Available online at www.sciencedirect.com

ScienceDirect

Ann. I. H. Poincaré – AN 34 (2017) 793–816

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
ANALYSE
NON LINÉAIRE

www.elsevier.com/locate/anihpc

Convex billiards on convex spheres

Pengfei Zhang

Department of Mathematics, University of Mississippi, Oxford, MS 38677, United States

Received 26 May 2015; received in revised form 17 February 2016; accepted 7 July 2016

Available online 26 July 2016

Abstract

In this paper we study the dynamical billiards on a convex 2D sphere. We investigate some generic properties of the convex billiards on a general convex sphere. We prove that C^∞ generically, every periodic point is either hyperbolic or elliptic with irrational rotation number. Moreover, every hyperbolic periodic point admits some transverse homoclinic intersections. A new ingredient in our approach is Herman's result on Diophantine invariant curves that we use to prove the nonlinear stability of elliptic periodic points for a dense subset of convex billiards.

MSC: 37D40; 37D50; 37C20; 37E40

Keywords: Convex billiards; Generic properties; Homoclinic intersections; Diophantine number; Nonlinearly stable

1. Introduction

The dynamical billiards, as a class of dynamical systems, were introduced by Birkhoff [5,6] in his study of Lagrangian systems with two degrees of freedom. A Lagrangian system with two degrees of freedom is isomorphic with the motion of a mass particle moving on a surface rotating uniformly about a fixed axis and carrying a fixed conservative field of force with it. If the surface is not rotating and the force vanishes, then the particle moves along geodesics on the surface. If the surface has boundary, then the resulting system is a billiard system.

The classical results of dynamical billiards are closely related to geometrical optics, which has a much longer history. For example, the discovery of the integrability of elliptic billiards, according to Sarnak [49], goes back at least to Boscovich in 1757. Surprisingly, the billiard dynamics is also related to the spectra property of Laplace–Beltrami operator on manifolds with a boundary. More precisely, Weyl's law in spectral theory gives the first order asymptotic distribution of eigenvalues of the Laplace–Beltrami operator on a bounded domain. Weyl's conjecture on the second order asymptotic distribution was proved by Ivrii [29] for any compact manifold with boundary, under the assumption that the measure of periodic points of billiard dynamics on that manifold is zero.

E-mail address: pzhang2@olemiss.edu.

http://dx.doi.org/10.1016/j.anihpc.2016.07.001

Current study of dynamical billiard systems mainly focuses on the Euclidean case. Birkhoff studied the dynamical billiards inside a convex domain on the plane. Birkhoff also conjectured that ellipses are the only integrable billiards. A weak version of this conjecture was proved by Bialy [3]. The dynamical billiards on a bounded domain with convex scatterers were introduced by Sinai in his study of Boltzmann Ergodic Hypothesis [50] on ideal gases. Sinai discovered the dispersing mechanism and proved that dispersing billiards are hyperbolic and ergodic. Since then, the mathematical and physical study of chaotic billiards has developed at a remarkable speed (see [14]), particularly after the various defocusing mechanisms discovered by Bunimovich [9,10], Wojtkowski [55], Markarian [33] and Donnay [21]. Very recently, the dynamics of some asymmetric lemon billiards are proved to be hyperbolic [12], for which the separation condition in the defocusing mechanism was strongly violated. See [53,30,28] for the study of chaotic billiards on general surfaces. The study of chaotic billiards also provides the key idea for the construction of hyperbolic geodesic flows on S^2, see [19,20,13].

Dynamical billiards on curved surfaces are related to the study of quantum magnetic confinement of non-planar 2D electron gases (2DEG) in semiconductors [25], where the effect of varying the curvature of the surface corresponds to a change in the potential energy of the system. The dynamical billiards can be viewed as a mathematical model for this system, and may be used to investigate the electron transport properties of the semiconductors. As mentioned in [28], the advances in semiconductor fabrication techniques allow to manufacture solid state (mesoscopic) devices where electrons are confined to curved surfaces.

In this paper we consider the convex billiards on convex spheres. Recall that the 2D sphere S^2 with a smooth Riemannian metric g is said to be (strictly) *convex*, if it has positive Gaussian curvature: $K_g(x) > 0$ for all $x \in S^2$. Given a tangent vector $\mathbf{v} \in T_x S^2$, the geodesic passing through x in the direction of $\mathbf{v}$ is defined by the exponential map $\gamma_{\mathbf{v}} : \mathbb{R} \to S^2$, $t \mapsto \exp_x(t\mathbf{v})$. For any two points $p, q \in S^2$, let $d(p,q)$ be the length of the shortest geodesics connecting p and q. Let $\mathrm{Inj}(S^2, g)$ be the injective radius of (S^2, g).

Example. Let S^2 be the unit sphere in $\mathbb{R}^3$ endowed with the round metric g_0. Then $K_0 \equiv 1$, and every geodesic on S^2 moves along a great circle. Let $p, q \in S^2$ be two points on the sphere, and α be the angle between the two position vectors $\mathbf{p}, \mathbf{q}$. Then the geodesic distance $d_0(p,q)$ between p and q is given by $d_0(p,q) = \alpha(\mathbf{p}, \mathbf{q})$, and $\cos\alpha = \langle \mathbf{p}, \mathbf{q} \rangle$. Therefore, $d_0(p,q) = \arccos\langle \mathbf{p}, \mathbf{q} \rangle$. Moreover, $\mathrm{Inj}(S^2, g_0) = \pi$. The dynamical billiards inside convex subsets of (S^2, g_0) have been studied recently in [8,4,16]. Regarding the Ivrii conjecture, it is proved in [7] that the set of periodic points of period 3 has zero measure for *any* billiard on the unit sphere.

Definition 1.1. Let (S^2, g) be a convex sphere. A closed subset $Q \subset S^2$ is said to be (geodesically) *convex*, if Q is simply connected, and for any two points $x, y \in Q$, there is a unique minimizing geodesic contained in Q connecting x and y. A convex domain Q is said to be *strictly convex*, if the interior of each minimizing geodesic is contained in the interior Q^o of Q.

Let $Q \subset S^2$ be a convex domain, s be the arc-length parameter of $\Gamma = \partial Q$, and $\kappa(s)$ be the geodesic curvature of Γ at $\Gamma(s)$. Note that $\kappa(s) \geq 0$ for all s. If Q is strictly convex, then $\kappa(s) > 0$ for all s (except on a closed set without interior). By definition, there are no conjugate points inside a convex domain Q. In the following we require that there are no conjugate points on the closed domain Q. A sufficient condition for nonexistence of conjugate point is that $\mathrm{diam}(Q) < \mathrm{Inj}(S^2, g)$.

The dynamical billiard on Q can be defined analogously to the planar case. That is, a particle moves along geodesics inside Q, and reflects elastically upon hitting the boundary ∂Q. Suppose the previous reflection happens at $\Gamma(s)$. Let θ be the angle measured from the (positive) tangent direction $\dot{\Gamma}(s)$ to the post-reflection velocity of that particle. Then the *billiard map* F sends (s, θ) to the next reflection (s_1, θ_1) with ∂Q. The *phase space* of the billiard map F on Q is given by $M = \Gamma \times (0, \pi)$. Note that the 2-form $\omega = \sin\theta\, ds \wedge d\theta$ is a symplectic form on M. Let μ be the smooth probability measure on M with density $d\mu = \frac{1}{2|\partial Q|} \sin\theta\, ds\, d\theta$.

Theorem 1. *Let (S^2, g) be a convex sphere and $Q \subset S^2$ be a strictly convex domain with C^r smooth boundary $\Gamma = \partial Q$. Then billiard map $F : M \to M$ is a symplectic twist map. In particular, F preserves the measure μ.*

It is well known that a twist map has periodic orbits of Birkhoff type (m, n) for all coprime pairs (m, n) [6,2]. It may (most likely will) have some non-Birkhoff periodic orbits.[1] We study some generic properties of general periodic points of dynamical billiards on a strictly convex domain Q on (S^2, g). To this end, we identify the boundary $\Gamma = \partial Q$ with the corresponding embedding function $f : \mathbb{T} \to S^2$. Let $r \geq 2$ (r could be ∞), $\Upsilon^r(S^2, g)$ be the set of C^r smooth embeddings $\Gamma \subset S^2$ such that the enclosed domains $Q = Q(\Gamma)$ are strictly convex. Then $\Upsilon^r(S^2, g)$ inherits a C^r topology from $C^r(\mathbb{T}, S^2)$.

Theorem 2. *There is a residual subset $\mathcal{R}^r \subset \Upsilon^r(S^2, g)$, such that for each $\Gamma \in \mathcal{R}^r$, the billiard map on Γ satisfies*

(1) *each periodic point is either hyperbolic, or elliptic with irrational rotation number;*
(2) *any two branches of invariant manifolds of hyperbolic periodic points either do not intersect, or they have some transverse intersections.*

Theorem 2 resembles the classical Kupka–Smale properties for dynamical billiards. The abstract Kupka–Smale property is proved by applying Thom Transversality Theorem, which requires the *richness* of local perturbations. However, dynamical billiards are known for the *lack* of local perturbations, since any perturbation of Γ results in a (semi)-global perturbation of the billiard map. See §4 for more details.

Given two hyperbolic periodic points p and q, these two points and their stable and unstable manifolds may be separated by some KAM-type invariant curves (which are persistent under small perturbations). So the existence of heteroclinic intersections may not be generic. The following theorem answers positively the generic existence of homoclinic intersections.

Theorem 3. *There is a residual subset $\mathcal{R}^r \subset \Upsilon^r(S^2, g)$, such that for each $\Gamma \in \mathcal{R}^r$, there exist transverse homoclinic intersections for each hyperbolic periodic point of the billiard map F induced by Γ.*

The proof of above theorem is based on Mather's characterization [35] (developed by Franks and Le Calvez in [26]) of the *prime-end extension* of diffeomorphisms on open surfaces. In his proof, Mather made an assumption that each *elliptic fixed point*, if exists, is Moser stable. To apply Mather's result, we have to study the elliptic periodic points first, although the hyperbolic periodic points are the ones we are interested in. The nonlinear stability is proved by one of Herman's results on Diophantine invariant curves. This property guarantees that there is no interaction between the hyperbolic and elliptic periodic points.

Note that there are plenty of periodic points for twist maps, and hyperbolic periodic points exist generically. So the transverse homoclinic intersections in above theorem do exist generically.

Corollary 1. *There is an open and dense subset $\mathcal{U}^r \subset \Upsilon^r(S^2, g)$, such that for each $\Gamma \in \mathcal{U}^r$, the billiard map on Γ has positive topological entropy.*

Angenent [1] proved that a twist map with zero topological entropy must have an invariant circle for each rotation number in its rotation interval. On the other hand, invariant curves with rational rotation numbers are fragile and can easily break up. Therefore, the majority of twist maps should have positive topological entropy. So Corollary 1 can be viewed as a special case of Angenent's result.

Entropy is an important quantity indicating how chaotic a dynamical system is. The mechanism that a transverse homoclinic intersection generates chaos was first realized by Poincaré when he came across certain nonconvergent trigonometric series during his study of the n-body problem [43]. This mechanism was developed later by Birkhoff for the existence of infinitely many periodic points, and by Smale for the formulation of hyperbolic sets (horseshoe). Poincaré conjectured that for a generic $f \in \mathrm{Diff}^r_\mu(M)$, and for every hyperbolic periodic point p of f,

(P1) $W^s(p) \cap W^u(p) \backslash \{p\} \neq \emptyset$ (weaker version);
(P2) $W^s(p) \cap W^u(p)$ is dense in $W^s(p) \cup W^u(p)$.

[1] Take a planar elliptic billiard for example. The periodic orbits with elliptic caustics are Birkhoff, while the periodic orbits with hyperbolic caustics are non-Birkhoff.

This is the so called Poincaré's connecting problem.[2] In the case $r = 1$, (P1) was proved by Takens in [52]; (P2) was proved in [52] on surfaces, and by Xia [56] in full generality. For $r \geq 2$, most results about this connecting problem are on surfaces. Pixton proved in [42] the property (P1) for planar surfaces, by extending Robinson's result [47] on fixed points. For $M = \mathbb{T}^2$, (P1) was proved by Oliveira [37]. For general surfaces, (P1) was proved by Oliveira [38] for those with irreducible homological actions; and by Xia in [57] for Hamiltonian diffeomorphisms. The proof of (P1) is still not complete for general surfaces, and there is almost no result on higher dimensions. The property (P2) is completely open even on surfaces. For planar convex billiards, (P1) was proved in [58].

Finally we make a few comments on the positive Gaussian curvature assumption of the Riemannian metric g on S^2. Suppose the curvature can be negative somewhere on the sphere. For example, one can put a small light bulb on the table Q as in [23, Fig. 2]. Then the *neck* of the light bulb will be a hyperbolic closed geodesic, and some geodesic on its unstable manifold will hit the boundary Γ of Q. Reversing the time, we get a billiard trajectory starting on Γ that will not collide with Γ in the future. In other words, the billiard map F is not defined on the whole phase space and is certainly not continuous. It seems that our method in this paper does not work (at least not directly).

2. Preliminaries

Let (S^2, g) be a convex sphere, and $Q \subset S^2$ be a strictly convex domain with C^r smooth boundary $\Gamma = \partial Q$. Let $M \subset T_\Gamma S^2$ be the set of unit tangent vectors $x = (p, \mathbf{v})$ based at points $p \in \Gamma$ that point to the interior of Q. Given a point $x \in M$, let $\gamma_x(t) = \exp_p(t\mathbf{v})$ be the geodesic on Q with initial condition $(\gamma(0), \dot{\gamma}(0)) = (p, \mathbf{v}) = x$. Let t_1 be the next hitting time of $\gamma(t)$ with Γ, $p_1 = \gamma(t_1) \in \Gamma$, and x_1 be the reflection of $\dot{\gamma}(t_1)$ with respect to the tangent line $T_{p_1}\Gamma \subset T_{p_1}S^2$. Then the billiard map F is defined as $M \to M$, $x \mapsto x_1$. It is convenient to introduce a coordinate system on M. That is, given $x = (p, \mathbf{v}) \in M$, let $s = s(p)$ be the arc-length parameter of Γ, $\theta = \theta(\mathbf{v})$ be the angle of $\mathbf{v}$ measured from the tangent direction $\dot{\Gamma}(s)$. In the following we will represent M via this coordinate system $\{(s, \theta) : s \in \Gamma, 0 < \theta < \pi\}$, and rewrite the billiard map F as $x = (s, \theta) \mapsto x_1 = (s_1, \theta_1)$.

2.1. Generating function of billiard map

The dynamical billiard has an alternative definition using the generating function. More precisely, let $s \mapsto \Gamma(s)$ be the arc-length parameter. We will write $s \in \Gamma$ by identifying s with $\Gamma(s)$ if there is no confusion. For example, we set $d_\Gamma(s_1, s_2) = d(\Gamma(s_1), \Gamma(s_2))$. Let $S(s_1, s_2) = -d_\Gamma(s_1, s_2)$, and $\partial_i S$ be the partial derivative of S with respect to s_i, $i = 1, 2$. We extend the generating function to an arbitrary finite segment $(s_m, \dots, s_n)$ with $s_k \in \Gamma$, $k = m, m+1, \dots, n$, and define the *action functional* $W(s_m, \dots, s_n) = \sum_{k=m}^{n-1} S(s_k, s_{k+1})$ along the segment $(s_m, \dots, s_n)$. Such a segment is said to be an orbit segment, if $\partial_{s_k} W = \partial_2 S(s_{k-1}, s_k) + \partial_1 S(s_k, s_{k+1}) = 0$ for each $k = m, \dots, n-1$.

