Upgrading the Local Ergodic Theorem for planar semi-dispersing billiards

N. Chernov[1] and N. Simányi[1]

June 4, 2009

Abstract

The Local Ergodic Theorem (also known as the 'Fundamental Theorem') gives sufficient conditions under which a phase point has an open neighborhood that belongs (mod 0) to one ergodic component. This theorem is a key ingredient of many proofs of ergodicity for billiards and, more generally, for smooth hyperbolic maps with singularities. However, the proof of that theorem relies upon a delicate assumption (Chernov-Sinai Ansatz), which is difficult to check for some physically relevant models, including gases of hard balls. Here we give a proof of the Local Ergodic Theorem for two dimensional billiards without using the Ansatz.

Keywords: Hard balls, Bolzmann-Sinai hypothesis, semi-dispersing billiards, ergodicity.

1 Introduction

In this work we make a step toward a complete solution (yet to be achieved) of the celebrated Boltzmann-Sinai ergodic hypothesis. The latter asserts [16] that every system of $n \geq 2$ hard balls on a torus of dimension $d \geq 2$ is ergodic (provided the trivial first integrals are eliminated). This model reduces to the motion of a billiard particle in a $d(n-1)$-dimensional torus bouncing off $n(n-1)/2$ cylindrical obstacles (the billiard particle hits a cylinder whenever

[1]Department of Mathematics, University of Alabama at Birmingham, AL, 35294, USA; Email: chernov@math.uab.edu and simanyi@math.uab.edu.

two balls collide). Billiards with cylindrical walls belong to a more general category of semi-dispersing billiards, where a particle moves in a container with concave (but not necessarily strictly concave) boundaries.

We remark that in the case $n = 2$ the cylinders actually become spheres, i.e. any system of 2 hard balls reduces to a billiard particle in a torus with a spherical obstacle. Such billiards belong to a more special class of dispersing billiards, where a particle moves in a container with strictly concave walls.

Dispersing billiards are always completely hyperbolic and ergodic [17], but for semi-dispersing billiards this may not be true. For example, a billiard in a 3-torus with a single cylindrical wall has zero Lyapunov exponents and is not ergodic; on the other hand, 2 transversal cylindrical walls within a 3-torus ensure hyperbolicity and ergodicity [6]. For the systems of $n \geq 3$ hard balls, one has to carefully explore the geometry of the cylindrical walls in order to derive hyperbolicity and ergodicity.

There are two complications in the study of hard balls (or more generally, semi-dispersing billiards). One is caused by the *singularities* of the dynamics – these happen during simultaneous multiple collisions of ≥ 3 balls and during grazing (tangential) collisions. In the phase space, singular points make submanifolds of codimension one. The other complication is caused by *non-hyperbolicity* (i.e. the existence of zero Lyapunov exponents) at some phase points. Such points make various structures, ranging from smooth submanifolds to Cantor-like subsets of the phase space.

Powerful techniques have been developed to handle these two complications separately (singularities and non-hyperbolicity), but the combination of the two still presents an unmanageable situation. More precisely, if non-hyperbolic sets and singularities intersect in a subset of positive $[2d(n-1)-2]$-dimensional measure, then modern proofs of ergodicity stall. On the other hand, such substantial overlaps between singularities and non-hyperbolic sets appear very unlikely (physically); they are regarded as 'conspiracy'.

To bypass this scenario in an early work, Ya. Sinai and N. Chernov [17] *assumed* that almost every point on the singularity manifolds (with respect to the intrinsic Lebesgue measure) was completely hyperbolic. Under this assumption (now referred to as Chernov-Sinai Ansatz) they proved the so-called Local Ergodic Theorem (also called 'Fundamental Theorem'), which later became instrumental in the proofs of ergodicity for various billiards [1, 7, 10]. It gives sufficient (and easily verifiable) conditions under which a phase point has an open neighborhood which belongs (mod 0) to one ergodic component.

A. Krámli, N. Simányi and D. Szász built upon the results of [17] and established the ergodicity for systems of $n = 3$ hard balls in any dimension [8] and for $n = 4$ hard balls in dimension $d \geq 3$ [9]; in particular they verified Chernov-Sinai Ansatz in these cases. However, their techniques could not be extended to $n \geq 5$. The situation called for novel approaches.

A partial breakthrough was made by Simányi and Szász when they invoked ideas of algebraic geometry to rule out various 'conspiracies' (at least for generic systems of hard balls), which were in the way of proving hyperbolicity and ergodicity. Precisely, they assumed that the balls had arbitrary masses $m_1, \dots, m_n$ (but the same radius r) and proved [11] complete hyperbolicity at a.e. phase point for generic vectors of 'external parameters' $(m_1, \dots, m_n, r)$; the latter needed to avoid some exceptional submanifolds of codimension one in $\mathbb{R}^{n+1}$, which remained unspecified and unknown. Later Simány used [13, 14] the same approach to prove Chernov-Sinai Ansatz and ergodicity for generic systems of hard balls (in the above sense). He also established hyperbolicity for systems of hard balls of arbitrary masses [12].

Thus the Boltzmann-Sinai ergodic hypothesis is now proved for typical, or generic, systems of hard balls. This seems to be a comforting settlement in both topological and measure-theoretic senses, but it falls short of solving physically relevant problems, as there is no way to check whether any particular system of hard balls is ergodic or not. Most notably, for the system of balls with all equal masses (which lies in the foundation of statistical mechanics) the ergodicity remains open.

In an attempt to extend his results to ALL gases of hard balls (without exceptions), Simanyi developed [15] a new approach based on purely dynamical (rather than algebro-geometric) ideas; this allowed him to derive ergodicity from Chernov-Sinai Ansatz for all hard ball systems. Thus the Boltzmann-Sinai hypothesis is now solved conditionally, modulo the Ansatz. It remains to prove Ansatz, or alternatively, derive Local Ergodic Theorem without Ansatz. (As a side remark, it is ironic that Ansatz, which originally seemed to be just a convenient and temporary technical assumption, now remains the only unresolved issue in the whole picture.)

