CHAPTER 1

Simple examples

We start with a few simple examples of mathematical billiards, which will help us introduce basic features of billiard dynamics. This chapter is for the complete beginner. The reader familiar with some billiards may safely skip it – all the formal definitions will be given in Chapter 2.

1.1. Billiard in a circle

Let $\mathcal{D}$ denote the unit disk $x^2+y^2 \le 1$. Let a point-like (dimensionless) particle move inside $\mathcal{D}$ with constant speed and bounce off its boundary $\partial\mathcal{D}$ according to the classical rule *the angle of incidence is equal to the angle of reflection*; see below.

Denote by $q_t = (x_t, y_t)$ the coordinates of the moving particle at time t and by $v_t = (u_t, w_t)$ its velocity vector. Then its position and velocity at time $t+s$ can be computed by

$$\begin{aligned} x_{t+s} &= x_t + u_t s \qquad & u_{t+s} &= u_t \\ y_{t+s} &= y_t + w_t s \qquad & w_{t+s} &= w_t \end{aligned} \tag{1.1}$$

as long as the particle stays inside $\mathcal{D}$ (makes no contact with $\partial\mathcal{D}$).

When the particle collides with the boundary $\partial\mathcal{D} = \{x^2+y^2=1\}$, its velocity vector v gets reflected across the tangent line to $\partial\mathcal{D}$ at the point of collision; see Fig. 1.1.

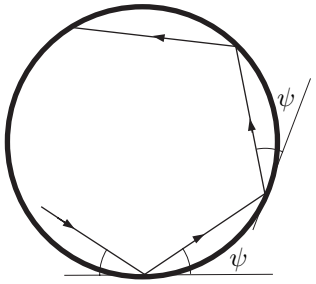

FIGURE 1.1. Billiard motion in a circle.

EXERCISE 1.1. Show that the new (postcollisional) velocity vector is related to the old (precollisional) velocity by the rule

$$v^{\text{new}} = v^{\text{old}} - 2\,\langle v^{\text{old}}, n\rangle\, n, \tag{1.2}$$

where $n = (x, y)$ is the unit normal vector to the circle $x^2+y^2=1$ and $\langle v, n\rangle = ux + wy$ denotes the scalar product.

After the reflection, the particle resumes its free motion (1.1) inside the disk $\mathcal{D}$, until the next collision with the boundary $\partial\mathcal{D}$. Then it bounces off again, and so on. The motion can be continued indefinitely, both in the future and the past.

For example, if the particle runs along a diameter of the disk, its velocity vector will get reversed at every collision, and the particle will keep running back and forth along the same diameter forever. Other examples of periodic motion are shown in Fig. 1.2, where the particle traverses the sides of some regular polygons.

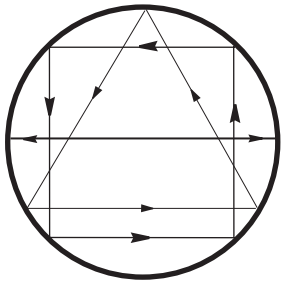

FIGURE 1.2. Periodic motion in a circle.

In the studies of dynamical systems, the primary goal is to describe the evolution of the system over long time periods and its asymptotic behavior in the limit $t \to \infty$. We will focus on such a description.

Let us parameterize the unit circle $x^2 + y^2 = 1$ by the polar (counterclockwise) angle $\theta \in [0, 2\pi]$ (since θ is a cyclic coordinate, its values 0 and 2π are identified). Also, denote by $\psi \in [0, \pi]$ the angle of reflection as shown in Fig. 1.1.

REMARK 1.2. We note that θ is actually an arc length parameter on the circle $\partial\mathcal{D}$; when studying more general billiard tables $\mathcal{D}$, we will always parameterize the boundary $\partial\mathcal{D}$ by its arc length. Instead of ψ, a reflection can also be described by the angle $\varphi = \pi/2 - \psi \in [-\pi/2, \pi/2]$ that the postcollisional velocity vector makes with the inward normal to $\partial\mathcal{D}$. In fact, all principal formulas in this book will be given in terms of φ rather than ψ, but for the moment we proceed with ψ.

For every $n \in \mathbb{Z}$, let θ_n denote the nth collision point and ψ_n the corresponding angle of reflection.

EXERCISE 1.3. Show that

$$\begin{aligned} \theta_{n+1} &= \theta_n + 2\psi_n \pmod{2\pi} \\ \psi_{n+1} &= \psi_n \end{aligned} \tag{1.3}$$

for all $n \in \mathbb{Z}$.

We make two important observations now:

- All the distances between reflection points are equal.
- The angle of reflection remains unchanged.

COROLLARY 1.4. *Let (θ_0, ψ_0) denote the parameters of the initial collision. Then*

$$\begin{aligned} \theta_n &= \theta_0 + 2n\psi_0 \qquad (\mathrm{mod}\ \ 2\pi) \\ \psi_n &= \psi_0. \end{aligned}$$

Every collision is characterized by two numbers: θ (the point) and ψ (the angle). All the collisions make the *collision space* with coordinates θ and ψ on it. It is a cylinder because θ is a cyclic coordinate; see Fig. 1.3. We denote the collision space by $\mathcal{M}$. The motion of the particle, from collision to collision, corresponds to a map $\mathcal{F}\colon \mathcal{M} \to \mathcal{M}$, which we call the *collision map*. For a circular billiard it is given by equations (1.3).

Observe that $\mathcal{F}$ leaves every horizontal level $\mathcal{C}_\psi = \{\psi = \text{const}\}$ of the cylinder $\mathcal{M}$ invariant. Furthermore, the restriction of $\mathcal{F}$ to $\mathcal{C}_\psi$ is a rotation of the circle $\mathcal{C}_\psi$ through the angle 2ψ. The angle of rotation continuously changes from circle to circle, growing from 0 at the bottom $\{\psi = 0\}$ to 2π at the top $\{\psi = \pi\}$ (thus the top and bottom circles are actually kept fixed by $\mathcal{F}$). The cylinder $\mathcal{M}$ is "twisted upward" ("unscrewed") by the map $\mathcal{F}$; see Fig. 1.3.

