AN INTRODUCTION TO DYNAMICAL BILLIARDS

SUN WOO PARK

Abstract. Some billiard tables in $\mathbb{R}^2$ contain crucial references to dynamical systems but can be analyzed with Euclidean geometry. In this expository paper, we will analyze billiard trajectories in circles, circular rings, and ellipses as well as relate their charactersitics to ergodic theory and dynamical systems.

Contents

Dynamical billiards exhibits crucial characteristics related to dynamical systems. Some billiard tables in $\mathbb{R}^2$ can be understood with Euclidean geometry. In this expository paper, we will analyze some of the billiard tables in $\mathbb{R}^2$, specifically circles, circular rings, and ellipses. In the first section we will present some preliminary background. In the second section we will analyze billiard trajectories in the aforementioned billiard tables and relate their characteristics with dynamical systems. We will also briefly discuss the notion of completely integrable billiard mappings and Birkhoff's conjecture.

1. Background

(This section follows Chapter 1 and 2 of Chernov [1] and Chapter 3 and 4 of Rudin [2])

In this section, we define basic concepts in measure theory and ergodic theory. We will focus on probability measures, related theorems, and recurrent sets on certain maps. The definitions of probability measures and *σ-algebra* are in Chapter 1 of Chernov [1].

1.1. Recurrence.

Definition 1.1. Let $(X,\mathcal{A},\mu)$ and $(Y,\mathcal{B},\upsilon)$ be measure spaces. A map $T\colon X \to Y$ is *measure-preserving* if it satisfies the following two properties:

(1) For each open set $B \in \mathcal{B}$, $T^{-1}(B) \in \mathcal{A}$
(2) $\mu(T^{-1}(B))=\upsilon(B)$

We will now prove the probabilistic version of Poincaré Recurrence Theorem.

Theorem 1.2 (*Poincaré Recurrence Theorem*)**.** *Given a measure space (X,$\mathcal{A}$,μ), let T: X $\to$ X be a measure-preserving map. Given A $\in$ $\mathcal{A}$, define A_0 as the set of points a $\in$ A such that $T^n(a) \in A$ for infinitely many n $\geq$ 0. Then $A_0 \in \mathcal{A}$ and $\mu(A_0)=\mu(A)$.*

Proof. Let $B_n := \{a \in \mathrm{A} \mid T^j(a) \notin \mathrm{A} \text{ for all } j \geq n\}$. Notice that

$$A_0 = A \setminus \cup_{n=1}^{\infty} B_n$$

Thus, it suffices to show that $B_n \in \mathcal{A}$ and $\mu(B_n)=0$ for every $n \geq 1$. Since

$$B_n = A \setminus \cup_{j \geq n} T^{-j}(A)$$

it is clear that $B_n \in \mathcal{A}$. This implies

$$B_n \subset \cup_{j \geq 0} T^{-j}(A) \setminus \cup_{j \geq n} T^{-j}(A)$$

Observe that

$$\cup_{j \geq 0} T^{-j}(A) = T^{-n}(\cup_{j \geq n} T^{-j}(A))$$

Since T is a measure-preserving map,

$$\mu(\cup_{j \geq 0} T^{-j}(A)) = \mu(\cup_{j \geq n} T^{-j}(A))$$

Thus, $\mu(B_n) = 0$. □

The above theorem shows that under a measure-preserving map almost every point of a measurable set is recurrent.

1.2. **Invariance and Ergodicity.**

Definition 1.3. Let (X,$\mathcal{A}$) be a measurable space and let T: $X \to X$ be a measurable map. A measure μ is *T-invariant* if for every $A \in \mathcal{A}$,

$$\mu(T^{-1}(A)) = \mu(A)$$

Definition 1.4. Let T be a measure preserving map on a probability space (X,$\mathcal{A}$,μ). T is *ergodic* if every T-invariant set has measure 0 or 1.

We now state one of the most important theorems in ergodic theory. The proof of the theorem is in Chapter 2 of Chernov [1].

Theorem 1.5 (*Birkhoff-Khinchin*)**.** *Given a probability space (X,$\mathcal{A}$,μ), let T: X $\to$ X be a measure preserving map. If f: X $\to \mathbb{R}$ is in $\mathcal{L}^1$(X,$\mathcal{A}$,μ), the limit*

$$\tilde{f}(x) := \lim_{n \to \infty} \frac{1}{n+1} \sum_{j=0}^{n} f(T^j(x))$$

exists for almost every x $\in$ X, and the function $\tilde{f(x)}$ is T-invariant, integrable, and

$$\int_X \tilde{f} \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu$$

The function $\tilde{f}$ is called the time average of f.

Using the theorem, we can prove the following useful proposition. The proof of the proposition is also in Chapter 2 of Chernov [1].

Proposition 1.6. *Given a probability space (X,$\mathcal{A}$,μ), the following statements are equivalent:*

(1) T is ergodic.
(2) If $f \in \mathcal{L}^p(X)$ for a positive integer p, then f is constant almost everywhere.
(3) For every A,B $\in \mathcal{A}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \mu(T^{-n}(A) \cap (B)) = \mu(A)\mu(B)$$

(4) For every $f \in \mathcal{L}^1(X)$ we have $\tilde{f}(x) = \int_X f \, d\mu$ almost everywhere.

We will use the proposition in understanding ergodicity of irrational rotations on the circle, one of the examples in dynamical systems which is important for understanding billiard trajectories in circles.

1.3. **Rotation.** We now state how ergodicity and invariance of measures can be used in rational and irrational rotations on the circle. Dynamical systems deal with understanding the rules of a time-dependent mathematical object. We can describe the time-dependence of an object by iterations of a map on the object.