Here we make another step toward a final solution of the classical ergodic hypothesis: we derive Local Ergodic Theorem without Ansatz for arbitrary semi-dispersing billiards in dimension two. Our method does not yet apply to higher dimensions, but we are working on this.

2 Statement of the result

A planar (two-dimensional) billiard is a dynamical system where a point q moves freely with unit velocity v, $\|v\| = 1$, in a bounded connected domain $Q \subset \mathbb{R}^2$ or $Q \subset \mathrm{Tor}^2$ and reflects off its boundary ∂Q by the rule

$$v^+ = v^- - 2\,\langle n(q), v^-\rangle\, n(q) \tag{1}$$

where v^+ and v^- denote the postcollisional and precollisional velocities, $n(q)$ is the inward unit normal vector to ∂Q at the collision point $q \in \partial Q$, and $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^2$.

As usual, ∂Q is a finite union of C^3 compact curves that can only intersect at common endpoints (which make corners of the table Q). Whenever the particle hits a corner point $q \in \partial Q$, there are two normal vectors to ∂Q, thus the rule (1) gives two possible continuations (two branches) of the billiard trajectory. Of course, this is an exceptional event (a singularity, see below).

A billiard table Q is semi-dispersing if every smooth component of ∂Q is convex (but not necessarily strictly convex) inward. We also suppose that the set of inflection points $q \in \partial Q$ (where the curvature of ∂Q vanishes) is a finite union of straight line segments (flat sides of Q) and some isolated points. A simple example is a polygon with one or several convex ovals removed from its interior. In semi-dispersing billiards, collisions cannot accumulate [4, 18], i.e. within any finite time period the particle experiences finitely many collisions, hence its trajectory is always well defined (though it might be multiply defined, due to corner points).

The phase space of the billiard system is a compact three dimensional manifold $\Omega = Q \times S^1$, and the billiard flow $\Phi^t \colon \Omega \to \Omega$ preserves a uniform measure μ on Ω. The collision space

$$\mathcal{M} = \{(q, v) \in \Omega \colon q \in \partial Q,\ \langle v, n(q)\rangle \geq 0\}$$

consists of all postcollisional velocity vectors at reflection points. We define the first collision time $\tau(x) = \min\{t > 0 \colon \Phi^t(x) \in \mathcal{M}\}$ and the (first) collision map $T(x) = \Phi^{\tau(x)+0}(x)$ that maps Ω onto $\mathcal{M}$; its restriction to $\mathcal{M}$ is called the billiard map (or collision map). Canonical coordinates on $\mathcal{M}$ are r and φ, where r is the arc length parameter on ∂Q and $\varphi \in [-\pi/2, \pi/2]$ is the angle between v and $n(q)$. The map $T \colon \mathcal{M} \to \mathcal{M}$ preserves the smooth measure $d\nu = \cos\varphi\, dr\, d\varphi$.

For every $x = (q, v) \in \Omega$ we put $-x = (q, -v)$; similarly for every $x = (q, v^+) \in \mathcal{M}$ we put $-x = (q, -v^-)$, where v^+ and v^- are related by (1).

If $Q \subset \mathrm{Tor}^2$, we may have an unpleasant case of 'infinite horizon', where $\sup_x \tau(x) = \infty$. In that case we enlarge $\mathcal{M}$ to make the horizon finite [17]. Suppose Tor^2 is obtained by identifying the opposite sides of the boundary of a rectangle K; then we add the set $\partial K \times S^1$ to $\mathcal{M}$. In other words, every time the particle crosses ∂K, we record a 'collision', though the particle keeps moving straight with the same velocity (we call ∂K a transparent wall). Now it is clear that $\sup_x \tau(x) < \infty$.

The billiard flow Φ^t is a suspension flow over the base map $T\colon \mathcal{M} \to \mathcal{M}$ under the ceiling function τ; it is ergodic if and only if T is.

The flow Φ^t and the map T are singular (non-differentiable) whenever the particle hits a corner of Q or makes a grazing (tangential) collision with ∂Q, i.e. whenever the next collision point belongs to $\mathcal{S}_0 = \{(q, v) \in \Omega : \langle v, n(q)\rangle = 0 \text{ or } q \in \Gamma^*\}$, where Γ^* denotes the set of corner points (observe that $\varphi = \pm\pi/2$ at grazing collisions). The singularity set $\mathcal{S}_1 = T^{-1}(\mathcal{S}_0)$ of T is a finite union of smooth compact curves in $\mathcal{M}$ (it is exactly to ensure its finiteness why we added the transparent wall to $\mathcal{M}$). Similarly, for each $n \neq 0$ the singularity set $\mathcal{S}_n = T^{-n}(\mathcal{S}_0)$ (which is part of the singularity set of the iterate T^n) is a finite union of smooth compact curves in $\mathcal{M}$.

In semi-dispersing billiards, the set $\mathcal{S}_n$ consists of increasing curves for $n < 0$ and of decreasing curves for $n > 0$; thus singularity curves always intersect each other transversally at some time in their lives. Points $x \in \mathcal{M}$ whose trajectories are singular in both future and past (the so called 'double-singularities') make a countable set, which can be easily neglected in the studies of ergodic properties of T. Accordingly, the singularities of the flow Φ^t are a countable union of hypersurfaces in Ω, and future singularities intersect past singularities transversally.

Next we describe hyperbolic properties of Φ^t and T. A local orthogonal manifold (LOM), also called wave front, denoted by $\Sigma \subset \Omega$, is a smooth oriented curve $\gamma \subset Q$ equipped with a family of unit normal vectors (note that there are exactly two such families). The Φ^t-image of a LOM is a finite union of LOMs (sometimes having common endpoints) in Ω.

If the map T is smooth on $\Sigma \subset \Omega$ then, slightly abusing notation, we call $\Sigma^c = T(\Sigma) \in \mathcal{M}$ a LOM as well. Given a LOM $\Sigma^c \subset \mathcal{M}$, we call $\Sigma \subset \Omega$ the corresponding flow-sync LOM (the latter is not unique of course).