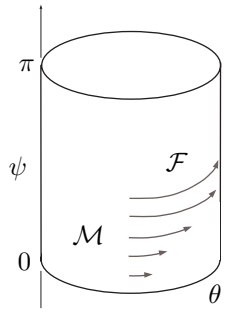

FIGURE 1.3. Action of the collision map $\mathcal{F}$ on $\mathcal{M}$.

Rigid rotation of a circle is a basic example in ergodic theory; cf. Appendix C. It preserves the Lebesgue measure on the circle. Rotations through rational angles are periodic, while those through irrational angles are ergodic.

EXERCISE 1.5. Show that if $\psi < \pi$ is a rational multiple of π, i.e. $\psi/\pi = m/n$ (irreducible fraction), then the rotation of the circle $\mathcal{C}_\psi$ is periodic with (minimal) period n, that is every point on that circle is periodic with period n, i.e. $\mathcal{F}^n(\theta, \psi) = (\theta, \psi)$ for every $0 \le \theta \le 2\pi$.

If ψ/π is irrational, then the rotation of $\mathcal{C}_\psi$ is ergodic with respect to the Lebesgue measure. Furthermore, it is *uniquely ergodic*, which means that the invariant measure is unique. As a consequence, for *every point* $(\psi, \theta) \in \mathcal{C}_\psi$ its images $\{\theta + 2n\psi, \quad n \in \mathbb{Z}\}$ are dense and uniformly distributed[1] on $\mathcal{C}_\psi$; this last fact is sometimes referred to as Weyl's theorem [**Pet83**, pp. 49–50].

[1]A sequence of points $x_n \in \mathcal{C}$ on a circle $\mathcal{C}$ is said to be uniformly distributed if for any interval $I \subset \mathcal{C}$ we have $\lim_{N\to\infty} \#\{n\colon 0 < n < N,\ a_n \in I\}/N = \text{length}(I)/\text{length}(\mathcal{C})$.

EXERCISE 1.6. Show that every segment of the particle's trajectory between consecutive collisions is tangent to the smaller circle $S_\psi = \{x^2 + y^2 = \cos^2\psi\}$ concentric to the disk $\mathcal{D}$. Show that if ψ/π is irrational, the trajectory densely fills the ring between $\partial\mathcal{D}$ and the smaller circle S_ψ (see Fig. 1.4).

Remark: One can clearly see in Fig. 1.4 that the particle's trajectory looks denser near the inner boundary of the ring (it "focuses" on the inner circle). If the particle's trajectory were the path of a laser ray and the border of the unit disk were a perfect mirror, then it would feel "very hot" there on the inner circle. For this reason, the inner circle is called a *caustic* (which means "burning" in Greek).

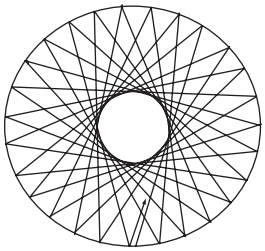

FIGURE 1.4. A nonperiodic trajectory.

EXERCISE 1.7. Can the trajectory of the moving particle be dense in the entire disk $\mathcal{D}$? (Answer: No.)

EXERCISE 1.8. Does the map $\mathcal{F}\colon \mathcal{M} \to \mathcal{M}$ preserve any absolutely continuous invariant measure $d\mu = f(\theta,\psi)\,d\theta\,d\psi$ on $\mathcal{M}$? Answer: Any measure whose density $f(\theta,\psi) = f(\psi)$ is independent of θ is $\mathcal{F}$-invariant.

Next, we can fix the speed of the moving particle due to the following facts.

EXERCISE 1.9. Show that $\|v_t\| = \text{const}$, so that the speed of the particle remains constant at all times.

EXERCISE 1.10. Show that if we change the speed of the particle, say we set $\|v\|_{\text{new}} = c\,\|v\|_{\text{old}}$ with some $c > 0$, then its trajectory will remain unchanged, up to a simple rescaling of time: $q_t^{\text{new}} = q_{ct}^{\text{old}}$ and $v_t^{\text{new}} = v_{ct}^{\text{old}}$ for all $t \in \mathbb{R}$.

Thus, the speed of the particle remains constant and its value is not important. It is customary to set the speed to one: $\|v\| = 1$. Then the velocity vector at time t can be described by an angular coordinate ω_t so that $v_t = (\cos\omega_t, \sin\omega_t)$ and $\omega_t \in [0, 2\pi]$ with the endpoints 0 and 2π being identified.

Now, the collision map $\mathcal{F}\colon \mathcal{M} \to \mathcal{M}$ represents collisions only. To describe the motion of the particle inside $\mathcal{D}$, let us consider all possible *states* (q, v), where $q \in \mathcal{D}$ is the position and $v \in S^1$ is the velocity vector of the particle. The space of all states (called the *phase space*) is then a three-dimensional manifold $\Omega\colon = \mathcal{D} \times S^1$, which is, of course, a solid torus (doughnut).

The motion of the billiard particle induces a continuous group of transformations of the torus Ω into itself. Precisely, for every $(q, v) \in \Omega$ and every $t \in \mathbb{R}$ the billiard particle starting at (q, v) will come to some point $(q_t, v_t) \in \Omega$ at time t.

Thus we get a map $(q, v) \mapsto (q_t, v_t)$ on Ω, which we denote by Φ^t. The family of maps $\{\Phi^t\}$ is a *group*; i.e. $\Phi^t \circ \Phi^s = \Phi^{t+s}$ for all $t, s \in \mathbb{R}$. This family is called the *billiard flow* on the phase space.