Define a mapping T: [0,1] $\to$ [0,1] such that $T_a(x) = x + a \pmod 1$ where $x \in$ [0,1] and $a \in \mathbb{R}$. Here x denotes the initial position and a denotes the angle of rotation along the circle. Then the n iterations of the map T can be expressed as follows: $T_a^n(x) = x + na \pmod 1 = m + r \pmod 1$ where $m \in \mathbb{Z}$ and $r \in$ [0,1]. It is clear that T is measure-preserving.

Proposition 1.7. *Every orbit of the mapping $T_a(x)$ is dense in [0,1] if $a \in \mathbb{R}\backslash\mathbb{Q}$.*

Proof. First we will show that each $T_a^n(x)$ is distinct. For two distinct natural numbers i and j assume $T_a^i(x) = T_a^j(x) + k$ for some positive integer k. In other words, $x + ia = x + ja + k$. So, $a(i - j)$ is a natural number. Since a is irrational, $i = j$, a contradiction.

Divide the interval [0,1] into n intervals, each with length $1/n$. Notice that the $n + 1$ terms x, $T_a(x)$, ... $T_a^n(x)$ are distinct. Then by pigeon hole principle there exists an interval that includes at least two of the terms $T_a^i(x)$ and $T_a^j(x)$ for $i \neq j$. Thus for every positive integer n there exist two distinct natural numbers i and j such that $|T_a^i(x) - T_a^j(x)| < 1/n$.

Now we prove that every neighborhood of an arbitrary point y in [0,1] contains the term $T_a^m(x)$. Without loss of generality we can show that every neighborhood of the initial term x contains the term $T_a^m(x)$. Notice that T_a is a measure-preserving map with respect to the lebesgue measure μ. In other words, the euclidean distance between two points of the interval is invariant under iterations of the map T_a; so, the following holds for natural numbers i, j, and n.

$$d(T_a^{i-j}(x), x) = d(T_a^j(T_a^{i-j}(x), T_a^j(x)) = d(T_a^i(x), T_a^j(x)) < 1/n$$

Thus every neighborhood of the initial term x contains the term $T_a^{i-j}(x)$. Hence every orbit of T_a is dense in [0,1]. □

Now we prove one of the important characteristics of T_a.

Theorem 1.8. *T is ergodic if and only if $a \in \mathbb{R}\backslash\mathbb{Q}$.*

Proof. Assume a is irrational. Define f to be an arbitrary T-invariant function in $\mathcal{L}^2(X)$ where X is the interval [0,1]. We claim that f is constant almost everywhere. The Fourier expansion of f in $\mathcal{L}^2(X)$ is as follows.

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi inx}$$

Similarly, the Fourier expansion of $f \circ T$ in $\mathcal{L}^2(X)$ is as follows.

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi in(x+a)} = \sum_{n=-\infty}^{\infty} c_n e^{2\pi ina} e^{2\pi inx}$$

Since f is T-invariant, $f \circ T$ is same as f, which implies

$$\sum_{n=-\infty}^{\infty} e^{2\pi inx} c_n (1 - e^{2\pi ina}) = 0$$

Since the elements of the series $e^{2\pi inx}$ form an orthonormal basis of the Hilbert space, all coefficients c_n(1-$e^{2\pi ina}$) are zero. Since a is irrational, 1-$e^{2\pi ina}$ is not 0. Thus, every Fourier series coefficient is equal to 0. This shows that f is constant. Hence, T is ergodic by proposition 1.6.

Now assume T is ergodic. We can prove by contrapositive that if a is rational, then T is not ergodic. Observe the Fourier expansion of $f - f \circ T$ in $\mathcal{L}^2(X)$. Since a is rational, 1-$e^{2\pi ina}$ is 0 for some integer n. Hence not all coefficients of the Fourier series are zero, which implies that there exists a function f in $\mathcal{L}^2(X)$ such that f is not constant. Thus, T is not ergodic by proposition 1.6. □

2. Dynamical Billiards

(This section follows Chapter 4 of Chernov [1] and Chapter 4 of Tabachnikov [3].)

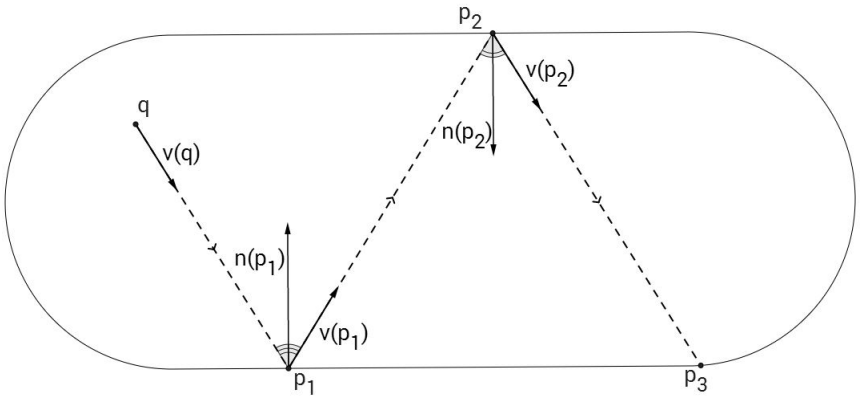
Dynamical Billiards focuses on the characteristics of billiard trajectory in respect to time. In this section we focus on some of the simple forms of smooth billiard tables on 2-dimensional euclidean space. We start the section with an overview of dynamical billiards.

Definition 2.1. A billiard table $Q \in \mathbb{R}^2$ is an open bounded connected domain such that its boundary ∂Q is a finite union of smooth compact curves.

