Ergodicity of Billiards in Polygons with Pockets

N. Chernov and S. Troubetzkoy
Department of Mathematics
University of Alabama at Birmingham
Birmingham, AL 35294, USA
E-mail: chernov@math.uab.edu, troubetz@math.uab.edu

October 22, 1997

Abstract

The billiard in a polygon is not always ergodic and never K-mixing or Bernoulli. Here we consider billiard tables by attaching disks to each vertex of an arbitrary simply connected, convex polygon. We show that the billiard on such a table is ergodic, K-mixing and Bernoulli.

1 Introduction

Consider the billiard problem in a polygon. Let P be a polygon in which a particle moves freely and bounces elastically off the boundary ∂P. Assuming the speed of the particle be unit, the phase space will be $TP = P \times S^1$. The flow $\phi_t : TP \to TP$ is called the billiard flow. It preserves the Liouville measure $d\mu = dq \times dv$, where dq and dv are uniform measures on P and S^1, respectively.

All the Lyapunov exponents of the billiard flow in any polygon are zero, its topological entropy [18] and Kolmogorov-Sinai entropy [1, 24] are zero as well. The ergodic properties of the billiard flow depend on the shape of the polygon P. On the one hand, billiards in the so called rational polygons, where each angle is a rational multiple of π, are never ergodic, their phase space TP foliates by compact invariant surfaces [30]. On the other hand, there is a 'topologically large', dense G_δ, subset in the space of all polygons consisting of those where the billiard flow is ergodic [19]. There are no known techniques to determine whether the billiard in a given polygon is ergodic, however. First explicit examples of polygons with ergodic billiard flows were found very recently [27]. It is widely believed that billiards in polygons are never strongly mixing, but they may be weakly mixing [17, 16]. It is known that they cannot be K-mixing or Bernoulli.

In order to ensure hyperbolicity (nonzero Lyapunov exponents) and better ergodic and mixing properties, one has to perturb the polygonal shape of the table by putting in bumps or pockets. Here we study one class of such perturbations.

Let P be a convex simply connected polygon. Assume that at every vertex of P a small pocket is attached to the table. The pockets are bounded by circular arcs that terminate on the sides adjacent to the vertex, see Fig. 1. We call the new billiard table (the union of P and the pockets) by P_ε, thinking of ε as the radius of the pockets, even though the pockets do not have to be of the same radius. We still denote by $\phi_t : TP_\varepsilon \to TP_\varepsilon$ the billiard flow.

Let $Q = \partial P_\varepsilon$ and $TQ = \{x = (q, v) \in TP_\varepsilon : q \in Q$ and v points inside $P_\varepsilon\}$. The flow ϕ_t induces the first return map $f : TQ \to TQ$ that is called the billiard ball map. It preserves a smooth measure, m, on TQ. The ergodicity of the flow $(TP_\varepsilon, \phi_t, \mu)$ is equivalent to that of the map (TQ, f, m).

The main result of the paper is the following.

Theorem 1.1 *Both the flow $(TP_\varepsilon, \phi_t, \mu)$ and the map (TQ, f, m) are hyperbolic and ergodic.*

The following is then standard [2, 3, 8, 23]:

Corollary 1.2 *Both the flow $(TP_\varepsilon, \phi_t, \mu)$ and the map (TQ, f, m) are K-mixing and Bernoulli.*

Remark. We consider circular pockets because this model is the most pictorial. Our results remain valid for small convex pockets of more general shape described in [29, 22, 10, 4], as well as concave bumps, see Fig. 1. It is important that pockets and/or bumps are attached to *every* vertex of the polygon P.

We now describe the main difficulty in the proof of Theorem 1.1.

Let $N \subset TQ$ be the set of points whose trajectories $\{\phi_t x : -\infty < t < \infty\}$ never get into pockets, i.e. which hit only straight sides of the table P_ε. We call them neutral trajectories. Obviously, the set N is f-invariant. The very first task on our way is to make sure that $m(N) = 0$, which was shown some time ago in Ref. [1].

Next, we have to show that the set N is 'slim' enough, so that it cannot separate two ergodic components of f. The slimness of N is based on very recent results described in the next section. After that we employ the standard machinery for proving ergodicity for hyperbolic systems with singularities.

