Lecture 1: Introduction to planar billiards

Let $D \subset \mathbb{R}^2$ be a compact planar domain bounded by a smooth curve $\partial D$. The dynamics of a billiard ball (which we assume to be a point mass) in $D$ is defined as follows: the ball is moving inside $D$ along a straight line with constant velocity, until it hits the boundary. Upon hitting the boundary, the ball is reflected from it according to the law the angle of reflection is equal to the angle of incidence, after which it continues moving along a straight line with constant velocity until it hits the boundary again. Then it is again reflected according to the same law, and so on. See Figure 1.

We would like to understand the dynamics of the ball, i.e. find, if possible, its position and velocity as a function of time, as well of the initial position and velocity. To begin with, we need to describe the *phase space* of the ball, that is the space of all its possible positions and velocities. There are two possible approaches to defining this space for the billiard system. The first one is to consider all possible positions,
including those strictly inside the domain $D$. However, in this approach, the dynamics is trivial on most of the phase space, because when the ball is strictly inside $D$, it is just moving along a straight line. The second approach is to only look at the positions of the ball at the boundary $\partial D$. Knowing the current position of the ball at the boundary, and also its current velocity, we can easily recover the subsequent dynamics of the ball, at least until it hits the boundary next time. So, we can only trace subsequent positions of the ball at the boundary, while all other positions are easily recovered. Also notice that since the magnitude of the velocity vector is preserved and does not affect the dynamics at all, we can assume that the velocity is given by a unit vector. Finally, notice that there is two kind of possible velocity vectors of the ball at the boundary: outward velocities, and inward velocities. Outward velocity means that the ball has just hit the boundary, and has not yet been reflected, while inward velocities are velocities after reflection.

**Remark 1.1.** Let $\gamma(t)$ be a parametrization of $\partial D$, corresponding to its counter-clockwise orientation. We say that a vector $v$ at $x = \gamma(t_0)$ is *inward* if the orientation of the frame $(\gamma'(t_0), v)$ is positive, and *outward* if the orientation of the frame $(\gamma'(t_0), v)$ is negative, see Figure 2. Also note that a velocity vector which is neither inward, nor outward, must be tangent to the boundary. Such velocities are only possible for billiards in non-convex domains, see Figure 3.

Since inward velocities and outward velocities are connected by the reflection law, it suffices to consider just one type of velocity vectors, for example the inward ones. So, we define the *phase space* of a billiard ball in $D$ as the space of inward unit vectors attached at points of $\partial D$. In other words, an element of the phase space is a pair $(x, v)$, where $x \in \partial D$, and $v \in T_x D$ is an inward tangent vector. Note that since $\partial D$ is homeomorphic to a circle, while an inward tangent vector $v$ is characterized by the angle $\alpha \in (0, \pi)$ it makes with the positive direction of $\partial D$ (see Figure 4), our phase space can be viewed as the cylinder $S^1 \times (0, \pi)$. As coordinates on this cylinder, we can take the angle $\alpha \in (0, \pi)$, as well as any coordinate on $S^1 = \partial D$. In what follows, as coordinate on $\partial D$ we will use the arc-length parameter $t$. So, the phase space of the billiard in $D$ is a cylinder with coordinates $(t, \alpha)$, where $t$ is a real number defined up to an integer multiple of the total length $L$ of $\partial D$, and $\alpha \in (0, \pi)$.

Let $M$ be the phase cylinder of the billiard in $D$. Define the billiard map $T : M \to M$ as follows: for an outward velocity vector $v$ at a boundary point $x \in \partial D$, its image under $T$ is the velocity of the

![Figure 2: $v$ is an inward velocity vector, and $w$ is an outward velocity vector.](image1)

![Figure 3: In a non-convex domain $D$, the velocity of a billiard ball may be tangent to the boundary.](image2)
Figure 4: An element of the phase space for the billiard in $D$ is a point $x \in D$ and an inward tangent vector at $v \in T_x D$. The latter vector can be characterized by the angle $\alpha \in (0, \pi)$ it makes with the positive direction of $\partial D$.

Figure 5: Definition of the billiard map $T$.

Figure 6: Iterations of the billiard map $T$.

Billiard ball with initial velocity $v$ immediately after it is reflected from the boundary for the first time, see Figure 5. Then the iterate $T^k$ computes the velocity of the ball after it hits the boundary for the $k$’th time (see Figure 6). So, the dynamics of the billiard ball is pretty much determined by the behavior of the billiard map $T$: $M \to M$ and its iterates. For this reason, most (if not all) people studying billiards study properties of this map.