Assume the moving particle in the billiard table $Q \in \mathbb{R}^2$ has position $q \in Q$ and velocity vector $v \in \mathbb{R}^2$. Then the curves comprising the boundary of the billiard table and the velocity vector of the moving particle satisfy the following conditions.

(1) The curves are disjoint but may have common endpoints.
(2) At point $q \in Q$ the billiard travels in a straight line parallel to the direction of the velocity vector at point q until it hits the boundary ∂Q. In other words the billiard always moves in a straight line.
(3) Define the billiard trajectory as the segment $\overline{p_1 p_2}$ where p_1 and p_2 are the points on ∂Q where the billiard consecutively hits the boundary.
(4) Define $n(p)$ as the inward pointing normal vector at point $p \in \partial Q$.
(5) Let p_1, p_2, and p_3 be three consecutive points the billiard contacts with the boundary of the billiard table. At point p_2 define the angle of incidence as the angle between the inward pointing normal vector $n(p_2)$ at point p_2 and the billiard trajectory $\overline{p_1 p_2}$. Similarly, define the angle of reflection as the angle between $n(p_2)$ and the billiard trajectory $\overline{p_2 p_3}$.

(6) At every point $p \in \partial Q$ where the billiard hits the boundary, the angle of incidence is same as the angle of reflection. This is an empirical fact in physics.



The picture above shows the initial billiard trajectories on Bunimovich stadium, a billiard table created by connecting two semicircles with segments tangent at the endpoints of the semicircles. The boundary of the stadium is a union of four smooth curves. In this paper, however, we will mainly focus on the billiard table whose boundary is a single smooth compact curve.

Definition 2.2. A phase space $\mathcal{M}$ of the billiard table Q is $\mathcal{M} = \bar{Q} \times S^1$ where $\bar{Q}$ is the closure of Q and S^1 is the unit circle of all velocity vectors. At ∂Q the velocity vector is always headed inwards. Given the phase space $\mathcal{M}$ we define the flow Φ^t as the set of all possible billiard trajectories with the related velocity vectors on $\mathcal{M}$ parametrized by time.

Notice that S^1 corresponds to possible directions the billiard trajectory could be heading in. The flow on the billiard table Q can be thought as a family of billiard trajectories on the closure of the billiard table and the velocity vectors of the trajectory along time flow t. Given the following definitions, we can now define the billiard trajectory as a form of a map.

Definition 2.3. Let the hypersurface M of the phase space $\mathcal{M}$ be defined as follows.

$$M = \{x = (p, v) \in \mathcal{M} | p \in \partial Q, \langle v, n(p) \rangle \geq 0\}$$

We define the *billiard map* T: $M \to M$ as $Tx = \Phi^{\tau(x)}x$ such that

$$\tau(x) = min\{t > 0 | \Phi^t x \in M\}$$

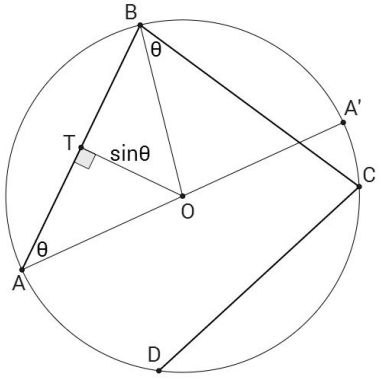
The significance of $\langle v, n(p) \rangle \geq 0$ is that the value $|v||n(p)|\cos\theta$ should be positive. Hypersurface requires the angle between the velocity vector and the normal vector to be between $-\pi/2$ and $\pi/2$. Also the definition of billiard mapping T reassures one of the assumptions of billiard table that the billiard trajectory is a straight line that minimizes the time interval between two consecutive billiard trajectory points on ∂Q. We may consider the elements (p,v) in M as (φ, θ) where φ denotes the position of a point on ∂Q and θ denotes the angle of incidence. In the subsequent sections we will observe how the billiard maps are defined in some of the billiard tables whose boundary can be considered as a single smooth compact curve.

2.1. **Circle.** We first consider a circular billiard table Q with radius 1. On the boundary of the billiad table Q, the angle of incidence has to be between $-\pi/2$ and $\pi/2$. Thus the hypersurface M is $\partial Q \times [-\pi/2, \pi/2]$, a cylinder with radius 1 and height π.

We claim that the billiard map T: $M \to M$ is in fact the rotation mapping defined in section 1.3.. This comes from the fact that the angle of incidence is preserved throughout the whole billiard trajectory as seen from the following proposition.

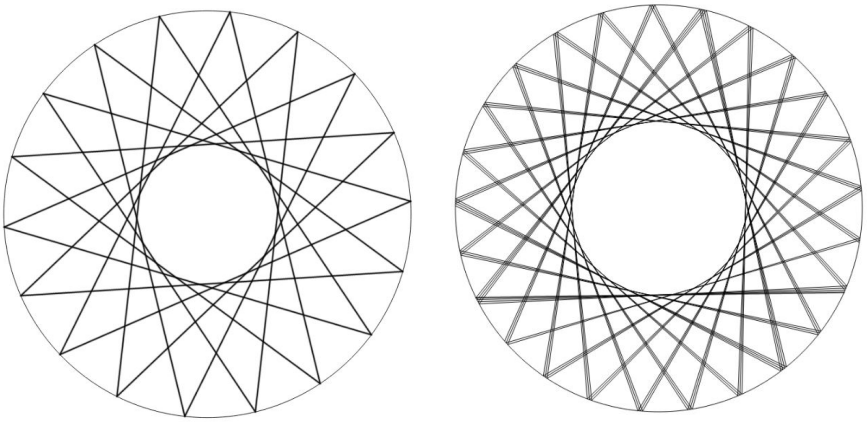
Proposition 2.4. *The billiard map T: $M \to M$ is given by*

$$T^n x = (\varphi + n(\pi - 2\theta), \theta)$$



Proof. Let the initial starting point of the billiard be A and let the subsequent points the moving particle contacts with the boundary of the circle be B, C, ... Let the initial position at point α be φ and let the initial angle be θ. Since $\overline{OA} = \overline{OB}$, $\angle$OAB = $\angle$OBA. The same relationship applies between two adjecent contact points. Hence, the angle of incidence is always θ. Notice that the location of the contact points move along the arc of the circle. Since $\angle$OAB = $\angle$OBA = θ, $\angle$BOA = π - 2θ. Thus through each billiard mapping the point moves along the arc the distance of π-2θ. □