Proof of Theorem 1. Given two points s_1 and s_2, let $\gamma_1(t)$ be the geodesic from $\gamma_1(0) = \Gamma(s_1)$ to $\gamma_1(d) = \Gamma(s_2)$, where $d = d_\Gamma(s_1, s_2)$. Let θ_1 be the angle from $\dot{\Gamma}(s_1)$ to $\dot{\gamma}_1(0)$, and θ_2 be the angle from $\dot{\Gamma}(s_2)$ to $\dot{\gamma}_1(d)$. At $\Gamma(s_2)$, γ_1 experiences an elastic reflection, and the new geodesic, say γ_2, starts from $\gamma_2(0) = \Gamma(s_2)$, such that the angle from $\dot{\Gamma}(s_2)$ to $\dot{\gamma}_2(0)$ equals θ_2. One can check that

$$\partial_1 S(s_1, s_2) = \cos\theta_1, \quad \partial_2 S(s_1, s_2) = -\cos\theta_2. \tag{2.1}$$

Therefore, $F(s_1, \theta_1) = (s_2, \theta_2)$ if and only if $\partial_1 S(s_1, s_2) = \cos\theta_1$ and $\partial_2 S(s_1, s_2) = -\cos\theta_2$. Rewriting (2.1) in total differential form, we get $dS = \cos\theta_1 ds_1 - \cos\theta_2 ds_2$. Taking exterior differential and using $d^2 S = 0$, we get $\sin\theta_2 ds_2 \wedge d\theta_2 = \sin\theta_1 ds_1 \wedge d\theta_1$. Therefore, the 2-form $\omega = \sin\theta ds \wedge d\theta$ is invariant under F, so is the probability measure $d\mu = \frac{1}{2|\Gamma|} \sin\theta ds d\theta$ on $M = \Gamma \times (0, \pi)$.

To show that F is a twist map on $M = \Gamma \times (0, \pi)$, let's consider the image of $M_s = \{s\} \times (0, \pi)$ under F. Let $\gamma_\theta(t)$ be the geodesic starting from $\Gamma(s)$ in the direction of θ, and $t_\theta > 0$ be the first moment that $\gamma_\theta(t)$ hits Γ. The hitting position is exactly $s_1(\theta) = p_1 \circ F(s, \theta)$. Since Q is a strictly convex domain on S^2, the map $s_1 : (0, \pi) \to \Gamma$ is monotonically increasing. Therefore, F is a symplectic twist map on M. □

[2] Poincaré also raised the closing problem about the denseness of periodic points, see [44,45].

Corollary 2.1. *Let $\Gamma \in \Upsilon^r(S^2, g)$, and F be the billiard map induced by Γ. Then for any coprime positive integers (p, q) with $q \geq 2$, there exists a periodic orbit $\mathcal{O}_{p,q}$ of period q that goes around the table p times after one period.*

Such an orbit $\mathcal{O}_{p,q}$ is called a Birkhoff periodic orbit of type (p, q). See [6,2] for more details. Note that there may be some periodic orbits of non-Birkhoff type.

2.2. Criterion of nondegenerate periodic orbits

Let $W(s_1, \dots, s_n) = \sum_{k=1}^{n} S(s_{k-1}, s_k)$ be the action on the space of the n-periodic configurations (s_k) in the sense that $s_{n+k} = s_k$ for all k. Then $x = (s, \theta) \in M$ is a periodic point with period n if and only if $\partial_k W(s_1, \dots, s_n) = 0$ for each $k = 1, \dots, n$, where $x_k = F^k x = (s_k, \theta_k)$ are the iterates of x under the billiard map. Given a critical n-periodic configuration (s_k), we let $D^2 W(s_1, \dots, s_n) = (\partial^2_{ij} W)$ be the $n \times n$ Hessian matrix of W at $(s_1, \dots, s_n)$.

Let $D_x F^n$ be the tangent map at x (counted to its period), which is a 2×2 matrix with determinant 1 (since F preserves the symplectic form ω). Then x is said to be *non-degenerate*, if 1 is not an eigenvalue of $D_x F^n$. The later condition is equivalent to $\mathrm{Tr}(D_x F^n) \neq 2$. Mackay and Meiss proved in [32] that the trace $\mathrm{Tr}(D_x F^n)$ is closely related to the Hessian $D^2 W$ of W at its critical path $(s_1, \dots, s_n)$.

Proposition 2.2. *Let $\{F^k x = (s_k, \theta_k)\}$ be a periodic orbit of period n, $W_2 = D^2 W(s_1, \dots, s_n)$ be the Hessian matrix of W at $(s_1, \dots, s_n)$. Then* $\mathrm{Tr}(D_x F^n) - 2 = (-1)^n \cdot \det(W_2) \cdot \left(\prod_{i=1}^{n} S_{12}(s_{i-1}, s_i) \right)^{-1}$.

Note that $\mathrm{Tr}(D_x F^n) = 2$ if and only if $\det(W_2) = 0$. So we have the following equivalent formulations:

(1) a periodic orbit $x = T^n x$ of the billiard map F is nondegenerate;
(2) a critical cycle $(s_1, \dots, s_n)$ of the action functional W is nondegenerate.

Birkhoff made the following observation in [6]. Let $(s_1, \dots, s_n)$ be an n-periodic configuration at where W attains its minimum. Assume the corresponding periodic orbit x is nondegenerate. Then $D^2 W(s_1, \dots, s_n)$ is positive definite, and $\mathrm{Tr}(D_x F^n) - 2 > 0$. So the periodic point x corresponding to each minimizer turns out to be a hyperbolic periodic point.

2.3. Curvature and focusing time of a tangent vector

Now we describe some geometrical features of the tangent map of a billiard map $F : M \to M$ on the configuration space S^2, see [53] for more details. We start with the coordinate system $\{(s, \theta) : s \in \Gamma, \theta \in (0, \pi)\}$ on M, where s is the arc-length parameter of the boundary $\Gamma = \partial Q$, and θ is the angle of a unit tangent vector $\mathbf{v} \in T_{\Gamma(s)} Q$ with the direction $\dot{\Gamma}(s)$. Let $x_0 = (s_0, \theta_0) \in M$, $\gamma_0(t)$ be the geodesic generated by x_0, $V = a\partial_s + b\partial_\theta \in T_{x_0} M$ be a tangent vector on the phase space M, and $m(V) = \frac{b}{a}$ be the slope of V with respect to the (s, θ)-coordinate. Let $c : (-\epsilon, \epsilon) \to M$ be a smooth curve passing through $c(0) = x_0$ such that $V = \dot{c}(0)$. Then for each $-\epsilon < u < \epsilon$, the point $c(u)$ will determine a geodesic on Q, say $\gamma_u(\cdot)$. Putting them together, we get a beam of geodesics around the geodesic γ_0. A curve $\rho : (-\epsilon, \epsilon) \to S^2$ with $\rho(0) = \Gamma(s_0)$ and $\rho(u) \in \gamma_u$ is called a *wave-front* corresponding to $V \in T_x M$, if $\rho(u)$ is perpendicular to each γ_u at $\rho(u)$. Let $\mathcal{B}(V)$ be the geodesic curvature of ρ at $\rho(0)$. Note that $\mathcal{B}(V)$ does not depend on the choices of curves c with $\dot{c}(0) = V$.

Convention. A wave-front has negative curvature if it is focusing, and has positive curvature if it is dispersing. Let $\mathcal{B}(V) = \infty$ if p itself is a focusing point.

Any (infinitesimal) wave-front of billiard trajectories on Q focuses at some point forward and some point backward on S^2 (not necessarily in Q), say p_+ and p_-. Let $f(V) = d(\Gamma(s_0), p_+)$ be the forward focusing distance (time) of the wavefront related to $V \in T_{x_0} M$. Set $f(V) = 0$ when $\Gamma(s_0)$ itself is a focusing point of the wavefront of V.

Note that $\mathcal{B}(V)$ and $f(V)$ can be defined via normal Jacobi fields. That is, let $\mathbf{J}(t) = \frac{d}{du}\Big|_{u=0} \gamma_u(t)$ be the Jacobi field generated by a beam of geodesics γ_u along γ_0. Jacobi fields are characterized by Jacobi equation: $\ddot{\mathbf{J}} + R(\mathbf{J}, \dot{\gamma}_0)\dot{\gamma}_0 = 0$, where R is the curvature tensor. A Jacobi field $\mathbf{J}$ is said to be *normal*, if $\mathbf{J}(t)$ is perpendicular to $\dot\gamma(t)$ for all t. In this case we can write $\mathbf{J}(t) = J(t)\mathbf{n}_t$ for some scalar function $J(t)$, where $\mathbf{n}_t$ is the unit normal vector field along $\gamma_0(t)$. The scalar Jacobi function $J(t)$ satisfies the scalar Jacobi equation $\ddot J + K_g \cdot J = 0$, where K_g is the Gaussian curvature of (S^2, g). Note that we have $\mathcal{B}(V) = \frac{\dot J(0)}{J(0)}$, $f(V) = \min\{t \ge 0 : J(t) = 0\}$. So the relation between $\mathcal{B}(V)$ and $f(V)$ is given by the solution of the Jacobi equation. For example, if $\mathcal{B}(V) = 0$ then the wavefront focuses at two *focal points* along the geodesic γ_x (one forward focal point, and one backward focal point), and these two focal points are conjugate along γ_x.

The wave-front of a vector V changes its curvature at the moment when the billiard orbit collides with the boundary Γ. More precisely, let $\mathcal{B}^{\pm}(V)$ be the curvature of the wavefront before and after the reflection with Γ, respectively. The relation between the curvature $\mathcal{B}^{\pm}(V)$ and the slope $m(V)$ is given by

$$m(V) = \mathcal{B}^-(V)\sin\theta - \kappa(s) = \mathcal{B}^+(V)\sin\theta + \kappa(s),$$

where $s = p_1(x)$ is the projection to the first coordinate of x.

Now let $x = (s, \theta) \in M$, $Fx = (s_1, \theta_1)$, $V \in T_xM$, $V_1 = DF(V) \in T_{x_1}M$, and ρ be a wavefront related to V. Let $\mathcal{B}_t(V)$ and $f_t(V)$ be the curvature and forward focusing time of the wavefront during the free flight time $0 < t < d_1 = d_\Gamma(s, s_1)$, $\mathcal{B}^{\pm}(V_1)$ and $f^{\pm}(V_1)$ be the curvature and focusing time right before/after the collision $t \to d_1 \pm 0$. Then we have

(1). $\mathcal{B}_t(V) = \frac{\dot J(t)}{J(t)}$, where $J(t)$ is the solution of Jacobi equation;

(2). $\mathcal{B}^+(V_1) = \mathcal{B}^-(V_1) - \dfrac{2\kappa(s_1)}{\sin\theta_1}$, where $\kappa(s_1) > 0$ is the curvature at $\Gamma(s_1)$.

Item (2) is the so called Mirror Formula for geometrical optics on surfaces. Note that $f_t(V) = f(V) - t$ when $t \le f(V)$. If $f(V) < d_\Gamma(s, s_1)$, then the wavefront focuses between two consecutive reflections, $\mathcal{B}_t(V)$ jumps from $-\infty$ to $+\infty$, and $f_t(V)$ jumps from 0 to the next focusing time.

Example. In the case that $g = g_0$ is the round metric on S^2, the quantities $\mathcal{B}(V)$, $f(V) = d(p, p_+)$ and $\hat f(V) = d(p, p_-)$ are related by the following formula:

$$f(V) + \hat f(V) = \pi, \quad \mathcal{B}(V) = -\cot f(V) = \cot \hat f(V). \tag{2.2}$$

Let $\mathcal{B}(V) = \cot\alpha_0$. Then $\mathcal{B}_t(V) = \cot(\alpha_0 + t)$ for all $0 \le t < d(s, s_1)$.

Proof of (2.2). Let's consider the circles L_α of latitude on S^2 surrounding the north pole, where α is the angle of the circle with the positive z-axis. Then the radius of L_α is $r(\alpha) = \sin\alpha$, and the geodesic curvature is $\kappa(\alpha) = \sqrt{1/r^2 - 1} = \cot\alpha$. Then the results follow from the observation that $d(p, p_+) = \alpha$ and $d(p, p_-) = \pi - \alpha$ (and the convention on the choices of signs of the curvature). □

2.4. Some generic properties of periodic orbits

Let (S^2, g) be a convex sphere, $Q \subset S^2$ be a strictly convex domain, and $F : M \to M$ be the induced billiard map on Q, where $M = \Gamma \times (0, \pi)$. Note that the geodesics on Riemannian manifolds are time-reversal invariant (this may not be true on general Finsler manifolds). Similarly, the billiard dynamics on a convex table $Q \subset S^2$ is time-reversal invariant. More precisely, let $\Theta : M \to M$, $(s, \theta) \mapsto (s, \pi - \theta)$ be the time-reversal map. Then $F \circ \Theta = \Theta \circ F^{-1}$. So if $\mathcal{O}$ is a periodic orbit of F, so is $\Theta(\mathcal{O})$; and these two orbits are distinct if $\pi/2 \notin p_2(\mathcal{O})$, where $p_2 : M \to (0, \pi)$ is the projection to the θ coordinate. Note that $\mathcal{O}$ and $\Theta(\mathcal{O})$ have the same dynamical characteristics. We only need to consider one of them when making perturbations.

Definition 2.1. Two different periodic orbits $\mathcal{O}_1$ and $\mathcal{O}_2$ are said to be *essentially different*, if $\mathcal{O}_2$ is not the time-reversal of $\mathcal{O}_1$.

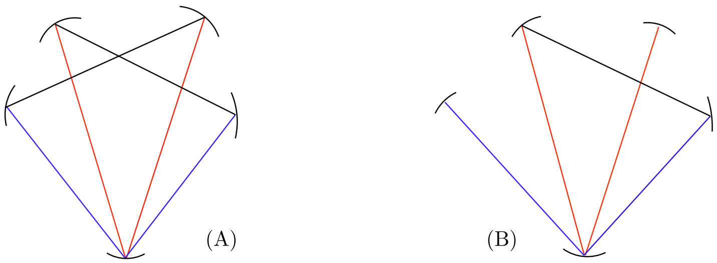

Fig. 1. Periodic orbits with positive defects. (A): nonsymmetric case; (B): symmetric case. This is merely a simplistic sketch, to illustrate the two types of defects of periodic orbits. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

There are some special features for the periodic orbits on the billiard map on Q (see [51]):

(1) it is possible that $|\mathcal{O}(p)| \neq |p_1(\mathcal{O}(p))|$: the orbit passes some reflection point more than once during a minimal period;
(2) it is possible that $|p_1(\mathcal{O}_1 \cup \mathcal{O}_2)| \neq |p_1(\mathcal{O}_1)| + |p_1(\mathcal{O}_2)|$: two essentially different periodic orbits have some common reflection points.

Take the round table on standard sphere for example: on each point $s \in \Gamma$, there exist periodic orbits of type (m, n) for all (m, n). This happens even among the orbits with the same period: the $(1, 5)$-orbit (pentagon) and the $(2, 5)$-orbit (pentagram).

Before giving the precise definition, we need to distinguish the following two cases: symmetric and nonsymmetric orbits. A periodic orbit $\mathcal{O}(p)$ is said to be *symmetric*, if $\theta_k = \pi/2$ for some k. Along such an orbit, the period $n = 2m$ is an even number, the right angle reflections happen exactly twice, and the orbit travels back and forth between these two reflection points. See [51]. A periodic orbit is said to be nonsymmetric, if it is not symmetric.