The first thing we would like to understand about this map $T$ is how regular it is. This very much depends on the properties of the domain $D$ and its boundary $\partial D$. In particular, if $D$ is not convex, then $T$ does not even have to be continuous. For example, in Figure 7, initially close vectors $v$ and $w$ are mapped by $T$ to vectors that are far away from each other.

For convex domains with smooth boundary, a routine application of the implicit mapping theorem show that $T$ is continuous, and, moreover, a diffeomorphism of the open phase cylinder $M = \partial D \times (0, \pi)$ to itself. If, moreover, the boundary of $D$ has strictly positive curvature, then we have the following result:
Figure 7: In a non-convex domain, the billiard map may be discontinuous.

**Theorem 1.2** (see e.g. [1]). If the boundary of the domain D is smooth and has strictly positive curvature, then the corresponding billiard map $T: M \to M$ extends to a diffeomorphism of the closed phase cylinder $\partial D \times [0, \pi]$ to itself.

In other words, for sufficiently nice boundaries, the billiard map stays well-defined and smooth when the initial velocity vector tends to a vector parallel to the boundary.

**Lecture 2: The billiard map is area-preserving**

Recall that the phase space $M$ of a billiard in a domain $D$ is the space of pairs $(x, v)$, where $x \in \partial D$ is a point at the boundary of $D$, and $v \in T_x D$ is an inward unit tangent vector at $x$. Since $\partial D$ is homeomorphic to a circle, and the space of unit inward tangent vectors at each point is homeomorphic to an open interval, the phase space $M$ is topologically a cylinder. As coordinates on $M$, we take the arc length parameter $t$ on $\partial D$, and the angle $\alpha$ which the inward vector $v$ makes with the positive direction of $\partial D$. Note that the coordinate $t$ is defined modulo the total length $L$ of $\partial D$, while $\alpha$ is well-defined real number in $(0, \pi)$. Consider the following 2-form on $M$:

$$\omega = \sin(\alpha) d\alpha \wedge dt.$$ 

Note that even though $t$ is not quite a function (it is defined up to an additive constant $L$, so it is in a fact a map to a circle), its differential $dt$ is a well-defined 1-form. So, the 2-form $\omega$ is well-defined as well. Furthermore, we have $d\alpha \wedge dt \neq 0$, because $\alpha$ and $t$ are, by construction, coordinates on $M$. Finally, notice that $\sin(\alpha) \neq 0$, since $\alpha \in (0, \pi)$. So, $\omega$ is a non-vanishing 2-form on a 2-manifold $M$, i.e. it is an area form (a particular case of a volume form, that is a non-vanishing $n$-form on an $n$-dimensional manifold).

**Theorem 2.1.** The area form $\omega$ on $M$ is preserved by the billiard map $T$: $T^* \omega = \omega$. In other words, for any domain $U \subset M$, we have

$$\int_U \omega = \int_{T(U)} \omega,$$

provided that at least one of these integrals is well-defined.

The proof is based on the following lemma:

**Lemma 2.2.** Let $\gamma$ be an arc length parametrized curve in $\mathbb{R}^2$. Let also $||\gamma(t') - \gamma(t)||$ be the Euclidian distance between the points $\gamma(t)$ and $\gamma(t')$. Finally, let $\alpha$ and $\alpha'$ be the angles between the chord joining $\gamma(t)$ and $\gamma(t')$ and the arc of $\gamma$ connecting those two points (see Figure 8). Then

$$\frac{\partial}{\partial t} ||\gamma(t') - \gamma(t)|| = -\cos(\alpha), \quad \frac{\partial}{\partial t'} ||\gamma(t') - \gamma(t)|| = \cos(\alpha').$$

(1)
Figure 8: The partial derivatives of the function \(|\gamma(t') - \gamma(t)|\) are given by \(\cos(\alpha')\) and \(\cos(\alpha)\).

**Figure 9:** The gradient of the function \(g(x) = ||x - \gamma(t)||\) at the point \(x = \gamma(t')\).

**Proof of the lemma.** The expression \(|\gamma(t') - \gamma(t)|\) is a symmetric function in \(t\) and \(t'\), so its partial derivatives with respect to these variables should be the same, up to replacing \(t\) with \(t'\). The reason for different signs of the partial derivatives in (1) is the asymmetric definition of the angles \(\alpha\) and \(\alpha'\): \(\alpha\) is the angle between the chord \(\gamma(t)\gamma(t')\) and the positive direction of \(\gamma\), while \(\alpha'\) is the angle between the same chord and the negative direction of \(\gamma\) (see Figure 8). So, it suffices to establish the second of formulas (1).