The above proposition clearly shows that the angle of incidence is T-invariant while the location of the contact points behaves the same as the rotation of the points on the boundary of circle with angle π-2θ. Thus, we proved our claim that the billiard mapping T is indeed the rotation mapping on the circle. Using the proof of the proposition 2.4, the distance between the center of the circle and the billiard trajectory, or $\overline{OT}$, is always $\sin(\theta)$. In fact every billiard trajectory in circular billiard table is tangent to the concentric circle with radius $\sin(\theta)$. The following two pictures, the left picture which shows the case with rational values of θ and the right picture which shows the case with irrational values of θ, graphically show that the billiard trajectories are tangent to the concentric circle.



We can now apply our findings from section 1.3. to billiards in circles. It is clear that T is measure-preserving. If $2\theta/\pi$ is rational, then the billiard returns to its initial position. Also, T is not ergodic for rational angle of incidence. If $2\theta/\pi$ is irrational, then T is ergodic. The billiard creates a dense trajectory inside the circle excluding the area of the concentric circle with radius $\sin(\theta)$. Observe that the trajectory is more dense in the vicinity of the concentric circle than other regions of the circle. Imagine the boundary of the billiard table is a mirror and the billiard trajectory is a laser ray. Then the boundary of the concentric circle will be significantly hotter than other regions. For the following reason, in mechanical terms, the concentric circle is called a *caustic* which also means *burning*. In other billiard tables we can think of the caustic as the region to which every billiard trajectory is tangent.

2.2. **Circular Ring.**

Definition 2.5. A billiard table R is called a *circular ring* if its domain is bounded by two concentric circles Q_1 and Q_2 with different radii.Let the radius of Q_1 be r_1 and radius of Q_2 be r_2 such that $r_1 > r_2$. The phase space M in the circular ring can be defined as

$$M = \Gamma \times [-\pi/2, \pi/2] = (\partial Q_1 \cup \partial Q_2) \times [-\pi/2, \pi/2]$$

where $\Gamma = \partial Q_1 \cup \partial Q_2$ is the boundary of the ring and $[-\pi/2, \pi/2]$ is the interval of the angle of incidence.

As same as the condition given in circular billiards, let the initial angle of the trajectory be θ. If $r_2 \leq r_1 \sin(\theta)$, then the trajectory of the billiard map is identical to the trajectory of the circular map. The set of trajectories will be dense in the area between the outer circle with radius r_1 and the circular caustic with radius $r_1 sin(\theta)$. However, in case of $r_2 > r_1 \sin(\theta)$ the trajectory is different from that in circular billiard tables. Given a circular ring R, let T: $M \to M$ be the billiard mapping on R. Define $\{a_n\}$ as the sequence of billiard trajectory points on ∂Q_1 and $\{b_n\}$ as the sequence of billiard trajectory points on ∂Q_2. Let a_0 be the starting point with initial angle as θ such that $r_2 > r_1 \sin(\theta)$. Denote the first point where the billiard contacts ∂Q_2 as b_0 with the angle of incidence as θ'. The billiard hits the circular boundaries in the following order: $\{a_0, b_0, a_1, b_1, \dots\}$.

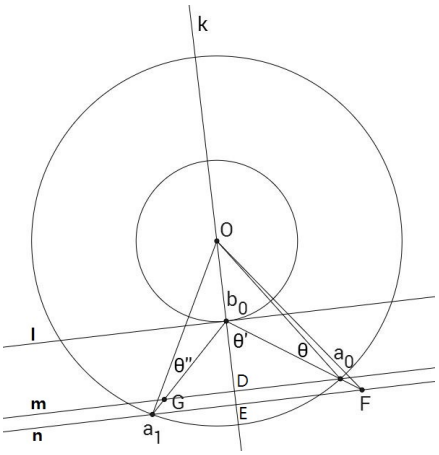
Proposition 2.6. *The following holds for every positive integer n for the billiard map T on R.*

(1)

$$T^{2n}((a_0,\theta)) = T^{2n-1}((b_0,\theta')) = (a_0 + 2nr_1(\theta' - \theta), \theta)$$

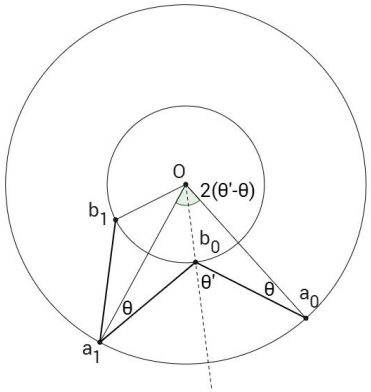
(2)

$$T^{2n+1}((a_0,\theta)) = T^{2n}((b_0,\theta')) = (b_0 + 2nr_2(\theta' - \theta), \theta')$$



Proof. We first claim that for every a_n the angle of incidence is θ. Given the conditions from the proposition, consider three consecutive points a_0, b_0, and a_1 on the ring. Denote the center of the circles as point O. Draw line k which passes through point O and point b_0 and line l which is tangent to the inner circle Q_2 at point b_0. Denote the angle of incidence at point a_1 as θ''.