Let us consider a modification of the circular billiard. Denote by $\mathcal{D}_+$ the upper half disk $x^2 + y^2 \le 1$, $y \ge 0$, and let a point particle move inside $\mathcal{D}_+$ and bounce off $\partial\mathcal{D}_+$. (A delicate question arises here: what happens if the particle hits $\partial\mathcal{D}_+$ at $(1, 0)$ or $(-1, 0)$, since there is no tangent line to $\partial\mathcal{D}_+$ at those points? We address this question in the next section.)

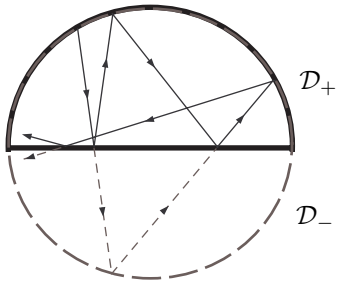

FIGURE 1.5. Billiard in the upper half circle.

A simple trick allows us to reduce this model to a billiard in the full unit disk $\mathcal{D}$. Denote by $\mathcal{D}_-$ the closure of $\mathcal{D} \setminus \mathcal{D}_+$, i.e. the mirror image of $\mathcal{D}_+$ across the x axis $L = \{y = 0\}$. When the particle hits L, its trajectory gets reflected across L, but we will also draw its continuation (mirror image) below L. The latter will evolve in $\mathcal{D}_-$ symmetrically to the real trajectory in $\mathcal{D}_+$ until the latter hits L again. Then these two trajectories will merge and move together in $\mathcal{D}_+$ for a while until the next collision with L, at which time they split again (one goes into $\mathcal{D}_-$ and the other into $\mathcal{D}_+$), etc.

It is important that the second (imaginary) trajectory never actually gets reflected off the line L; it just crosses L every time. Thus it evolves as a billiard trajectory in the full disk $\mathcal{D}$ as described above. The properties of billiard trajectories in $\mathcal{D}_+$ can be easily derived from those discussed above for the full disk $\mathcal{D}$. This type of reduction is quite common in the study of billiards.

EXERCISE 1.11. Prove that periodic trajectories in the half-disk $\mathcal{D}_+$ correspond to periodic trajectories in the full disk $\mathcal{D}$. Note, however, that the period (the number of reflections) may differ.

EXERCISE 1.12. Investigate the billiard motion in a quarter of the unit disk $x^2 + y^2 \le 1$, $x \ge 0$, $y \ge 0$.

1.2. Billiard in a square

Here we describe another simple example – a billiard in the unit square $\mathcal{D} = \{(x, y) \colon 0 \le x, y \le 1\}$; see Fig. 1.6. The laws of motion are the same as before, but this system presents new features.

First of all, when the moving particle hits a vertex of the square $\mathcal{D}$, the reflection rule (1.2) does not apply (there is no normal vector n at a vertex). The particle

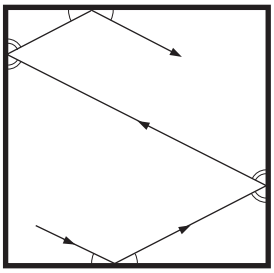

FIGURE 1.6. Billiard in a square.

then stops and its trajectory terminates. We will discuss this exceptional situation later. First we consider regular trajectories that never hit the vertices.

Let $v_t = (u_t, w_t)$ denote the velocity vector of the moving particle at time t (in the x, y coordinates). If it hits a vertical side of $\mathcal{D}$ at time t, then u_t changes sign ($u_{t+0} = -u_{t-0}$) and w_t remains unchanged. If the particle hits a horizontal side of $\mathcal{D}$, then w_t changes sign ($w_{t+0} = -w_{t-0}$) and u_t remains unchanged. Thus,

$$u_t = (-1)^m u_0 \quad \text{and} \quad w_t = (-1)^n w_0, \tag{1.4}$$

where m and n denote the number of collisions with vertical and, respectively, horizontal sides of $\mathcal{D}$ during the time interval $(0, t)$.

EXERCISE 1.13. Show that if $u_0 \neq 0$ and $w_0 \neq 0$ (and assuming the particle never hits a vertex), then all four combinations $(\pm u_0, \pm w_0)$ appear along the particle's trajectory infinitely many times.

Next we make use of the trick shown in Fig. 1.5. Instead of reflecting the trajectory of the billiard particle in a side of $\partial\mathcal{D}$, we reflect the square $\mathcal{D}$ across that side and let the particle move straight into the mirror image of $\mathcal{D}$. If we keep doing this at every collision, our particle will move along a straight line through the multiple copies of $\mathcal{D}$ obtained by successive reflections (the particle "pierces" a chain of squares; see Fig. 1.7). This construction is called the *unfolding* of the billiard trajectory. To recover the original trajectory in $\mathcal{D}$, one *folds* the resulting string of adjacent copies of $\mathcal{D}$ back onto $\mathcal{D}$.

We denote the copies of $\mathcal{D}$ by

$$\mathcal{D}_{m,n} = \{(x, y)\colon\ m \leq x \leq m+1,\ n \leq y \leq n+1\}. \tag{1.5}$$

EXERCISE 1.14. Show that if m and n are even, then the folding procedure transforms $\mathcal{D}_{m,n}$ back onto $\mathcal{D} = \mathcal{D}_{0,0}$ by translations $x \mapsto x - m$ and $y \mapsto y - n$, thus preserving orientation of both x and y. If m is odd, then the orientation of x is reversed (precisely, $x \mapsto m + 1 - x$). If n is odd, then the orientation of y is reversed (precisely, $y \mapsto n + 1 - y$). Observe that these rules do not depend on the particular trajectory that was originally unfolded.

The squares $\mathcal{D}_{m,n}$ with $m, n \in \mathbb{Z}$ tile, like blocks, the entire plane $\mathbb{R}^2$. Any regular billiard trajectory unfolds into a directed straight line on the plane, and any directed line (which avoids the sites of the integer lattice) folds back into a billiard trajectory. A trajectory hits a vertex of $\mathcal{D}$ iff the corresponding line runs into a site of the integer lattice.