The first one then follows by symmetry.

To prove second of formulas (1), fix \(t\) and consider the function \(g(x) = ||x - \gamma(t)||\), where \(x \in \mathbb{R}^2\). The level set of this function through the point \(\gamma(t')\) is a circle centered at \(\gamma(t)\). Therefore, the gradient of \(g(x)\) at \(\gamma(t')\) is orthogonal to that circle and hence collinear to the vector \(\gamma(t') - \gamma(t)\), see Figure 9 (this can also be seen from the definition of the gradient as the direction of fastest increase). Furthermore, the derivative of the function \(g(x) = ||x - \gamma(t)||\) at \(x = \gamma(t')\) in the direction \(\gamma(t') - \gamma(t)\) is clearly equal to 1, so the gradient of \(g(x)\) at \(\gamma(t')\) is a unit vector positively collinear to \(\gamma(t') - \gamma(t)\) (here we use that the magnitude of the gradient is equal to the derivative of the function in the direction of the gradient, and also that a function must increase, not decrease in the direction of its gradient).

Now we have that \(|\gamma(t') - \gamma(t)| = g(\gamma(t'))\), so the partial derivative of \(|\gamma(t') - \gamma(t)|\) with respect to \(t'\) is equal to the derivative of the function \(g(x)\) along the velocity vector \(\gamma'(t')\) of the curve \(\gamma\) at the point \(\gamma(t')\). Therefore, we have

\[
\frac{\partial}{\partial t'} ||\gamma(t') - \gamma(t)|| = \langle \text{grad} g(\gamma(t')), \gamma'(t') \rangle,
\]

where \(\text{grad} g(\gamma(t'))\) is the gradient of \(g(x)\) at \(x = \gamma(t')\). Furthermore, we already saw that \(\text{grad} g(\gamma(t'))\) is a unit vector, while \(\gamma'(t')\) is a unit vector because the curve \(\gamma\) is arc length parametrized. So, the inner product between these two vectors is equal to the cosine of the angle between them. To complete the proof, it suffices to notice that this angle is precisely \(\alpha'\) (see Figure 9).

**Proof of Theorem 2.4.** Let \((t', \alpha')\) be the coordinates of the image of the point \((t, \alpha)\) under the billiard map \(T\). Then \(t', \alpha'\) are smooth functions of \(t, \alpha\), and hence smooth functions on \(M\). Let also \(\gamma\) be the

\[\text{Here we distinguish between the derivative in the direction } v, \text{ and the derivative along } v. \text{ By definition, the derivative in the direction } v \text{ is the derivative along } v/||v||.\]
Figure 10: The billiard map \( T \) takes the point \( (t, \alpha) \) to the point \( (t', \alpha') \).

arc length parametrized boundary of the billiard table \( D \). Then \( \alpha' \) can be defined as the angle between the chord \( \gamma(t)\gamma(t') \) and the negative direction of the curve \( \gamma \), see Figure 10. Now, consider the function \( f = ||\gamma(t') - \gamma(t)|| \). Since \( t \) and \( t' \) are smooth functions on \( M \), so is \( f \). Furthermore, by Lemma 2.2 we have

\[
df = \cos(\alpha')dt' - \cos(\alpha)dt.
\] (2)

Consider now the 1-form \( \xi \) on \( M \) given by

\[
\xi = -\cos(\alpha)dt.
\]

Notice that

\[
d\xi = d(-\cos(\alpha)) \wedge dt = \sin(\alpha)d\alpha \wedge dt = \omega.
\]

Furthermore, by (2) we have

\[
df = \xi - T^*\xi.
\] (3)

Any \( f \) satisfying this equation for some \( \xi \) such that \( d\xi = \omega \) is called a generating function of the map \( T \). A map possessing a generating function is automatically area-preserving. Indeed, taking the differential of both sides in (3), we get

\[
0 = d\xi - dT^*\xi = d\xi - T^*d\xi = \omega - T^*\omega,
\]

so

\[
T^*\omega = \omega,
\]

as desired. Thus, the billiard map \( T \) has a generating function given by \( ||\gamma(t') - \gamma(t)|| \) and hence is area-preserving, q.e.d.

Lecture 3: Poincaré’s recurrence theorem and billiard in a disk

**Theorem 3.1** (Poincaré’s recurrence theorem for billiards). The billiard map \( T: M \to M \) has the following recurrence property: for any open subset \( U \subset M \), almost all points in \( U \) will eventually return, under the action of iterations of \( T \), to the set \( U \). In other words, for almost any \( v \in U \) there exists a natural number \( n > 0 \) such that \( T^n(v) \in U \).