We distinguish divergent, convergent, and flat LOMs, as determined by the curvature of its carrier $\gamma \subset Q$. In semi-dispersing billiards, future images of divergent LOMs are always divergent and their sizes keep growing in time; this is the cause of hyperbolicity. On the other hand, images of flat LOMs

remain flat as long as they collide with flat sides of Q; but they become divergent immediately after a collision with a curved side of Q.

We assume that ∂Q has non-zero curvature in at least one point; otherwise Q is a polygon and there are no hyperbolic points. Billiards in generic polygons are ergodic [5] (though it is hard to construct explicit examples [19]), but they are never hyperbolic.

For $x \in \Omega$ and $a < b$, a trajectory segment $\Phi^{[a,b]}(x)$ of the point x is said to be sufficient if there is a collision at some time $a < t < b$ with a curved side of Q (the curvature of ∂Q must be different from zero at the collision point); if the segment $\Phi^{[a,b]}(x)$ passes through singular points and branches out, then every branch must hit a curved side of Q. A point $x \in \Omega$ is sufficient in the future (past) if its semitrajectory $\Phi^{[0,\infty)}(x)$ (resp., $\Phi^{(-\infty,0]}(x)$) is sufficient. If a nonsingular point $x \in \mathcal{M}$ is sufficient (future or past), then in a vicinity U_x of x almost every point $y \in U_x$ is hyperbolic (this follows from the Poincaré theorem); thus sufficiency guarantees (local) hyperbolicity.

Chernov-Sinai Ansatz. Almost every point $x \in \mathcal{S}_1$ (with respect to the one-dimensional Lebesgue measure on $\mathcal{S}_1$) is past sufficient or, equivalently, almost every point $x \in \mathcal{S}_{-1}$ is future sufficient.

Here is our main result:

Theorem 1. *Let $x_0 \in \mathcal{M}$ be a point whose entire trajectory $\Phi^{(-\infty,\infty)}(x)$ passes through at most one singularity and is sufficient. Then there exists an open neighborhood U_0 of x_0 that belongs (mod 0) to one ergodic component of the map T.*

We note that the base neighborhood $U_0 = U_0(x_0)$ in this theorem is any open neighborhood U_0 of x_0 for which

(i) U_0 is a subset of the neighborhood $U_{\varepsilon_1}(x_0)$ of x_0 featuring Theorem 3.6 in [7], where $0 < \varepsilon_1 < 1$ is any fixed number, and

(ii) U_0 admits a family

$$\mathcal{G}^\delta = \left\{ G_i^\delta \,\middle|\, i = 1, 2, \ldots, I(\delta) \right\} \quad (0 < \delta < \delta_0)$$

of regular coverings with a small enough threshold $\delta_0 > 0$ as explained in [7].

All the existing proofs of the Local Ergodic Theorem [17, 7] assume the Ansatz, and we relax that assumption.

Given a particular 2d semi-dispersing billiard, one can verify its ergodicity by showing that the set of sufficient points is connected and has full measure. We note, however, that it is unknown if every semi-dispersing billiard (excluding polygons) is ergodic (or even completely hyperbolic), and our result will not solve this open problem, because we cannot yet control the measure of insufficient points. Proving that a.e. phase point in any semi-dispersing billiard is sufficient amounts to showing that in every polygonal billiard a.e. trajectory is dense, but this is an old (and notoriously hard) open problem.

An interesting result in this direction was obtained in [3]: it was shown that billiards in any polygon where a 'bump' or a pocket is attached at every vertex are hyperbolic and ergodic. But in that case the verification of Ansatz was trivial, as every non-sufficient trajectory was periodic.

3 Proof of the result

We begin with a helpful geometric fact that gives a sufficient condition under which two nearby phase points (points in Q equipped with unit velocity vectors) belong to one divergent local orthogonal manifold.

Lemma 1. *Let* (q_1, v_1), $(q_2, v_2) \in \mathbb{R}^2 \times \mathbb{R}^2$, $\|v_i\| = 1$, $\|q_1 - q_2\| < \varepsilon_0$, $\|v_1 - v_2\| < \varepsilon_0$, $\langle q_1 - q_2, v_1 - v_2\rangle \geq 0$, *with some fixed constant* $\varepsilon_0 \ll 1$. *We claim that there are reals* $\tau_1, \tau_2 \in \mathbb{R}$, $|\tau_i| < 10000\varepsilon_0$, *such that the phase points* $(q_1 + \tau_1 v_1, v_1)$ *and* $(q_2 + \tau_2 v_2, v_2)$ *can be included in a divergent LOM* $\Sigma \subset \mathbb{R}^2 \times S^1$.

Proof. We assume the strict inequality $\langle q_1 - q_2, v_1 - v_2\rangle > 0$. The general result then follows by simply passing to the limit.

Let O be the point of intersection of the lines $l_1 = \{q_1 + tv_1 \,|\, t \in \mathbb{R}\}$ and $l_2 = \{q_2 + tv_2 \,|\, t \in \mathbb{R}\}$. We may and shall assume that $O \in \mathbb{R}^2$ is the origin of the plane $\mathbb{R}^2$. Let $q_1 = t_1 v_1$, $q_2 = t_2 v_2$. The assumed inequality says that $\langle q_1 - q_2, v_1 - v_2\rangle = (t_1 + t_2)(1 - \langle v_1, v_2\rangle) > 0$, thus $t_1 + t_2 > 0$, so by symmetry we may assume that $t_1 > 0$. We distinguish between two cases:

Case 1: $t_1 \geq 5000\varepsilon_0$. We take $q_2 + \tau_2 v_2 = t_1 v_2$, i. e. $\tau_1 = 0$ and $\tau_2 = t_1 - t_2$. Clearly, both (q_1, v_1) and $(q_2 + \tau_2 v_2, v_2)$ are elements of the (outer unit normal field of) the circle Σ defined by the equation $\|x\| = t_1$.