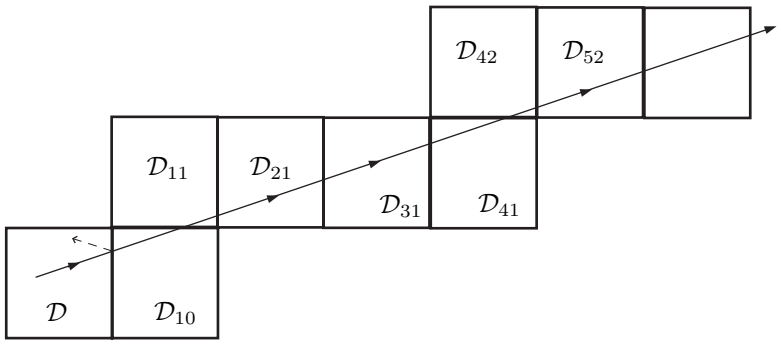

FIGURE 1.7. Unfolding a billiard trajectory.

The structure of blocks $\mathcal{D}_{m,n}$ with the respective folding rules is clearly periodic, in which the 2×2 square

$$\mathbb{K}_2 = \{(x, y) \colon 0 \le x, y \le 2\}$$

plays the role of a fundamental domain – the entire plane is covered by parallel translations of $\mathbb{K}_2$. Thus the standard projection of $\mathbb{R}^2$ onto $\mathbb{K}_2$ transforms unfolded trajectories into directed straight lines on the 2×2 torus Tor^2 (the latter is obtained by identifying the opposite sides of the square $\mathbb{K}_2$). The billiard in the unit square $\mathcal{D}$ thus reduces to the simple linear flow on a flat 2×2 torus Tor^2, in which points move with constant (unit) velocity vectors.

The linear flow on a flat torus is one of the standard examples in ergodic theory; cf. Appendix C and [**KH95, Pet83, Sin76**]. Its main properties are these:

- if a trajectory has rational slope $dy/dx \in \mathbb{Q}$, then it is periodic (it runs along a closed geodesic);
- if a trajectory has irrational slope $dy/dx \notin \mathbb{Q}$, then it is dense (its closure is the whole torus).

This translates into the following alternative for regular billiard trajectories in the unit square $\mathcal{D}$:

COROLLARY 1.15. *If $w_0/u_0 \in \mathbb{Q}$, then the corresponding regular billiard trajectory in the unit square $\mathcal{D}$ is periodic. If $w_0/u_0 \notin \mathbb{Q}$, then the corresponding regular billiard trajectory is dense.*

EXERCISE 1.16. Extend this result to the billiard in a rectangle $\mathcal{R}$ with sides a and b. Answer: a regular billiard trajectory in $\mathcal{R}$ is periodic iff $(aw_0)/(bu_0) \in \mathbb{Q}$; otherwise it is dense. Hint: Transform the rectangle into the unit square by scaling the coordinates: $(x, y) \mapsto (x/a, y/b)$. Argue that the billiard trajectories in $\mathcal{R}$ will be thus transformed into those in $\mathcal{D}$.

EXERCISE 1.17. Extend the above result to billiards in the following polygons: an equilateral triangle, a right isosceles triangle, a right triangle with the acute angle $\pi/6$, and a regular hexagon. What is common about these polygons? (Note that the billiard in a hexagon does not reduce to a geodesic flow on a torus. Does it reduce to a geodesic flow on another manifold?)

The phase space of the billiard system in the unit square $\mathcal{D}$ is the three-dimensional manifold $\Omega = \mathcal{D} \times S^1$; cf. the previous section. The billiard flow Φ^t is defined for all times $-\infty < t < \infty$ on regular trajectories. On exceptional trajectories (which hit a vertex of $\mathcal{D}$ at some time), the flow is defined only until the trajectory terminates in a vertex.

EXERCISE 1.18. Show that the set of exceptional trajectories is a countable union of 2D surfaces in Ω.

We see that the set of exceptional trajectories is negligible in the topological and measure-theoretic sense (it has zero Lebesgue measure and is an F_σ set, i.e. a countable union of nowhere dense closed subsets), but still its presence is bothersome. For the billiard in a square, though, one can get rid of them altogether by extending the billiard flow by continuity.

EXERCISE 1.19. Show that the flow Φ^t can be uniquely extended by continuity to all exceptional trajectories. In that case every trajectory hitting a vertex of $\mathcal{D}$ will simply reverse its course and run straight back; see Fig. 1.8.

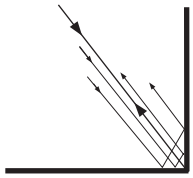

FIGURE 1.8. Extension of the flow near a vertex.

The above extension defines the billiard flow Φ^t on the entire phase space Ω and makes it continuous everywhere. We will assume this extension in what follows. We remark, however, that in generic billiards such nice extensions are rarely possible; see Section 2.8.

Now, the action of the flow Φ^t on the phase space Ω can be fully described as follows. For every unit vector $v_0 = (u_0, w_0) \in S^1$, consider the set

$$\mathcal{L}_{v_0} = \{(q, v) \in \Omega\colon\ q \in \mathcal{D},\ v = (\pm u_0, \pm w_0)\}$$

(the two signs are, of course, independent). Due to (1.4), each set $\mathcal{L}_{v_0}$ remains invariant under the flow Φ^t.

Suppose first that $u_0 \neq 0$ and $v_0 \neq 0$; then $\mathcal{L}_{v_0}$ is the union of four squares obtained by "slicing" Ω at the four "levels" corresponding to the vectors $(\pm u_0, \pm w_0)$; see Fig. 1.9.

EXERCISE 1.20. Check that the four squares constituting the set $\mathcal{L}_{v_0}$ can be glued together along their boundaries, and obtain a smooth closed surface without boundary (a 2×2 torus) $\mathbb{T}^2_{v_0}$ on which the billiard flow will coincide with the linear flow along the vector v_0 (i.e. the flow on $\mathbb{T}^2_{v_0}$ will be defined by differential equations $\dot{x} = u_0$, $\dot{y} = w_0$). Hint: The assembly of the torus $\mathbb{T}^2_{v_0}$ from the squares of $\mathcal{L}_{v_0}$ is very similar to the reduction of the billiard dynamics in $\mathcal{D}$ to the geodesic flow on the 2×2 torus described above (in fact, these two procedures are equivalent).

