Dynamical Billiards

All the complexity of Hamiltonian systems, from integrability to chaotic motion, without the difficulties of integrating the equations of motion.

Shihan Kanungo

Euler Circle, Palo Alto, CA 94306

Abstract

In this expository paper, we investigate the topic of *dynamical billiards.* The field of dynamical billiards analyzes the dynamics of the motion of a ball bouncing in a billiard table, which is bounded by a smooth closed curve. The movement of the ball satisfies the properties that it always moves in a straight line, and the angle of incidence with the boundary equals the angle of reflection. This second property is an empirical fact in physics. In this paper, we look at the dynamics of some billiard boards in $\mathbb{R}^2$, and we use Euclidean geometrical methods to understand and classify the ergodicity of these billiards. In particular, we classify the ergodicity of billiards in circles and circular rings. Then, we state some results about elliptic billiards. Then, we give some examples of chaotic billiards, where a billiard is chaotic if after many bounces, the starting point has no relation to the current point; in other words, a slight deviation in the starting point can create absolutely divergent trajectories. We conclude with a physical application of billiards.

1 Introduction

Dynamical billiards study the dynamics of an idealized billiard ball in a billiard table. The billiard table is a smooth closed curve, and the path of the ball is the union of straight line segments, which satisfy the following property: *the angle of incidence equals the angle of reflection.* In particular, dynamical billiards looks at the ergodicity and other properties of different billiards. In this paper, we present some results for specific billiard tables, and prove them using techniques from Euclidean geometry.

The main object of study in ergodic theory is that of a measure-preserving transformation.

Definition 1. Let $(X, \mathcal{A}, \mu)$ be a measure space. A function $T : X \to X$ is said to be a *measure-preserving transformation* if for all $A \in \mathcal{A}$, $T^{-1}(A) \in \mathcal{A}$, and $\mu(T^{-1}(A)) = \mu(A)$. By $T^{-1}(A)$, we mean $\{x \in X : T(x) \in A\}$; this is the preimage of A.

Here is a typical example of a measure-preserving transformation that we will refer to again in the following sections.

Example 1 (Rotations on the unit circle)**.** Let $X = [0, 1)$, and let $\mathcal{A}$ be the Lebesgue σ-algebra on X. Let μ be the Lebesgue measure on X. Let θ be a real number, and consider the transformation $R_\theta : X \to X$, defined by $R_\theta(x) = (x + \theta) \pmod 1$, or equivalently $R_\theta(x) = x + \theta - \lfloor x + \theta \rfloor$. Let's check that this transformation is measure-preserving. First, $R_\theta^{-1}(A) = (A - \theta) \pmod 1$, so it's pretty clear that $R_\theta^{-1}(A) \in \mathcal{A}$ whenever $A \in \mathcal{A}$. As for the measure-preserving property, if $A \in \mathcal{A}$, then $\mu(R_\theta^{-1}(A)) = \lambda(A - \theta) = \lambda(A) = \mu(A)$, where λ is the Lebesgue measure on $\mathbb{R}$. Thus R_θ is measure-preserving.

One good way of thinking about this transformation R_θ is by thinking of X as a circle: take the half-open interval $[0, 1)$, and connect the missing endpoint 1 to the present endpoint 0, so that it wraps around. Then R_θ is a rotation by θ on the circle, where θ is the proportion of the circumference of the circle. This is often a good way of parametrizing the circle, so that the full circle has measure 1.

In the future, we'll write $\mathbb{R}/\mathbb{Z}$ for X: it consists of the real numbers, except that two real numbers are considered equal if they differ by an integer.

We now introduce a key result concerning rational and irrational rotations.

2 Rational and Irrational Rotations

In many of the billiard tables we will investigate, we can reduce the dynamics to a measure-preserving transformation on $\mathbb{R}/\mathbb{Z}$, namely rational and irrational rotations.

