray "strip" of parallel trajectories is also illustrated.

**Lemma 2.2.** A trajectory perpendicular to the hypotenuse in a right angled triangle is periodic as long as it hits the other two sides before returning.

**Proof.** Assume a triangle $ABC$ and that $\angle ACB = 90^\circ$ and that a trajectory leaves the hypotenuse perpendicular at a point $P$. Without loss of generality we can assume it reflects in $CB$ then in $AC$ in point $Q$ and $R$ respectively. That means $\angle CQR = \angle PQB$ and $\angle QRC = \angle ARS$. Triangle $QBP$ is similar to $ABC$ which makes $QRC$ similar to $ABC$ which makes $ARS$ similar to $ABC$ so $\angle RSA = 90^\circ$ which means $PQRSRQ$ is a periodic trajectory.

3 The Zemlyakov-Katok construction

3.1 Rational angles

Let $P$ be an $n$-gon with edges $e_1, e_2, ..., e_n$ counter-clockwise. Whenever the particle is reflected in an edge its speed vector undergoes a transformation by an element in $O(2)$. Denote reflection in the edge $e_i$ by $r_i \in O(2)$. $O(2)$ is the orthogonal group of dimension 2 i.e. rotations and reflections in the plane. Let $\Gamma \subset O(2)$ be the set generated by the $e_i$’s. $\Gamma$ is the set of possible transformations of the velocity we may have during any trajectory.

**Theorem 3.1.** $\Gamma$ is finite if and only if the angles of the polygon $P$ is rational.

**Proof.** Assume $\Gamma$ is finite. $r_kr_{k+1}$ is a rotation by $2\angle e_k e_{k+1}$ clockwise $k \in 1, ..., n-1$. Because $\Gamma$ is finite there is some $m \in \mathbb{N}$ such that $(r_k r_{k+1})^{k+m} = (r_k r_{k+1})^k$ which means $(r_k r_{k+1})^m = 1$ (identity). Thus there is an $m$ such that $m \cdot 2\angle e_k e_{k+1} = n \cdot 2\pi$ and $\angle e_k e_{k+1}$ is rational. Replace $e_k$ and $e_{k+1}$ with $e_n$ and $e_1$ and repeat this argument.

Assume $\angle e_k e_{k+1} = \alpha_k$ and $\angle e_n e_1 = \alpha_n$ is rational for all $k \in \{1, 2, ..., n-1\}$. Extend $e_1$ and let it be our $x$-axis. All of the $e_i$’s will be parallel or make a rational angle with the $x$-axis, denote this angle $\frac{\pi}{q_i}$. Let $Q = \text{LCM}(q_1, q_2, ..., q_n)$. The matrix representation for reflection in the edges will be the same as for reflections in the the dihedral group of order $Q$. Thus $\Gamma \subset D_Q$ and $\Gamma$ is finite.
3.2 Unfoldning

Instead of reflecting the particle when we can reflect the table instead in a process called unfolding. This makes us able to study trajectories as straight lines instead.

Assume we have a rational polygon $P$ that we want to unfold. Whenever we unfold $P$ in an edge $e_i$ we transform $P$ with some element $r_i \in \Gamma$ and then translate it with some translation $c_i$. So what we are doing when we unfold is transforming $P$ with some $\rho_i = r_i + c_i$, an element in the group of Euclidian isometries. Because $\Gamma$ is finite we our constructed surface will consist of translations of all possible combinations of transformations of $P$ by $\Gamma$. Instead of continuing this unfolding process we can associate sides such that lines will keep their direction. Thus we achieve a finite flat area on which geodesic flow corresponds to billiard flow.

Figure 13: An unfolding of a triangle with angles $\pi/2, \pi/4, \pi/4$. It turns out it have the same geometry as a flat torus.
Translation surfaces

The Zemlyakov-Katok construction is an example of a translation surface. A translation surface is constructed by gluing parallel sides of polygons in the plane. We require edges to have opposite inward normals, otherwise direction will not be preserved. On these surfaces we can study trajectories as geodesic flows.

It may be tempting to try to define what a reflection at a vertex would look like as the geodesic flow is defined for all points in this particular unfolding. This is not in general possible, at vertices we may have cone points of negative curvature. The angle at these points is $2\pi n, n \in \{2, 3, ...\}$ and the geodesic flow is not defined at these points.

4.1 The first return map

Let $I$ be a closed horizontal segment and $\phi_t$ a vertical flow on a translation surface $S$. The first return map to $I$ is a map
\[
f : I \to I
\]
\[
p \mapsto \phi_{t_0}(p)
\]
where $t_0 \neq 0$ is the smallest $t$ for which $\phi_t(p) \in I$. But this map is not defined for all points.