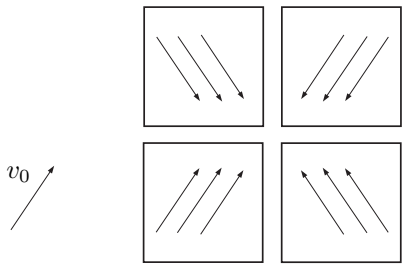

FIGURE 1.9. Four squares constituting $\mathcal{L}_{v_0}$.

Now, it is a standard fact in ergodic theory (cf. Appendix C) that the linear flow on a 2D torus defined by $\dot{x} = u_0$, $\dot{y} = w_0$ is periodic if $w_0/u_0 \in \mathbb{Q}$ and ergodic (furthermore, uniquely ergodic) if $w_0/u_0 \notin \mathbb{Q}$. In the latter case every trajectory is dense and uniformly distributed[2] on the torus.

In the two remaining cases (first $u_0 = 0$, and second $w_0 = 0$) the set $\mathcal{L}_{v_0}$ consists of just two squares. We leave their analysis to the reader as an easy exercise.

This fully describes the action of the flow $\Phi^t \colon \Omega \to \Omega$ for the billiard in the unit square.

1.3. A simple mechanical model

As a motivation for the study of billiards, one usually describes a simple model of two moving particles in a one-dimensional container. It reduces to a billiard in a right triangle, which is similar to a billiard in a square. We describe this model here; see also [**CFS82, CM03**].

Consider a system of two point particles of masses m_1 and m_2 on a unit interval $0 \le x \le 1$. The particles move freely and collide elastically with each other and with the 'walls' at $x = 0$ and $x = 1$. Let x_1 and x_2 denote the positions of the particles and u_1 and u_2 their velocities. Since the particles collide upon contact, their positions remain ordered; we assume that $x_1 \le x_2$ at all times.

FIGURE 1.10. Two particles in a unit interval.

Next we describe collisions. When a particle hits a wall, it simply reverses its velocity. When the two particles collide with each other, we denote by u_i^- the precollisional velocity and by u_i^+ the postcollisional velocity of the ith particle, $i = 1, 2$. The law of elastic collisions requires the conservation of the total momentum, i.e.

$$m_1 u_1^+ + m_2 u_2^+ = m_1 u_1^- + m_2 u_2^-$$

[2] A line x_t on a 2D torus Tor^2 is said to be uniformly distributed if for any rectangle $R \subset \mathrm{Tor}^2$ we have $\lim_{T\to\infty} \mathbf{m}\big(\{t \colon 0 < t < T,\ x_t \in R\}\big)/T = \mathrm{area}(R)/\mathrm{area}(\mathrm{Tor}^2)$. Here $\mathbf{m}$ is the Lebesgue measure on $\mathbb{R}$.

and the total kinetic energy, i.e.

$$m_1[u_1^+]^2 + m_2[u_2^+]^2 = m_1[u_1^-]^2 + m_2[u_2^-]^2. \tag{1.6}$$

Solving these equations gives

$$u_1^+ = u_1^- + \frac{2m_2}{m_1+m_2}(u_2^- - u_1^-)$$

and

$$u_2^+ = u_2^- + \frac{2m_1}{m_1+m_2}(u_1^- - u_2^-)$$

(we recommend the reader derive these formulas for an exercise). Note that if $m_1 = m_2$, then the particles simply exchange their velocities: $u_1^+ = u_2^-$ and $u_2^+ = u_1^-$.

The variables x_i and u_i are actually inconvenient, so we will work with new variables defined by

$$q_i = x_i\sqrt{m_i} \qquad \text{and} \qquad v_i = dq_i/dt = u_i\sqrt{m_i} \tag{1.7}$$

for $i = 1, 2$. Now the positions of the particles are described by a point $\mathbf{q} = (q_1, q_2) \in \mathbb{R}^2$ (it is called a *configuration point*). The set of all configuration points (called the *configuration space*) is the right triangle

$$\mathcal{D} = \{\mathbf{q} = (q_1, q_2) : 0 \le q_1/\sqrt{m_1} \le q_2/\sqrt{m_2} \le 1\}.$$

The velocities of the particles are described by the vector $\mathbf{v} = (v_1, v_2)$. Note that the energy conservation law (1.6) implies that $\|\mathbf{v}\| = \text{const}$; thus we can set $\|\mathbf{v}\| = 1$.

The state of the system is described by a pair $(\mathbf{q}, \mathbf{v})$. The configuration point $\mathbf{q}$ moves in $\mathcal{D}$ with velocity vector $\mathbf{v}$. When the first particle collides with the wall ($x_1 = 0$), the configuration point hits the left side $q_1 = 0$ of the triangle $\mathcal{D}$. When the second particle collides with the wall ($x_2 = 1$), the point $\mathbf{q}$ hits the upper side $q_2/\sqrt{m_2} = 1$ of $\mathcal{D}$. When the particles collide with each other, the point $\mathbf{q}$ hits the hypotenuse $q_1/\sqrt{m_1} = q_2/\sqrt{m_2}$ of $\mathcal{D}$.

Exercise 1.21. Prove that the velocity vector $\mathbf{v}$ changes at collisions so that it gets reflected at $\partial\mathcal{D}$ according to the law 'the angle of incidence is equal to the angle of reflection'.

Thus, the motion of the configuration point $\mathbf{q}$ is governed by the billiard rules. Hence the evolution of the mechanical model of two particles in a unit interval reduces to billiard dynamics in a right triangle.

If $m_1 = m_2$, we obtain a billiard in a right isosceles triangle, which readily reduces to a billiard in a square; see Exercise 1.17. For generic mass ratio m_1/m_2, we obtain a billiard in a generic right triangle, which may be rather complicated (such billiards are not covered in our book).