2 Neutral trajectories

Let $\pi : TP \to P$ be the natural projection. In [12] the following theorem was proven.

Theorem 2.1 *For an arbitrary convex simply connected polygon P and for any $x \in TP$ either the orbit is periodic or the closure of the set $\pi\{\phi_t x : t \geq 0\}$ contains at least one vertex of the polygon.*

Remark. Convexity was not assumed in [12], but is needed for corollary 2.2.

We review some notation from [12], see also [13],[14] and [26]. We call a set $S \subset TP$ a strip if, for each $n \geq 0$ the set $f^n S$ consists of parallel vectors whose base points form an interval. If we code billiard orbits to the sequence of sides they hit, then a strip codes to a single sequence. We call a strip periodic if each $x \in S$ is periodic. A periodic strip consists of a union of periodic orbits of the same period and length (here we must traverse twice any periodic orbit which makes an odd number of reflections). Periodic orbits in polygons always come in strips. A maximal width strip is bounded by one or more generalized diagonals, i.e. orbits segments which connect a vertex to a vertex. The number of generalized diagonal is countable, thus the number of (maximal width) period strips is also countable. Since P is convex, having a vertex in the closure of the set $\pi\{\phi_t x : t \geq 0\}$ implies that the orbit of x must hit a circular pocket in the boundary of P_ε (note: this is the only place that we use convexity of P). Thus, as a corollary to theorem 2.1 we have:

Corollary 2.2 *The set of points $x \in Q$ whose future semitrajectories never hit any pocket in the boundary of P_ε is an at most countable union of periodic strips.*

Recently several specialists in polygonal billiards noticed that this corollary can be strengthened [9],[15].

Theorem 2.3 *The set of points $x \in Q$ whose future semitrajectories never hit any pocket in the boundary of P_ε is a* finite *union of periodic strips*

Remark. If the polygon P has a periodic orbit then this set is nonempty for sufficiently small ε. It is unknown if there is a polygon without periodic orbits.

For completeness we sketch a proof here.

Proof: Suppose, by way of contradiction, that the above mentioned union is countable, i.e. there is a countable number of strips of periodic orbits that never hit a pocket. Label the maximal width strips S_i, and an (arbitrary) point on the central orbit of S_i by x_i. Consider the set Z of limit points of the x_i. This set is ε-separated from the vertices, thus by a strengthened version of the Birkhoff recurrence theorem [11] the dynamical system (Z, f) has a uniformly recurrent point x, i.e. for each neighborhood U of x there exists a constant $C > 0$ such that the return time sequence m_i defined by $f^{m_i} x \in U$ satisfies $m_{i+1} - m_i < C$. Fix a $\delta > 0$ and consider the maximal width strip S containing x together with its δ-neighborhood S_δ, see Fig. 2. By maximality, the trajectory of S's boundary points come arbitrarily close to some vertices of the polygon, thus some vertices fall into S_δ. Since x is uniformly recurrent, the left most and right most boundary points of S are also uniformly recurrent, thus vertices fall with uniformly bounded gaps into each of the two components of $S_\delta - S$. By going to a subsequence we can assume that $x_i \to x$. As $i \to \infty$ we consider the intersection of the strips S_i with $S_\delta - S$. Because the gaps between the vertices that fall into $S_\delta - S$ are uniformly bounded, a vertex will eventually appear in the interior of the strip S_i, see Fig. 3. This is a contradiction.□

3 Hyperbolicity

Billiard tables whose boundary consists of straight segments and convex circular arcs were introduced by Bunimovich [2, 3]. He discovered the defocusing mechanism, see below, and studied the hyperbolic and ergodic properties of such billiards. His results have been extended to wide classes of billiards with other convex (focusing) components of the boundary [29, 22, 10, 4]. We only recall here necessary definitions and properties.