Case 2: $0 < t_1 < 5000\varepsilon_0$. We take $q_1 + \tau_1 v_1 = 5000\varepsilon_0 v_1$, $q_2 + \tau_2 v_2 = 5000\varepsilon_0 v_2$, i. e. $\tau_1 = 5000\varepsilon_0 - t_1$, $\tau_2 = 5000\varepsilon_0 - t_2$. The (unit normal bundle

of the) circle Σ containing $(q_1 + \tau_1 v_1,\, v_1)$ and $(q_2 + \tau_2 v_2,\, v_2)$ is now defined by the equation $\|x\| = 5000\varepsilon_0$ in $\mathbb{R}^2$. $\square$

Next we turn to the proof of Local Ergodic Theorem (without using Ansatz). We follow the lines and notation of [7] that presents one of the clearest and most complete proofs of that theorem. For the given sufficient point x_0 we consider a small enough open neighborhood $U_0 = U_0(x_0)$ of x_0, as described right after Theorem 1.

Given a divergent LOM $\Sigma \subset \Omega$ with a carrier $\gamma \subset Q$, we use the metric on it generated by the distance along the curve γ, and denote by $\|\cdot\|$ the corresponding norm in its tangent space $\mathcal{T}\Sigma$. For LOMs $\Sigma \subset \mathcal{M}$ we use the norm and metric on the corresponding flow-sync LOM's in Ω (constructed right at the given point $x \in \Sigma$).

For any $x \in \Sigma \subset \mathcal{M}$ denote by $D^n_{x,\Sigma}$ the Jacobian of the map T^n restricted to Σ at x, in the above norm. If Σ is a divergent LOM (in our terminology the word "divergent" always means "not necessarily strictly divergent", and a similar convention applies to convergent LOMs), then $D^n_{x,\Sigma} \geq 1$ for every $n \geq 1$. Denote

$$\kappa_{n,0}(x) = \inf_{\Sigma} D^n_{-T^n x,\Sigma}$$

where the infimum is taken over all divergent LOMs through $-T^n x$; this quantity is the minimal expansion of divergent LOMs on their way from $-T^n x$ back to $-x$ (we note that the inf is actually attained at the flat LOM, cf. [2]). Given $\delta > 0$ we denote

$$\kappa_{n,\delta}(x) = \inf_{\Sigma} \inf_{y \in \Sigma} D^n_{y,\Sigma}$$

where the infimum is taken over all divergent LOMs Σ through $-T^n x$ such that T^n is smooth on Σ and $\operatorname{dist}(-x, \partial T^n \Sigma) \leq \delta$ (of course, for any LOM Σ, the boundary $\partial\Sigma$ consists of the two endpoints of that LOM). We observe that $1 \leq \kappa_{n,\delta}(x) \leq \kappa_{n,0}(x)$, and both $\kappa_{n,0}(x)$ and $\kappa_{n,\delta}(x)$ are non-decreasing functions of n.

For $x \in \mathcal{M}$ we denote

$$z_{\text{tub}}(x) = \sup_{\Sigma} \{\operatorname{dist}(x, \partial\Sigma) \colon T \text{ is smooth on } \Sigma\}$$

where the supremum is taken over all flat LOMs $\Sigma \subset \operatorname{int}\mathcal{M}$ through x; this is the so-called radius of the maximal tubular neighborhood of the billiard link joining x with Tx.

Note that $z_{\mathrm{tub}}(-Tx) = z_{\mathrm{tub}}(x)$.

We denote by $\Sigma^u(x)$ and $\Sigma^s(x)$ the unstable and stable manifolds through x; the former is a divergent LOM and the latter a convergent one. We also put $r^\alpha(x) = \operatorname{dist}(x, \partial\Sigma^\alpha(x))$ for $\alpha = u, s$. It is known [7, Lemma 5.4] that for every semi-dispersing billiard table Q there exists a constant $c_3 > 0$ (using the notation of [7]) such that if

$$x \in U^{\mathrm{g}} = U^{\mathrm{g}}(\delta) = \{y \in \mathcal{M}\colon \quad \forall n > 0 \;\; z_{\mathrm{tub}}(-T^n y) \geq (\kappa_{n,c_3\delta}(y))^{-1} c_3\delta\}$$

then $r^s(x) \geq c_3\delta$. Thus the points of U^{g} ('good set') have stable manifolds of order δ. A similar property holds for unstable manifolds. The set of points with shorter stable manifolds ('bad set') must be carefully analyzed. We put

$$\begin{aligned} U^{\mathrm{b}} &= U^{\mathrm{b}}(\delta) = U_0 \setminus U^{\mathrm{g}} = \cup_{n\geq 1} U_n^{\mathrm{b}} \\ U_n^{\mathrm{b}} &= U_n^{\mathrm{b}}(\delta) = \{y \in U^{\mathrm{b}}\colon z_{\mathrm{tub}}(-T^n y) < (\kappa_{n,c_3\delta}(y))^{-1} c_3\delta\}. \end{aligned}$$

A crucial fact in the proof of Local Ergodic Theorem is the following tail bound: for any function $F(\delta) \to \infty$ as $\delta \to 0$ the set

$$U_\omega^{\mathrm{b}} = U_\omega^{\mathrm{b}}(\delta) = \cup_{n > F(\delta)} U_n^{\mathrm{b}}(\delta)$$

has measure

$$\nu(U_\omega^{\mathrm{b}}) = o(\delta). \tag{2}$$

In fact, the derivation of the Local Ergodic Theorem from the tail bound does not require the Ansatz, so we will not repeat it here, see [7, Section 5]. In what follows, we prove the tail bound.

First we need a few additional constructions. Denote

$$\begin{aligned} \hat{U}_n^{\mathrm{b}}(\delta) = \big\{ x \in U_0 \big| \; &\exists \text{ a divergent LOM } \Sigma, -T^n x \in \partial\Sigma, \; \partial\Sigma \cap \mathcal{S}_0 = \emptyset, \\ &T^n \text{ and } T^{-1} \text{ are smooth on } \operatorname{int}\Sigma, \; T^{-1}\Sigma \text{ is also divergent}, \\ &T^{-1} \text{ is not smooth at the endpoint } x' \in \partial\Sigma \text{ other than } -T^n x, \\ &\operatorname{dist}(-T^n x, \, x') \leq \kappa_{n,c_3\delta}(x)^{-1} c_3\delta, \text{ and } T^n x' \in U_0 \big\}. \end{aligned}$$

For any $x \in \hat{U}_n^{\mathrm{b}}(\delta) = \hat{U}_n^{\mathrm{b}}$ and Σ as above, we denote

$$\begin{aligned} &z\left(T^n x, \, \Sigma\right) = \operatorname{dist}(-T^n x, x') \big| \; x' \in \partial\Sigma, \; x' \neq -T^n x, \\ &T^{-1} \text{ is not smooth at } x' \quad \left(\leq \kappa_{n,c_3\delta}(x)^{-1} c_3\delta\right). \end{aligned} \tag{3}$$

We note that $z_{\text{tub}}(T^n x) \le z\,(T^n x,\, \Sigma)$.