**Remark 3.2.** As one can see from the proof, this theorem is true for any volume-preserving map of an \( n \)-dimensional manifold, provided that the total volume of the manifold is finite. Furthermore, the same result holds for any measure-preserving map of any measure space, again provided that the measure of the whole space is finite. This is the way this theorem is usually stated.
Proof of Theorem 3.1. Let $V \subset U$ be the set of points which do not return to $U$ under the iterations of $T$. This means that $V$ consists of points $v \in U$ such that $T^n(v) \notin U$ for any natural number $n > 0$. We need to show that the set $V$ has area 0. To that end, notice that for any natural $n > 0$, the set $T^n(V)$ is disjoint from $V$. Indeed, by definition of $V$, the set $T^n(V)$ does not intersect $U$ and hence $V$. So, we have that $T^n(V) \cap V$ is an empty set. Furthermore, since $T$ is invertible, it follows that for any integer $l \geq 0$ the set $T^{n+l}(V) \cap T^l(V)$ is also empty (if $v \in T^{n+l}(V) \cap T^l(V)$, then it must be that $T^{-l}(v) \in T^n(V) \cap V$, which is impossible). And since $n > 0$ and $l \geq 0$ are arbitrary, it follows that the sets $V, T(V), T^2(V), \ldots$ are all pairwise disjoint. So, their total area cannot exceed the total area of the phase cylinder $M$, which is

$$\int_0^L \int_0^\pi \sin(\alpha) d\alpha dt = 2L,$$

i.e. twice the perimeter of the billiard table. At the same time, since $T$ is area-preserving, all the sets (4) have the same area and thus may have finite total area only if each of them has area zero. Thus, the theorem is proved.

Remark 3.3. There is one step that we skipped in the proof: one actually needs to explain why the set $V$ has well-defined area at all. This can be done by showing that $V$ is a difference of two open sets and hence measurable.

Example 3.4. Consider a billiard table $D$ of an arbitrary shape, and let $I \subset \partial D$ be a (possibly very small) open subset of the boundary. Consider billiard trajectories that start at points of $I$ and make an angle with $\partial D$ which is between $89^\circ$ and $91^\circ$. Then, according to Poincaré’s theorem (applied to the subset $U = I \times (89^\circ, 91^\circ) \subset M$), almost all of these trajectories will eventually hit $I$ again, and, moreover, the angle of incidence will again be between $89^\circ$ and $91^\circ$. Of course, it may (and, generally speaking, will) take a long long time for this event to occur: Poincaré’s recurrence theorem is saying that for almost all initial data $v$ in our open subset $U$, there is $n > 0$ such that $T^n(v) \in U$, but it is aying nothing about what the value of $n$ is.

We now turn to studying concrete examples of billiards. We will start with the simplest case of billiards in disks. Clearly, it is sufficient to consider the case when $D$ is a unit disk: billiards in disks of different radii behave in the same way. Let $A$ be a point in the unit circle. Consider the billiard trajectory in the unit disk which starts at $A$ and makes an angle $\alpha$ with the positively oriented unit circle (see Figure 11).
Then this trajectory will meet the unit circle again at a point $B$, and the angle of incidence is also equal to $\alpha$ (because a chord intersects a circle at the same angle at both intersection points). Therefore, after the trajectory is reflected at $B$, it will again make an angle $\alpha$ with the positive direction of the unit circle. This means that billiard map $T$ does not affect the $\alpha$-coordinate on the phase cylinder $M$ at all: $\alpha' = \alpha$. Further, denote the $t$-coordinate of $A$ by $t$, and the $t$-coordinate of $B$ by $t'$. Then the increment $t' - t$ is equal to the measure of the arc of the circle that goes from $A$ to $B$ (in the counter-clockwise direction). This measure is equal to the angle $\angle AOB$ (where $O$ is the center of the circle), which, in turn, is equal to $2\alpha$ (by the tangent-chord theorem combined with the inscribed angle theorem). Therefore, in coordinates, the billiard map $T$ in the unit disk is given by

$$
\begin{cases}
t' = t + 2\alpha, \\
\alpha' = \alpha,
\end{cases}
$$

where addition in the formula for $t'$ is understood modulo the total length of the unit circle, which is $2\pi$. This is a linear map (at least if we forget about the periodicity of the $t$-coordinate), and the corresponding dynamics (i.e. the behavior of iterates of this map) is easy to understand. Namely, we have the following explicit formulas for the iterate $T^n$:

$$
\begin{cases}
t' = t + 2n\alpha, \\
\alpha' = \alpha,
\end{cases}
$$

where, again, addition in the formula for $t'$ is understood modulo $2\pi$. We, however, would like to understand what the dynamics looks like qualitatively. To that end, notice that since $\alpha$ is preserved by the map, we can just fix it and study the dynamics, of the $t$ variable, which is given by

$$
t \mapsto t + 2\alpha.
$$

Recall that $t$ can be thought of a as a point in a circle, $t \in \mathbb{R}/2\pi\mathbb{Z}$. The map (5) is then a rotation of that circle by angle $2\alpha$. This rotation behaves differently depending on whether the number $\alpha/\pi$ is rational or irrational. If $\alpha/\pi = m/n$ is rational, then $2n\alpha = 2m\pi$, and thus the $n$'th power of the map (5) is the identity map. So, for $\alpha/\pi \in \mathbb{Q}$, the map (5) is periodic. In terms of the billiard, this means that any trajectory with such $\alpha$ eventually closes up. Figure 12 shows two 5-periodic trajectories corresponding to $\alpha = \pi/5$ and $\alpha = 2\pi/5$.

Now, consider the case $\alpha/\pi \notin \mathbb{Q}$. In this case, we can still approximate $\alpha/\pi$ with a rational, with arbitrary precision (the denominator of that rational may be large, though). So, even though the dynamics in this case is not periodic, it is close to periodic. Such dynamics is known as almost periodic or quasi-periodic (the periodic case can also be considered a particular case of the quasi-periodic one). Note that in this case $t + 2n\alpha \neq t + 2m\alpha$ modulo $2\pi$ unless $m = n$, which means that the orbit of any $t \in S^1$ under

![Figure 12: Periodic billiard trajectories in a disk.](image-url)
iterations of the map $[5]$ is infinite. One can in fact quite easily show that any orbit is dense in the circle. Figure [13] shows what a piece of a non-periodic billiard trajectory in a disk, corresponding to $\alpha / \pi / \not\in \mathbb{Q}$, looks like. We in fact generated this as a periodic trajectory with large period. In practice, one does not really see the difference between quasiperiodic dynamics and periodic dynamics with large period.

We conclude this discussion by noticing that Poincaré recurrence clearly holds for the billiard in a disk (as it should by Theorem [3.1]). Moreover, since the dynamics is quasi-periodic, every point eventually returns to any its small neighborhood.

**Lecture 4: Billiards in ellipses**

We now turn our attention to billiards in ellipses. Most of the results we will obtain also apply to hyperbolas. The case of a hyperbola is, however, somewhat more sophisticated, because, in contrast to ellipses, hyperbolas do not bound compact domains. So we will stick with ellipses. That being said, hyperbolas will still be appearing from time to time in our study.

Consider an ellipse given in a Cartesian coordinate system $(x_1, x_2)$ by

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} = 1.$$  

As known from linear algebra, any ellipse can be written in this form in a suitable orthogonal coordinate system.

Given a point $(x, v)$ in the phase cylinder $M$ for the billiard in the ellipse, consider the function

$$J = -\left( \frac{x_1 v_1}{a_1^2} + \frac{x_2 v_2}{a_2^2} \right),$$  

where $(x_1, x_2)$ are components of $x$, and $(v_1, v_2)$ are components of $v$.

**Proposition 4.1.** The function $J$ is invariant under the billiard map $T: M \to M$, i.e. $J(T(x, v)) = J(x, v)$.

**Remark 4.2.** A function invariant under a map is called a conserved quantity, a first integral, or just an integral of the map. This particular function $J$ is known as the Joachimsthal integral. Note that this integral can also be rewritten as

$$J = -\frac{1}{2} \langle v, \text{grad } f(x) \rangle,$$

where

$$f = \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2}$$
is the defining function of the ellipse. This explains the negative sign in the definition: since \( \nabla f(x) \) is an outward normal to the ellipse, and \( v \) is an inward vector, the angle between \( \nabla f(x) \) and \( v \) must be obtuse, and \( \langle v, \nabla f(x) \rangle < 0 \). Therefore, our function \( J \) is always positive. Of course, the function 
\[-J = \frac{x_1v_1}{a_1^2} + \frac{x_2v_2}{a_2^2} \]
is also preserved by the billiard map, it is just somewhat more pleasant to work with positive functions than with negative ones.