Proposition 2. *The mapping R_a is ergodic if and only if a is irrational.*

Proof. To show that that rational rotations are not ergodic, suppose $a = p/q$ is rational, with $\gcd(p, q) = 1$. Then $e^{2\pi qx}$ is a non-constant measurable function that is invariant under R_a. Thus R_a is not ergodic. Let A' and B' be sets of positive measure. There exist dyadic intervals I and J such that

$$\lambda(A' \cap I) > \tfrac{3}{4}\lambda(I) \quad \text{and} \quad \lambda(B' \cap J) > \tfrac{3}{4}\lambda(J).$$

Furthermore, we may assume that I and J are of the same measure (suppose if J is bigger than I, then at least one of the two halves of J must be $\frac{3}{4}$-full of B'; continue in this way until obtaining a subinterval of J of the same measure as I that is $\frac{3}{4}$-full of B', finally rename it J). Write

$$A = A' \cap I \;\text{ and }\; B = B' \cap J.$$

Suppose $I = [a, b)$ is to the left of $J = [c, d)$ in $\mathbb{R}/\mathbb{Z}$, i.e., $a \leqslant c$. As the orbit of b under $T := R_a$ is dense, there is an integer $n > 0$ such that

$$d - \tfrac{1}{4}(d - c) < T^n(b) < d.$$

Therefore $\lambda(T^n(I) \cap J) > \frac{3}{4}\lambda(J)$. Thus,

$$\begin{aligned} \lambda(T^n(A) \cap B) &\geqslant \lambda(T^n(I) \cap J) - \lambda(I \backslash A) - \lambda(J \backslash B) \\ &> \tfrac{3}{4}\lambda(J) - \tfrac{1}{4}\lambda(I) - \tfrac{1}{4}\lambda(J) > 0. \end{aligned}$$

Therefore T is ergodic. □

3 Billiards

The field of dynamical billiards analyzes the dynamics of the motion of a ball bouncing in a billiard table, which is bounded by a smooth closed curve. We start by defining dynamical billiards, and then we dive into some examples.

To start, we first need a table.

Definition 3. 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.

Once we have the table, we can add a ball, and we can define how the ball moves. We want these to match what we see on an actual billiard table, which motivates the following conditions.

The curves comprising the boundary of the billiard table and the velocity vector of the moving particle satisfy the following conditions.

(1) The path of the ball (represented as moving point) is the union of line segments (i.e. the ball always travels in a straight line), with consecutive segments sharing one common endpoint.

(2) Let $n(p)$ be the inward pointing normal vector at a point p on the boundary of Q (denoted as ∂Q). Define the billiard trajectory to be the segment $\overline{p_1 p_2}$, where p_1, p_2 are the points where the billiard consecutively hits the boundary. Then, *the angle of incidence equals the angle of reflection*, i.e. the angle between $\overline{p_1 p_2}$ and $n(p_2)$ (this is the angle of incidence) is the same as the angle between $n(p_2)$ and $\overline{p_2 p_3}$ (this is the angle of reflection).

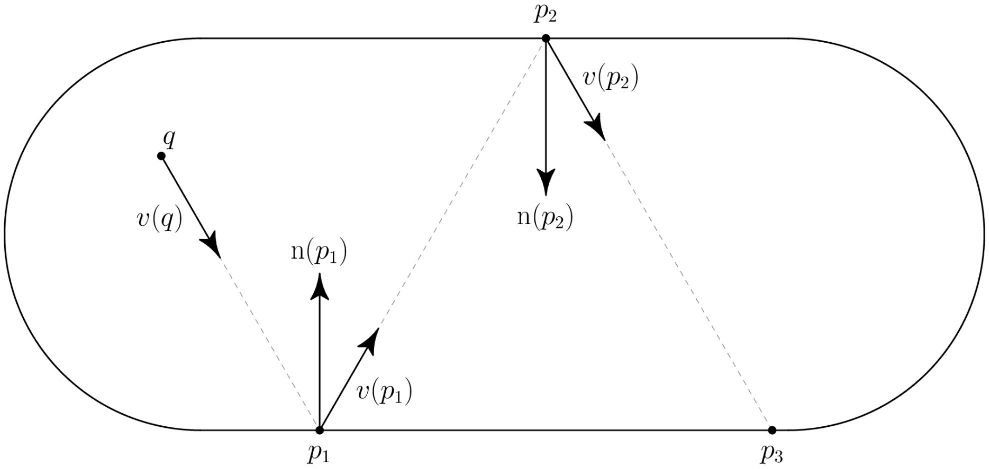

Fig. 1. Billiard trajectories on Bunimovich stadium

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.