Definition. Let $B \subset \mathbb{R}^2$ be a connected billiard table, not a perfect disk, and the boundary ∂B consist of a finite number of straight segments and convex circular arcs, the latter denoted by $\Gamma_1, \ldots, \Gamma_r$. Every Γ_i is an arc of a circle, C_i, that bounds a disk, D_i. Assume that $D_i \subset B$ for all $1 \le i \le r$. Such tables are called *Bunimovich-type billiards with pockets.*

Let $F = \partial B$ and $f : TF \to TF$ be the billiard ball map in B, see Introduction. The map f is piecewise C^∞. Denote by S_- the singularity set for f, it consists of points mapped into the corners of the billiard table B (their further iterations are not defined). Let S_+ be the singularity set for f^{-1}. For $n \ge 1$ denote by $S_{+,n} = S_+ \cup f(S_+) \cup \cdots \cup f^{n-1}(S_+)$ the singularity set for f^{-n}, and $S_{+\infty} = \cup S_{+,n}$. Likewise, put $S_{-,n} = S_- \cup f^{-n}(S_-) \cup \cdots \cup f^{-n+1}(S_-)$ and $S_{-\infty} = \cup S_{-,n}$. Let $SS = S_{+\infty} \cap S_{-\infty}$ be the set of points whose trajectories terminate (hit corners) both in the future and the past. It is known that $S_{+,n}$ and $S_{-,n}$ are finite unions of smooth curves [2, 3, 5, 6].

The main defocusing property of billiards with pockets is the following. Let $q_0 \in \partial B$ and let v_0 be a unit inward velocity vector attached to q_0. Let Σ_0 be an infinitesimal bundle of rays leaving ∂B in the vicinity of q_0, containing v_0 on one of the rays and going into B. Let γ be the orthogonal cross section of the bundle Σ_0 passing through q_0, see Fig. 4, and χ_0 be the signed curvature of γ at the point q_0. The sign of χ_0 is set to be positive if the bundle Σ_0 is diverging and negative if Σ_0 is converging (focusing), as in Fig. 4.

At the time the bundle Σ_0 reaches ∂B again it reflects in ∂B and a new bundle of rays, Σ_1, goes back into B. Let τ_0 be the travel time, $q_1 = q_0 + \tau_0 v_0$ the point of reflection and v_1 the reflected velocity vector at q_1. The new bundle Σ_1 has a certain curvature at q_1, call it χ_1. It is an easy consequence of the mirror equation [29] that

$$\chi_1 = -\frac{2\kappa_1}{\cos\varphi_1} + \frac{1}{\tau_0 + \frac{1}{\chi_0}} \tag{3.1}$$

where φ_1 is the angle between the vector v_1 and the inward normal vector to ∂B at q_1, and $\kappa_1 \ge 0$ is the curvature of ∂B at the point q_1.

The bundle Σ_0 is said to be *unstable* (at q_0) if either
(i) the point q_0 lies on a straight segment in ∂B and $\chi_0 \ge 0$, or
(ii) the point q_0 lies on a circular arc Γ_i of radius R_i, and $\chi_0 \le -(R_i \cos\varphi_0)^{-1}$, where φ_0 is the angle between v_0 and the inward normal vector to Γ_i at q_0.

Theorem 3.1 ([2, 3]) *If Σ_0 is unstable, then so is Σ_1.*

Proof: it is a direct calculation based on (3.1).

In the language of the theory of dynamical systems [28], unstable bundles specify an invariant family of unstable cones, C^u_x, $x \in TF$, for the billiard ball map $f : TF \to TF$.

In the important case (ii) above, the unstable bundle Σ_0 focuses before it reaches the midpoint between the collisions. After that it defocuses and becomes divergent. When it hits ∂B again, at q_1, it already gets *wider* than it was near the point q_0. Obviously, in the case (i), Σ_1 is also wider than Σ_0. The expansion of the bundle between the collisions (with respect to the width measured in the direction perpendicular to the rays) is the main property of unstable bundles. The factor of expansion is $L = 1 + \tau_0\chi_0$ in the case (i) and $L = -1 - \tau_0\chi_0$ in the case (ii), in both cases $L \geq 1$.

The width of unstable bundles specifies a metric, ρ, in the unstable cones. It does not correspond to any metric on TF, so we will call ρ *a pseudometric.* Note that it is monotone under the action of f, i.e. Df expands every unstable vector.