Assume that θ and θ'' are not the same. Draw two lines parallel to line l, one - line m - which passes through a_0 and the other - line n - which passes through a_1. Clearly the two lines do not intersect each other. Denote the intersection of line m and line k as point D and the intersection of line n and line k as point E. Now draw two lines which are the extension of the segment $\overline{b_0a_0}$ and $\overline{b_0a_1}$. Call each line i and j respectively. Denote the intersection of line n and line i as point F and the intersection of line m and line j as point G. Notice that $\triangle \mathrm{F}b_0\mathrm{E} \equiv \triangle a_1b_0\mathrm{E}$, which implies $\triangle \mathrm{OF}b_0 \equiv \triangle \mathrm{O}a_1b_0$. Then $\overline{OF} = \overline{Oa_1}$, so point E should be on the boundary of the outer circle Q_1. This is a contradiction since line m and line n become the same line. Thus the angle of incidence is θ for every a_n and θ' for every b_n.



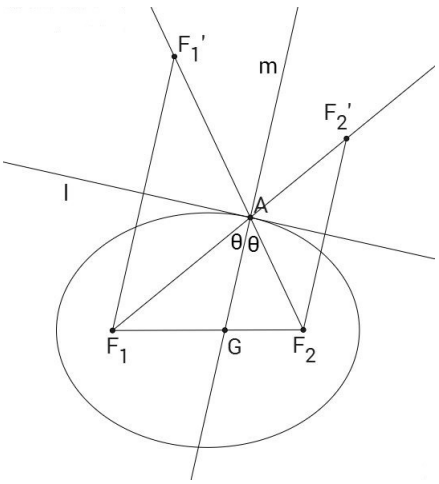
Observe that $\angle a_0\mathrm{O}a_1 = \angle b_0\mathrm{O}b_1 = 2(\theta'\text{-}\theta)$. It is clear that for sequence $\{a_n\}$ the billiard mapping shifts the points along ∂Q_1 by $2r_1(\theta'\text{-}\theta)$ while for sequence $\{b_n\}$ the billiard mapping shifts the points along ∂Q_2 by $2r_2(\theta'\text{-}\theta)$. □

The proof for the above proposition shows that circular ring billiard tables share the same properties with circular billiard tables. If $(\theta'\text{-}\theta)/2\pi$ is rational, the billiard trajectory will return to its starting point after finitely contacting the two boundaries of the circular ring. Also the billiard mapping T is not ergodic for both boundaries ∂Q_1 and ∂Q_2. If $(\theta'\text{-}\theta)/2\pi$ is irrational, the billiard trajectory will be dense in the circular ring. The mapping T is ergodic in the billiard table.

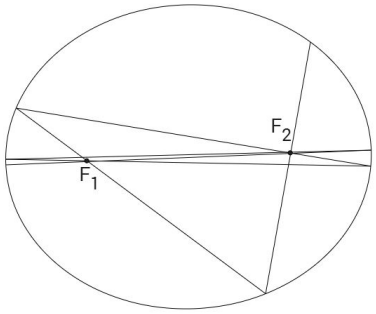
2.3. **Ellipse.** A circular billiard table is a special case of an elliptical billiard table in $\mathbb{R}^2$ such that the two foci of the ellipse are the same. In elliptical billiard table we consider the billiard trajectory in three cases - if the trajectory crosses the foci, if it crosses the line segment between the two foci, and if it does not cross the line segment between the two foci. Here is the theorem which will be useful in this subsection.

Theorem 2.7. *Let Q be an ellipse with foci F_1 and F_2. Let A be a point on the boundary of the ellipse. Draw line l which is tangent to the ellipse at point A. Then the segment $\overline{F_1A}$ and $\overline{F_2A}$ make the same angle with line l.*

Proof. Draw rays of the two segments $\overline{F_1A}$ and $\overline{F_2A}$ outside the ellipse, both in the direction from the foci to point A. On the ray $\overrightarrow{F_1A}$ set a point F_2' such that the length of the segment $\overline{F_2'A}$ is equal to $\overline{F_2A}$. Similarly, on the ray $\overrightarrow{F_2A}$ set a point F_1' such that the length of the segment $\overline{F_1'A}$ is equal to $\overline{F_1A}$. Then $\triangle \mathrm{A}F_1F_1'$ and $\triangle \mathrm{A}F_2F_2'$ are isosceles triangles which are similar to each other. Now draw a line m which passes through point A and bisects $\angle F_1\mathrm{A}F_2$. Denote the intersection between the line m and the segment $\overline{F_1F_2}$ as point G. It is clear that $\angle F_1\mathrm{A}F_2 = 2\angle F_2'F_2\mathrm{A}$, which implies $\angle \mathrm{GA}F_2 = \angle F_2'F_2\mathrm{A}$. Thus line m and segment $\overline{F_2F_2'}$ are parallel. Denote the intersection between line l and segment $\overline{F_2F_2'}$ as point H. Notice that line l bisects $\angle F_2\mathrm{A}F_2$'. Thus the segment $\overline{F_1A}$ and $\overline{F_2A}$ make the same angle with line l. □

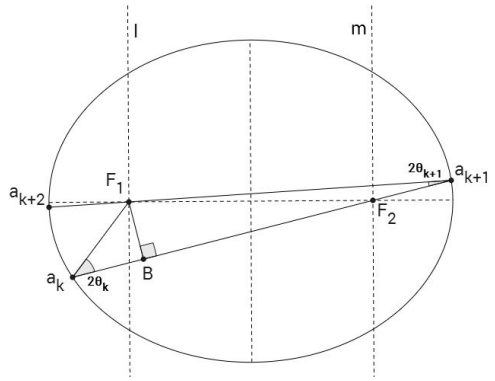


The above theorem explains that the trajectory on an elliptical billiard table which passes through one of the foci always passes through the other foci in the consecutive trajectory. However the above theorem does not imply that the angle of incidence is invariant to the billiard mapping. As seen from the picture below, the angle of incidence varies throughout the billiard mapping. In fact we can notice that the billiard trajectory converges to the major axis of the ellipse, which is one of the exercises in Chapter 4 of Tabachnikov [3].