**Remark 4.3.** Also notice that in the case of a unit circle the norm of the vector \( \nabla f(x) \) is equal to 2, so 
\[ J = -\frac{1}{2} \langle v, \nabla f(x) \rangle = -\cos(\alpha + \pi/2) = \sin(\alpha), \]
where we also used that \( v \) is a unit vector, and that its angle with the outward normal \( \nabla f(x) \) is equal to \( \alpha + \pi/2 \), with \( \alpha \) being the angle between \( v \) and the positive direction of the unit circle. So, in the circle case the conserved quantity \( J \) essentially coincides with the conserved quantity \( \alpha \) which we found in the previous lecture (more precisely, \( J = \sin(\alpha) \), but saying that \( \alpha \) is preserved is more or less the same as saying that \( \sin(\alpha) \) is preserved). So, we already know that Proposition 4.1 is true for the circle.

**Proof of Proposition 4.1.** Let \((x, v) \in M\) be a point in the phase cylinder, and let \((x', v') = T(x, v)\) be its image under the billiard map. Then, by definition of the billiard map, the vectors \( v \) and \( v' \) make the same angle with the (tangent line to the) ellipse at \( x' \). Since those are unit vectors, this is equivalent to saying that the vector \( v + v' \) is tangent to the ellipse at \( x' \), see Figure 14. Using also that the gradient \( \nabla f(x') \) is orthogonal to the ellipse at \( x' \), this gives
\[ \langle v + v', \nabla f(x') \rangle = 0, \]
so 
\[ J(x', v') = -\frac{1}{2} \langle v, \nabla f(x') \rangle = \frac{1}{2} \langle v, \nabla f(x') \rangle. \]
We need to show that this is equal to 
\[ J(x, v) = -\frac{1}{2} \langle v, \nabla f(x) \rangle, \]
which is equivalent to proving that 
\[ \langle v, \nabla f(x) + \nabla f(x') \rangle = 0. \]
Also notice that \( v \) is collinear to \( x' - x \) (by definition of the billiard map), so we need to show that 
\[ \langle x - x', \nabla f(x) + \nabla f(x') \rangle = 0. \]
To show this, we will use that $f$ is a homogeneous quadratic function. Write $f$ in the form $f = \langle x, Ax \rangle$, where

$$A = \begin{pmatrix} 1/a_1^2 & 0 \\ 0 & 1/a_2^2 \end{pmatrix}.$$ 

Notice that $\text{grad} f(x) = 2Ax$. So, the left hand side of (7) can be rewritten as

$$2\langle x - x', Ax + Ax' \rangle = 2 \left( \langle x, Ax \rangle - \langle x', Ax' \rangle + \langle x, Ax' \rangle - \langle x', Ax \rangle \right).$$

But $\langle x, Ax \rangle = f(x) = 1$, since $x$ lies on the ellipse. Likewise, $\langle x', Ax' \rangle = 1$. Finally, notice that $\langle x, Ax' \rangle = \langle x', Ax \rangle$ since $A$ is symmetric. So, we conclude that (7) indeed holds, as desired.

We now want to find a geometric interpretation of the Joachimsthal integral $J$. To that end, we will need a more geometric definition of an ellipse:

**Definition 4.4.** Let $f_1, f_2$ be two points in the Euclidian plane $\mathbb{R}^2$, and let $l > 0$ be a positive real number. Then the set of points $\{x \in \mathbb{R}^2 \mid ||xf_1|| + ||xf_2|| = l\}$ is called an ellipse with foci $f_1, f_2$. Here $||xf_i||$ stands for the Euclidian distance between $x$ and $f_i$. Similarly, the set of points $\{x \in \mathbb{R}^2 \mid ||xf_1|| - ||xf_2|| = l\}$ is called a hyperbola with foci $f_1, f_2$.

We now want to obtain an analytic description of ellipses and hyperbolas with given foci $f_1, f_2$. We will assume that $f_1 = (-a, 0)$, $f_2 = (a, 0)$. By doing so we do not loose any generality because any pair of points has such coordinates in a suitable orthogonal coordinate system.

**Proposition 4.5.** The equation of an ellipse/hyperbola with foci $f_1 = (-a, 0), f_2 = (a, 0)$ and given parameter $l > 0$ is

$$\frac{x_1^2}{l^2/4} + \frac{x_2^2}{l^2/4 - a^2} = 1.$$  \hspace{1cm} (8)

**Remark 4.6.** The reader may be confused by the fact that an ellipse and a hyperbola with the same parameters are described by the same equation. But there is no contradiction here. The point is that by the triangle inequality we must always have $l > 2a$ for an ellipse and $l < 2a$ for a hyperbola. So for any $l$ there may exist either an ellipse, or a hyperbola with such $l$, but not both at a time.