The unstable subspace E^u_x for every $x \in TF$ is defined, as usual, by $E^u_x = \cap_{n\geq 0} Df^n C^u_{f^{-n}x}$. This subspace corresponds to the unstable bundle with the curvature

$$\chi^u_0 = \cfrac{1}{\tau_{-1} + \cfrac{1}{\cfrac{2\kappa_0}{\cos\phi_0} + \cfrac{1}{\tau_{-2} + \cfrac{1}{\cfrac{2\kappa_{-1}}{\cos\phi_{-1}} + \cfrac{1}{\tau_{-3} + \cdots}}}}} \tag{3.2}$$

Here the quantities τ_{-n}, κ_{-n}, and ϕ_{-n} correspond to the point $x_{-n} = f^{-n}x$, $n \geq 1$. This continuous fraction converges whenever $\sum_{n\geq 1}\tau_{-n} = \infty$, i.e. whenever the past semitrajectory of the point x is defined, i.e. for all $x \notin S_{+\infty}$. Hence, E^u_x exists for all $x \in TF \setminus S_{+\infty}$. It also depends continuously on x.

Denote by $L^u_x = |1 + \tau_0\chi^u_0|$ be the factor of expansion of the unstable subspace E^u_x under Df. For $n \geq 1$, denote by $L^u_x(n) = L^u_x L^u_{fx} \cdots L^u_{f^{n-1}x}$ the factor of expansion of E^u_x under Df^n. The factor $L^u_x(n)$ is bounded away from unity only when the trajectory leaves an arc, Γ_i, at time 0 and lands on another arc Γ_j, $j \neq i$, at time n (possibly, with some reflections at straight sides in between). Every time this happens we say that the trajectory experiences an essential transition. During long series of consecutive reflections at straight segments or at the same arc Γ_i (i.e., without essential transitions), the expansion of unstable vectors is weak, so that the expansion factor $L^u_x(n)$ over the entire series of n reflections with no essential transitions grows at most linearly in n, no matter how large n is, cf. [5, 6].

By reversing the time, one can similarly define stable bundles of rays, stable cones C^s_x with a pseudometric ρ, stable subspaces E^s_x, and the expansion factors $L^s_x(n) \geq 1$ of E^s_x under Df^{-n}, $n \geq 1$, for all $x \in TF \setminus S_{-\infty}$. Stable and unstable cones C^u_x and C^s_x never overlap but may have common boundaries.

Definition. A point $x \in TF$ is said to be sufficient if there exists $A > 1$ and two integers $n < m$, such that $f^n x$ and $f^m x$ are defined, and a neighborhood V of the point x such that $L^u_{f^n y}(m-n) > A$ and $L^s_{f^m y}(m-n) > A$ for all $y \in V$.

Definition. A point $x \in TF$ is said to be u-essential if for any $A > 1$ there is an $n \geq 1$, such that $f^n x$ is defined, and a neighborhood V of the point x such that $L^u_y(n) > A$ for all $y \in V$. Similarly, s-essential points are defined (by replacing $L^u_y(n)$ with $L^s_y(n)$ and $f^n x$ with $f^{-n} x$).

The following immediately follows from the previous observations.

Proposition 3.2 *A point x is sufficient if its trajectory (whenever defined) experiences at least one essential transition. A point x is u-essential (or s-essential) if its future (resp. past) semitrajectory is entirely defined and experiences an infinite number of essential transitions. Furthermore, if the subspaces E^u_x and E^s_x are characteristic subspaces with a positive and, respectively, negative Lyapunov exponent, then essential transitions in the entire trajectory of x occur with a positive frequency.*

For the class of billiard tables P_ε we can completely characterize the sets of points $x \in TQ$ that fail to be sufficient or essential. The future semitrajectory $\{f^n x : n \geq 0\}$ of a point $x \notin S_{-\infty}$ experiences at least one essential transition unless (i) $x \in N$, or (ii) the trajectory of x is periodic with all its reflection points lying on one arc, Γ_i. Denote by $G \subset TQ$ the set of points of type (ii). Obviously, it consists of a finite number of disjoint segments in TQ such that the angle of reflection is constant on every of those segments. Put $NG = N \cup G$. We then obtain the following.