For any point $x \in \hat{U}_n^{\text{b}}$ as above, we choose a phase point $x_{\varepsilon_1} = -\Phi^{\varepsilon_1}(T^n x)$ with a suitably selected ε_1, $0 < \varepsilon_1 < \tau(T^n x)$.

Note. From now on we will be recycling the notation ε_1 that appeared earlier in the closed formula in Theorem 1. We think that this action should not be the source of any confusion.

For an additional condition on how to select ε_1, see below. For any Σ featuring the definition of $\hat{U}_n^{\text{b}}(\delta)$ and (3) let $\hat{\Sigma}$ denote the flow-sync version of Σ containing the point $x_{\varepsilon_1} = (q_{\varepsilon_1},\, v_{\varepsilon_1}) = -\Phi^{\varepsilon_1}(T^n x)$, see Fig. 1. Now $x_1 = (q_1, v_1) \in \partial\hat{\Sigma}$ is the projection (by the flow) of the point $x' \in \partial\Sigma$ defined above, with the property

$$\operatorname{dist}(-T^n x,\, x') = z(T^n x,\, \Sigma).$$

The other endpoint of the curve $\hat{\Sigma}$ is $x_{\varepsilon_1} = (q_{\varepsilon_1},\, v_{\varepsilon_1})$, see Fig. 1.

While selecting the time ε_1 above, we try to make it sure that x_1 be a post-singularity phase point, i.e. $T(-x_1) \in \mathcal{S}_0$ and $Tx_1 \notin \mathcal{S}_0$.

Here we first consider the case when such a synchronization of $\hat{\Sigma}$ is possible. After that, right before exposing (6), we explain how to modify the following argument if the required synchronization is not feasible.

Consider the line segment

$$\mathcal{H} = \left\{ q\left(\Phi^t(q,\, v_{\varepsilon_1})\right) \,\middle|\, (q,\, v_{\varepsilon_1}) \in \mathcal{S}_0^+,\ 0 < t < \tau(q,\, v_{\varepsilon_1}) \right\}$$

in the domain Q, see Fig. 1. (Note that the point q_1 does not belong to $\mathcal{H}$, since $v_1 \ne v_{\epsilon_1}$ for any strictly convex $\hat{\Sigma}$.) The configuration component $q_3 \in \mathcal{H}$ of the phase point $x_3 = (q_3,\, v_{\varepsilon_1})$ is defined as the orthogonal projection of q_{ε_1} onto the line $\mathcal{H}$. According to this definition, the line segment

$$\left\{ \lambda q_{\varepsilon_1} + (1-\lambda) q_3 \,\middle|\, 0 \le \lambda \le 1 \right\} \tag{4}$$

is perpendicular to $\mathcal{H}$ at q_3.

Lemma 2. *The scalar product condition of Lemma 1 holds true for the pair of phase points* $(x_3,\, x_1)$ *and* $(x_3,\, x_{\varepsilon_1})$, *i. e.*

$$\begin{aligned} &\langle q_1 - q_3,\, v_1 - v_{\epsilon_1} \rangle \ge 0 \\ &\langle q_{\varepsilon_1} - q_3,\, v_{\varepsilon_1} - v_{\epsilon_1} \rangle \ge 0, \end{aligned}$$

see also Fig. 1.

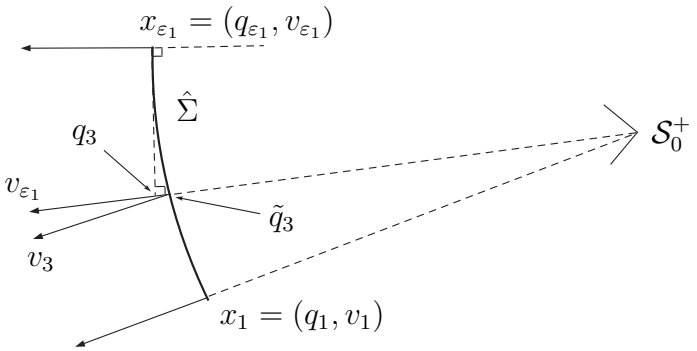

Figure 1: Illustration to Lemma 2.

Proof. Denote the point of intersection of the line segment $\mathcal{H}$ and the carrier of the LOM $\hat{\Sigma}$ by $\tilde{q}_3$ and the outer unit normal vector to $\hat{\Sigma}$ at $\tilde{q}_3$ by v_3, see Fig. 1. Then, by the convexity of $\hat{\Sigma}$, $q_3 = \tilde{q}_3 + \eta v_{\epsilon_1}$ with some small scalar $\eta > 0$, $\langle q_1 - \tilde{q}_3,\, v_3 - v_{\epsilon_1}\rangle \geq 0$, and $\langle q_1 - \tilde{q}_3,\, v_1 - v_3\rangle \geq 0$. Thus we obtain the chain of inequalities

$$\begin{aligned}&\langle q_1 - q_3,\, v_1 - v_{\epsilon_1}\rangle = \langle q_1 - \tilde{q}_3,\, v_1 - v_{\epsilon_1}\rangle + \eta\langle v_{\epsilon_1},\, v_{\epsilon_1} - v_1\rangle\\ &\geq \langle q_1 - \tilde{q}_3,\, v_1 - v_{\epsilon_1}\rangle = \langle q_1 - \tilde{q}_3,\, v_1 - v_3\rangle + \langle q_1 - \tilde{q}_3,\, v_3 - v_{\epsilon_1}\rangle \geq 0,\end{aligned}$$