These conditions result in the trajectory of an ideal billiard on a physical billiard table; the key point being condition (2).

Now we introduce some definitions to formalize billiards.

Definition 4. A *phase space* $\mathcal{M}$ of the billiard table Q is $\mathcal{M} = \overline{Q} \times S^1$ where $\overline{Q}$ is the closure of Q and S^1 is the unit circle of all velocity vectors (we are not concerned with the magnitude of the velocity in this paper). 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.

Note that S^1 is the set of all possible directions of the billiard trajectory. 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 5. Let the *hypersurface* M of the phase space $\mathcal{M}$ be defined as follows.

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

The inner product $\langle -, - \rangle$ is the standard inner product on $\mathbb{R}^2$. In other words, M consists of the set of tuples of points on the boundary of M, together with the possible angles of incidence. We define the *billiard map* $T : M \to M$ as $Tx = \Phi^{\tau(x)}x$ such that

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

In other words, this is the tuple consisting of the next point the ball hits the boundary, and the corresponding angle of incidence.

The significance of $\langle v, n(p)\rangle \geqslant 0$ is that the value $|v||n(p)|\cos\theta$ should be positive, i.e. the angle of incidence is between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. We put in this condition because otherwise the trajectory would go outside of M. We may consider the elements (p, v) in $\mathcal{M}$ 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.

3.1. Circular Billiards

We start by looking at the simplest dynamical billiard: the circle. We first consider a circular billiard table Q with radius 1. On the boundary of the billiard table Q, the angle of incidence has to be between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Thus the hypersurface $\mathcal{M}$ is $\partial Q \times [-\frac{\pi}{2}, \frac{\pi}{2}]$, a cylinder with radius 1 and height π.

We claim that the billiard map $T : \mathcal{M} \to \mathcal{M}$ 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.

Theorem 6. *The billiard map* $T : \mathcal{M} \to \mathcal{M}$ *is given by* $T^n x = (\varphi + n(\pi - 2\theta), \theta)$.

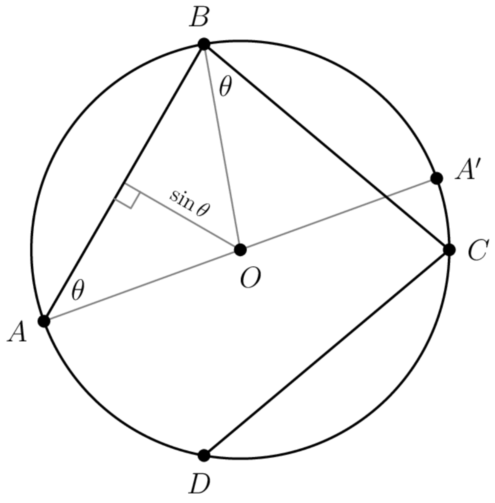

Fig. 2. Billiard trajectory in a circle.

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 $OA = OB$, $\angle OAB = \angle OBA$. The same relationship applies between two adjacent 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 $\angle OAB = \angle OBA = \theta$, $\angle BOA = \pi - 2\theta$. Thus through each billiard mapping the point moves along the arc the distance of $\pi - 2\theta$. □

This shows that billiards in the circle is equivalent to rotations; specifically rotation by $\frac{2\theta}{\pi}$. If this number is irrational, the billiards is ergodic, while it is periodic otherwise.

Below are some pictures of billiard trajectories for rational and irrational values of $\frac{2\theta}{\pi}$.