Proposition 3.3 *Every point $x \in TQ \setminus (SS \cup NG)$ is sufficient. Every point $x \in TQ \setminus (S_{-\infty} \cup NG)$ is u-essential. Every point $x \in TQ \setminus (S_{+\infty} \cup NG)$ is s-essential.*

The last known fact we need is this [5]: the tangent line to any smooth singularity curve in $S_{+,n}$ lies strictly inside an unstable cone, and the tangent line to any curve in $S_{-,n}$ lies strictly inside a stable cone.

4 Ergodicity

Theorem 4.1 *Let $x \in TQ \setminus (SS \cup NG)$. Then there is a neighborhood $U(x) \subset TQ$ that belongs (mod 0) in one ergodic component of f.*

Proof. This theorem is a version of the local ergodic theorem (or 'fundamental theorem') in the theory of hyperbolic billiards. It was first developed in ref. [25] for gases of hard balls, then generalized in ref. [20] to semi-dispersing billiards (in any dimension) and in ref. [21] to Hamiltonian systems with invariant cone families under certain conditions. The most general and convenient for our purposes version of that theorem was proved in ref. [7]. It requires the verification of the following five properties:

Property 1 (double singularities). For any $n \geq 1$ the set $S_{+,n} \cap S_{-,n}$ consists of a finite number of isolated points.

Property 2 (thickness of neighborhoods of singularities). For any $\delta > 0$ let $U_\delta(S_+ \cup S_-)$ be the δ-neighborhood[1] of the set $S_+ \cup S_-$. Then $m(U_\delta(S_+ \cup S_-)) \leq \text{const} \cdot \delta$.

Property 3 (continuity). The families of stable and unstable subspaces E^s_x and E^u_x are continuous on their domains. Furthermore, the limit spaces $\lim_{y \to x} E^u_y$ and $\lim_{y \to x} E^s_y$ are always transversal at every sufficient point x, even if E^u_x or E^s_x does not exist.

Property 4 ("ansatz"). Almost every point of S_+ (with respect to the Lebesgue length on it) is u-essential, and almost every point of S_- is s-essential.

Property 5 (transversality). At almost every point $x \in S_+$ the subspace E^s_x is defined and transversal to S_+, and at almost every point $x \in S_-$ the subspace E^u_x is defined and transversal to S_-.

The property 1 follows from the last remark in the previous section. The property 2 is based on certain direct but rather delicate calculations, which are described in detail in Refs. [5, 6].

The property 3 follows from the last remark in the previous section and the fact that for any sufficient point $x \in TF$ at least one of the spaces E^u_x, E^s_x lies strictly inside the corresponding cone.

Next, observe that the sets $S_+ \cap S_{-\infty}$ and $S_- \cap S_{+\infty}$ are countable, and $(S_+ \cup S_-) \cap NG = \emptyset$. So, all the points $x \in S_+ \setminus S_{-\infty}$ are u-essential, and all the points $x \in S_- \setminus S_{+\infty}$ are s-essential. This proves 4 and 5.

Now the theorem proved in ref. [7] ensures that every sufficient point, i.e. every point $x \in TQ \setminus (SS \cup NG)$, has a neighborhood that belongs (mod 0) in one ergodic component. $\square$.

We now prove our main theorem 1.1. The set SS is countable. The set NG consists of a finite number of disjoint parallel segments in TQ. Therefore, the set $TQ \setminus (SS \cup NG)$ of points satisfying the assumptions of Theorem 4.1 is an *arcwise connected* set of full measure. This proves Theorem 1.1.

Remark: If we use Corollary 2.2 instead of Theorem 2.3 then we can prove a slightly weaker proposition which is still enough to conclude the ergodicity of f, namely that the set $TQ \setminus (SS \cup NG)$ has an arcwise connected subset of full measure.

References

[1] C. Boldrighini, M. Keane, and F. Marchetti *Billiards in polygons* Ann. Prob. **6** (1978) 532–540.