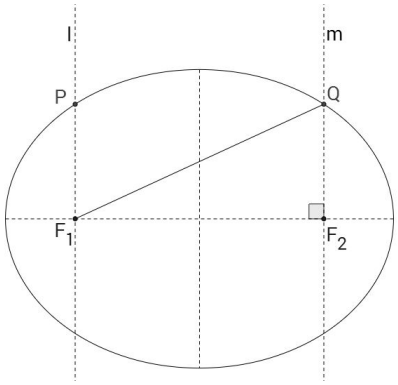
Theorem 2.8. *Let Q be an elliptical billiard table in $\mathbb{R}^2$. Let $\{a_n\}$ be the sequence of points on ∂Q where the billiard contacts the boundary of Q while for every non-negative integer n the segment $\overline{a_n a_{n+1}}$ crosses the two foci. Then the trajectory of the billiard converges to the major axis of the ellipse.*

Proof. Divide the boundary of the elliptical billiard table into three sections as follows. Draw two lines l and m parallel to the minor axis, each passing through the focus of the ellipse. Label the closed boundary left to the line l as ∂Q_1, the boundary between the two lines as ∂Q_2, and the boundary right to the line m as ∂Q_3. We will prove the theorem in two steps. First we show that the the set of trajectories W_1 which starts from ∂Q_1 and passes through F_2 converges to the major axis. Then we show that proving the first step is equivalent to proving the theorem.



Observe that the angle of incidence is 0 if and only if the trajectory is along the major axis of the ellipse. Then it is suffice to show that the sequence of angle of incidence of the set W_1, denoted as $\{\theta_{n_1}\}$, converges to 0. We claim that the sequence $\{\theta_{n_1}\}$ is monotonically decreasing. Let a_k, a_{k+1}, and a_{k+2} be three consecutive elements of the sequence $\{a_n\}$. Assume that a_k is on ∂Q_1. Let line

l be perpendicular to the segment $\overline{a_k a_{k+1}}$ which passes through point F_1 and let point B be the intersection between the line l and the segment $\overline{a_k a_{k+1}}$. Denote the segment $\overline{a_k F_1}$ as x, the segment $\overline{a_{k+1} F_1}$ as y, and the segment $\overline{F_1 B}$ as z.



Observe that $x \leq y$. As shown in the picture above, denote one of the intersections between the ellipse and line l as point P and denote one of the intersections between the ellipse and line m as point Q. Notice that the segment $\overline{PF_1}$ is the longest among the set of segments $\overline{SF_1}$ for S $\in \partial Q_1$. The segment $\overline{QF_1}$ is the shortest among the set of segments $\overline{TF_1}$ for T $\in \partial Q_3$. Label the length of the minor axis as $2a$, the major axis as $2b$, $\overline{QF_1}$ as α, $\overline{PF_1}$ as β, and $\overline{F_1F_2}$ as $2f$. Since $\overline{PF_1}$ $= \overline{QF_2}$ and $\triangle QF_1F_2$ is a right triangle, the following holds.

$$\alpha + \beta = 2b$$

$$\alpha^2 - \beta^2 = 4f^2 = 4(b^2 - a^2)$$

Solving the equation we get $\alpha = 2b - (a^2/b)$ and $\beta = a^2/b$. Notice that $\alpha \geq \beta$ and the equality holds when the ellipse is a circle. Thus for arbitrary points S $\in \partial Q_1$ and T$\in \partial Q_3$ the following holds.

$$x = \overline{SF_1} \leq \overline{PF_1} \leq \overline{QF_1} \leq \overline{TF_1} = y$$

Hence $x \leq y$. This implies

$$sin(2\theta_k) = \frac{z}{x} \geq \frac{z}{y} = sin(2\theta_{k+1})$$

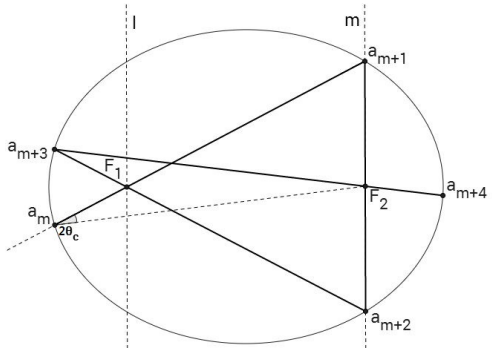
Since sin(θ) is an increasing function on the interval [-π/2,π/2], $\theta_k \geq \theta_{k+1}$ for arbitrary k. Thus the sequence is monotonically decreasing. Since the sequence is bounded by the interval [-π/2,π/2], the sequence converges, which implies the billiard trajectory also converges to a segment in the ellipse.

We observe that the sequence converges to 0. Assume that the sequence converges to a non-zero value θ. Notice that the billiard trajectory has to pass through either of the two foci. Then the trajectory has to oscillate between two parallel segments, each passing through one of the foci. This is clearly a contradiction since the billiard trajectory has to converge. Even if we assume that the trajectory converges to one of the two segments, the consecutive trajectory has angle of incidence less than θ, another contradiction. Thus the billiard trajectories in the set W_1 converge to the major axis of the ellipse.