finishing the proof of the lemma. □

Next, consider the two-dimensional "tube" (actually, a strip)

$$\mathcal{T} = q\left(\left\{\Phi^t w \middle|\, w \in \hat{\Sigma},\ 0 \leq t \leq \tau_{n+1}(w)\right\}\right), \tag{5}$$

where $q(x)$ denotes the natural projection of Ω onto Q, and $\tau_{n+1}(w)$ is the time of the $(n+1)$st collision on the forward orbit of w. Recall that T^{n+1} is smooth on $\hat{\Sigma}$ (where $T^{n+1}w$ is defined as $\Phi^{\tau_{n+1}(w)}(w)$), the endpoints of $T^{n+1}(\hat{\Sigma})$ belong to the base neighborhood U_0 by the construction of $\hat{\Sigma}$, and the curves $T^{n+1}(\hat{\Sigma})$ are monotonic in the canonical $(r,\, \phi)$ (arc-length, angle of reflection) coordinates, thus all the "landing points" $\Phi^{\tau_{n+1}(w)}w = T^{n+1}w$ $(w \in \hat{\Sigma})$ are in the base neighborhood U_0, hence these points w are all sufficient.

Lemma 3. *The footpoint $q_3 = q(x_3)$ belongs to the strip $\mathcal{T}$.*

Proof. Drop a perpendicular line l from q_3 to the supporting curve $q(\hat{\Sigma})$ of $\hat{\Sigma}$. It follows from the convexity of $\hat{\Sigma}$ that the intersection point q_4 of l and $q(\hat{\Sigma})$ lies on the arch connecting q_{ϵ_1} and $\tilde{q}_3$, and $x_3 = \Phi^\eta(q_4, v_4)$ with some small $\eta > 0$ and $x_4 = (q_4, v_4) \in \hat{\Sigma}$. □

According to Lemma 1, some small time shifts of the phase points x_3 and $x_1 \in \partial\hat{\Sigma}$ are contained by a divergent LOM, thus they are not mapped to the same foot point by any positive iterate of the flow, as long as that iterate is smooth on the mentioned LOM. (In other words, the billiard flow, being semi-dispersive, lacks focal points.)

The same statement can be made regarding the pair of phase points $(x_3,\, x_{\varepsilon_1})$. Therefore, the orbit segment $q\left(\Phi^{[0,\,\tau_{n+1}(x_3)]}x_3\right)$ of x_3 does not intersect the boundary of the strip $\mathcal{T}$ of (5), so the orbit segment $q\left(\Phi^{[0,\,\tau_{n+1}(x_3)]}x_3\right)$ cannot escape from $\mathcal{T}$. As a consequence, its endpoint $\Phi^{\tau_{n+1}(x_3)}x_3 = T^{n+1}x_3$ lies in U_0, hence the point x_3 is future sufficient.

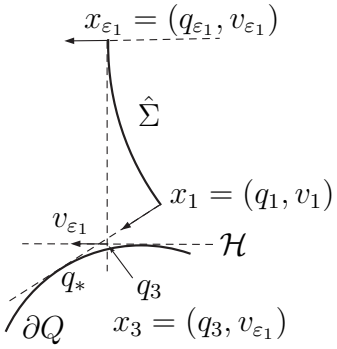

Figure 2: Illustration to the argument below.

As we said earlier, there is the possibility that the required synchronization of $\hat{\Sigma}$ (with a small enough $\varepsilon_1 > 0$, so that x_1 becomes a post-singular phase point) is not feasible. This phenomenon can only happen if the collision of $T^n x$ takes place very close to a corner of the configuration space $\mathbf{Q}$, and very soon after this collision the orbit segment $[T^n x,\, T^{n+1}x]$ flies near a tangency, so that no matter how small $\varepsilon_1 > 0$ one chooses, the endpoint x_1 of $\hat{\Sigma}$ (other than x_{ε_1}) is always a pre-tangency phase point, see Fig. 2. In this case the proof can proceed with such a $\hat{\Sigma}$ and x_1 as follows:

The forward orbit of the pre-tangency phase point $x_1 = (q_1,\, v_1)$ touches $\partial\mathbf{Q}$ at q_*. Let $\mathcal{H}$ be the tangent line to $\partial\mathbf{Q}$ near q_* that is parallel to the

velocity v_{ε_1} of x_{ε_1} (analogously to the previous case), and let again q_3 be the perpendicular projection of q_{ε_1} onto $\mathcal{H}$, and $x_3 = (q_3,\, v_{\varepsilon_1})$, see Fig. 2. It is clear from the picture that the forward orbit of x_3 enters the trip $\mathcal{T}$ soon after time zero. Furthermore, this forward orbit is bound to stay in $\mathcal{T}$ until it reaches U_0 by the same reasoning as before: Both pairs $(x_3,\, x_{\varepsilon_1})$ and $(x_3,\, x_1)$ can be embedded in a divergent LOM by Lemma 1, thus the footpoints $q\,(S^t x_3)$ of $S^t x_3$ $(t > 0,\ S^t x_3$ is in the closure of the strip $\mathcal{T})$ cannot be equal to any of the footpoints $q\,(S^\tau x_{\varepsilon_1})$ or $q\,(S^\tau x_1)$ $(\tau > 0)$, so the forward orbit of x_3 is unable to escape from the strip $\mathcal{T}$. Since reaching the neighborhood U_0 with its forward orbit, the phase point x_3 proves to be future sufficient.