[2] L.A. Bunimovich, *On billiards close to dispersing*, Math. USSR Sbornik **23** (1974), 45–67.

[1] This must be measured in a *monotone* (pseudo)metric, in which the expansion of unstable vectors and the contraction of stable vectors is monotone. Our pseudometric ρ is exactly such.

[3] L.A. Bunimovich, *On the ergodic properties of nowhere dispersing billiards*, Comm. Math. Phys. **65** (1979), 295–312.

[4] L.A. Bunimovich, *On absolutely focusing mirrors*, Lect. Notes Math., **1514**, Springer, New York, 1990.

[5] L.A. Bunimovich, Ya.G. Sinai and N.I. Chernov, *Markov partitions for two-dimensional hyperbolic billiards*, Russ. Math. Surv. **45** (1991), 97–134.

[6] L.A. Bunimovich, Ya.G. Sinai and N.I. Chernov, *Statistical properties of two-dimensional hyperbolic billiards*, Russ. Math. Surv. **46** (1991), 47–106.

[7] N.I. Chernov, *On local ergodicity in hyperbolic systems with singularities*, Functs. Anal. Applic. **27** (1993), 51–54.

[8] N.I. Chernov, C. Haskell, *Nonuniformly hyperbolic K-systems are Bernoulli*, Ergodic Theory and Dynamical Systems **16** (1996), 19–44.

[9] C. Delman, G. Galpein, and S. Troubetzkoy, *Billiards in a polygon with pockets*, manuscript.

[10] V. Donnay, *Using integrability to produce chaos: billiards with positive entropy*, Comm. Math. Phys. **141** (1991), 225–257.

[11] H. Furstenberg, *Recurrence in ergodic theory and combinatorial number theory,* Princeton NJ, Princeton Univ. Press 1981.

[12] G. Galperin, T. Krüger, and S. Troubetzkoy, *Local instability of orbits in polygonal and polyhedral billiards,* Comm. Math. Phys. **169** (1995) 463–473.

[13] E. Gutkin, *Billiard in polygons,* Physica D **19** (1986) 311-333.

[14] E. Gutkin, *Billiard in polygons: survey of recent results,* J. Stat. Phys., **83** (1996) 7–26.

[15] E. Gutkin, *oral communication.*

[16] E. Gutkin and A. Katok, *Weakly mixing billiards* in *Holomorphic dynamics,* LNM **1345**, Springer-Verlag, Berlin, (1989) 163–176.

[17] A. Katok, *Interval exchange transformations and some special flows are not mixing*, Isr. J. Math. **35** (1980) 301–310.

[18] A. Katok, *The growth rate for the number of singular and periodic orbits for a polygonal billiard*, Comm. Math. Phys. **111** (1987), 151–160.

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

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

[21] C. Liverani, M. Wojtkowski, *Ergodicity in Hamiltonian systems*, Dynamics Reported **4** (1995), 130–202.

[22] R. Markarian, *Billiards with Pesin region of measure one*, Comm. Math. Phys. **118** (1988), 87–97.

[23] D. Ornstein and B. Weiss, *On the Bernoulli nature of systems with some hyperbolic structure*, manuscript, 1995.

[24] Ya.G Sinai, *Introduction to ergodic theory*, Princeton University Press, Princeton (1977).

[25] 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.

[26] S. Tabachnikov, *Billiards,* "Panoramas et Syntheses", Soc. Math. France (1995).

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

[28] M.P. Wojtkowski, *Invariant families of cones and Lyapunov exponents*, Ergod. Th. Dynam. Sys. **5** (1985), 145–161.

[29] M.P. Wojtkowski, *Principles for the design of billiards with nonvanishing Lyapunov exponents*, Commun. Math. Phys. **105** (1986), 391–414.

[30] A. Zemlyakov and A. Katok, *Topological transitivity of billiards in polygons*, Math. Notes **18** (1976), 760–764.

Figure captions

1. A triangle with two pockets and one ‘bump’.

2. The strip S together with it’s δ-neighborhood S_δ.

3. The intersection of S_i with $S_\delta \setminus S$ is shaded.

4. An unstable focusing bundle Σ_0 gets wider at the next reflection near the point q_1.