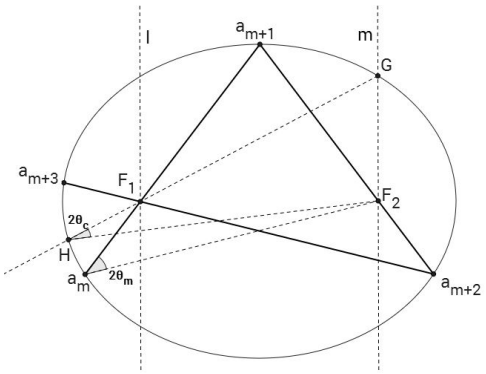
We now show that proving the first step is equivalent to proving the theorem. Except for the case in which the billiard trajectory is on the major axis or the

minor axis of the ellipse, at some point the set of trajectories has to intersect with ∂Q_2. Assume that the set of billiard trajectories W_2 starts from a point on ∂Q_1 but crosses the focus F_1. Notice that observing W_2 in time is equivalent to observing W_1 backwards in time. Since the sequence $\{\theta_{n_1}\}$ monotonically decreases, the sequence $\{\theta_{n_2}\}$, which is the sequence of angles of incidence in W_2, monotonically increases as long as W_2 intersects at the interiors of ∂Q_1 and ∂Q_3.

Let point G be one of the intersections between line m and ∂Q_3 and let point H be the intersection between the ray GF_1 and ∂Q. Define $\angle F_1\mathrm{H}F_2$ as the critical angle of incidence, denoted as $2\theta_c$. It is clear that there exists a finite number m such that the angle of incidence θ_{m_2} at point a_m of the set W_2 eventually equals or exceeds θ_c. Without loss of generality assume a_m is in ∂Q_1.



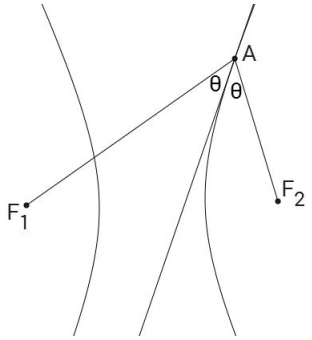
If θ_{m_2} equals θ_c (or a_m is identical to point H and a_{m+1} is identical to point G), then a_{m+3} is on ∂Q_1 while the trajectory $\overline{a_{m+3}a_{m+4}}$ passes through F_2. Thus W_2 follows the trajectory W_1 after finite $m+3$ iterations of billiard mapping.



If θ_{m_2} is greater than θ_c, then a_{m+1} is on ∂Q_2, which implies the trajectory $\overline{a_{m+2}a_{m+3}}$ holds the following condition: Either a_{m+2} is in ∂Q_1 and $\overline{a_{m+2}a_{m+3}}$ crosses through F_2 or a_{m+2} is in ∂Q_3 and $\overline{a_{m+2}a_{m+3}}$ crosses through F_1. Thus, every trajectory in elliptical billiard table is equivalent to the trajectory W_1 after finite $m+2$ iterations of billiard mapping. □

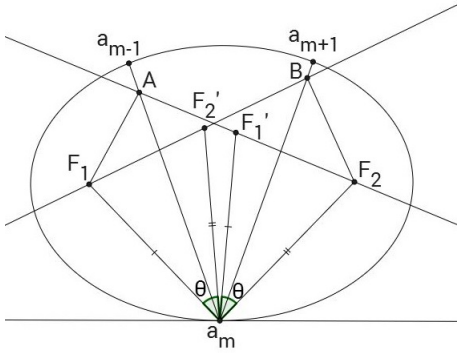
Next we consider the case in which the trajectory crosses the line segment between the two foci. Before we consider the case here is the theorem on hyperbolas similar to theorem 2.7. We skip the proof of the theorem because it uses the same procedure to proving theorem 2.7.

Theorem 2.9. *Let Q be a hyperbola with foci F_1 and F_2. Let A be a point on the boundary of the hyperbola. Draw line l which is tangent to the hyperbola at point A. Then the segment $\overline{F_1A}$ and $\overline{F_2A}$ make the same angle with line l.*



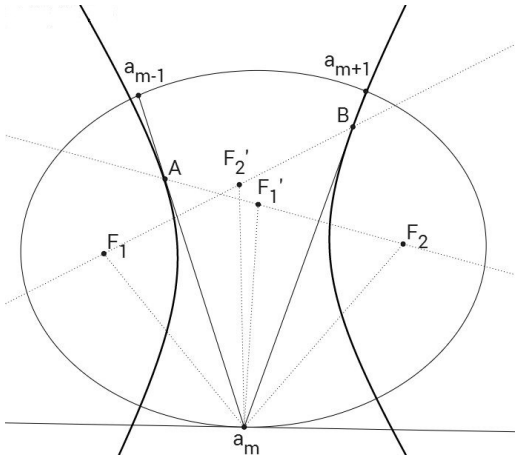
Now we prove the following theorem which explains the second case where the billaird trajectory crosses the line segment between the two foci of the ellipse.