According to the previous constructions, the relevant set for the "upgraded" version of Local Ergodic Theorem is defined as follows:

$$
\begin{aligned}
\tilde{U}_n^{\mathrm{b}}(\delta) = \Big\{ & x \in U_0 \big|\ \exists \varepsilon_1,\ 0 < \varepsilon_1 < \tau(T^n x), \exists \text{ a past sufficient point} \\
& y \in \mathrm{int}\Omega,\ Ty \in \mathcal{S}_0,\ v(y) = v(T^n x) = v\left(\Phi^{\varepsilon_1}(T^n x)\right), \\
& \Delta q = q(y) - q\left(\Phi^{\varepsilon_1}(T^n x)\right) \perp v(T^n x),\ ||\Delta q|| \le \kappa_{n,c_3\delta}(x)^{-1} c_3 \delta \Big\}, \qquad (6) \\
\tilde{U}_{n,m}^{\mathrm{b}} = \tilde{U}_{n,m}^{\mathrm{b}}(\delta) = & \left\{ x \in \tilde{U}_n^{\mathrm{b}} \big|\ \Lambda^m \le \kappa_{n,c_3\delta}(x) < \Lambda^{m+1} \right\},
\end{aligned}
$$

see the definition of the corresponding set $U_{n,m}^{\mathrm{b}}$ before Lemma 6.3 in [7].

Note. For any point $x \in \hat{U}_n^{\mathrm{b}}$ the existence of a suitable point y (required by (6)) is shown by taking $y = -x_3 = (q_3,\, -v_{\varepsilon_1})$. Hence $\hat{U}_n^{\mathrm{b}} \subset \tilde{U}_n^{\mathrm{b}}$.

Now, with the above constructions, we are ready to complete the proof of the tail bound (2). It goes along the same lines as in [7, Section 6], but at several points the argument needs modifications in order to avoid using the Ansatz. We describe these modifications in detail.

1. First of all, in construction of local stable manifolds [7, Lemma 5.4], we can restrict ourselves to the (limits of the) inverse images $\Sigma_0^t(y) = \Phi^{-t}\left(\Sigma_t^t(y)\right)$ of strictly concave, local orthogonal manifolds $\Sigma_t^t(y)$ containing the phase point $\Phi^t y = y_t$. Indeed, if necessary, the partially flat, concave, local orthogonal manifolds $\Sigma_t^t(y)$ may be slightly curved to make them strictly concave. These arbitrarily small perturbations of the manifolds $\Sigma_t^t(y)$, obviously, produce no effect on the limiting process and the overall proof of the Local Ergodic Theorem.

2. The "n-step bad set" $U_n b = U_n^{\mathrm{b}}(\delta)$ defined above can be replaced with our new $\hat{U}_n^{\mathrm{b}} = \hat{U}_n^{\mathrm{b}}(\delta)$, for what it makes a phase point $y \in U_0$ "n-step bad", i.e. $z(T^n y) < \kappa_{n,c_3\delta}(y)^{-1} c_3\delta$, is not that this inequality holds true, rather the fact that the construction of the stable manifold at the phase point y breaks down at the $(n+1)$-st iteration of the billiard map, due to hitting a nearby singularity, just as precisely described in the definition of the set $\hat{U}_n^{\mathrm{b}}$ above. We note that, due to this change, the sets $U^{\mathrm{b}} = \cup_{n=1}^{\infty} U_n^{\mathrm{b}}$ and $U^{\mathrm{g}} = U_0 \setminus U^{\mathrm{b}}$ will change, accordingly.

3. Now, in the tail bound (2), the measures of the sets $\hat{U}_n^{\mathrm{b}}$ and $\hat{U}_\omega^{\mathrm{b}} = \cup_{n>F(\delta)} \hat{U}_n^{\mathrm{b}}$ must be estimated from above. Furthermore, these sets may be replaced by the larger sets $\tilde{U}_n^{\mathrm{b}}$ of (6) and $\tilde{U}_\omega^{\mathrm{b}} = \cup_{n>F(\delta)} \tilde{U}_n^{\mathrm{b}}$.

4. The centerpiece estimate in the proof of the tail bound [7, Section 6] is that for any given $m \in \mathbb{N}$ the ν measure of the set

$$\bigcup_{n \geq N_\eta} T^n \tilde{U}_{n,m}^{\mathrm{b}} \subset (\mathcal{S}_1 \setminus K_\eta)^{[c_3\delta]} \tag{7}$$

(featuring (6.10) in [7]) is bounded above by $c_\eta \delta$ if δ is small enough, where the constant $c_\eta > 0$ can be made arbitrarily small by choosing the parameter $\eta > 0$ small enough. Here the compact subset K_η of the singularity manifold $\mathcal{S}_1$ almost exhausts $\mathcal{S}_1$, that is,

$$\mathbf{m}_{\mathcal{S}_1} (\mathcal{S}_1 \setminus K_\eta) < \eta, \tag{8}$$

where $A^{[c_3\delta]}$ denotes the open $(c_3\delta)$-neighborhood of a subset $A \subset \mathcal{S}_1$ inside $\mathcal{M}$, $\mathbf{m}_{\mathcal{S}_1}$ is the Lebesgue measure on $\mathcal{S}_1$, and $N_\eta \nearrow \infty$ (as $\eta \to 0$) is some threshold function. We note that, both in the original proof of the Fundamental Theorem and in the current one, the set K_η consists of sufficient points only, and – in the original proof – exhausting $\mathcal{S}_1$ in the sense of (8) was made possible by the Ansatz! On the other hand, in the current scenario, lacking the Ansatz, we can only say that $K_\eta \subset \mathrm{Suff}\,(\mathcal{S}_1)$, where $\mathrm{Suff}\,(\mathcal{S}_1)$ denotes the set of all sufficient points of $\mathcal{S}_1$, and

$$\mathbf{m}_{\mathcal{S}_1} (\mathrm{Suff}\,(\mathcal{S}_1) \setminus K_\eta) < \eta.$$

We denote the set $\mathrm{Suff}\,(\mathcal{S}_1) \setminus K_\eta$ by B_η.