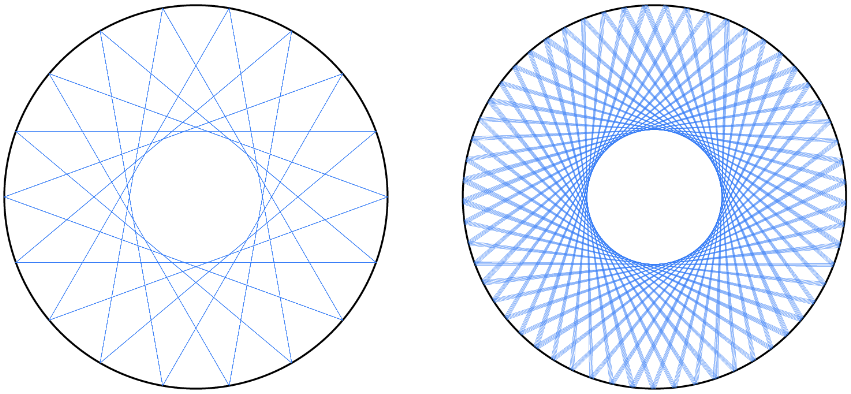

Fig. 3. Trajectories for rational and irrational values of $2\theta/\pi$

As we can see, there is a small circle inside the trajectory that all of the segments are tangent to. Note that the trajectory is more dense in the vicinity of this circle, so if we imagine the billiard is a laser beam and the boundary of the table is a mirror, this region will be significantly hotter than other areas. We call this circle the *caustic*, which comes from the Greek *kaiein*, which means *to burn*.

3.2. Circular Ring

Now we add a smaller circle inside the original circle, forming the circular ring.

Definition 7. 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

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

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

For the most part, i.e. when the trajectory does not hit the inner ring, billiards on the circular ring is the same as that on the circle. But even otherwise we can still map this to rotations in $\mathbb{R}/\mathbb{Z}$.

Theorem 8. *Suppose we have a trajectory that hits the inner ring. Let $\{a_n\}$ be the sequence of points on ∂Q_1 and $\{b_n\}$ be the sequence of points on ∂Q_2. Then, let θ be the initial angle of incidence; i.e. the angle of incidence at a_0 (note that $r_2 > r_1 \sin\theta$). Also define θ' to be the angle of incidence at b_0. Note that the ball hits the boundary in the order $a_0, b_0, a_1, b_1, \ldots$. We have, for every positive integer n for the billiard map T on R:*

$$T^{2n}((a_0,\theta)) = T^{2n-1}((b_0,\theta')) = (a_0 + 2nr_1(\theta' - \theta), \theta) \tag{3.1}$$

$$T^{2n+1}((a_0,\theta)) = T^{2n}((b_0,\theta')) = (b_0 + 2nr_2(\theta' - \theta), \theta') \tag{3.2}$$