Definition 2.2. If a periodic orbit $\mathcal{O}(p)$ is nonsymmetric, then the defect of p is defined by the difference $d(p) = |\mathcal{O}(p)| - |p_1(\mathcal{O}(p))|$. If $\mathcal{O}(p)$ is symmetric, then the defect of p is defined by $d(p) = \frac{1}{2}|\mathcal{O}(p)| + 1 - |p_1(\mathcal{O}(p))|$.

See Fig. 1 for a schematic sketch of (planar) periodic orbits with positive defect: (A) for nonsymmetric case, and (B) for symmetric case.

Proposition 2.3. *Let $P_n(\Gamma)$ be the set of points fixed by F^n. There is a residual subset $\mathcal{S}_n \subset \Upsilon^r(S^2, g)$, such that the following hold for the billiard map of each $\Gamma \in \mathcal{S}_n$,*

(1) *every periodic orbit in $P_n(\Gamma)$ has zero defect;*
(2) *two essentially different periodic orbits in $P_n(\Gamma)$ have no common reflection point.*

Note that the periodic orbits of period 2 always have zero defect. So $\mathcal{S}_2 = \Upsilon^r(S^2, g)$.

For billiards in the Euclidean domain, Proposition 2.3 has been proved by Stojanov [51]. Note that the following two statements are equivalent for a given Γ:

- every periodic orbit has zero defect;
- any periodic path $s_0, s_1, \ldots, s_n = s_0$ with positive defect is not a billiard orbit.

Then Proposition 2.3 is proved by showing that the second statement holds generically. The proof for billiards on S^2 follows the same idea, and is sketched in the Appendix.

Remark 2.1. Let $\mathcal{S} = \bigcap_{n \geq 1} \mathcal{S}_n$, which contains a residual subset of $\Upsilon^r(S^2, g)$. Then for each $\Gamma \in \mathcal{S}$,

(a) every periodic orbit of F has zero defect;
(b) two different periodic orbits of F do not pass any common reflection point.

One would expect that $\mathcal{S}_n$ could be open and dense, not just residual. However, this may not be true for general domains. In next section we will prove that the properties (a) and (b) do hold on an open and dense subset of the convex domains in $\Upsilon^r(S^2, g)$.

Remark 2.2. The following properties are obtained in [39–41] for billiard systems on a generic connected domain in $\mathbb{R}^d$:

(I) the set of points fixed by F^n is finite;
(II) the eigenvalue of each periodic point fixed by F^n is not in $\mathcal{A}$,

where $\mathcal{A}$ is any countable subset of $\mathbb{R}$ given in advance. The 2D version has been obtained by Lazutkin [31]. We will prove that these properties hold on an open and dense subset of convex billiards, and the sets of points fixed by F^n actually vary continuously. This continuity plays a key role in the study of homoclinic and heteroclinic intersections.

2.5. Parametric Transversality Theorem

Let M and N be two manifolds, $K \subset M$ be a subset and $V \subset N$ be a submanifold. A smooth map $f : M \to N$ is said to be *transverse* to V at $x \in K$ if one of the following holds:

- $fx \notin V$;
- $y = fx \in V$ and $D_x f(T_x M) + T_y V = T_y N$.

Then f is said to be *transverse* to V along K, denoted by $f \pitchfork_K V$, if f is transverse to V at each $x \in K$. Note that for a diffeomorphism $f \in \mathrm{Diff}^r(M)$, a periodic point x of period k is nondegenerate if and only if the map $(\mathrm{Id}, f^n) : M \to M \times M$ is transverse to the diagonal $\Delta \subset M \times M$.

Let M be a smooth manifold, $D \subset \mathbb{R}^q$ be an open subset, and $\rho : D \to C^r(M, M \times M)$ be a continuous map for some $r \geq 1$. The evaluation of ρ, denoted by $\rho^{\mathrm{ev}} : D \times M \to M \times M$ is given by $(v, x) \mapsto \rho(v)(x)$. Now we can state the Parametric Transversality Theorem (see [48]).

Theorem 2.4. *Suppose $\rho : D \to C^r(M, M \times M)$ is continuous and $\rho^{\mathrm{ev}} : D \times M \to M \times M$ is C^r. Let $K \subset M$ be a compact subset such that ρ^{ev} is transverse to Δ along $D \times K$. Then the set $\{v \in D : \rho(v) \pitchfork_K \Delta\}$ is open and dense in D.*

An intuitive description is also given in [48]: if there are enough parameters with which to make the necessary perturbations at *one point at a time*, then the above theorem implies that the function can be approximated by one which is transverse at *all points in the same time*.

3. Perturbations of periodic points of billiard systems

There are various types of perturbation techniques in the study of dynamical systems. One of the widely used technique is *Franks' Lemma*, which allows us to manipulate the derivatives along a periodic orbit. The perturbations for billiard dynamics are very limited, since one cannot perturb the billiard map F directly, while the perturbation of the underlining table changes the dynamics (semi)-globally. See Visscher's thesis [54] for several results on Franks's lemma in geometric contexts (geodesics flows and billiards). In [18] the effect of the perturbation of a planar billiard system is computed explicitly via a step by step induction. It is difficult to generalize their approach to dynamical billiards on surfaces with non-constant curvature. In this section we present another proof, which uses the geometric features of the tangent vectors of the phase space M on the configuration space S^2.

We first give some basic definitions. Let p be a periodic point of F of period n, $D_p F^n : T_p M \to T_p M$ be the tangent map, which can be viewed as a matrix in $\mathrm{SL}(2, \mathbb{R})$. Let λ_p be an eigenvalue of $D_p F^n$. Then p is said to be *hyperbolic* if $|\lambda_p| \neq 1$, be *parabolic* if $\lambda_p = \pm 1$, and be *elliptic* otherwise. Recall that a periodic point p is said to be *degenerate* if $\lambda_p = 1$, and be *nondegenerate* if it is not degenerate.

Let $\tau(p)$ be the trace of D_pF^n. Then we have the following equivalent definition: p is said to be hyperbolic if $|\tau(p)| > 2$, be parabolic if $|\tau(p)| = 2$, be elliptic if $|\tau(p)| < 2$, be degenerate if $\tau(p) = 2$, and be nondegenerate if $\tau(p) \neq 2$. All nondegenerate periodic points persist under small perturbations.

3.1. Useful perturbations of billiard systems

The following perturbations have been widely used in the study of generic properties of billiards.

Definition 3.1. Let $s_0 \in \Gamma$, and $I \subset \Gamma$ be a neighborhood of s_0. Then a *normal perturbation* Γ_ϵ of Γ at s_0 supported on I is a convex curve on S^2 that satisfies $\Gamma_\epsilon(s) = \Gamma(s)$ for $s = s_0$ and for $s \notin I$, $\dot\Gamma_\epsilon(s_0) = \dot\Gamma(s_0)$, while the curvature changes to $\kappa_\epsilon(s_0) = \kappa(s_0) + \epsilon$.

The normal perturbations are essentially the only types of perturbations that preserve the orbit $\mathcal{O}(p)$, in the meanwhile, change the derivatives of DF^n at p. However, a degenerate periodic point may be *robustly degenerate* under normal perturbations.

Example. Let γ be a geodesic starting at a point $\mathbf{p} \in S^2$, and $\mathbf{q}$ be a conjugate point of $\mathbf{p}$ along γ. Let $Q \subset S^2$ be a convex domain containing the geodesic segment γ from $\mathbf{p}$ to $\mathbf{q}$ as a diameter. Then there is a periodic orbit of period 2 traveling along γ back and forth. Let $p = (\mathbf{p}, \pi/2)$ be the corresponding point on the phase space M. Then the wavefront leaving $\mathbf{p}$ as a focusing point will bounce back and forth between these two reflection points $\mathbf{p}$ and $\mathbf{q}$, and focus at each reflection. If p is a degenerate periodic point for F, then the degeneracy of p persists under normal perturbations.

Proof. Our proof actually works for any period. This general formulation will be used later. Let p be a periodic point such that there is no multiple reflections at its base point s_0, Γ_ϵ be a normal perturbation of Γ at s_0. Then for each $V \in T_pM$, the total effect of $D_pF^n_\epsilon$ on V is a shift of the curvature of the returning wave-front of $D_pF^n(V)$: $\mathcal{B}^+(D_pF^n_\epsilon(V)) = \mathcal{B}^+(D_pF^n(V)) - \frac{2\epsilon}{\sin\theta}$, and a shift of the slope $m(D_pF^n_\epsilon(V)) = m(D_pF^n(V)) - \epsilon$. Therefore, $D_pF^n_\epsilon = \pm\begin{bmatrix} 1 & 0 \\ -\epsilon & 1 \end{bmatrix} \circ D_pF^n$. Then the sign is positive, since Γ_ϵ is a small perturbation of Γ.

In the setting of the above example, we denote $D_pF^2 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then $b = 0$ since the line $\langle \partial_\theta \rangle$ is invariant, and $a = d = 1$, since $a + d = 2$ (degeneracy assumption) and $ad = 1$ (symplectic property). Therefore, $D_pF^2 = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}$ and $D_pF^2_\epsilon = \begin{bmatrix} 1 & 0 \\ c-\epsilon & 1 \end{bmatrix}$. This implies that p is degenerate for any normal perturbation. □

This type of persistence of degeneracy of periodic orbits (with higher periods) may happen for the convex billiards on S^2 and for planar billiards. To overcome this difficulty, we need to consider another type of perturbations, which shift the base point s_0 along the normal direction at $\Gamma(s_0)$. It is very likely that, after the shifting perturbation, the orbit passing through p is not even closed. Luckily for us, such a shift is only needed when the reflection at p is the right angle, and there is no multiple reflections at its base point s_0 within one period of p. In (and only in) this case, the periodic orbit $\mathcal{O}(p)$ stays the same after the shift of Γ along the normal direction at $\Gamma(s_0)$.

3.2. Perturbations of periodic points

Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{R})$, and $A_\epsilon = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix} \circ A$. Then $\mathrm{Tr}(A) = a + d$ and $\mathrm{Tr}(A_\epsilon) = a + d + \epsilon b$. Given a periodic point p of period n, we let $D_pF^n = \begin{bmatrix} a_p & b_p \\ c_p & d_p \end{bmatrix}$, and denote $\tau(p) := \mathrm{Tr}(D_pF^n) = a_p + d_p$.

Note that the dynamics near a hyperbolic periodic point is topologically conjugate to the linearized map D_pF^n (by Hartman–Grobman Theorem) and is well understood. However, the dynamics surrounding the degenerate and elliptic ones are quite complicated, very sensitive to the arithmetic properties of the linearization of F^n at p, and depend on the nonlinear part of F.

Proposition 3.1. *Let $\Gamma \in \Upsilon^r(S^2, g)$, and let p be a periodic point of the billiard map F with zero defect. Suppose p is not hyperbolic. Then there is a C^r small perturbation Γ_ϵ of Γ such that the trace $\tau_\epsilon(p) \neq \tau(p)$.*

In other words, we have the following qualitative descriptions:

(1) if p is degenerate, then after the perturbation, it is either hyperbolic or elliptic;
(2) if p is elliptic, then the rotation number of p can be shifted continuously under the perturbation.

Proof. Let p be a periodic point with period n. Let Γ_ϵ be a C^r small normal perturbation of Γ at $s_0 = p_1(p)$ which increases the curvature at s_0 by ϵ^r. Then we have

$$\tau_\epsilon(p) = \mathrm{Tr}(D_pF_\epsilon^n) = a_p + d_p - \epsilon^r \cdot b_p.$$

If $b_p \neq 0$, then we are done. In the following we assume $b_p = 0$.

If $b_p = 0$, then we have $a_p \cdot d_p = 1$, which implies $|a_p + d_p| \geq 2$. Note that p is assumed to be non-hyperbolic. So we actually have $|a_p + d_p| = 2$, and $D_pF^n = \pm \begin{bmatrix} 1 & 0 \\ c_p & 1 \end{bmatrix}$. Then the line $\langle \partial_\theta \rangle_p$ is fixed by D_pF^n. Equivalently, the corresponding wavefront ρ_p focuses at $s_0 = p_1(p)$, and will focus at s_0 again when it returns after one period. So we only need to show that a small perturbation can destroy the last property (for some point on the orbit) of p.

Case 1. The orbit of p is not symmetric. Then the zero defect property implies that there is no multiple reflection along the orbit $\mathcal{O}(p)$.

Case 1a. $c_p \neq 0$. In this case, $\langle \partial_\theta \rangle_p$ is the only line fixed by D_pF^n, and ρ_p is the only invariant wavefront at p and along the whole orbit of p. Clearly this wavefront does not focus at $s_1 = p_1(Fp)$, since there is no conjugate point on Γ. Therefore $\langle \partial_\theta \rangle_{Fp}$ is not fixed by $D_{Fp}F^n$, since the wavefront corresponding to $\langle \partial_\theta \rangle_{Fp}$ focuses at s_1 (hence is not invariant). This implies $b_{Fp} \neq 0$, and a normal perturbation Γ_ϵ of Γ is performed at s_1. Then $\tau_\epsilon(Fp) = \tau(Fp) - \epsilon^r \cdot b_{Fp}$, and the proposition follows since $\tau(Fp) = \tau(p)$ and $\tau_\epsilon(Fp) = \tau_\epsilon(p)$.

Case 1b. $c_p = 0$. In this case $D_pF^n = \pm I_2$. We first make a normal perturbation at s_0, and get $D_pF_\epsilon^n \neq \pm I_2$. Then we do another perturbation given as in Case 1a.

Case 2. Now we assume that the orbit of p is symmetric. Without loss of generality we assume $n > 2$. Note that there always exist multiple reflections (even though there is no defect). More precisely, there are exactly two simple reflections among the orbit $\mathcal{O}(p)$, and all other reflections happen exactly twice (forward and backward). Moreover, the orbit has perpendicular reflections at these two simple reflections.

Case 2a. There is a wavefront that focuses exactly at those two ends. In this case we make a small shift of Γ along the normal direction at one end, denote the resulting table by $\hat{\Gamma}$, so that the focused wavefront is not invariant any more. Then one automatically gets $\hat{b}_p \neq 0$ and $\tau_\epsilon(p) = \tau(\hat{p}) - \epsilon^r \cdot \hat{b}_p$.

Case 2b. No wavefront focuses exactly at those two ends. In this case we make a normal perturbation at one of the two ends as we did for Case 1. This completes the proof. □

4. Kupka–Smale properties for convex billiards

Let $Q \subset S^2$ be a strictly convex domain with C^r smooth boundary Γ, $M = \Gamma \times (0, \pi)$ be the phase space of the billiard map F induced on Q. For each $n \geq 2$, let $P_n(\Gamma) \subset M$ be the set of points fixed by F^n. In the following we will show that there is an open and dense subset $\mathcal{U}_n$ such that for each $\Gamma \in \mathcal{U}_n$, $P_n(\Gamma)$ is finite and depends continuously on Γ.

Denote by $A = \mathbb{T} \times (0, \pi)$ the open annulus and by $\bar{A} = \mathbb{T} \times [0, \pi]$ the closed annulus. Let $\mathcal{E}(\bar{A})$ be the set of positive twist homeomorphisms on $\bar{A}$ that *fix every point* on the two boundary circles $\mathbb{T} \times \{0, \pi\}$. Let $f \in \mathcal{E}(\bar{A})$, and f_A be the restriction of f on the open annulus A. The following proposition shows that $P_n(f_A)$ cannot accumulate to the boundary $\Gamma \times \{0, \pi\}$.

