BI-ASYMPTOTIC BILLIARD ORBITS INSIDE PERTURBED ELLIPSOIDS

S. Bolotin[1], **A. Delshams**[2], **Yu. Fedorov**[1] **and R. Ramírez-Ros**[2]

[1]Department of Mathematics and Mechanics (MSU), Moscow, Russia
[2]Departament de Matemàtica Aplicada I (UPC), Barcelona, Spain

1 Introduction

Seventy-five years ago, the *billiard ball problem* was introduced by G. Birkhoff to describe the motion of a billiard ball inside a convex billiard table [1, §VI.6]. Since then, billiards have become paradigmatic models for many questions in dynamical systems. The monographs [9] and [17] present a good overview of the current state-of-the-art in billiard problems.

Integrability is a classical subject in dynamical systems. Integrable billiards seem to be very rare. Indeed, in the famous *Birkhoff's conjecture* it is stated that among all the convex smooth billiard tables, only ellipses are integrable [17, §2.4]. Several attempts have been made to prove this conjecture, but, so far, it remains open.

One of these attempts relies on the phenomenon of the *splitting of separatrices*, discovered a century ago by Henry Poincaré in his celebrated memoir on the three-body problem [14]. In our context, it can be described as follows. The major axis of an ellipse is a hyperbolic 2-periodic trajectory whose stable and unstable invariant curves coincide, giving rise to several bi-asymptotic connections (called *separatrices*) between the two points of the periodic trajectory. Although the hyperbolic periodic trajectory persists under small perturbations of the ellipse, its stable and unstable invariant curves generically do not coincide, but give rise to a complicated web whose existence is an obstruction for the integrability. From a more dynamical point of view, according to the *Birkhoff-Smale theorem* [15], the billiard is *chaotic* if the invariant curves have transverse intersections. Chaotic means that the restriction of the billiard to some invariant Cantor set is conjugated to a transitive topological Markov chain.

Several authors have analyzed the splitting of separatrices in that frame. This approach began with the works of Tabanov [16], Lomelí [11], and Levallois [10]. For instance, Tabanov proved that the stable and unstable invariant curves become transverse under the quartic perturbation

$$x = a\cos\phi \qquad y = b(1+\epsilon\cos^2\phi)\sin\phi$$

for small enough ϵ. This implies that the perturbed billiard is non-integrable and chaotic.

Up to our knowledge, the best result on this problem was obtained by two of the authors (A.D. and R.R.-R.) in [6], where it was shown that the perturbation

$$x = a\cos\phi \qquad y = b(1+\epsilon\eta(\phi))\sin\phi$$

becomes non-integrable for any non-constant entire π-periodic function $\eta : \mathbb{R} \to \mathbb{R}$. Moreover, if the unperturbed ellipse is narrow enough ($b \ll a$), the invariant curves become transverse for any non-constant analytic π-periodic function $\eta : \mathbb{R} \to \mathbb{R}$. The basic tool to prove these results is a discrete version of the Melnikov method.

From a physical point of view, it is natural to consider spatial billiards instead of planar ones. That is, to study the motion of a particle inside regions enclosed by closed convex surfaces of the three-dimensional Euclidean space.

Billiards inside ellipsoids are the only known examples of integrable billiards inside convex smooth surfaces, so there is a spatial version of the Birkhoff conjecture. The major axis of a generic ellipsoid is also a hyperbolic periodic trajectory whose stable and unstable invariant surfaces coincide. These invariant surfaces form a bi-asymptotic set with a richer topology than in the planar case. It is a CW-complex[1] with two zero-dimensional cells (the periodic points), eight one-dimensional cells (called *loops*) and eight two-dimensional cells (called *squares*), see figure 2.

The goal of this lecture is to review some recent results obtained by the authors on billiards inside perturbed ellipsoids. The results have a similar flavor to those above-mentioned on billiards inside perturbed ellipses: non-integrability, splitting of separatrices, chaotic behavior, persistence of bi-asymptotic orbits, etc. Full details and many additional results can be found in [5] and [4].