We want to divide $I$ into a finite amount of open subintervals $I_j$ where the first return map is defined and continuous. Points for which the first return map is not defined are points where the vertical flow hits a cone point or does not return to $I$ at all. The number of vertical trajectories that can hit a cone point is small, a cone point with angle $2\pi n$ only have $n$ vertical trajectories hitting it forward in time, it have $n$ trajectories hitting it backward in time. The first return map have discontinuities at points where the vertical flow hits the endpoints of the interval $I$, two points close to each other may find themselves on opposite sides of a point whose vertical flow hits the boundary of $I$. Now we can divide $I$ into a finite amount of segments $I_j$ where the points stay together under the vertical flow until their first return to $I$.

**Lemma 4.1.** The first return map is defined for all $I_j$'s.

**Proof.** Assume there is some $p \in I_k$ for which the first return map is not defined. Because we have removed all cone points this means that the vertical flow from $p$ doesn’t return to $I$. This must be true for all points in $I_k$ as we have ensured, by removing all points of discontinuity, they must travel together. The area swept by $I_k$ under the vertical flow is finite as the area of our translation surface is finite and it must overlap with itself. That is there is some $0 < t_1 < t_2$ such that $\phi_{t_1}(I_k) \cap \phi_{t_2}(I_k) \neq \emptyset$ chose $p \in \phi_{t_1}(I_k) \cap \phi_{t_2}(I_k)$, then there is $p_1, p_2$ such that $\phi_{t_1}(p_1) = \phi_{t_2}(p_2) = p$. Now consider $\phi_{-t_1}(p) = \phi_{t_2-t_1}(p_2) = p_1$. This contradict our initial statement that there is no point in $I_k$ that return to $I_k$. \hfill \square

4.2 Boundaries of geodesics

With the help of the first return map we shall prove what have have been illustrated in previous figures.

**Theorem 4.2.** Every closed geodesic is part of a family of closed geodesics of the same length contained in an open cylinder whose boundary is a union of saddle connections.
Note: a saddle connection is a trajectory hitting singular points forward and backward in time.

**Proof.** Assume there is a point \( p \) with a closed geodesic passing through it. Choose as the vertical direction the direction of the geodesic and choose an appropriate horizontal interval \( I \) such that \( P \in I \). The geodesic through \( p \) will have some period \( n \in \mathbb{N} \) with respect to the first return map of \( I \), i.e. how many times it returns to \( I \) before returning to \( p \). Let \( B \) be the set of points for which the first return map is not defined and points that hit the endpoints of \( I \). The set 

\[
B = f^{-1}(B) \cup f^{-2}(B) \cup \ldots \cup f^{-n}(B)
\]

divides \( I \) into open intervals \( I_k \) that travel together until at least the \( n \)-th time they reach \( I \). If \( p \in I_a \) then all points in \( I_a \) have period \( n \). \( I_a \) is bounded by points \( s, t \) that forward in time hit cone points. If \( s, t \) does not hit a cone point backwards in time they must be periodic as they travel together with \( p \), which is a contradiction. \( s, t \) must hit cone points backwards in time and are saddle connections. If \( s \) hits different cone points \( s_1, s_2 \) forward and backward there is a geodesic parallel to the periodic geodesics that goes between \( s_1, s_2 \) or goes from \( s_1 \) to a new cone point \( s_3 \), in the latter case we can again say that there is a geodesic parallel to the periodic ones between \( s_3, s_2 \) or between \( s_3 \) and a new \( s_4 \). The amount of cone points is finite so this cannot happen an infinite amount of times.

**Theorem 4.3.** Given a non-periodic vertical trajectory \( \phi_t(p) \) in a translation surface \( S \), the closure of \( \{ \phi_t(p) | t \in \mathbb{R} \} \) is a subsurface of \( S \) whose boundary is a union of saddle connections.

**Proof.** Assume there is a boundary point \( q \), if the closure \( A \) of \( \{ \phi_t(p) | t \in \mathbb{R} \} \) contains no boundary points then \( A = S \) and the theorem holds. It is possible to construct a horizontal segment \( I \) such that \( A \cap I = q \) because \( q \) is a boundary point of the closed set \( A \).

Assume that \( q \) does not hit a singular point. \( q \) will be the endpoint of one of the \( I_j \)'s on which the first return map is defined and continuous. Thus \( f(q) \) is defined and \( q \) must return to \( I \). But \( A \) is invariant under the vertical flow so \( f(q) \in A \) thus \( f(q) = q \) and the vertical trajectory from \( q \) is periodic. By theorem 4.2 we then know \( q \in C \) where \( C \) is some open cylinder but \( \{ \phi_t(p) \} \cap C = \emptyset \) as \( \{ \phi_t(p) \} \) would be periodic otherwise. This means \( \{ \phi_t(p) \} \) is not approaching \( q \) wich is a contradiction and thus \( q \) must hit a singular point. We can then repeat the same argument to see that \( q \) hits a singular point backwards in time.