Proposition 4.1. *Let $n \geq 2$ and $f \in \mathcal{E}(\bar{A})$. There exist a compact set $K \subset A$ and a small neighborhood $\mathcal{W}$ of f in the C^0 topology, such that $P_n(g_A) \subset K$ for each $g \in \mathcal{W}$.*

Remark 4.1. In the previous version of the paper Proposition 4.1 is formulated for convex billiards, and its proof relies heavily on some geometric feature of billiard systems. The current formulation of Proposition 4.1 and its proof were provided to the author in the report of one of referees. It is clear that the current formulation is better and could be useful in other situations. Moreover, the proof is much shorter and easier to follow than the previous version. The contribution from the anonymous referee is kindly acknowledged by the author.

Proof. Let $n \geq 2$ and $f \in \mathcal{E}(\bar{A})$ be given. Let $\tilde{f}$ be the unique lift of f to $\mathbb{R} \times [0, \pi]$ that fixes all the points of $\mathbb{R} \times \{0\}$. There exists a neighborhood $V_0 \supset \mathbb{T} \times \{0\}$ such that for any $z \in V_0$, $0 \leq p_1(\tilde{f}\tilde{z}) - p_1(\tilde{z}) < \frac{1}{3n}$, where $\hat{z}$ is a lift of z and p_1 is the projection from $\mathbb{R} \times [0, \pi]$ to its first coordinate. Then there exists a small neighborhood $\mathcal{V}_0$ of f in $\mathcal{E}(\bar{A})$ such that for each $g \in \mathcal{V}_0$, the lift $\tilde{g}$ satisfies $0 \leq p_1(\tilde{g}\tilde{z}) - p_1(\tilde{z}) < \frac{1}{2n}$ for any lift $\tilde{z}$ of a point $z \in V_0$.

Pick a smaller neighborhood $W_0 \subset V_0$ of $\mathbb{T} \times \{0\}$ whose closure is contained in $\bigcap_{0 \leq k < n} f^{-k} V_0$. Then there exists a smaller neighborhood $\mathcal{W}_0 \subset \mathcal{V}_0$ of f in $\mathcal{E}(\bar{A})$ such that $W_0 \subset \bigcap_{0 \leq k < n} g^{-k} V_0$ for each $g \in \mathcal{W}_0$. Let $g \in \mathcal{W}_0$. Then we have $0 \leq p_1(\tilde{g}^n \tilde{z}) - p_1(\tilde{z}) < \frac{1}{2}$ for each lift $\tilde{z}$ of $z \in W_0$. Therefore, $P_n(g_A) \cap W_0 = \emptyset$ for each $g \in \mathcal{W}_0$.

Similarly we construct W_1 and $\mathcal{W}_1$ for the other boundary component $\mathbb{T} \times \{\pi\}$, and show that $P_n(g_A) \cap W_1 = \emptyset$ for each $g \in \mathcal{W}_1$. Then let $K = A \backslash (W_0 \cup W_1)$, and $\mathcal{W} = \mathcal{W}_0 \cap \mathcal{W}_1$. This completes the proof. □

4.1. *Finiteness of $P_n(\Gamma)$ for most Γ*

A periodic point x is said to be *non-degenerate*, if 1 is not an eigenvalue of $D_x F^{m(x)} : T_x M \to T_x M$, where $m(x)$ is the minimal period of x. The minimal period of a periodic point $x \in P_n(\Gamma)$ satisfies $m(x)|n$, and may be strictly less than n. Then x is said to be *non-degenerate under F^n*, if 1 is not an eigenvalue of $D_x F^n : T_x M \to T_x M$.

Let $\mathcal{U}_n \subset \Upsilon^r(S^2, g)$ be the set of strictly convex domains $\Gamma \in \Upsilon^r(S^2, g)$ such that every periodic point $x \in P_n(\Gamma)$ is non-degenerate under F^n.

Lemma 4.2. *The set $P_n(\Gamma)$ is a finite set for each $\Gamma \in \mathcal{U}_n$, and the map $\Gamma \mapsto P_n(\Gamma)$ is continuous on $\mathcal{U}_n$.*

The proof is omitted since it is a classical application of the local inversion theorem in the study of dynamical systems. Here we use the local uniform compactness of the set $P_n(\Gamma)$ proved in Proposition 4.1.

Note that there is no bifurcation of periodic points in $\mathcal{U}_n$. So we have the following corollary.

Corollary 4.3. *The cardinal map $\Gamma \in \mathcal{U}_n \to |P_n(\Gamma)|$ is locally constant.*

Now we state the first main result of this section.

Proposition 4.4. *Let $2 \leq r < \infty$. Then the set $\mathcal{U}_n$ is an open and dense subset of $\Upsilon^r(S^2, g)$.*

The proof of Proposition 4.4 consists of two parts: the openness and the denseness of $\mathcal{U}_n$. The proof of openness of $\mathcal{U}_n$ is quite standard and follows easily from Lemma 4.2. The proof of the denseness of $\mathcal{U}_n$ is quite long and involved. We first give a direct proof for $n = 2$ to illustrate the idea of the proof. The proof for the general case is given after that.

Now we start to prove the denseness of $\mathcal{U}_2$. First we introduce a useful notation. Given an open interval $I = (a - \epsilon, a + \epsilon)$, the subinterval $I^r = (a - \epsilon^r, a + \epsilon^r)$ will be called the core of I.

Proof of the denseness of $\mathcal{U}_2$. Let $\Gamma \in \Upsilon^r(S^2, g)$ be parameterized by $\mathbb{T} \to S^2$. Given $\epsilon > 0$, pick a sequence of open intervals $I_i = (s_i - \epsilon, s_i + \epsilon)$, $1 \le i \le q_2$ such that the union of their cores $\bigcup_{i=1}^{q_2} I_i^r$ covers $\mathbb{T}$. We pick ϵ small enough such that each geodesic started on Γ with $\theta = \pi/2$ hits each arc I_i no more than once. Then we cover the central line $M_{\pi/2} := \Gamma \times \{\pi/2\} \subset M$ by finitely many open subsets $B_j \subset M$, $1 \le j \le m_2$, such that for each $j \in \{1, \dots, m_2\}$ and each $k = 0, 1$, there exists $i = i(j, k)$ such that $p_1(F^k B_j) \subset I^r_{i(j,k)}$. From now on we fix a set $B_j \subset M$ and the corresponding index $i = i(j, 0)$.

By taking ϵ smaller if necessary, we may assume that there exists a local coordinate map $\phi_i : B(0, 2) \to S^2$ around $\Gamma(I_i)$ such that $\phi_i([-1, 1]) \supset \Gamma(I_i)$, where $B(0, 2) \subset \mathbb{R}^2$ is the 2D disk of radius 2. Clearly ϕ_i does not reflect the curvature of Γ. Given $(s, \alpha) \in \mathbb{R}^2$, let $\Gamma_i(s, \alpha)$ be a C^r-small and C^∞-smooth perturbation of Γ supported on I_i (viewed in the local coordinate system $\phi_i : B(0, 2) \to S^2$) that

a). shifts I_i^r s distance along the geodesic passing through s_i in the direction of $\theta = \pi/4$,
b). then rotates the shifted piece of I_i^r around its center by an angle α,
c). the complement $\Gamma \backslash I_i$ stays unchanged.

Then we use a C^∞ bump function to connect the two pieces I_i^r and $\Gamma \backslash I_i$. Note that the exact number $\theta = \pi/4$ in Step a) is not important, as along as $\theta \ne \pi/2$. Since $\Upsilon^r(S^2, g)$ is open, there exists an open disk $D_i \subset \mathbb{R}^2$ around $(s, \alpha) = (0, 0)$, such that $\Gamma_i(s, \alpha) \in \Upsilon^r(S^2, g)$ and it is (C^r, ϵ)-close to Γ for each $(s, \alpha) \in D_i$. Let $F_{i,s,\alpha}$ be the billiard map induced by $\Gamma_i(s, \alpha)$. This gives rise to a map $\zeta_i : (s, \alpha) \in D_i \to F_{i,s,\alpha}$, and an evaluation map $\zeta_i^{\mathrm{ev}} : D_i \times M \to M$, $(s, \alpha, x) \mapsto F_{i,s,\alpha}(x)$.

Note that $\{F_{i,s,0}(x) : (s, 0) \in D_i\}$ and $\{F_{i,0,\alpha}(x) : (0, \alpha) \in D_i\}$ are two smooth curves passing through $F_{i,0,0}(x) = Fx$. Let γ_s be the geodesic generated by $F_{i,s,0}(x)$, and η_α be the geodesic generated by $F_{i,0,\alpha}(x)$, respectively. Then $\{\gamma_s : (s, 0) \in D_i\}$ and $\{\eta_\alpha : (0, \alpha) \in D_i\}$ are two beams of geodesics surrounding the geodesic generated by Fx. Let $\mathbf{J}_1 = \frac{d}{ds}\Big|_{s=0} \gamma_s$ and $\mathbf{J}_2 = \frac{d}{d\alpha}\Big|_{s=\alpha} \eta_\alpha$ be the corresponding Jacobi fields. It follows from the construction of the perturbations $\Gamma_i(s, \alpha)$ that $\mathbf{J}_1$ and $\mathbf{J}_2$ are two linearly independent solutions of the Jacobi equation. In particular, the two curves $F_{i,\cdot,0}(x)$ and $F_{i,0,\cdot}(x)$ are transverse to each other at Fx. Therefore,

$$D_{(0,0,x)}\zeta_i^{\mathrm{ev}}(\mathbb{R}^2 \times \{0_x\}) = T_{Fx}M. \tag{4.1}$$

Note that $F_{i,s,\alpha} \equiv F$ on the set of points not based on I_i. Therefore, $F^2_{i,s,\alpha}x = F \circ F_{i,s,\alpha}x$ for any x based on I_i and for any $(s, \alpha) \in D_i$. Then let us consider the graph map $\rho_{2,i}$ of ζ_i and the corresponding evaluation map $\rho_{2,i}^{\mathrm{ev}}$, which are given by

$$\rho_{2,i} : D_i \mapsto C^{r-1}(M, M \times M), (s, \alpha) \mapsto (\mathrm{Id}, F^2_{i,s,\alpha}),$$
$$\rho_{2,i}^{\mathrm{ev}} : D_i \times M \to M \times M, (s, \alpha, x) \mapsto (x, F^2_{i,s,\alpha}(x)).$$

Let $\Delta \subset M \times M$ be the diagonal. We need to show that $\rho_{2,i}^{\mathrm{ev}}$ is transverse to Δ along $(0, 0) \times B_j$. The map $\rho_{2,i}^{\mathrm{ev}}$ is certainly transverse to Δ at the places that they do not intersect. In the following we assume that they do intersect, and let $x \in B_j$ be a point such that $(x, F^2x) \in \Delta \cap \rho_{2,i}^{\mathrm{ev}}((0, 0) \times B_j)$. In particular this implies $F^2x = x$. Note that

$$D_{(0,0,x)}\rho_{2,i}^{\mathrm{ev}}(\mathbb{R}^2 \times \{0_x\}) = T_xM \times \{0_x\}, \quad D_{(0,0,x)}\rho_{2,i}^{\mathrm{ev}}(\{(0, 0)\} \times T_xM) = T_{(x,x)}\Delta. \tag{4.2}$$

Clearly $T_xM \times \{0_x\}$ and $T_{(x,x)}\Delta$ span $T_{(x,x)}(M \times M)$. Therefore, the image of $\rho_{2,i}^{\mathrm{ev}}$ is transverse to $\Delta \subset M \times M$ along $(0, 0) \times B_j$.

Now we combine all the pieces together and define a new map

$$\zeta : (s_i, \alpha_i)_{i=1}^{q_2} \in \prod_{1 \le i \le q_2} D_i \mapsto F_{(s_i,\alpha_i)_{i=1}^{q_2}}, \tag{4.3}$$

such that $F_{(0^{2i-2}, s_i, \alpha_i, 0^{2q_2-2i})} = F_{i,s_i,\alpha_i}$. Note that the combined perturbations may be destructively large and even destroy the convexity of Γ. However there does exist a small open neighborhood of $\mathbf{0} = 0^{2q_2}$ in $\prod_{1 \le i \le q_2} D_i$, say D^{2q_2}, such that $F_{(s_i,\alpha_i)_{i=1}^{q_2}}$ is well defined and C^{r-1} close F. Once again, let $\zeta^{\mathrm{ev}} : D^{2q_2} \times M \to M$ be the evaluation map of ζ, let ρ_2 be the map from D^{2q_2} to $C^{r-1}(M, M \times M)$ such that $\rho_2\big((s_i, \alpha_i)_{i=1}^{q_2}\big) = \big(\mathrm{Id}, F^2_{(s_i,\alpha_i)_{i=1}^{q_2}}\big)$. We claim that the evaluation map ρ_2^{ev} is transverse to $\Delta \subset M \times M$ along $\mathbf{0} \times M$. This is clear since for each intersection $(x, F^2x) \in \rho_2^{\mathrm{ev}}(\mathbf{0} \times M) \cap \Delta$, we have $F^2x = x$ and hence $x \in M_{\pi/2}$, that is, an orbit bouncing back and forth between two points on Γ. Then $x \in B_j$ for some $1 \le j \le m_2$. Let $i = i(j, 0)$ be given such that $p_1(B_j) \subset I_i^r$. Then we have

$$D_{(\mathbf{0},x)}(\rho_2^{\mathrm{ev}})(\mathbb{R}^{2q_2} \times \{0_x\}) \supset D_{(0,0,x)}(\rho_{2,i}^{\mathrm{ev}})(\mathbb{R}^2 \times \{0_x\}) = T_xM \times \{0_x\},\ D_{(\mathbf{0},x)}(\rho_2^{\mathrm{ev}})(\mathbf{0} \times T_xM) = T_{(x,x)}\Delta. \quad (4.4)$$

Then there exists an open neighborhood $D \subset D^{q_2}$ of $\mathbf{0}$, such that the combined evaluation map ρ_2^{ev} is transverse to Δ along $D \times M$. Then by Theorem 2.4, there is a dense set subset $E \subset D$ such that for each $(s_i, \alpha_i)_{i=1}^{q_2} \in E$, the map $\big(\mathrm{Id}, F^2_{(s_i,\alpha_i)_{i=1}^{q_2}}\big)$ is transverse to Δ along $M_{\pi/2}$. In other words, $\Gamma_{(s_j,\alpha_j)_{i=1}^{q_2}} \in \mathcal{U}_2$ for each $(s_i, \alpha_i)_{i=1}^{q_2} \in E$, and Γ lies in the closure of $\{\Gamma_{(s_j,\alpha_j)_{i=1}^{q_2}} : (s_j, \alpha_j)_{i=1}^{q_2} \in E\} \subset \mathcal{U}_2$. This shows that $\mathcal{U}_2$ is dense in $\Upsilon^r(S^2, g)$. □

One advantage for the proof of the denseness of $\mathcal{U}_2$ is that $P_2(\Gamma) \subset M_{\pi/2}$ for any $\Gamma \in \Upsilon^r(S^2, g)$. So there is no interference when making perturbations. For periodic orbits of higher periods, there may exist some interference within its own orbit, since there can be some intermediate returns to the same region on Γ (with different directions). So we need to take care of the possible interferences when proving the denseness of $\mathcal{U}_n$ for $n \ge 3$. We will argue by Strong Induction. Suppose that we have demonstrated the C^r-denseness of the open subset $\mathcal{U}_k$ for each $2 \le k < n$. In the following we will prove that the set $\mathcal{U}_n$ is also C^r-dense.