The proof of Proposition 4.5 is a straightforward algebraic verification of the equivalence between equation (8) and Definition 4.4.

**Lecture 5: Geometric meaning of the Joachimsthal integral**

In this lecture we will reveal the geometric meaning of the Joachimsthal integral (6) for the billiard map in an ellipse. Consider once again an ellipse given by

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} = 1.$$  \hspace{1cm} (9)

Without loss of generality, we may assume that this ellipse is “horizontal”, i.e. $a_1 \geq a_2$ (this can be always achieved by rotating the coordinate system if necessary). Furthermore, since we already studied the circle case in detail, we will assume that our ellipse is not a circle, which means that $a_1 > a_2$.

**Proposition 5.1.** The foci of the ellipse (9) are the points $(\pm a, 0)$, where $a = \sqrt{a_1^2 - a_2^2}$. 

11
Remark 5.2. We assumed that $a_1 \geq a_2$ to have foci on the horizontal axis. For $a_1 \leq a_2$, the foci are on the vertical axis (see Figure 15).

Proof of Proposition 5.1. Equation (9) coincides with (8) if we set $a = \sqrt{a_1^2 - a_2^2}$ and $L = 2a_1$.

It follows that the equations of conics (i.e. ellipses and hyperbolas) which are confocal with (i.e. have the same foci as) our ellipse (9) is

$$\frac{x_1^2}{l^2/4} + \frac{x_2^2}{l^2/4 - a_1^2 + a_2^2} = 1.$$ 

This can be written in a more symmetric way if we define $\lambda = l^2/4 - a_1^2$. Then the above equation becomes

$$\frac{x_1^2}{a_1^2 - \lambda} + \frac{x_2^2}{a_2^2 - \lambda} = 1. \quad (10)$$

This is a standard equation of a confocal family. For an arbitrary value of $\lambda \neq a_1^2, a_2^2$, this equation defines a conic, and all these conics have the same foci as the ellipse (9). And conversely, any conic confocal with the ellipse (9) is contained in the family (10). Figure 16 shows an example of a confocal family.

We will now be interested in the following question: given a line in $\mathbb{R}^2$, how many, if any, conics from the family (10) are tangent to that line? The following lemma says that there is almost always exactly one, except for a couple of cases when there is none:

Lemma 5.3. Consider the line $x + tv$ through the point $x = (x_1, x_2)$ with direction $v = (v_1, v_2)$. Then

1. This line is tangent to at most one conic from the family (10).
2. If such a tangent conic exists, then its parameter $\lambda$ which distinguishes it in the family (10) is given by

$$\lambda = \frac{a_1^2 v_2^2 + a_2^2 v_1^2 - (x_1 v_2 - x_2 v_1)^2}{v_1^2 + v_2^2}. \quad (11)$$

3. Such tangent conic does not exist in the following two cases:

(a) The number $\lambda$ given by formula (11) is equal to $a_1^2$, in which case the line $x + tv$ coincides with the minor axis of the ellipse (9).

(b) The number $\lambda$ given by formula (11) is equal to $a_2^2$, in which case the line $x + tv$ is passing through one of the foci of the ellipse (9).

Proof. The line $x + tv$ is tangent to the conic (10) when the equation

$$\frac{(x_1 + tv_1)^2}{a_1^2 - \lambda} + \frac{(x_2 + tv_2)^2}{a_2^2 - \lambda} = 1$$

for their intersection points has exactly one solution in terms of $t$; that is, when its discriminant is equal to 0. Equating the discriminant to 0 and solving for $\lambda$, we get formula (11). The expression on the right-hand side of (11), however, may be equal to $a_1^2$ or $a_2^2$, which does not correspond to any conic in the family (10). In that case, there is no conic in the family (10) tangent to the line $x + tv$. In all other cases, such a conic exists, is unique, and corresponds to $\lambda$ given by (11).

To complete the proof it now suffices to obtain a geometric interpretation of the cases $\lambda = a_1^2$ and $\lambda = a_2^2$. First assume that $\lambda = a_1^2$. Then (11) gives

$$(a_1^2 - a_2^2)v_1^2 = -(x_1 v_2 - x_2 v_1)^2.$$ 

Notice the left-hand side of this equation is non-negative (since $a_1 > a_2 > 0$), while the right-hand side is non-positive. Therefore, this equation holds if and only if both sides are equal to 0, which is equivalent to $v_1 = 0$ and $x_1 = 0$. But these two conditions together hold precisely when the line $x + tv$ coincides with the vertical coordinate axis or, which is the same, with the minor axis of the ellipse (9).