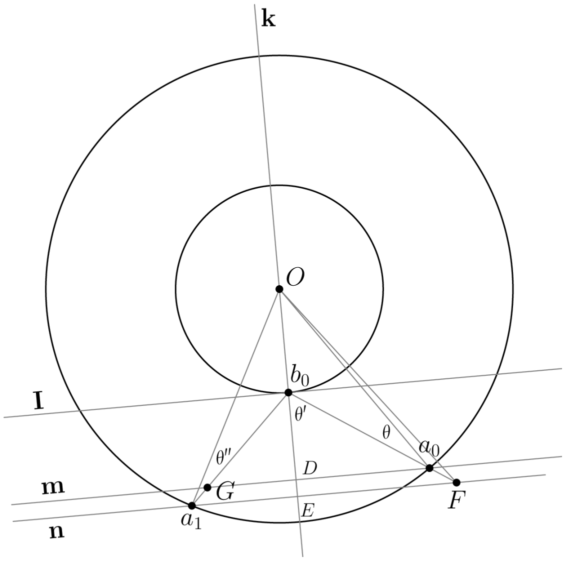

Fig. 4. Billiards on a circular ring

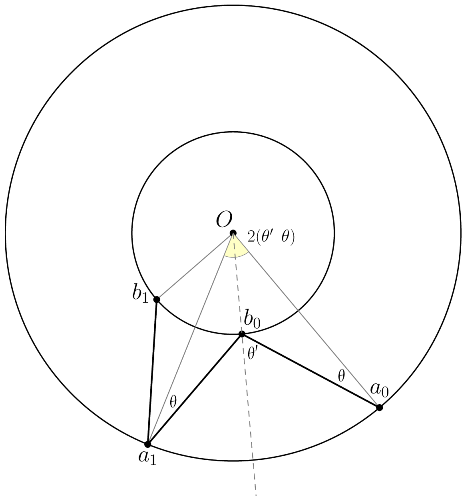

Fig. 5.

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; line m which passes through a_0, and 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 b_0a_0 and 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

$$\triangle Fb_0E \equiv \triangle a_1b_0E,$$

which implies $\triangle \mathrm{O}Fb_0 \equiv \triangle \mathrm{O}a_1b_0$. Then $\mathrm{O}F = \mathrm{O}a_1$, 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 (Fig. 5.)

$$\angle a_0\mathrm{O}a_1 = \angle b_0\mathrm{O}b_1 = 2(\theta' - \theta).$$

It is clear that for sequence $\{a_n\}$ the billiard mapping shifts the points along ∂Q_1 by $2r_1(\theta' - \theta)$ while for sequence $\{b_n\}$ the billiard mapping shifts the points along ∂Q_2 by $2r_2(\theta' - \theta)$.

This completes the proof. □

This shows that circular rings are essentially the same as circles, even when the trajectory hits the inner circle: if $(\theta' - \theta)/2\pi$ is rational, the trajectory is periodic, and otherwise the trajectory is ergodic.

3.3. Ellipses and Other Billiards

Finally, we state a set of results for ellipses; the proofs are rather long, so we will omit them. The interested reader can refer to [2] for complete proofs.

Theorem 9. *Let Q be an elliptical billiard table.*

(1) *Suppose the trajectory hits one of the foci of the ellipse. Then, the trajectory will converge to the semimajor axis of the ellipse.*

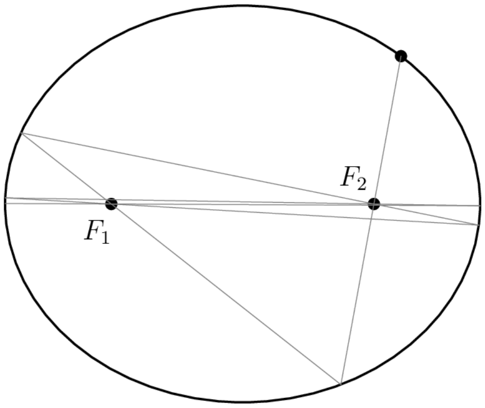

Fig. 5. Trajectory through one of the foci

(2) *Suppose the trajectory intersects the line segment connecting the two foci. Then, the resulting caustic will be a hyperbola with the same foci as the ellipse.*

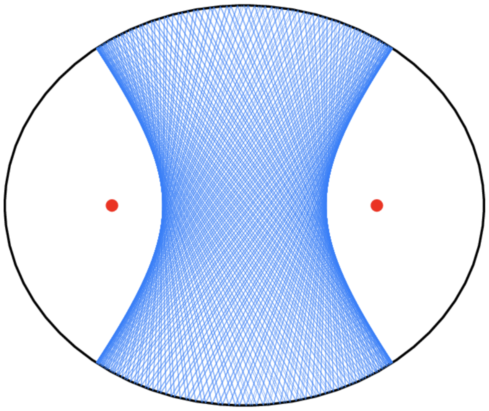

Fig. 6. A “hyperbolic” caustic

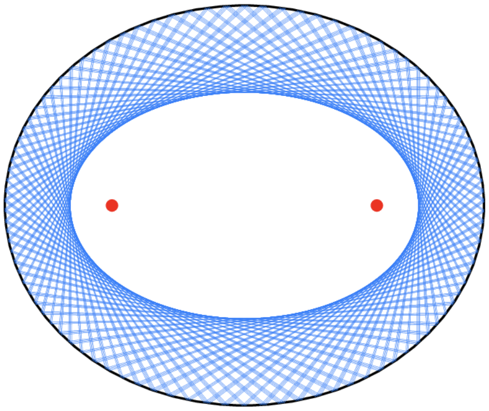

Fig. 7. An “elliptic” caustic

(3) *Suppose the trajectory never intersects the line segment connecting the two foci. Then, the resulting caustic will be an ellipse with the same foci as the ellipse.*

Finally, we include a brief introduction to chaotic billiards.