Let $P_n^*(\Gamma)$ be those periodic points in $P_n(\Gamma)$ with minimal period less than n, and $\bar{P}_n(\Gamma)$ be those with period exactly equal n. We deal with these two parts separately. Although a periodic point in $P_k(\Gamma)$ for $\Gamma \in \mathcal{U}_k$ is non-degenerate under F^k, it may be degenerate under F^n. The following lemma reduces the possible interferences from periodic orbits of lower periods.

Lemma 4.5. *Let $k < n$ with $k|n$. Then there is an open and dense subset $\mathcal{U}_{k,n} \subset \mathcal{U}_k$, such that for each $\Gamma \in \mathcal{U}_{k,n}$, all periodic points in $P_k(\Gamma)$ are non-degenerate under F^n.*

Proof. It follows from the definition that $\mathcal{U}_{k,n}$ is open in $\mathcal{U}_k$. So we only need to show the denseness of $\mathcal{U}_{k,n}$ in $\mathcal{U}_k$. Pick $\Gamma \in \mathcal{U}_k \cap \mathcal{S}_k$. Then we perturb one reflection point on each periodic orbit $\mathcal{O}(x)$, say $\Gamma_{\epsilon,x}$ such that the rotation number $\rho_\epsilon(x)$ of that orbit changes (see Lemma 3.1). Note that the new rotation number depends continuously on the size of the perturbation. By choosing $\epsilon(x)$ properly, we can assume the new rotation number is irrational. Note that $|P_k(\Gamma)|$ is locally constant and $P_k(\Gamma)$ varies continuously with respect to $\Gamma \in \mathcal{U}_k$. After a finite steps of perturbations, the new table is in $\mathcal{U}_{k,n}$. □

Proof of the denseness of $\mathcal{U}_n$ for $n \ge 3$. Let $\mathcal{S}_n$ be the open and dense subset given by Proposition 2.3, and $\mathcal{U}_{k,n} \subset \mathcal{U}_k$ be the open and dense subset given by Lemma 4.5 for each $k < n$ and $k|n$. Let $\mathcal{U} = \mathcal{S}_n \cap \bigcap_{k<n:k|n} \mathcal{U}_{k,n}$, which is also open and dense in $\Upsilon^r(S^2, g)$. It suffices to show that $\mathcal{U}_n$ is dense in $\mathcal{U}$. Now let fix $\Gamma \in \mathcal{U}$. We will show that Γ can be approximated by a sequence of $\Gamma_i \in \mathcal{U}_n$. The whole discussion below will be restricted to a small neighborhood $\mathcal{W}$ of Γ given by Lemma 4.1. In particular, let K_n be the uniform compact subset K given there.

It is important to notice that, each periodic point $x \in P_n^*(\Gamma)$ is nondegenerate under F^n (since we choose $\Gamma \in \mathcal{U}$), and is isolated in $P_n(\Gamma)$. So we can pick an open neighborhood $U \supset P_n^*(\Gamma)$, such that $P_n(\hat{\Gamma}) \cap \overline{U} = P_n^*(\hat{\Gamma}) \subset U$ for all $\hat{\Gamma}$ close to Γ. Then the function (Id, $\hat{F}^n$) is already transverse to Δ along $\overline{U}$ for all nearby $\hat{\Gamma}$. Hence we only need to consider the part $K_n \backslash U$.

Let $p_2 : M \to (0, \pi)$ be the projection to the second coordinate. Then $p_2(K_n)$ is a compact subset of $(0, \pi)$. Without loss of generality we assume $p_2(K_n) \subset [2\theta_n, \pi - 2\theta_n]$ for some $\theta_n \in (0, \pi/4)$. The perturbations used later in this proof will be shiftings in the direction of θ_n. Recall that for $n = 2$, $P_2(\Gamma) \subset M_{\pi/2}$ for any Γ, and we chose $\pi/4$ in the proof.

Let $x \in M$, and $\mathcal{O}_n(x) = \{x, Fx, \cdots, F^{n-1}x\}$ be an orbit segment of x of length n. Let $s_n(x)$ be the minimal separation of the set $\{p_1(x), p_1(Fx), \ldots, p_1(F^{n-1}x)\}$ on Γ. For example, $s_n(x) = 0$ if F^ix and F^jx are reflected on the same point on Γ for some $0 \le i < j \le n-1$. Clearly $s_n(x) > 0$ for each $x \in \bar{P}_n(\Gamma)$, since $\Gamma \in \mathcal{S}_n$ and every periodic orbit in $P_n(\Gamma)$ has zero defect. Therefore, $s_n(x)$ can be viewed as a quantitative version of the zero defect phenomenon.

Claim 1. $s_n(\Gamma) := \inf\{s_n(x) : x \in \bar{P}_n(\Gamma)\} > 0$.

Proof of Claim 1. Suppose on the contrary that there exists $x_k \in \bar{P}_n(\Gamma)$ with $s_n(x_k) \to 0$. Passing to a subsequence if necessary, we assume $x_k \to x$, which implies $F^n x = x$ and x is degenerate under F^n. Since every periodic point in $P_n^*(\Gamma)$ is nondegenerate under F^n, we must have $x \in \bar{P}_n(\Gamma)$ with $s_n(x) = 0$. This implies that the orbit of x has positive defect, contradicts the choice of $\Gamma \in \mathcal{S}_n$. □

The next claim follows directly from the fact that $s_n(x)$ depends continuously on x.

Claim 2. *There exists an open neighborhood V_n of $\bar{P}_n(\Gamma)$, such that $s_n(x) \ge s_n(\Gamma)/2$ for any $x \in V_n$.*

Let $s_n(\Gamma)$ be given as above, and $\epsilon \in (0, 1/5)$ be a positive number. Pick a sequence of open intervals $I_i = (s_i - \epsilon \cdot s_n(\Gamma), s_i + \epsilon \cdot s_n(\Gamma))$, $1 \le i \le q_n$, such that the union of their cores $I_i^r = (s_i - \epsilon^r \cdot s_n(\Gamma), s_i + \epsilon^r \cdot s_n(\Gamma))$ covers Γ. Then we can cover $K_n \backslash U$ by much smaller balls $\{B_j : j = 1, \dots, m_n\}$ such that $p_1(F^k B_j) \subset I_i^r$ (for some $i = i(j,k) \in \{1, \cdots, q_n\}$), for each $k \in \{0, \dots, n-1\}$ and for each $j \in \{1, \cdots, m_n\}$. Note that here one cannot require $i(j,k) \ne i(j,0)$ for every $k \in \{1, \cdots, n-1\}$. We will fix a set B_j and the corresponding index $i = i(j,0)$ for a moment and let j and i vary at the final step of the proof.

The perturbations we need here are similar to those we used for proving the denseness of $\mathcal{U}_2$, just here we shift I_i^r in the direction of θ_n, where $\theta_n \in (0, \pi/4)$ is given such that $p_2(K_n) \subset [2\theta_n, \pi - 2\theta_n]$. More precisely, for each $(s, \alpha) \in \mathbb{R}^2$, let $\Gamma_i(s, \alpha)$ be the perturbation supported on I_i that shifts the core part I_i^r along the θ_n direction, and then rotates the shifted I_i^r around its center by an angle α. There is an open neighborhood D_i of $(0, 0)$ such that $\Gamma_i(s, \alpha) \in \Upsilon^r(S^2, g)$ for any $(s, \alpha) \in D_i$. Note that $F_{i,s,\alpha} \equiv F$ on the set of points not based on I_i. Therefore $F_{i,s,\alpha}^k(x) = F^{k-1} \circ F_{i,s,\alpha}(x)$ for any $k \ge 2$ until next reflection of orbit with the segment I_i. Similarly we can define

(1) the evaluation map $\zeta_i^{\text{ev}} : D_i \times M \to M$, $(s, \alpha, x) \mapsto F_{i,s,\alpha}^n(x)$,
(2) the graph map $\rho_{n,i} : D_i \to C^{r-1}(M, M \times M)$, $(s, \alpha) \mapsto (\text{Id}, F_{i,s,\alpha}^n)$, and
(3) the related evaluation map $\rho_{n,i}^{\text{ev}} : D_i \times M \to M \times M$, $(s, \alpha, x) \mapsto (x, F_{i,s,\alpha}^n x)$.

We need to show that $\rho_{n,i}^{\text{ev}}$ is transverse to the diagonal Δ along $(0, 0) \times B_j$. The proof of this part is slightly different from the case $n = 2$, since here we may have intermediate returns: $p_1(F^k B_j) \cap I_i \ne \emptyset$ for some $1 \le k \le n-1$. The map $\rho_{n,i}^{\text{ev}}$ is certainly transverse to Δ at the places that they do not intersect. In the following we assume that they do intersect, and let $x \in B_j$ be a point such that $(x, F^n x) \in \Delta \cap \rho_{n,i}^{\text{ev}}((0, 0) \times B_j)$. In particular this implies $F^n x = x$, and $x \in \bar{P}_n(\Gamma)$. Let $V_{n,j} = V_n \cap B_j$, where V_n is given by Claim 2. Note that $V_{n,j}$ contains a small neighborhood of x in M, and $s_n(y) \ge s_n(\Gamma)/2$ for any $y \in V_{n,j}$. Therefore, $p_1(F^k V_{n,j}) \cap I_i = \emptyset$ for each $1 \le k \le n-1$, where $i = i(j, 0)$ is fixed at the beginning of the proof. This implies that $F_{i,s,\alpha}^n(y) = F^{n-1} \circ F_{i,s,\alpha}(y)$ for any $y \in V_{n,j}$. In this way we exclude the possible interference of intermediate returns. Then the same argument as in the proof of the case $n = 2$ shows that $\rho_{n,i}^{\text{ev}}$ is transverse to the diagonal Δ at each of their intersection points, see Eq. (4.1) and (4.2). Therefore, $\rho_{n,i}^{\text{ev}}$ is transverse to the diagonal Δ along $(0, 0) \times B_j$.

Now we combine all the pieces together and define a new map

$$\zeta : (s_i, \alpha_i)_{i=1}^{q_n} \in \prod_{1 \le i \le q_n} D_i \to F_{(s_i, \alpha_i)_{i=1}^{q_n}}. \tag{4.5}$$

Again there exists a small neighborhood of $\mathbf{0} = 0^{2q_n}$, say D^{2q_n}, such that $F_{(s_i,\alpha_i)_{i=1}^{q_n}}$ is well defined and C^r close to F for each $(s_i, \alpha_i)_{i=1}^{q_n} \in D^{2q_n}$. In the same way we define the combined map $\rho_n : D^{2q_n} \to C^{r-1}(M, M \times M)$ and its evaluation $\rho_n^{\text{ev}} : D^{2q_n} \times M \to M \times M$. Note that any point x in the intersection $(x, F^n x) \in \Delta \cap \rho_n^{\text{ev}}(\mathbf{0} \times M)$ satisfies $F^n x = x$, and hence $x \in P_n(\Gamma) \subset K_n$. Moreover, if $x \in P_n^*(\Gamma)$ has minimal period less than n, then it is already nondegenerate with respect to F^n. So we are left with the case that $x \in \bar{P}_n(\Gamma)$ has minimal period exactly n.

In this case $x \in B_j$ for some j. Then using the same argument as in Eq. (4.4), we see that ρ_n^{ev} is transverse to Δ at $(\mathbf{0}, x)$. Letting x vary, we have that the map ρ_n^{ev} is transverse to Δ along $\mathbf{0} \times K_n$. By the openness property of transverse intersection, there is an open neighborhood $D \subset D^{2q_n}$ of $\mathbf{0}$ such that the map ρ_n^{ev} is transverse to Δ along $D \times K_n$. Then Theorem 2.4 implies that there exists a dense subset of parameters $E \subset D$ such that for each $(s_j, \alpha_j)_{i=1}^{q_n} \in E$, the map $\left(\mathrm{Id}, F^n_{(s_j,\alpha_j)_{i=1}^{q_n}}\right)$ is transverse to the diagonal Δ along K_n. In other words, $\Gamma_{(s_j,\alpha_j)_{i=1}^{q_n}} \in \mathcal{U}_n$ for each $(s_j, \alpha_j)_{i=1}^{q_n} \in E$, and Γ lies in the closure of $\{\Gamma_{(s_j,\alpha_j)_{i=1}^{q_n}} : (s_j, \alpha_j)_{i=1}^{q_n} \in E\} \subset \mathcal{U}_n$. This proves the denseness of $\mathcal{U}_n$ in $\mathcal{U}$ and hence in $\Upsilon^r(S^2, g)$. □

In the previous part of this section, we fix the regularity $r \geq 2$ and use the notation $\mathcal{U}_n$. Now we switch to $\mathcal{U}_n^r$ to indicate the dependence of $\mathcal{U}_n$ on the regularity r. Let $\mathcal{R}^r = \bigcap_{n\geq 2} \mathcal{U}_n^r$. Recall that a periodic point is said to be *elementary*, if it is either hyperbolic, or elliptic with irrational rotation number. Then we have

Theorem 4.6. *There exists a residual subset $\mathcal{R}^r$ of $\Upsilon^r(S^2, g)$, such that for each $\Gamma \in \mathcal{R}^r$, every periodic point of the billiard map induced on Γ is elementary.*

Remark 4.2. The proof of Theorem 4.6 among the abstract space $\mathrm{Diff}^r_\mu(M)$ was given in [46]. Robinson's proof is based on some version of transversality theorem. The proof using Parametric Transversality Theorem was given later in his book [48]. Generally speaking, the transversality result applies if the perturbation space is rich enough. This is the difficult part in the study of dynamical billiards, since the perturbations of the billiard map F can only be made via deformations of the billiard table Γ.

Note that the proof of the denseness of Proposition 4.4 does not apply to the case $r = \infty$ (at least not directly). The dynamical nature guarantees that the genericity holds also in C^∞ category.

Theorem 4.7. *There is a residual subset $\mathcal{R}^\infty \subset \Upsilon^\infty(S^2, g)$, such that for each $\Gamma \in \mathcal{R}^\infty$, every periodic point of the billiard map induced on Γ is elementary.*

Proof. Consider the set $\mathcal{U}_n^\infty = \left(\bigcup_{r\geq n} \mathcal{U}_n^r\right) \cap \Upsilon^\infty(S^2, g)$: this set is open in $\Upsilon^\infty(S^2, g)$ and C^r dense for each $r \geq n$. Therefore $\mathcal{U}_n^\infty$ is open and C^∞ dense in $\Upsilon^\infty(S^2, g)$. Let $\mathcal{R}^\infty = \bigcap_{n\geq 2} \mathcal{U}_n^\infty$. □

Let $\mathcal{V}_n \subset \Upsilon^r(S^2, g)$ be the set of strictly convex domains $Q \subset S^2$ such that

(a). each periodic orbit in $P_n(\Gamma)$ has zero defect;
(b). any two periodic orbits in $P_n(\Gamma)$ have no common reflection points.

The following proof is based on our understanding of the properties of the billiard maps in the open and dense subset $\mathcal{U}_n$ in $\Upsilon^r(S^2, g)$.