Similarly, if $\lambda = a_2^2$, then

$$(a_2^2 - a_1^2)v_2^2 = -(x_1 v_2 - x_2 v_1)^2,$$

which is equivalent to

$$x_1 v_2 - x_2 v_1 = \pm av_2,$$

with $a = \sqrt{a_1^2 - a_2^2}$. The geometric meaning of this equation is that the vectors $(x_1, x_2) - (\pm a, 0)$ and $(v_1, v_2)$ are collinear, which is the same as to say that the line $x + tv$ is passing through one of the points $(a, 0), (-a, 0)$. But those points are precisely the foci of the ellipse (9), hence the result.

We now show that Joachimsthal integral computed at a point $(x, v)$ of the phase cylinder for the billiard in an ellipse is closely related to the parameter $\lambda$ of the confocal conic to which the line $x + tv$ is tangent:

**Proposition 5.4.** The quantity $\lambda$ given by formula (11), regarded as a function on the phase cylinder $M$ of the billiard in the ellipse (9), is related to the Joachimsthal integral $J$ by the formula

$$\lambda = a_1^2 a_2^2 J^2. \quad (12)$$
Proof. For a unit vector $v$, formula (11) becomes

$$\lambda = a_1^2v_2^2 + a_2^2v_1^2 - (x_1v_2 - x_2v_1)^2 = (a_1^2 - x_1^2)v_2^2 + (a_2^2 - x_2^2)v_1^2 + 2x_1x_2v_1v_2.$$  

(13)

Furthermore, since the point $(x_1, x_2)$ lies on the ellipse (9), we have

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} = 1 \implies \frac{x_1^2 - a_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} = 0 \implies a_1^2 - x_1^2 = \frac{a_2^2x_2^2}{a_1^2}.$$

Similarly, we have

$$a_2^2 - x_2^2 = \frac{a_2^2x_1^2}{a_2^2},$$

so (13) can be written as

$$\lambda = \frac{a_1^2x_2^2}{a_1^2} + \frac{a_2^2x_1^2}{a_2^2} + 2x_1x_2v_1v_2 = a_1^2a_2^2\left(\frac{x_1v_1}{a_1^2} + \frac{x_2v_2}{a_2^2}\right)^2 = a_1^2a_2^2J^2,$$

as desired.

**Corollary 5.5.**

1. Assume that a segment of a billiard trajectory in the ellipse (9) is tangent to some confocal conic (10). Then all segments of that trajectory are tangent to that conic.

2. Assume that a segment of a billiard trajectory in the ellipse (9) belongs to the minor axis. Then all segments belong to the minor axis.

3. Assume that a segment of a billiard trajectory in the ellipse (9) is passing through one of the foci. Then all segments are passing through one of the foci.

**Proof.** From (12) and preservation of $J$ it follows that the function $\lambda$ given by (11) is preserved by the billiard map. The above three statements are the particular cases of this result corresponding to, respectively, generic values of $\lambda$, $\lambda = a_1^2$, and $\lambda = a_2^2$.

**Remark 5.6.** In view of Proposition 5.4, Corollary 5.5 is equivalent to preservation of the Joachimsthal integral and thus can be viewed as a geometric form of the latter. Furthermore, all statements of Corollary 5.5 can be obtained geometrically. The second statement is particularly straightforward: since the minor axis is orthogonal to the ellipse, a billiard ball moving along that axis will continue doing so after any number of reflections (see Figure 17). Of course, the same is true for the major axis, but the major axis is not distinguished by any specific value of $\lambda$: its $\lambda$ is the same as for any other line passing through one of the foci.

Also note that the third statement of the corollary can be strengthened as follows: if a segment of a billiard trajectory in the ellipse is passing through one of the foci, then the next segment must pass through the other focus. Indeed, a generic billiard trajectory through one of the foci is not orthogonal
Figure 18: A segment of a billiard trajectory through a focus of an ellipse.

to the ellipse (see Figure 18) and hence cannot be reflected to the same focus. So, almost all trajectories through one of the foci get reflected to the other focus, and by continuity it must be true for all such trajectories.

The fact that a billiard trajectory through one of the foci is reflected to the other focus is known as the optical property of the ellipse. It can be reformulated by saying that all light rays starting at one of the foci get reflected, by an elliptic mirror, to the other focus. The optical property can be proved quite easily without using the Joachimsthal integral, see e.g. [2, Lemma 4.2]. See also [2, Theorem 4.4] for an independent geometric proof of the first result of Corollary 5.5.

References