## 5 A discussion of tetrahedrons

Translation surfaces and the theorems we proved concerning closed and open geodesics are not restricted to the study of billiards. In this chapter we will use a what we proved in the previous chapter together with linear transformations of the translation surfaces.

**Theorem 5.1.** The closure of any open geodesic on a tetrahedron with each face congruent covers the whole tetrahedron.

**Proof.** Assume for some tetrahedron there is a open geodesic \( \phi_t(p) \) whose closure does not cover the entire tetrahedron. We can unfold this tetrahedron to create a translation surface in the shape of a parallelogram. For any given flow in the parallelogram there is a linear transformation that preserve direction and relative distance of the flow that also transform the parallelogram into a square. Using theorem 4.3 we know the closure of \( \{ \phi_t(p) | t \in \mathbb{R} \} \) will be bounded by saddle connections. Looking at the renormalization in we realize that the slope of any saddle connection, with respect to an orthonormal coordinate system, is rational. But geodesics with
Figure 14: Given any triangle we can construct a tetrahedron such that all faces of the tetrahedron are congruent to the triangle.

Figure 15: To the left is a translation surface of the tetrahedron. To the right is a linear transformation of said translation surface. This have the geometry of a flat torus.

A rational slope on a flat torus is clearly periodic. This is a contradiction as \( \phi_t(p) \) will have the same slope as its saddle connections and is not periodic.

The technique of linearly transforming the translation surface is very useful indeed. When we study billiard trajectories and geodesics in the translation surface we study them one at a time. Linearly transforming the translation surface is not going to affect the behaviour of the geodesics. If it is periodic it will remain periodic, if it is open and is dense on the surface it will remain so.

This concludes this short introduction. The subject of billiards is indeed a very rich subject and here we have only seen a fraction of a subgroup of the polygonal ones.

6 Some thoughts on billiards

In this section I will discuss questions, ideas and approaches which didn’t really work out or that I had to abandon.
6.1 Moving forward and backwards in time

When we have analyzed trajectories we have propagated them both forward and backwards in time. A natural question rises, "Are there trajectories such that the behaviour forwards in time is different from that backwards in time?". First we need to answer what we mean by "different behaviour", as an example there are many trajectories that terminate in one direction (they hit a non-smooth piece of the boundary) but not in the other. If we don’t take these trajectories in consideration, are there billiards with trajectories such that they leave a region and never return? Here I will give a convoluted example of a billiard with such trajectories.

Figure 16: A billiard consisting of two tangent circles. Instead of reflecting trajectories hitting their common point we can instead let it continue in the other circle.

If we have a non-periodic trajectory starting in the left circle and it hits the circles common point we will never return to the circle we was in. This convoluted example is not really that interesting, it is not a proper billiard and the measure of the trajectories that hit their common point is zero. I have not found an example of this behaviour can occur in proper billiards.

6.2 Generalising the L-shaped billiard

Figure 17: From the perspective of the gray triangle the red and blue trajectories behave as if they were reflections in the dashed line.

In the introduction I gave an example of a billiard with non-periodic trajectories that was not dense on the billiard. Using that billiard as an inspiration it may seem we can construct more billiards with that same property. From Figure 17 it may seem that we can, starting from any rational triangular billiard $P$, easily construct a billiard with non-periodic trajectories whose
closure is not dense on the billiard. Imagine we have a non-periodic trajectory in the gray triangle, it seems plausible that by taking all possible directions the trajectory have during the motion and designing a special rectangle that have periodic trajectories in the same directions we could use it 'simulate' reflections by attaching it to one of the sides of the triangle. It turns out it is impossible to construct a rectangle for a general triangle.

The direction the trajectory have is dependent on the starting velocity and $\Gamma$. $\Gamma$ consist of rotations of rational multiples of $\pi$. We have seen that the periodic trajectories in a rectangle have a slope of $k$ where $k$ is a rational multiple of $\frac{\pi}{n}$ where $a$ is the base and $b$ the height. For any two angles that differ by a rational multiple of $\pi$ it is not always possible to find two different $k_1,k_2$ such that the lines with slope $k_1,k_2$ differ by that angle.

Related to this question is if we can for any given a trajectory on a billiard elongate the billiard in some way to get a polyhedra, as in the introduction, with the same behaviour. In general this is not possible again due to the dynamics of the rectangle.

References