3.4. Chaotic Billiards

For an ordinary rectangular billiard table, if two paths start very close together, the deviation will eventually result in a significant change in the long run, but the growth of this deviation is a linear function in time. A chaotic billiard, on the other hand, is characterized by exponential growth in this deviation. One example of a chaotic billiard is the Sinai billiard (created by putting a circle into the center of a square).

The picture on the left (of the diagram below) shows how two paths that start close together remain together on a ordinary square billiard table, and the picture on the right shows how quickly two paths that started close together diverge on the Sinai billiard table.

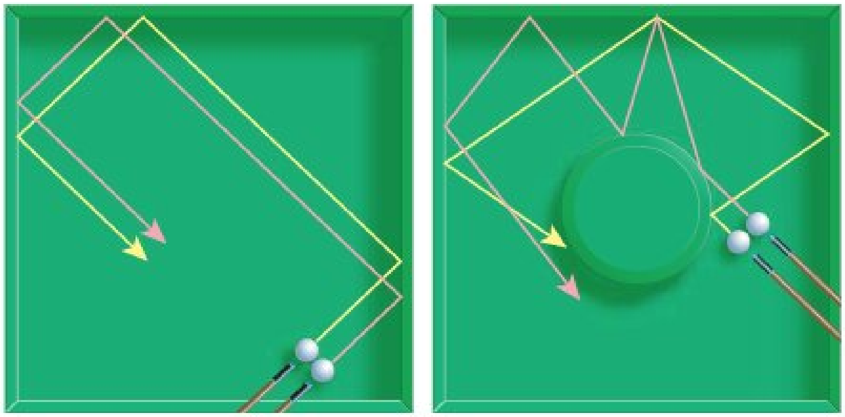

Fig. 8. "Sinai billiard", created by putting a circle into the center of a square

For a long time, it was believed that a concave shape was necessary for chaotic behavior. The reasoning behind this was that a concave shape acts as a dispersing mechanism: the curvature amplifies the angle between two slightly different trajectories. Thus, it was believed that a convex billiard table could never be chaotic.

In 1974, Leonid Bunimovich showed this was not true. He proved that the Bunimovich stadium, described in the beginning of this section, was chaotic even though the boundary was completely convex. The main idea is that even though the convex semicircles act as a focusing mechanism, after passing the focusing points, the trajectories in fact diverge, which is allowed by the length of the straight part of the table. Thus the focusing mechanism actually acts in reverse as a defocusing mechanism.

However, this reasoning is not sufficient to show that a billiard table is chaotic. In 1973, Lazutkin showed that any convex table with a differentiable boundary was ergodic. In fact, he proved it for any convex table whose boundary has 553 continuous derivatives! This number was later revised to 6 by Douady in 1982, and he conjectured that 4 was enough. See the blog post [3] for further discussion.

4 A Physical Application

Billiards have many applications to real-world problems. One example is the Sinai billiard, mentioned in the previous section. The Sinai billiard is a good model for the behavior of molecules in an ideal gas. In the model of an ideal gas, we consider many tiny molecules bouncing inside a square, and off each other. The Sinai billiard gives a simplified, but a very good illustration of this model.

We can also use a billiard table with one ball to model a physical setup involving two molecules.