5. Keeping in mind definition (6) of the sets $\tilde{U}_n^{\mathrm{b}}$, some open, tubular neighborhood V_0 of $\mathcal{S}_1$ in $\mathcal{M}$ has a uniquely defined foliation

$$\Psi \colon \mathcal{S}_1 \times (-\varepsilon_0,\, \varepsilon_0) \xrightarrow{\cong} V_0$$

with the properties that the section $\Psi\left(\mathcal{S}_1 \times \{0\}\right)$ is equal to $\mathcal{S}_1$ (more precisely, $\Psi(y,0) = y$ for all $y \in \mathcal{S}_1$), and on the one-dimensional foliae (curves) $\Psi\left(\{y_0\} \times (-\varepsilon_0,\, \varepsilon_0)\right)$ (here $y_0 \in \mathcal{S}_1$) the phase points have a constant velocity vector

$$v\left(\Psi(y_0,s)\right) = v\left(\Psi(y_0,0)\right) = v(y_0) \tag{9}$$

for $|s| < \varepsilon_0$, and the curve $\Psi(y_0,s)$ has a time-sync version

$$\gamma_0(s) = \Phi^{\tau(s)}\left(\Psi(y_0,s)\right)$$

(where $0 < \tau(s) < \tau\left(\Psi(y_0,s)\right)$ and $|s| < \varepsilon_0$) such that the carrier $q(\gamma_0(s)) \subset Q$ is linear and orthogonal to the flow-invariant hull

$$\left\{ q\left(\Phi^{\tau}(y_0)\right) \,\middle|\, 0 < \tau < \tau(y_0) \right\},$$

and $\left\| \dfrac{d}{ds} q\left(\gamma_0(s)\right) \right\| = 1$ for all $y_0 \in \mathcal{S}_1$ and $|s| < \epsilon_0$. (So s is an arc-length parameter.) Furthermore, according to (6), the crucial set

$$\cup_{N \geq N_\eta} T^n \tilde{U}^{\mathrm{b}}_{n,m}$$

in (7) is a subset of

$$\Psi\left(B_\eta \times [-c_3\delta,\, c_3\delta]\right)$$

with $\mathbf{m}_{\mathcal{S}_1}(B_\eta) < \eta$, so we have that

$$\nu\left(\cup_{N \geq N_\eta} T^n \tilde{U}^{\mathrm{b}}_{n,m}\right) \leq c_2 c_3 \eta \delta$$

with an absolute constant $c_2 > 0$. The existence of such a constant c_2 is attributed to the facts that

(a) the flow invariant measure μ is uniform on Ω, i.e. $d\mu = \mathrm{const} \cdot dq\, d\theta$, where θ is the angular coordinate of the unit velocity vector;

(b) the T-invariant measure ν on $\mathcal{M}$ is the projection of μ onto $\mathcal{M}$ via the flow;

(c) the above curves $\gamma_0(t)$ are uniformly transversal to the forward invariant hull $\cup_{t>0}\Phi^t(\mathcal{S}_0)$ of the singularity manifold $\mathcal{S}_0$.

As a matter of fact, this constant c_2 is exactly the same as the constant appearing in [17, Lemma 2], and the proof of its existence goes along the same lines, see also [7, Lemma 4.10].

Since the multiplier $c_2 c_3 \eta$ of δ can be made arbitrarily small by choosing the number $\eta > 0$ small enough, we finish the proof of the tail bound without having used the Ansatz.

References

[1] P. Balint, N. Chernov, D. Szasz, and I. P. Toth, *Multi-dimensional semi-dispersing billiards: singularities and the fundamental theorem*, Ann. H. Poincaré **3** (2002), 451–482.

[2] N. Chernov and R. Markarian, *Chaotic Billiards*, Mathematical Surveys and Monographs, **127**, AMS, Providence, RI, 2006. (316 pp.)

[3] N. Chernov and S. Troubetzkoy, *Ergodicity of billiards in polygons with pockets*, Nonlinearity, **11** (1998), 1095–1102.

[4] G. Galperin, *On systems of locally interacting and repelling particles moving in space*, Trudy MMO **43** (1981), 142–196.

[5] S. Kerckhoff, H. Masur, and J. Smillie, *Ergodicity of billiard flows and quadratic differentials*, Annals of Math. **124** (1986), 293–311.

[6] A. Krámli, N. Simányi, and D. Szász, *Ergodic properties of semi-dispersing billiards. I. Two cylindric scatterers in the 3D torus.* Nonlinearity **2** (1989), 311–326.

[7] A. Krámli, N. Simányi, and D. Szász, *A "transversal" fundamental theorem for semi-dispersing billiards*, Comm. Math. Phys., **129** (1990), 535–560.

[8] A. Krámli, N. Simányi, and D. Szász, *The K–Property of Three Billiard Balls*, Ann. Math., **133** (1991), 37–72

[9] A. Krámli, N. Simányi, and D. Szász, *The K-property of four billiard balls*, Comm. Math. Phys. **144** (1992), 107–142.

[10] C. Liverani and M. Wojtkowski, *Ergodicity in Hamiltonian systems*, Dynamics reported, Dynam. Report. Expositions Dynam. Systems (N.S.) **4**, Springer, Berlin (1995), 130–202.

[11] N. Simányi and D. Szász, *Hard ball systems are completely hyperbolic*, Ann. Math. **149** (1999), 35–96.

[12] N. Simányi, *The Complete hyperbolicity of cylindric billiards*, Ergod. Th. Dynam. Syst. **22** (2002), 281–302.

[13] N. Simányi, *Proof of the Boltzmann-Sinai ergodic hypothesis for typical hard disk systems*, Invent. Math. **154** (2003), 123–178.

[14] N. Simányi, *Proof of the ergodic hypothesis for typical hard ball systems*, Ann. H. Poincaré **5** (2004), 203–233.

[15] N. Simányi, *Conditional Proof of the Boltzmann-Sinai Ergodic Hypothesis*, Invent. Math. Online First Publications, DOI: 10.1007/s00222-009-0182-x

[16] Ya. G. Sinai, *On the Foundation of the Ergodic Hypothesis for a Dynamical System of Statistical Mechanics*, Dokl. Akad. Nauk SSSR, **153** (1963), 1261–1264.

[17] Ya. G. Sinai and N. I. Chernov, *Ergodic properties of some systems of 2-dimensional discs and 3-dimensional spheres*, Russ. Math. Surv. **42** (1987), 181–207.

[18] L. N. Vaserstein, *On Systems of Particles with Finite Range and/or Repulsive Interactions*, Commun. Math. Phys. **69** (1979), 31–56.

[19] Ya. B. Vorobets, *Ergodicity of billiards in polygons: explicit examples*, Uspekhi Mat. Nauk **51** (1996), 151–152.