One complication arises when a billiard trajectory hits a corner point of $\mathcal{D}$. Hitting the vertex of the right angle corresponds to an event when both particles simultaneously collide with opposite walls. Then their further motion is clearly well defined; thus the billiard trajectory can easily be continued (cf. Exercise 1.19).

However, hitting the vertex of an acute angle of $\mathcal{D}$ corresponds to an event when both particles simultaneously collide with the *same* wall (either $x = 0$ or $x = 1$). In this case, for generic m_1 and m_2, the billiard flow cannot be extended by continuity, as nearby trajectories hitting the two adjacent sides in different order will come back to $\mathcal{D}$ along different lines; see Fig. 1.11.

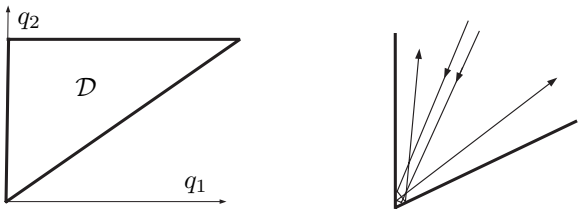

FIGURE 1.11. The right triangle $\mathcal{D}$; hitting the vertex of an acute angle.

In mechanical terms, hitting the vertex of an acute angle of $\mathcal{D}$ corresponds to a *multiple collision*. Such exceptional events usually cannot be resolved by the laws of classical mechanics.

1.4. Billiard in an ellipse

We proceed to yet another simple example that admits a completely elementary analysis – the billiard in an ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

with some $a > b > 0$. In fact, it was this example that Birkhoff described in the very first book on mathematical billiards in 1927 [**Bi27**, Chapter VIII].

We denote by $\mathcal{D}$ the domain bounded by the ellipse (it will be our billiard table). Let F_1 and F_2 denote the foci of the ellipse, and observe that they lie on the x axis. The ellipse is the locus of points $A \in \mathbb{R}^2$ such that

$$\operatorname{dist}(A, F_1) + \operatorname{dist}(A, F_2) = \text{ const.}$$

EXERCISE 1.22. Let $A \in \partial\mathcal{D}$ and L denote the tangent line to the ellipse at A. Prove that the segments AF_1 and AF_2 make equal angles with L. (This fact is known in projective geometry as the Poncelet theorem.) Hint: Reflect the point F_2 across the tangent line L and show that its image will lie on the line AF_1.

Thus, if a billiard trajectory passes through one focus, then it reflects at a point $A \in \partial\mathcal{D}$ on the ellipse and runs straight into the other focus. Such a trajectory will then pass through a focus after every reflection; see Fig. 1.12.

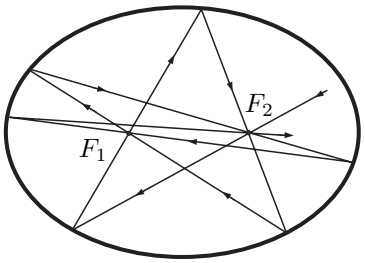

FIGURE 1.12. A trajectory passing through the foci.

EXERCISE 1.23. Show that every trajectory passing through the foci F_1 and F_2 converges to the major axis of the ellipse (the x axis).

By the way, the major and the minor axes of the ellipse are clearly two periodic trajectories – they run back and forth between their endpoints.

In Section 1.1 we used the coordinates ψ and θ to describe collisions in a circular billiard, and the cyclic coordinate θ was actually the arc length parameter on the circle (Remark 1.2). Here we use two coordinates ψ and r, where ψ is the same angle of reflection as in Section 1.1 and r is an arclength parameter on the ellipse. We choose the reference point $r = 0$ as the rightmost point $(a, 0)$ on the ellipse and orient r counterclockwise. Note that $0 \le r \le |\partial\mathcal{D}|$ and $0 \le \psi \le \pi$.

The collision space $\mathcal{M}$ is again a cylinder whose base is the ellipse and whose height is π. It is shown in Fig. 1.13 as a rectangle $[0, |\partial\mathcal{D}|] \times [0, \pi]$, but we keep in mind that the left and right sides of this rectangle must be identified. The motion of the billiard particle, from collision to collision, induces the collision map $\mathcal{F}\colon \mathcal{M} \to \mathcal{M}$.

EXERCISE 1.24. Verify that the trajectories passing through the foci lie on a closed curve on the surface $\mathcal{M}$. Determine its shape. Answer: It is the ∞-shaped curve in Fig. 1.13 that separates the white and grey areas.

Thus, the trajectories passing through the foci make a special (one-dimensional) family in $\mathcal{M}$.

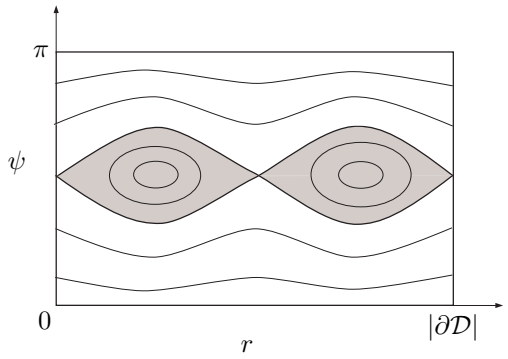

FIGURE 1.13. The collision space of an elliptic billiard.

EXERCISE 1.25. Show that if the trajectory of the billiard particle crosses the segment F_1F_2 joining the foci, then it reflects at $\partial\mathcal{D}$ and crosses this segment again. Similarly, if the trajectory crosses the major axis beyond the segment F_1F_2, say to the left of it, then after one or more reflections at $\partial\mathcal{D}$ it will cross the major axis to the right of this segment, etc.

The previous exercise shows that there are trajectories of two types: those crossing the inner segment F_1F_2 of the major axis after every reflection (we call them *inner* trajectories) and those going around this segment (we call them *outer* trajectories).