Consider the following setup. There are two molecules, moving in the interval $[0,1]$, colliding elastically with the walls and also with themselves. Let m_1, m_2 be the masses of the left and right molecules, respectively, and let x_1, x_2 be the positions of the molecules. Note that $0 \leqslant x_1 \leqslant x_2 \leqslant 1$ always. Consider a collision between the two molecules. Let v_1, v_2 be the velocities before the collision and let w_1, w_2 be the velocities after the collision. Conservation of momentum and energy give

$$m_1 v_1 + m_2 v_2 = m_1 w_1 + m_2 w_2 \tag{momentum}$$

$$\tfrac{1}{2} m_1 v_1^2 + \tfrac{1}{2} m_2 v_2^2 = \tfrac{1}{2} m_1 w_1^2 + \tfrac{1}{2} m_2 w_2^2. \tag{energy}$$

Consider the triangle shaped billiard table with vertices at $(0,0)$, $(0, \sqrt{m_2})$, and $(\sqrt{m_1}, \sqrt{m_2})$.

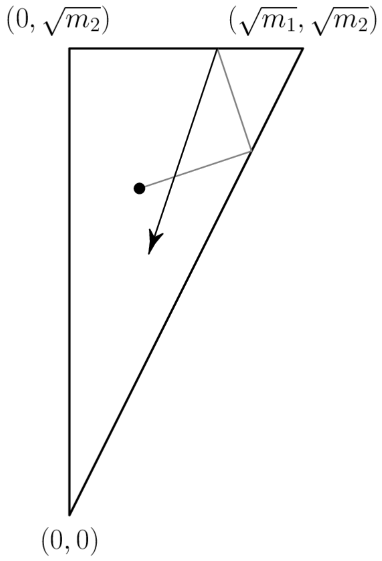

Fig. 9. Triangle shaped billiard table

Then we identify the point (x,y) with the state where the first molecule is at position $x_1 = \frac{x}{\sqrt{m_1}}$ and the second is at position $x_2 = \frac{y}{\sqrt{m_2}}$. Note that the shape of the triangle ensures that $0 \leqslant x_1 \leqslant x_2 \leqslant 1$. Solving the momentum and energy equations, we get

$$w_1 = \frac{2m_2 v_2 + (m_1 - m_2) v_1}{m_1 + m_2},$$

$$w_2 = \frac{2m_1 v_1 + (m_2 - m_1) v_2}{m_1 + m_2}.$$

Thus $v_2 - v_1 = w_1 - w_2$. Reflection along the hypotenuse of the billiard table gives that

$$(\sqrt{m_1}v_1, \sqrt{m_2}v_2) \cdot (-\sqrt{m_2}, \sqrt{m_1}) = -(\sqrt{m_1}w_1, \sqrt{m_2}w_2) \cdot (-\sqrt{m_2}, \sqrt{m_1})$$

which is equivalent to $-v_1 + v_2 = w_1 - w_2$. Thus, this triangle billiard completely describes the physical system of two molecules bouncing in the interval $[0, 1]$.

5 Other applications

Billiards have been applied in several areas of physics to model quite diverse real world systems. Examples include ray-optics, lasers, acoustics, optical fibers (e.g. double-clad fibers), or quantum-classical correspondence. One of their most frequent application is to model particles moving inside nanodevices, for example quantum dots, *pn*-junctions, antidot superlattices, among others. The reason for this broadly spread effectiveness of billiards as physical models resides on the fact that in situations with small amount of disorder or noise, the movement of e.g. particles like electrons, or light rays, is very much similar to the movement of the point-particles in billiards. In addition, the energy conserving nature of the particle collisions is a direct reflection of the energy conservation of Hamiltonian mechanics.

References

[1] Rubinstein-Salzedo, Simon (2023). *Ergodic Theory.* Lecture notes for Euler Circle, Winter 2023.

[2] Sun Woo Park (2014). *An Introduction to Dynamical Billiards.* math.uchicago.edu

[3] John Baez (2016). *Bunimovich Stadium, Visual Insight.* blogs.ams.org

[4] Arne B. Sletsjøe (2014). *Dynamical billiard.* abelprize.no

[5] Wikipedia contributors (2024). *Dynamical billiards.* In Wikipedia, The Free Encyclopedia.