Proposition 4.8. *The set $\mathcal{V}_n$ contains an open and dense subset of $\Upsilon^r(S^2, g)$.*

Proof. The denseness follows from Proposition 2.3. It suffices to show the openness of $\mathcal{V}_n$ in $\mathcal{U}_n$. Let $p_1 : M \to \Gamma$ be the projection to the first coordinate, $s_n(\Gamma)$ be the minimal separation between the points in $p_1(P_n(\Gamma)) \subset \Gamma$. Then $s_n(\Gamma) > 0$ for each $\Gamma \in \mathcal{U}_n \cap \mathcal{R}_0$. Pick a small open neighborhood $\mathcal{U} \subset \mathcal{U}_n$ on which $|P_n(\cdot)|$ is constant and $P_n(\cdot)$ varies continuously. Then there exists a smaller neighborhood $\mathcal{V} \subset \mathcal{U}$ of Γ, such that $s_n(\hat{\Gamma}) > 0$ for each $\hat{\Gamma} \in \mathcal{V}_n$. Therefore, $\mathcal{V}_n$ is open in $\mathcal{U}_n$. This completes the proof. □

4.2. Transverse heteroclinic intersections

Given a hyperbolic periodic point p of F on M, the stable manifold of p, $W^s(p)$ consists of points $x \in M$ that $d(F^n x, F^n p) \to 0$ as $n \to \infty$. Similarly we define the unstable manifold $W^u(p)$ of p. Note that both stable and unstable manifolds are immersed curves passing through p. Let $W^{s,u}_{\pm}(p)$ be the branches of $W^{s,u}(p)\backslash\{p\}$. Let $\mathcal{W}_n \subset \Upsilon^r(S^2, g)$ be the set of convex domains $\Gamma \in \Upsilon^r(S^2, g)$, such that for each pair of hyperbolic periodic points $p, q \in P_n(\Gamma)$, either $W^s(p)_\pm \cap W^u_\pm(q) = \emptyset$, or $W^s_\pm(p) \pitchfork_x W^u_\pm(q)$ for some $x \in W^s_\pm(p) \cap W^u_\pm(q)$.

Proposition 4.9. *The set $\mathcal{W}_n$ contains an open and dense subset of $\Upsilon^r(S^2, g)$.*

To prove this result, we need the following perturbation result of Donnay [22].

Lemma 4.10. *Let $\Gamma \in \Upsilon^r(S^2, g)$. For each $i = \pm 1$, let $x_i = F^i x_0$, $c_i : (-\epsilon, \epsilon) \to M$ be a smooth curve with $c_i(0) = x_i$ such that Fc_{-1} does not focus at $s_0 = p_1(x_0)$, and is tangent to $F^{-1}c_1$ at x_0. Then there is a C^r small perturbation of Γ at the base point s_0 such that $\hat{F}c_{-1}$ and $\hat{F}^{-1}c_1$ are transverse at x_0.*

Proof. We consider the normal perturbations $\hat{\Gamma}$ with $\hat{\Gamma}(s_0) = \Gamma(s_0)$, $\hat{\Gamma}'(s_0) = \Gamma'(s_0)$ and $\hat{\kappa}(s_0) = \kappa(s_0) + \epsilon$. If the perturbation is localized at $s_0 = p_1(x_0)$, then one always has $x_i = \hat{F}^i x_0$, and hence $x_0 \in \hat{F}c_{-1} \cap \hat{F}^{-1}c_1$.

The nonfocusing assumption of Fc_{-1} means that $\mathcal{B}^-(DF\dot{c}_{-1}(0)) \neq \infty$, and tangency assumption means that $\mathcal{B}^-(DF\dot{c}_{-1}(0)) = \mathcal{B}^-(DF^{-1}\dot{c}_1(0))$. Suppose $\hat{\kappa}(s_0) \neq \kappa(s_0)$ after the perturbation. First note that $\mathcal{B}^-(D\hat{F}\dot{c}_{-1}(0))$ and $\mathcal{B}^+(D\hat{F}^{-1}\dot{c}_1(0))$ stay unchanged, since these quantities do not depend on the reflection with $\hat{\Gamma}(s_0)$. Then according to the Mirror Formula,

$$\mathcal{B}^+(D\hat{F}\dot{c}_{-1}(0)) = \mathcal{B}^-(DF\dot{c}_{-1}(0)) - \frac{2\hat{\kappa}(s_0)}{\sin\theta_0} = \mathcal{B}^+(DF\dot{c}_{-1}(0)) - \frac{2\epsilon}{\sin\theta_0}.$$

Therefore $m(D\hat{F}\dot{c}_{-1}(0)) = m(D\hat{F}^{-1}\dot{c}_1(0)) - \epsilon$, and the intersection is transverse at x_0. □

Proof of Proposition 4.9. We will show that $\mathcal{W}_n$ contains an open and dense subset of $\mathcal{V}_n$. Pick a small open set $\mathcal{V} \subset \mathcal{V}_n$ on which $|P_n(\cdot)|$ is constant and $P_n(\cdot)$ is continuous. It suffices to show that $\mathcal{W}_n$ contains an open and dense subset in every such $\mathcal{V}$.

We enumerate $P_n(\Gamma)$ as $\{y_i(\Gamma) : 1 \leq i \leq I\}$. Given $1 \leq i, j \leq I$, $\alpha, \beta \in \{+, -\}$, let $\mathcal{W}_{ij\alpha\beta}$ be those $\Gamma \in \mathcal{W}$ such that either $W^s_\alpha(y_i) \cap W^u_\beta(y_j) = \emptyset$, or $W^s_\alpha(y_i) \pitchfork_x W^u_\beta(y_j)$ for some $x \in W^s_\alpha(y_i) \cap W^u_\beta(y_j)$. It suffices to show each $\mathcal{W}_{ij\alpha\beta}$ contains an open and dense subset in $\mathcal{V}$, since $\bigcap\{\mathcal{W}_{ij\alpha\beta} : 1 \leq i, j \leq I, \alpha, \beta \in \{+, -\}\}$ is contained in $\mathcal{W}_n$. In the following we will fix ij and $\alpha\beta$.

Note that there is a simple dichotomy for $\Gamma \in \mathcal{V}$:

(1) either there exist $\Gamma_k \to \Gamma$ such that $W^s_\alpha(y_i(k))$ and $W^u_\beta(y_j(k))$ intersect at some point, say x_k,
(2) or there is a smaller neighborhood of Γ among which $W^s_\alpha(y_i)$ and $W^u_\beta(y_j)$ do not intersect.

It suffices to show the intersections in the first alternative can be perturbed to be transverse. From now on we fix Γ_k such that $W^s_\alpha(y_i(k))$ and $W^u_\beta(y_j(k))$ intersect non-transversely at x_k, and drop the dependence on k safely.

Note that the minimal separation $s_n(\Gamma) > 0$, and the orbit $F^k x$ approximate y_i (or y_j) exponentially fast as $k \to +\infty$ (or $k \to -\infty$, respectively). By taking some iterates of x if necessary, we can assume that there exists an open interval $I \subset \Gamma$ of $s_0 = p_1(x)$ such that all other iterates of x stay out of I. Now we consider the wavefront at x generated by the stable and unstable branches. Note that there is no conjugate point in Q. So no wavefront can focus at x and fx simultaneously. Without loss of generality we assume they do not focus at x. In particular, it implies the stable and unstable branches are not tangent to the direction $\langle\partial_\theta\rangle$ and hence project down to an open interval on Γ, say I. Then we can make a very small perturbation of Γ supported on I, such that $W^s_\alpha(y_i)$ and $W^u_\beta(y_j)$ intersect transversely at x (see Lemma 4.10). Note that transverse intersection, once created, is robust under perturbations. Therefore $\mathcal{W}_{ij\alpha\beta}$ contains an open and dense subset in $\mathcal{V}$. This completes the proof. □

Let $\mathcal{R}^r_{KS} = \bigcap_{n\geq 2} \mathcal{W}_n$, which contains a residual subset of $\Upsilon^r(S^2, g)$.

Theorem 4.11. *There is a residual subset $\mathcal{R}^r_{KS}$ of $\Upsilon^r(S^2, g)$, such that for each $\Gamma \in \mathcal{R}^r_{KS}$,*

(1) *every periodic point of F is elementary;*

(2) *for any two hyperbolic branches $W^s_\alpha(p)$ and $W^u_\beta(q)$,*
  (2a) *either $W^s_\alpha(p) \cap W^u_\beta(p) = \emptyset$,*
  (2b) *or $W^s_\alpha(p) \pitchfork_x W^u_\beta(q)$ for some $x \in W^s_\alpha(p) \cap W^u_\beta(q)$.*

The case $r = \infty$ can be obtained in the same way as we did for Theorem 4.7.

Remark 4.3. The properties of the above theorem resemble the Kupka–Smale properties for convex billiards. However, the above theorem does not claim that $W^s_\alpha(p)$ and $W^u_\beta(q)$ are transverse (see [15]), neither that $W^s_\alpha(p)$ and $W^u_\beta(q)$ have nontrivial intersection. In general, $W^s_\alpha(p)$ and $W^u_\beta(q)$ may be separated by some (KAM) invariant curves, and this separation is persistent under perturbations. In next section we will study the case when $p = q$ and prove the generic existence of homoclinic intersections.

5. Homoclinic intersections for hyperbolic periodic points

In this section we study the existence of homoclinic intersections of hyperbolic periodic points of convex billiards on (S^2, g). Our main result is the following.

Proposition 5.1. *There is an open and dense subset $\mathcal{X}_n \subset \Upsilon^r(S^2, g)$ such that for each $\Gamma \in \mathcal{X}_n$, there exist transverse homoclinic intersections for each hyperbolic periodic point $p \in P_n(\Gamma)$.*

It suffices to show such $\mathcal{X}_n$ is open and dense in $\mathcal{W}_n$ (see Proposition 4.9 for the set $\mathcal{W}_n$). Note that $P_n(\Gamma)$ is finite and depends continuously for $\Gamma \in \mathcal{V}_n$, and the existence of transverse intersections is an open condition. Then $\mathcal{X}_n$ is automatically open in $\mathcal{W}_n$. So it suffices to show the C^r denseness of $\mathcal{X}_n$ in $\mathcal{W}_n$.

Remark 5.1. A simple fact that we will use repeatedly in this section is that $\Upsilon^\infty(S^2, g)$ is C^r dense in $\Upsilon^r(S^2, g)$ for any $r \geq 2$. For example, the perturbations constructed in Sect. 4 are *always* C^∞, although they are *only* C^r*-small*. Therefore, we only need to show that the C^∞ smooth ones in $\mathcal{X}_n$ are already C^r dense in $\mathcal{W}_n$. So in the following all the convex tables will be assumed to be C^∞, and the perturbations will always be C^∞ smooth although they are only C^r small in topology.

Before giving the proof, we need some preparations to cut off the connections between the elliptic periodic points and the hyperbolic periodic points of F.

5.1. *Nonlinear stability of elliptic periodic points*

Let $f \in \mathrm{Diff}^\infty_\mu(M)$ and p be a fixed point of f. An elliptic fixed point is also said to be *linearly stable*. Then a fixed point p is said to be (nonlinearly) *stable*, if there are nesting closed disks $\{D_n\}$ with $p \in D_{n+1} \subset D^o_n$ such that $\bigcap_{n\geq 1} D_n = \{p\}$ and $f|_{\partial D_n}$ is transitive. Note that stable fixed points are isolated from the dynamics, and any invariant rays either coincide with some of those ∂D_n, or are disjoint from ∂D_n.

Moser proved in [36] his Twist Map Theorem, which says that an elliptic fixed point p is stable, if there exists $n \geq 1$ such that the eigenvalue of $D_p f$ satisfies $\lambda^i_p \neq 1$ for each $1 \leq i \leq q$, and $a_j(f^n, p) \neq 0$ for some $1 \leq j \leq [n/2] - 1$, where a_k, $k \geq 1$ are the coefficients of Birkhoff normal form around p. In this case, p is also said to be Moser stable. By perturbing the Birkhoff normal form and then applying Moser twist map theorem, Robinson proved in [46] that generically, each elliptic periodic point is Moser stable.

It is expected that a small perturbation of the billiard table will change the coefficients of Birkhoff normal form around an elliptic periodic point, and turn that point into nonlinearly stable one. However, it is quite difficult (if not impossible) to compute the Birkhoff normal form for convex billiard dynamics on a convex sphere with non-constant curvature, since we do not know too much about the explicit form around an elliptic periodic point, and the dependence of $a_k(f^n, p)$ is quite involved (see [17,11] for the planar case).

In the following we will take a different (simpler) approach to improve the stability of an elliptic periodic points. For an elliptic periodic point p, the rotation number ρ of p is given by the rotation number of projective action

$[D_pF^n]$ on the projective space $\mathbb{P}^1$. Then p is said to have Diophantine rotation number, if ρ is Diophantine. That is, there exist positive numbers c, τ such that

$$\left|\rho - \frac{m}{n}\right| \geq \frac{c}{|n|^{2+\tau}}, \text{ for all rational numbers } \frac{m}{n}. \tag{5.1}$$

The following is the so called Herman's *Last Geometric Theorem*, which states that an elliptic fixed point with Diophantine rotation number is nonlinearly stable [59]. See [24] for the history and a complete proof of Herman's LGT.

Proposition 5.2. *Let $f \in \mathrm{Diff}^\infty_\mu(M)$ and p be an elliptic fixed point of f with rotation number ρ. If ρ is Diophantine, then p is stable.*

See [27] for some applications of Herman's LGT to the study of the stability of Lagrangian equilibrium solutions of circular restricted three body problems.

Proposition 5.3. *There is a C^r-dense subset $\mathcal{D}_n \subset \mathcal{W}_n \cap \Upsilon^\infty(S^2, g)$, such that for each $\Gamma \in \mathcal{D}_n$, all elliptic periodic points in $P_n(\Gamma)$ are stable.*

Proof. Given a convex domain $\Gamma \in \mathcal{W}_n \cap \Upsilon^\infty(S^2, g)$, pick a sufficiently small neighborhood $\mathcal{U} \subset \mathcal{W}_n$ of Γ such that $P_n : \hat{\Gamma} \in \mathcal{U} \mapsto P_n(\hat{\Gamma})$ has the same (finite) cardinality and varies continuously. Note that each periodic point $p \in P_n(\Gamma)$ has zero defect. We make a C^r-small and C^∞-smooth perturbation of Γ around one point p from each elliptic periodic orbit $\mathcal{O}(p)$ in $P_n(\Gamma)$, say the resulting domain $\hat{\Gamma}(\epsilon)$, such that the rotation number ρ_ϵ of p respecting the billiard map on $\hat{\Gamma}(\epsilon)$ is different from the initial rotation number, see Proposition 3.1. Note that the set of Diophantine numbers has full measure on the interval (ρ, ρ_ϵ) Picking a smaller size if necessary, we can assume ρ_ϵ is already Diophantine.

Any two periodic orbits in $P_n(\Gamma)$ have no common reflection points. So the above perturbation can be localized at one reflection point and they have disjoint supports on Γ. In particular the Diophantine rotation numbers of the already perturbed ones are preserved by the subsequent perturbations.