EXERCISE 1.26. Verify that the inner trajectories fill the white area in Fig. 1.13, and outer trajectories fill the grey area.

The following is the most important property of elliptic billiards:

THEOREM 1.27. *For every outer trajectory there is an ellipse with foci F_1 and F_2 that is tangent to each link of that trajectory. For every inner trajectory there is a hyperbola with foci F_1 and F_2 that is tangent to each link (or its linear extension) of that trajectory.*

PROOF. We prove only the first statement (about the outer trajectories); the proof of the second is similar. The argument is pretty elementary and illustrated in Fig. 1.14. Here A_1A and A_2A are two successive links of an outer trajectory. The points B_1 and B_2 are obtained by reflecting the foci F_1 and F_2 across the lines A_1A and A_2A, respectively. The four angles $\angle B_1AA_1$, $\angle A_1AF_1$, $\angle F_2AA_2$ and $\angle A_2AB_2$ are equal. Hence the triangles AB_1F_2 and AB_2F_1 are congruent, in particular $|B_1F_2| = |B_2F_1|$. Therefore

$$|F_1C_1| + |F_2C_1| = |F_1C_2| + |F_2C_2|,$$

where C_1 and C_2 are the points of intersection of A_1A with B_1F_2 and A_2A with B_2F_1, respectively. Thus, the points C_1 and C_2 belong to the same ellipse with foci F_1 and F_2, and the lines A_1A and A_2A are tangent to that ellipse. □

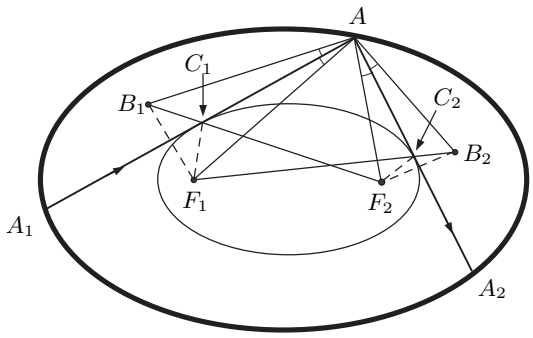

FIGURE 1.14. Proof of Theorem 1.27.

If every link of a billiard trajectory is tangent to a certain given curve, then that curve is called a *caustic*. Fig. 1.15 shows an elliptic caustic for an outer trajectory and a hyperbolic caustic for an inner trajectory. The term 'caustic' is borrowed from optics, where it means a curve on which light rays focus after being reflected off a mirror (we have seen caustics in circular billiards in Section 1.1). Fig. 1.15 demonstrates the concentration of rays on caustics (compare it to Fig. 1.4).

All the trajectories tangent to one elliptic caustic lie on a closed curve in the collision space $\mathcal{M}$. Such curves are shown as 'horizontal waves' in the white area in Fig. 1.13 (remember that the left and right sides of the rectangle need to be identified). Every such curve is obviously invariant under the map $\mathcal{F}$.

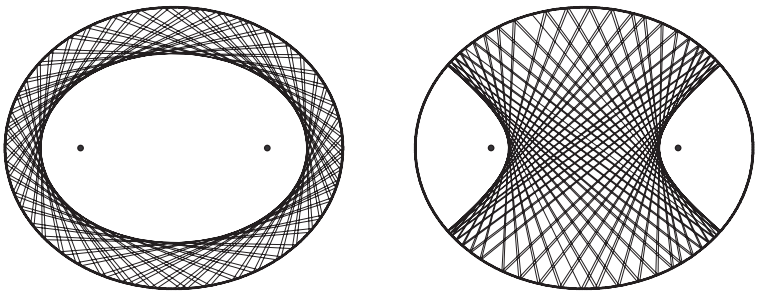

FIGURE 1.15. Elliptic and hyperbolic caustics in the elliptic billiard.

EXERCISE 1.28. On each invariant curve the map $\mathcal{F}$ is conjugate to a rigid circle rotation through some angle (that angle is called the *rotation number*). Show that the rotation number changes continuously and monotonically with the invariant curve. Hint: Consider two outer trajectories starting at the same point $A_0 \in \partial\mathcal{D}$ but with distinct elliptical caustics. Denote by A'_n the reflection points of the trajectory whose elliptical caustic is smaller and by A''_n those of the other trajectory. Observe that the sequence $\{A'_n\}$ will move along the ellipse faster than $\{A''_n\}$ does; see Fig. 1.16.

The action of the map $\mathcal{F}$ on each invariant curve can be analyzed explicitly and the rotation number can be computed analytically (see [**Be01**, Sections 2.5 and 3.2]) but we will not go that far.

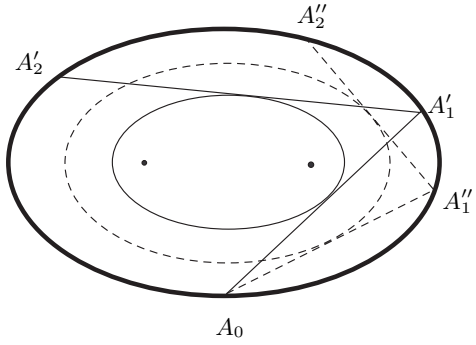

FIGURE 1.16. Exercise 1.28.

Next, all the trajectories tangent to one hyperbolic caustic lie on two closed curves in $\mathcal{M}$, one inside each half of the ∞-shaped grey domain. Such curves appear as ovals in Fig. 1.13. The map $\mathcal{F}$ transforms each oval into an identical oval within the other half of the ∞-shaped grey domain. Thus the union of the two identical (symmetric) ovals will be invariant under $\mathcal{F}$, and each oval separately will be invariant under $\mathcal{F}^2$.

Therefore, the collision space $\mathcal{M}$ of an elliptical billiard is completely foliated by invariant curves. In this sense, the elliptical billiard is similar to those in a circle and in a square. In physics, such models belong to a special class: if the phase

space of a system is foliated by one-dimensional invariant submanifolds, the system is said to be *integrable*; the dynamics in such a system is completely regular. Thus, billiards in circles, squares and ellipses are completely regular.