Theorem 2.10. *Let Q be an elliptical billiard table in $\mathbb{R}^2$. Let $\{a_n\}$ be the sequence of points on ∂Q where the billiard contacts the boundary of Q while for every non-negative integer n the segment $\overline{a_n a_{n+1}}$ intersects with the line segment between the two foci of the ellipse. Then the trajectory of the billiard is tangent to the hyperbola which shares the same foci with the ellipse Q. In other words, the trajectory has a caustic which is a hyperbola sharing the same foci with the ellipse.*



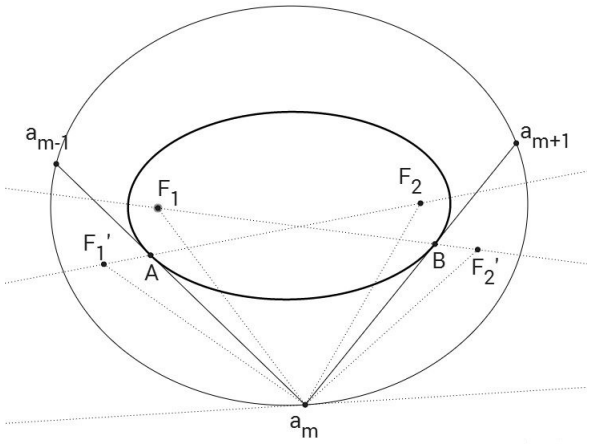
Proof. Assume we have three consecutive elements of the sequence $\{a_n\}$, namely a_{m-1}, a_m, and a_{m+1}. Let point F_1' be the reflection point of F_1 with respect to the segment $\overline{a_{m-1}a_m}$ and let point F_2' be the reflection point of F_2 with respect to the segment $\overline{a_m a_{m+1}}$. Let point A be the intersection of the line $\overleftrightarrow{F_1'F_2}$ and $\overline{a_{m-1}a_m}$ and let point B be the intersection of the line $\overleftrightarrow{F_2'F_1}$ and $\overline{a_m a_{m+1}}$. Since F_1' is a reflection point of F_1, $\angle a_{m-1}\mathrm{A}F_1' = \angle a_{m-1}\mathrm{A}F_1$, which implies $\angle a_{m-1}\mathrm{A}F_1 = \angle a_m \mathrm{A}F_2$. By theorem 2.9, A is the point where the segment $\overline{a_{m-1}a_m}$ is tangent to the hyperbola Q_1 with two foci F_1 and F_2. Similarly, B is the point where the segment $\overline{a_m a_{m+1}}$ is tangent to the hyperbola Q_2 with two foci F_1 and F_2.

Now we claim that $Q_1 = Q_2$. We can show this by showing $|\overline{F_1A}\text{-}\overline{F_2A}| = |\overline{F_1B}\text{-}\overline{F_2B}|$, or equivalently, $\overline{F_1'F_2} = \overline{F_1F_2'}$. By theorem 2.7, $\angle F_1' a_m F_2 = \angle F_1 a_m F_2'$, which implies $\triangle F_1' a_m F_2 \equiv \triangle F_1 a_m F_2'$. This clearly shows $\overline{F_1'F_2} = \overline{F_1F_2'}$. □



The third case in which the billiard trajectory intersects with the line segment between the two foci is similar to the second case. The following theorem shares the same proof with theorem 2.10.

Theorem 2.11. *Let Q be an elliptical billiard table in $\mathbb{R}^2$. Let $\{a_n\}$ be the sequence of points on ∂Q where the billiard contacts the boundary of Q while for every non-negative integer n the segment $\overline{a_n a_{n+1}}$ does not intersect with the line segment between the two foci of the ellipse. Then every trajectory of the billiard is tangent to the ellipse which shares the same foci with the ellipse Q. In other words, the trajectory has a caustic which is an ellipse confocal to the elliptical billiard table.*



2.4. **Completely Integrable.**

Definition 2.12. If a smooth dynamical system T: $M \to M$ on a manifold M admits a smooth nonconstant function F invariant under T, then F is *first integral* and T is said to be *integrable*.

Definition 2.13. If the manifold M can be foliated by one-dimensional T-invariant submanifolds or curves, then T is *completely integrable*.

In other words, if the first integral functions of T are one-dimensional, then T is completely integrable. In Chernov [1] Chapter 4 there is a brief explanation

of why circles and ellipses are completely integrable. In case of circular billiard tables, the function F defined by $F(\varphi,\theta) = \theta$ is invariant under the billiard map T. Hence, F is the first integral function of T. Then the hypersurface $M_p = \{F(x) = p\}$ such that $p \in \mathbb{R}$ is T-invariant. Thus, the manifold M can be foliated by T-invariant hypersurfaces $\{M_p\}$. Notice that the first integral function $F(x) = p$ is one-dimensional. Since M can be foliated by one-dimensional T-invariant curves, circles are completely integrable.

In case of elliptical billiard tables, the trajectories creating an elliptical caustic lie on a T-invariant curve on the manifold M, a deformation of the first integral function found in circular billiard table. The trajectories creating a hyperbolic caustic lie on a T^2 invariant curve on the manifold M. The trajectories passing through the foci make closed curves on the manifold M.

Then, are there any billiard tables other than circles and ellipses that are completely integrable? We end this paper by stating Birkhoff's conjecture, which has not been proved yet. Many mathematicians believe that the conjecture is true.

Conjecture 2.14. *The only completely integrable billiards are circles and ellipses.*

Acknowledgments. It is a pleasure to thank my mentor, Clark Butler, for introducing me into the field of dynamical systems, for guiding me on writing this paper, and for advising me in mathematics consistently and effectively. I also sincerely thank professor Peter May and other professors for giving me this wonderful opportunity to delve in mathematics, for organizing the REU, and for giving enlightening lectures.

References

[1] Nikolai Chernov and Roberto Markarian. Introduction to the Ergodic Theory of Chaotic Billiards. http://people.cas.uab.edu/ mosya/papers/rbook.pdf
[2] Walter Rudin. Real and Complex Analysis. McGraw-Hill International Editions. 1987.
[3] Serge Tabachnikov. Geometry and Biliards. http://www.math.psu.edu/tabachni/Books/billiardsgeometry.pdf