After a finite steps (at most $|P_n(\Gamma)|$) C^r-small and C^∞-smooth of perturbations, we arrive at some $\hat{\Gamma} \in \mathcal{U} \cap \Upsilon^\infty(S^2, g)$ such that $P_n(\hat{\Gamma}) = P_n(\Gamma)$, $\hat{F} = F$ on $P_n(\hat{\Gamma})$ and $\rho(p, \hat{F})$ is Diophantine for each $p \in P_n(\hat{\Gamma})$. Then Proposition 5.2 guarantees that each elliptic periodic point in $P_n(\hat{\Gamma})$ is stable. Such a perturbation $\hat{\Gamma}$ can be made arbitrarily C^r-close to Γ. Therefore, $\mathcal{D}_n$ is C^r-dense in $\mathcal{W}_n$. □

5.2. *Homoclinic intersections*

Now we study the hyperbolic periodic points in $P_n(\Gamma)$. Although each point $x \in P_n(\Gamma)$ is fixed by F^n, the two branches of the stable (and unstable) manifolds x may be switched by F^n. However, F^{2n} does fix each branch of the invariant manifolds of hyperbolic periodic points in $P_n(\Gamma)$. When studying $P_n(\Gamma)$, we actually consider the $2n$-th iteration F^{2n} of those $\Gamma \in \mathcal{D}_{2n}$. For simplicity we denote $f = F^{2n}$.

Let L be a branch of the unstable manifold $W^u(p)\backslash\{p\}$. Then for any $x \in L$, the segment $L[x, fx]$ can be viewed as a fundamental domain of L with respect to $f = F^{2n}$. As $k \to +\infty$, $f^{-k}L[x, fx]$ converges to p, while $f^kL[x, fx]$ may have various limiting behaviors. Denote by $\omega(L)$ the limit set of $f^kL[x, fx]$ as $k \to +\infty$. Similarly we define the ω-set[3] of stable branches (with respect to f^{-k}). There is a dichotomy for the branches of invariant manifolds (see [37]):

- either $\omega(L) \supset L$, or $\omega(L) \cap L = \emptyset$.

A stronger dichotomy was obtained in [58].

Proposition 5.4. *Let $f \in \mathrm{Diff}_\mu(M)$ such that each fixed point is nondegenerate, and each elliptic fixed point is stable. Let L be a branch of invariant manifolds of a hyperbolic fixed point p. Assume $fL = L$. Then*

[3] Technically, one should say the α-set of a stable branch. We use the same notation for stable and unstable branches just to unify the presentation of this paper.

- *either $\omega(L) \supset L$,*
- *or $\omega(L) = \{q\}$ is a singleton, where q is a hyperbolic fixed point.*

The branch L with $\omega(L) = \{q\}$ is called a saddle connection. A saddle connection is said to be a homoclinic (heteroclinic, respectively) connection if $q = p$ ($q \neq p$, respectively).

Proof. We sketch the main idea of the proof. See [58] for details. Let L be a branch of the unstable manifold of p. Suppose $\omega(L) \not\supset L$. Then $\omega(L) \cap L = \emptyset$. Let K be the closure of L, and U be a connected component of $M \backslash K$ attached to L. Let $\hat{U}$ be the prime-end compactification of U, whose boundary consists of finitely many circles. One of the circles, say C_p, contains the prime point $\hat{p}$ of p. The restriction of $\hat{f}$ on C_p is a circle diffeomorphism, and admits $\hat{p}$ as an expanding fixed point. So there is at least one more point on C_p fixed by $\hat{f}$, say $\hat{q}$. Let q be the underlining point of $\hat{q}$ on the closure $\overline{U}$, which must be fixed by f. Such a point cannot be elliptic, since elliptic ones are stable and cannot be approached by invariant curves outside D_n. Then q must be a hyperbolic fixed point, and L forms a branch of the stable manifold of q. Therefore, $\omega(L) = \{q\}$. □

As a corollary, we obtain the following result due to Mather [34]. Our formulation is slightly stronger. See also [58, Corollary 3.4].

Proposition 5.5. *Let $f \in \mathrm{Diff}_\mu(M)$ such that each fixed point of f is either hyperbolic, or elliptic with Diophantine rotation number. Let p be hyperbolic fixed point such that all four branches of $W^{s,u}_{\pm}(p)$ are fixed by f. Then either one of the branch forms a saddle connection, or all four branches have the same closure.*

Proof. Pick a local coordinate system $(U, (x, y))$ around p such that the branches leave p along the two axes. Suppose none of the four branches is a saddle connection. Then each branch is recurrent, and its ω-set contains the branch itself and at least one of the branches adjacent to it. If the ω-set of a branch L does not contain the other adjacent branch, say K, then we have $K \cap \overline{L} = \emptyset$. Consider the component V of $M \backslash \overline{L}$ containing K. Then there exists an smaller open neighborhood $W \subset U$ of p, such that $\partial V \cap W$ consists of the two pieces of the invariant manifolds of p. One of the two pieces is from L, the other piece must be from K_- (the other branch of the invariant manifold containing K). On the other hand, we have $\partial V \cap W \subset L$. Therefore $K_- = L$ forms a homoclinic loop, which contradicts the hypothesis we started with. □

Proposition 5.6. *Let $\Gamma \in \mathcal{D}_{2n}$. Then for each hyperbolic periodic point $x \in P_n(\Gamma)$, there exist transverse homoclinic intersections between each branch of the stable manifold and each branch of the unstable manifold of x.*

The proof mainly use the fact that the (algebraic) intersection number between two simple closed curves on M must be 0. This kind of arguments also appeared in [47,42,37,58].

Proof. Let $\Gamma \in \mathcal{D}_{2n}$, F be the induced billiard map on $M = \Gamma \times (0, \pi)$. Note that there is no saddle connection between any hyperbolic periodic points in $P_{2n}(\Gamma)$ (by the definition of $\mathcal{W}_{2n}$, since $\mathcal{D}_{2n} \subset \mathcal{W}_{2n}$), and each elliptic periodic point in $P_{2n}(\Gamma)$ is stable (by Proposition 5.3, since $\mathcal{D}_{2n} \subset \Upsilon^\infty(S^2, g)$).

Let p be a hyperbolic periodic point in $P_n(\Gamma)$, L be a branch of the unstable manifold of p, and K be a branch of the stable manifold of p. Then both L, K are fixed by F^{2n}, are recurrent, and they have the same closure (by Proposition 5.5). Pick a local coordinate system $(U, (x, y))$ around p such that L leaves p along the positive x-axis, and L approximates p through the first quadrant. Let $S_\epsilon = \{(x, y) \in U : 0 < x, y \leq 1, xy \leq \epsilon\}$, and q be the first moment on L that hits S_ϵ. Let C be the closed curve that starts from p, first travels along L to the point q, and then the segment $\overline{qp}$ from q to p. Then C is a simple closed curve. See Fig. 2.

Since the closure of K contains L, K also intersects S_ϵ. Let $\hat{C}$ be the corresponding simple closed curve by closing the first intersection $\hat{q}$ of K with S_ϵ. Then we see that C and $\hat{C}$ cross each other at p, and the two open segments (p, q) and $(p, \hat{q})$ do not intersect (by the entrance–exit analysis, see [37,58]). Clearly $L(p, q) \cap (p, \hat{q}) = \emptyset$ and $K(p, \hat{q}) \cap (p, q) = \emptyset$.

However, the algebraic intersection number between any two closed curves on M must be 0. So C and $\hat{C}$ have to cross each other at some point beside p, say y, and that intersection must happen between $L(p, q)$ and $K(p, \hat{q})$.

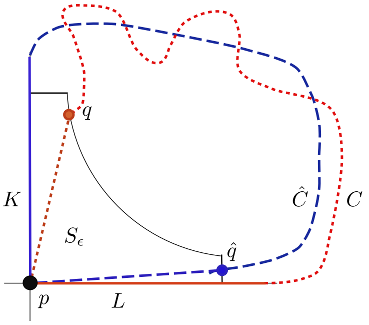

Fig. 2. The closing curves C (red) and $\hat{C}$ (blue) when K leaves along the positive y-axis. The case that K leaves along the negative y-axis is similar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Therefore, there is a homoclinic intersection between K and L. The intersection at y is a topological crossing, but may not be transverse. However, transverse homoclinic intersections do exist, since $\mathcal{D}_{2n} \subset \mathcal{W}_{2n}$. □

Note that no perturbation is needed for the proof of the above proposition.

Proof of Proposition 5.1. As we discussed right after stating Proposition 5.1, $\mathcal{X}_n$ is open in $\mathcal{W}_n$. Let $\mathcal{D}_{2n}$ be the dense subset of $\mathcal{W}_{2n}$ given by Lemma 5.3. Then Proposition 5.6 shows that $\mathcal{D}_{2n} \subset \mathcal{X}_n$. Therefore, $\mathcal{X}_n$ is open and dense in $\Upsilon^r(S^2, g)$. This completes the proof. □

Proof of Theorem 3. Let $\mathcal{R}^r = \bigcap_{n\geq 1} \mathcal{X}_n$. Then $\mathcal{R}^r$ is a residual subset of $\Upsilon^r(S^2, g)$. For each $f \in \mathcal{R}$, and each hyperbolic periodic point p, its stable and unstable manifolds admit some transverse intersections. This completes the proof. □

The case $r = \infty$ can be proved in the same way as we did for Theorem 4.7.

5.3. Positive topological entropy

Corollary 5.7. *There is an open and dense subset $\mathcal{U} \subset \Upsilon^r(S^2, g)$ such that for each $\Gamma \in \mathcal{U}$, the billiard map has transverse homoclinic intersections and positive topological entropy.*

Proof. Let $\mathcal{D}_2$ be the dense subset given in Lemma 5.3. Let $\Gamma \in \mathcal{D}_2$. Then each point $x \in P_2(\Gamma)$ is non-degenerate. Let $W(s_1, s_2) = S(s_0, s_1) + S(s_1, s_2)$ be the action along the 2-periodic configuration (s_k) on Γ. Let (s_k) be an 2-periodic configuration at where W attains its minimum, and x the corresponding periodic point of period 2. Then $D^2W(s_1, s_2)$ is positive definite, and $\mathrm{Tr}(D_x F^2) > 2$ (see Proposition 2.2). So x is hyperbolic. Moreover, each branch of the invariant manifolds of x is fixed by F^2, since both eigenvalues are positive (the double period iterate F^{2n} is *not* needed for minimizers). Then the proof of Proposition 5.6 shows that there exist transverse homoclinic intersections of the stable and unstable branches of x. Transverse intersections are robust. So there exists an open set $\mathcal{U} \supset \mathcal{D}_2$ such that each $\Gamma \in \mathcal{U}$ has transverse homoclinic intersections and positive topological entropy. □

Conflict of interest statement

There is no conflict of interest.

Acknowledgements

The author thanks Leonid Bunimovich, Victor Donnay, Alex Grigo, Saša Kocić, Jeff Xia and Hong-Kun Zhang for many valuable comments and useful discussions. The author is very grateful to the anonymous referees for many useful comments and suggestions, which helped him to improve the presentation of the paper significantly.

Appendix A. Zero defect for generic convex billiards

In this section, we give a proof of Proposition 2.3. Let $\Upsilon^r(S^2, g) \subset C^r(\mathbb{T}, S^2)$ be the set of convex curves. We will use $f : \mathbb{T} \to S^2$ to emphasize the role of f as an embedding function, and use $\Gamma = f(\mathbb{T})$ only for its image. Let $f : \mathbb{T} \to S^2$ be a simple closed curve enclosing a strictly convex domain Q, p be a nonsymmetric periodic point of the billiard map F with period $n = |\mathcal{O}(p)| \geq 3$. Let $F^k p = (s_k, \theta_k)$, and $\{y_1, \dots, y_t\} \subset \Gamma$ be the set of reflections of $\mathcal{O}(p)$ on Γ. Suppose p has positive defect: $d(p) = n - t > 0$, and $(y_{w(1)}, \dots, y_{w(n)})$ be the ordered reflection sequence of $\mathcal{O}(p)$. This gives rise to an onto map $w : \{1, \dots, n\} \mapsto \{1, \dots, t\}$. Such a map w is said to be the pattern of the orbit $\mathcal{O}(p)$. Without loss of generality we assume $w(1) = 1$ and $\{1 \leq j \leq n : w(j) = 1\} = \{j_1, \dots, j_r\}$ (with $j_1 = 1$) for some $r \geq 2$.

Now let $\mathcal{O}(p)$ be a symmetric periodic orbit of period n. Then $n = 2m$ is an even number, and there are exactly two reflections of right angle with Γ. Suppose p has positive defect, and t be the number of distinct reflection points on Γ. Let $w : \{1, \dots, m, m+1\} \to \{1, \dots, t\}$ be the pattern of $\mathcal{O}(p)$ such that $y_{w(1)}$ and $y_{w(m+1)}$ are the two reflection points on Γ with right angle. Note that $w(1) \neq w(m+1)$. We first study the nonsymmetric case in details. The symmetric case need some minor modifications, and will be given at the end of the proof.

Now we generalize above notations to any closed path of type w on S^2. Let $t \geq 3$ be given. Then a map $w : \mathbb{Z} \to \{1, \dots, t\}$ is said to be of period n if $w(k+n) = w(k)$ for all k; is said to be admissible if $w(k) \neq w(k+1)$ for all k. There are only finitely many admissible patterns of period n, and we will fix such a pattern from now. Let $\mathbb{T}^{(t)} \subset \mathbb{T}^t$ be the set of points $(s_1, \dots, s_t)$ with $s_i \neq s_j$ for all $i \neq j$. Then for each $\mathbf{x} \in \mathbb{T}^{(t)}$ and $\mathbf{y} = f^{(t)}(\mathbf{x})$, we have that $\{y_{w(k)}\}$ is a closed path of type w.

Let $\mathbf{y} = (y_1, \dots, y_t)$ be a collection of t distinct points on S^2. Define the perimeter of the geodesic polygon with the ordered corners at $\{y_{w(k)}\}$ as

$$H_w(\mathbf{y}) = \sum_{k=1}^{n} d(y_{w(k)}, y_{w(k+1)}).$$

Similarly, given $f : \mathbb{T} \to S^2$ and $\mathbf{x} \in \mathbb{T}^{(t)}$, let $H_w(f^{(t)}\mathbf{x})$ be the perimeter of the corresponding geodesic polygon with corners $(f(s_i))$ and pattern w.

Let $J^1(\mathbb{T}, S^2)$ be the 1-jet bundle, and $J^1_t(\mathbb{T}, S^2)$ be the t-fold jet bundle. For each $f \in C^r(\mathbb{T}, S^2)$, we have a section map $j_t f : \mathbf{x} \in \mathbb{T}^{(t)} \mapsto (jf(s_1), \dots, jf(s_t))$. Let V_w be the set of those $\tau = (jf_1(s_1), \dots, jf_t(s_t))$ such that

(1) $f_j(s_j) \neq f_i(s_i)$ for each $j \neq i$,
(2) $f_i'(s_i) \neq 0$ for every $i = 1, \dots, t$, and
(3) the polygon generated by $(f_1(s_1), \dots, f_t(s_t))$ is convex with t vertices.