1.5. A chaotic billiard: pinball machine

The simple examples were given in the previous sections for the sake of introducing some basic features of billiards to the novice reader. But they should not be taken as typical; in fact their dynamical characteristics are quite special and in a sense opposite to those of chaotic billiards which will be covered in the rest of the book. Here we will take a glimpse at something that happens in chaotic billiards.

Imagine you are playing a pinball machine. A small ball shoots from a cannon in the right bottom corner of a rectangular table; then it bounces off the edges until it either hits the target (then you win) or falls through an opening in the bottom (goes down the drain; then you lose). The target might be a special figure on the table that registers the hit when the ball touches it. To prevent direct hits, assume the target is screened from the cannon and hitting the screen is forbidden by the rules. Then the ball has to bounce off the edges before reaching the target; see Fig. 1.17. It is quite an unusual pinball machine, but for us it is a good starting example.

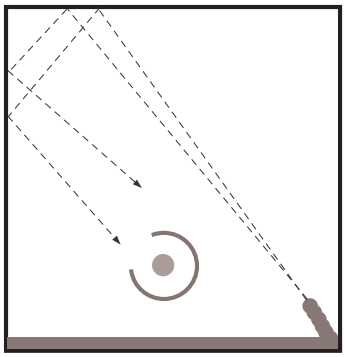

FIGURE 1.17. A pinball machine. The target is the grey disk screened from the cannon by a dark grey arc.

Suppose you can rotate the cannon to change the angle at which the ball shoots out. After you miss once, you can adjust the shooting angle and send the ball more accurately into the target. This is a relatively easy task (illustrated in Fig. 1.17), as the trajectory of the ball (in a rectangular billiard) is very simple and predictable. Also, you do not need to aim with absolute precision.

EXERCISE 1.29. Suppose the target is a disk of radius r and the moving ball is a point particle. Let L denote the distance covered by the ball from the cannon to the target. Show that if the shooting angle is off by less than r/L (radians), then the ball still hits the target.

Now let us make the task more realistic and challenging by installing some bumpers (round pillars) all over the table (see Fig. 1.18) so that our moving ball

will bounce between the bumpers on its way to the target (or down the drain). Anyone who has played real pinball machines can easily imagine such a process.

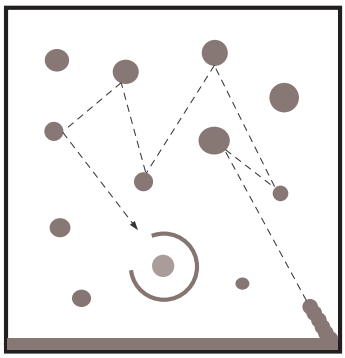

FIGURE 1.18. A pinball machine with bumpers (dark grey disks).

Would it be easy to adjust the cannon on this new table? Obviously not. The route of the ball is complicated and almost unpredictable, as it may bounce off different bumpers. Even finding the right sequence of bumpers that the ball needs to hit before it reaches the target (and avoids the screen) is not a simple task. Professional billiard players solve a similar problem when trying the send a ball to a pocket, so that it hits one or more other balls.

Furthermore, the cannon must be aimed with almost ultimate precision, as a tiny error in the shooting angle may send the ball rolling down along a completely wrong path. This is illustrated in Fig. 1.19, where just two successive bounces are shown. It is quite clear that the instability of the ball's motion increases with every subsequent reflection off a bumper. Again, professional billiard players know that if their ball needs to hit more than one other ball before sending one of them into a pocket, their task is very difficult. Furthermore, if the ball must hit three or more other balls, the task is almost impossible.

A rectangular table with round bumpers installed on it is a classical example of a chaotic billiard. The motion of the billiard particle on such a table is complicated and unpredictable. To the naked eye, it may look like a wild dance between the walls of the table, without any pattern or logic (that is why pinball machines are so attractive!). The lack of predictability is characteristic of chaotic billiards.

Furthermore, slight changes in the initial position and/or velocity of the particle quickly lead to large deviations (such as that in Fig. 1.19), so after just a few collisions with bumpers two trajectories, initially very close together, will separate and move far apart from each other, as if they were unrelated. This instability (also known as *sensitivity to initial conditions*) is another characteristic feature of chaotic billiards (and chaotic dynamics in general).

In practical terms, the best thing the player can do in our game is to shoot randomly and watch the ball running all over the table, bouncing around between bumpers – there will always be a chance that it hits the target 'by accident'. This is essentially a game of chance, just like flipping a coin, rolling a die, or playing cards. We will see in Chapters 6 and 7 that the motion in a chaotic billiard is indeed essentially random and is best described in terms of probability theory.

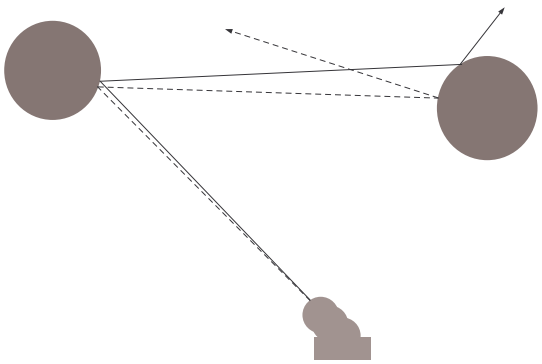

FIGURE 1.19. A ball bouncing off two bumpers: a slight error in the initial shooting angle results in a dramatic deflection in the end.

Our toy example actually has a lot in common with a classical model of statistical physics, called the Lorentz gas. In that model, a small ball (electron) bounces between large fixed disks (molecules) that make a regular periodic (crystalline) structure. We will present it in Chapter 5.6.

We will not attempt to go beyond this very informal introduction to the realm of chaotic billiards, leaving all formalities till future chapters. Interested readers may find a more extensive description of chaotic billiards, including computer illustrations, in [**Be01**, Section 1.1].