Let $\alpha : J^1(\mathbb{T}, S^2) \to \mathbb{T}$ be the source map, and $\beta : J^1(\mathbb{T}, S^2) \to S^2$ be the target map, $W := (\alpha^t)^{-1}(\mathbb{T}^{(t)}) \cap V_w$. Clearly W is an open submanifold of $J^1_t(\mathbb{T}, S^2)$. Given $\tau \in W$, there are neighborhoods $U_i \subset \mathbb{T}$ of s_i and $V_i \subset S^2$ of $f_i(U_i)$ with $U_i \cap U_j = \emptyset$ and $V_i \cap V_j = \emptyset$ whenever $1 \leq i < j \leq t$, such that

$$\Omega := W \cap \left(\prod_{i=1}^{t} J^1(U_i, V_i) \right)$$

is an open neighborhood of τ. Consider the coordinate map

$$\theta : \Omega \mapsto \prod_{i=1}^{t} U_i \times T_{V_i} S^2 \simeq \prod_{i=1}^{s} U_i \times V_i \times \mathbb{R}^2,$$

with $\theta(\tau) = (\mathbf{u}, \mathbf{v}, A)$, where $\mathbf{u} = (u_1, \dots, u_t)$ is the source of τ, $\mathbf{v} = (v_1, \dots, v_t)$ is the target of τ, and $A = \big(f_1'(u_1), \dots, f_i'(u_i), \dots, f_t'(u_t)\big) = \begin{pmatrix} f'_{1,1}(u_1) & \dots & f'_{i,1}(u_i) & \dots & f'_{t,1}(u_t) \\ f'_{1,2}(u_1) & \dots & f'_{i,1}(u_i) & \dots & f'_{t,2}(u_t) \end{pmatrix}$.

In the following we separate the role of s_1 from s_k, $2 \le k \le t$. For each $l = 1, \dots, r$, let $a = w(j_l - 1)$ and $b = w(j_l + 1)$, and η_a be the tangent direction of the shortest geodesic from y_1 to y_a, and similarly define η_b. Let $\mathbf{t}_{y_1} = f_1'(u_1)$ and $\mathbf{n}_{y_1}$ be the unit tangent and normal directions at y_1. Then we decompose $\eta_a + \eta_b$ as

$$\eta_a + \eta_b = \xi_l(\mathbf{y})\mathbf{t}_{y_1} + \zeta_l(\mathbf{y})\mathbf{n}_{y_1},$$

where $\xi_l(\mathbf{y}) = \langle \eta_a + \eta_b, \mathbf{n}_{y_1} \rangle$ and $\zeta_l(\mathbf{y}) = \langle \eta_a + \eta_b, \mathbf{t}_{y_1} \rangle$. Then it follows from the basic properties of billiard maps that

(1a). $\zeta_l(\mathbf{y}) = 0$ if $(y_{w(t)})_{t=1}^n$ is a periodic orbit;
(1b). $\zeta_l(\mathbf{y}) \ne 0$ if (y_a, y_1, y_b) does not describe a reflection.
(2a). $\partial_{s_k} H \circ f^{(t)}(\mathbf{x}) = 0$ for orbit paths;
(2b). $\partial_{s_k} H \circ f^{(t)}(\mathbf{x})$ may not be zero for non-orbit paths.

Let $\Sigma_w \subset M$ be those $\tau = (jf_1(s_1), \dots, jf_t(s_t)) \in M$ so that $\zeta_l(\mathbf{y}) = 0$ for each $1 \le l \le r$, and $\partial_{s_k} F_w \circ (f_1, \dots, f_t)(\mathbf{x}) = 0$ for each $2 \le k \le t$. We first estimate the codimension of Σ_w. Let $\tau \in \Sigma_w \subset M$ be given, and $\theta : \tau \mapsto (\mathbf{u}, \mathbf{v}, A)$ be the coordinate system around τ given as above. Define a function

$$\mathcal{K} : \theta(\Omega) \to \mathbb{R}^{r+t-1}, \quad \chi \mapsto (\phi_1(\chi), \dots, \phi_r(\chi); \psi_2(\chi), \dots, \psi_t(\chi)),$$

where

(1) $\phi_l : \theta(\Omega) \to \mathbb{R}$, $l = 1, \dots, r$, is defined by

$$\chi = (\mathbf{u}, \mathbf{v}, A) \mapsto \zeta_l(\mathbf{y}) = \langle \eta_a + \eta_b, \mathbf{t}_{y_1} \rangle,$$

where $a = w(j_l - 1)$, $b = w(j_l + 1)$, and $\mathbf{t}_{y_1}$ is the unit tangent direction along $f_1(u_1)$.
(2) $\psi_k : \theta(\Omega) \to \mathbb{R}$, $\chi \mapsto \langle \nabla_{y_k} H, \mathbf{t}_{y_k} \rangle$, for each $k = 2, \dots, t$.

Note that $\mathcal{K}(\tau) = 0$ for each $\tau \in \Sigma_w \cap \Omega$. We claim that $\mathcal{K}$ is a submersion at each point in Ω. The verification of the submersion is pretty simple for convex billiards: by pushing the point y_a along the normal direction of $f_a(s_a)$ (for $a = w(j_l - 1)$, while fixing all other y_k, $k \ne a$), we see that ϕ_l changes linearly (since $\mathbf{t}_{y_1}$ is fixed); by rotating the tangent direction $\mathbf{t}_{y_k}$ of y_k along $f_k(s_k)$ (while fixing all y_k) we see that ψ_k changes linearly (since $\nabla_{y_k} F$ is a fixed nonzero vector); and all these variations are independent.

Therefore, the map $\mathcal{K}$ is a submersion at each point in Ω. So the codimension of Σ_w in $\Omega \subset W$ is at least $\dim(\mathrm{Im}(\mathcal{K})) = r + t - 1 \ge t + 1$, which is larger than $\dim \mathbb{T}^{(t)} = t$. Then by Multi-jet Transversality Theorem, we have that $j_t f \cap \Sigma_w = \emptyset$ for a residual subset of convex tables. Similarly, we define $\Sigma_{w'}$ for any n-periodic admissible pattern $w' : \mathbb{Z} \to \{1, \dots, t\}$, and then for any $t = 2, \dots, n - 1$. This completes the proof for nonsymmetric periodic orbits.

For symmetric periodic orbits, the proof is almost the same. The only difference is that when $a := w(j_l - 1) = w(j_l + 1)$, and the collision from y_a to y_1 is at the right angle. In this case, we still have that ϕ_l changes linearly by pushing y_a along the normal direction of $f_a(s_a)$ (since $\mathbf{t}_{y_1}$ is fixed). Then the rest of the proof is the same. Putting together these results, we get that, for a residual subset of convex tables, each periodic orbits with period n has zero defect. This completes the proof of the genericity of zero defect.

For the second part of Proposition 2.3, we note that in the proof given above, we used the property that each folding of the path at $y_{w(k)}$ is a reflection; but we did not use any property that $\{y_{w(k)}\}$ is on a single orbit. In particular, one can take the union of the two periodic orbits and then study the paths with that joint pattern. Therefore the same analysis applies to the case that two orbits have some common reflection point. Then we conclude that, there is a residual subset of convex tables, for which any two periodic orbits with no common geodesic segment has no common reflection point. However, note that the orbit obtained by the time-reversal of one orbit has exact the same geodesic segments, and this does not count as positive defects.

References

[1] S.B. Angenent, A remark on the topological entropy and invariant circles of an area preserving twistmap, in: Twist Mappings and Their Applications, in: IMA Vol. Math. Appl., vol. 44, 1992, pp. 1–5.

[2] V. Bangert, Mather sets for twist maps and geodesics on tori, in: Dynamics Reported, vol. 1, Wiley, Chichester, 1988, pp. 1–56.

[3] M. Bialy, Convex billiards and a theorem by E. Hopf, Math. Z. 214 (1993) 147–154.

[4] M. Bialy, Hopf rigidity for convex billiards on the hemisphere and hyperbolic plane, Discrete Contin. Dyn. Syst. 33 (2013) 3903–3913.

[5] G. Birkhoff, Dynamical systems with two degrees of freedom, Trans. Am. Math. Soc. 18 (1917) 199–300.

[6] G. Birkhoff, Dynamical Systems, Colloq. Publ. – Am. Math. Soc., vol. IX, Amer. Math. Soc., Providence, RI, 1966; original, 1927.

[7] V. Blumen, K.Y. Kim, J. Nance, V. Zharnitsky, Three-period orbits in billiards on the surfaces of constant curvature, Int. Math. Res. Not. 21 (2012) 5014–5024.

[8] S.V. Bolotin, Integrable billiards on surfaces of constant curvature, Mat. Zametki 51 (1992) 20–28 (in Russian); translation: Math. Notes 51 (1992) 117–123.

[9] L.A. Bunimovich, On ergodic properties of certain billiards, Funct. Anal. Appl. 8 (1974) 254–255.

[10] L.A. Bunimovich, On absolutely focusing mirrors, in: Ergodic Theory and Related Topics, in: Lect. Notes Math., vol. 1514, Springer, Berlin, 1992, pp. 62–82.

[11] L. Bunimovich, A. Grigo, Focusing components in typical chaotic billiards should be absolutely focusing, Commun. Math. Phys. 293 (2010) 127–143.

[12] L. Bunimovich, H-K. Zhang, P. Zhang, On another edge of defocusing: hyperbolicity of asymmetric lemon billiards, Commun. Math. Phys. 341 (2016) 781–803.

[13] K. Burns, M. Gerber, Real analytic Bernoulli geodesic flows on S^2, Ergod. Theory Dyn. Syst. 9 (1989) 27–45.

[14] N. Chernov, R. Markarian, Chaotic Billiards, Math. Surv. Monogr., vol. 127, AMS, Providence, RI, 2006.

[15] G. Contreras-Barandiaran, G. Paternain, Genericity of geodesic flows with positive topological entropy on S^2, J. Differ. Geom. 61 (2002) 1–49.

[16] L. Coutinho dos Santos, S. Pinto-de-Carvalho, Oval billiards on surfaces of constant curvature, preprint, 2014.

[17] M.J. Dias Carneiro, S. Oliffson Kamphorst, S. Pinto-de-Carvalho, Elliptic islands in strictly convex billiards, Ergod. Theory Dyn. Syst. 23 (2003) 799–812.

[18] M.J. Dias Carneiro, S. Oliffson Kamphorst, S. Pinto-de-Carvalho, Periodic orbits of generic oval billiards, Nonlinearity 20 (2007) 2453–2462.

[19] V. Donnay, Geodesic flow on the two-sphere, part I: positive measure entropy, Ergod. Theory Dyn. Syst. 8 (1989) 531–553.

[20] V. Donnay, Geodesic flow on the two-sphere, part II: ergodicity, in: Dynamical Systems, in: Lect. Notes Math., vol. 1342, 1988, pp. 112–153.

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

[22] V. Donnay, Creating transverse homoclinic connections in planar billiards, J. Math. Sci. 128 (2005) 2747–2753.

[23] V. Donnay, Destroying ergodicity in geodesic flows on surfaces, Nonlinearity 19 (2006) 149–169.

[24] B. Fayad, R. Krikorian, Herman's last geometric theorem, Ann. Sci. Éc. Norm. Supér. 42 (2009) 193–219.

[25] C.L. Foden, M.L. Leadbeater, J.H. Burroughes, M. Peper, Quantum magnetic confinement in a curved two-dimensional electron gas, J. Phys. Condens. Matter 6 (1994) L127.

[26] J. Franks, P. Le Calvez, Regions of instability for non-twist maps, Ergod. Theory Dyn. Syst. 23 (2003) 111–141.

[27] Y. Hua, Z. Xia, Stability of elliptic periodic points with an application to Lagrangian equilibrium solutions, Qual. Theory Dyn. Syst. 12 (2013) 243–253.

[28] B. Gutkin, U. Smilansky, E. Gutkin, Hyperbolic billiards on surfaces of constant curvature, Commun. Math. Phys. 208 (1999) 65–90.

[29] V. Ivrii, Second term of the spectral asymptotic expansion of the Laplace–Beltrami operator on manifolds with boundary, Funct. Anal. Appl. 14 (1980) 98–106.

[30] A. Kramli, N. Simanyi, D. Szasz, Dispersing billiards without focal points on surfaces are ergodic, Commun. Math. Phys. 125 (1989) 439–457.

[31] V.F. Lazutkin, Convex Billiards and Eigenfunctions of the Laplace Operator, Leningrad Univ., Leningrad, 1981 (in Russian).

[32] R. Mackay, J. Meiss, Linear stability of periodic orbits in Lagrangian systems, Phys. Lett. A 98 (1983) 92–94.

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

[34] J. Mather, Invariant subsets of area-preserving homeomorphisms of surfaces, Adv. Math. Suppl. Stud. 7B (1981) 531–562.

[35] J. Mather, Topological proofs of some purely topological consequences of Caratheodory's theory of prime ends, in: Th.M. Rassias, G.M. Rassias (Eds.), Selected Studies, 1982, pp. 225–255.

[36] J. Moser, Stable and Random Motions in Dynamical Systems, Ann. Math. Stud., vol. 77, Princeton University Press, Princeton, NJ, 1973.

[37] F. Oliveira, On the generic existence of homoclinic points, Ergod. Theory Dyn. Syst. 7 (1987) 567–595.

[38] F. Oliveira, On the C^∞ genericity of homoclinic orbits, Nonlinearity 13 (2000) 653–662.

[39] V. Petkov, L. Stojanov, Spectrum of the Poincare map for periodic reflecting rays in generic domains, Math. Z. 194 (1987) 505–518.

[40] V. Petkov, L. Stojanov, Periods of multiple reflecting geodesics and inverse spectral results, Am. J. Math. 109 (1987) 619–668.

[41] V. Petkov, L. Stojanov, On the number of periodic reflecting rays in generic domains, Ergod. Theory Dyn. Syst. 8 (1988) 81–91.

[42] D. Pixton, Planar homoclinic points, J. Differ. Equ. 44 (1982) 365–382.

[43] H. Poincare, Les methodes nouvelles de la mecanique celeste (New Methods of Celestial Mechanics), Gauthier-Villars, Paris, 1892 (vol. 1); 1893 (vol. 2); 1899 (vol. 3) (in French).

[44] C. Pugh, The closing lemma, Am. J. Math. 89 (1967) 956–1021.

[45] C. Pugh, C. Robinson, The C^1 closing lemma, including Hamiltonians, Ergod. Theory Dyn. Syst. 3 (1983) 261–313.

[46] C. Robinson, Generic properties of conservative systems, Am. J. Math. 92 (1970) 562–603.

[47] C. Robinson, Closing stable and unstable manifolds in the two-sphere, Proc. Am. Math. Soc. 41 (1973) 299–303.

[48] C. Robinson, Dynamical Systems. Stability, Symbolic Dynamics, and Chaos, Stud. Adv. Math., CRC Press, 1995.

[49] P. Sarnak, Recent progress on the quantum unique ergodicity conjecture, Bull. Am. Math. Soc. 48 (2011) 211–228.

[50] Ya.G. Sinai, Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards, Russ. Math. Surv. 25 (1970) 137–189.

[51] L. Stojanov, Generic properties of periodic reflecting rays, Ergod. Theory Dyn. Syst. 7 (1987) 597–609.

[52] F. Takens, Homoclinic points in conservative systems, Invent. Math. 18 (1972) 267–292.

[53] A. Vetier, Sinai billiard in potential field (construction of stable and unstable fibers), in: Limit Theorems in Probability and Statistics, in: Colloq. Math. Soc. János Bolyai, vol. 36, North-Holland, Amsterdam, 1984, pp. 1079–1146.

[54] D. Visscher, Franks' lemma in geometric contexts, Thesis, Northwestern University, 2012.

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

[56] Z. Xia, Homoclinic points in symplectic and volume-preserving diffeomorphisms, Commun. Math. Phys. 177 (1996) 435–449.

[57] Z. Xia, Homoclinic points for area-preserving surface diffeomorphisms, arXiv:math/0606291.

[58] Z. Xia, P. Zhang, Homoclinic points for convex billiards, Nonlinearity 27 (2014) 1181–1192.

[59] Y. Yoccoz, Travaux de Herman sur les tores invariants (Works of Herman on invariant tori), in: Seminaire Bourbaki, Astérisque 206 (1992) 311–344.