Billiards inside perturbed ellipsoids are significantly harder than billiards inside perturbed ellipses. Before to tackle their study, it has been necessary to solve some technical problems and to develop new tools. The most important prerequisites of this work are listed below:

- To linearize explicitly the billiard dynamics on the bi-asymptotic set by means of a suitable parameterization. Such a parameterization (called *natural*) was found recently by one of the authors (Yu.F.) in terms of tau-functions [8].
- To obtain some high-dimensional symplectic discrete versions of the *Melnikov method*. This was accomplished independently by Lomelí [12] for twist maps defined on the cotangent bundle of a torus and by two of the authors (A.D. and R.R.-R.) for exact symplectic maps defined on exact symplectic manifolds [7]. For billiards inside perturbed generic ellipsoids, these methods can deal with the squares, but, at a first glance, they could not with the loops.
- To find a way to study the loops. The variational ideas contained in the works of one of the authors (S.B)—see [2] and [3]—have been essential for this point.

We finish this introduction with the organization of the paper. We first need to introduce convex billiards in section 2. Afterwards, in section 3 we present the main properties of billiards inside generic ellipsoids. The set formed by the orbits bi-asymptotic to the diameter inside a generic ellipsoid is studied from a dynamical,

[1] This CW-complex resembles the ones listed in the topological classification of the energy levels of saddle points of four-dimensional integrable Hamiltonians obtained by Lerman and Umanskiĭ [13].

geometrical and topological point of view. Next, a Melnikov method is applied to very general perturbations of generic ellipsoids in section 4. Some results about the splitting of separatrices, non uniform integrability, and chaotic behavior are briefly commented. Section 5 is devoted to the persistence of (not necessarily transverse) bi-asymptotic orbits. These persistence results have nothing to do with Melnikov methods. The last section deals with a degenerate case: the study of prolate ellipsoids. Finally, in the appendix it is stated a theorem on the persistence of bi-asymptotic orbits for twist maps, which is the key point of the persistence results obtained in this work. This result generalizes a previous one by Xia [18], whose proof is only valid when the unperturbed invariant manifolds are *completely doubled* (see the appendix for the definition).

2 Billiards inside convex surfaces

Let Q be a closed convex smooth surface of $\mathbb{R}^3$. A material point moves inside Q and collides elastically with Q. This billiard motion can be modeled by means of a diffeomorphism f defined on a phase space M consisting of positions q on the surface Q and unitary velocities p directed outward Q at q:

$$M = \left\{ m = (q,p) \in Q \times \mathbb{S}^2 : p \text{ is directed outward } Q \text{ at } q \right\}.$$

The *billiard map* $f : M \to M$, $f(q,p) = (q',p')$, is defined as follows:

- The new velocity p' is the reflection of p with respect to the tangent plane T_qQ.
- The new impact point $q' \in Q$ is determined by $p' = (q'-q)/|q'-q|$.

A *billiard orbit* is a bi-infinite sequence $(m_k)_k \in M^{\mathbb{Z}}$ such that $f(m_k) = m_{k+1}$. A *billiard configuration* is a bi-infinite sequence of impact points $(q_k)_k \in Q^{\mathbb{Z}}$ such that $f(q_k,p_k) = (q_{k+1},p_{k+1})$ for $p_{k+1} = (q_{k+1}-q_k)/|q_{k+1}-q_k|$. Billiard orbits and billiard configurations are in one-to-one correspondence.

It is well known that $f : M \to M$ is a twist map with *Lagrangian*

$$l : \{(q,q') \in Q^2 : q \neq q'\} \to \mathbb{R} \qquad l(q,q') = |q-q'|$$

so that the billiard configurations are just the critical points of the (formal) *action*

$$Q^{\mathbb{Z}} \ni (q_k)_k \mapsto \sum_{k \in \mathbb{Z}} l(q_k, q_{k+1}) \in \mathbb{R}.$$

Of course, this series can be divergent, but there are some kinds of orbits (for instance, heteroclinic orbits between hyperbolic periodic orbits) for which it makes sense.

3 Billiards inside generic ellipsoids

We explain here some properties of billiards inside generic ellipsoids following [5]. An ellipsoid is called *generic* when its three axes are different.

Let $f : M \to M$ be the billiard map associated to the generic ellipsoid

$$Q = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \right\} \qquad a > b > c > 0.$$

To begin with, we recall some basic concepts concerning ellipsoids. Let

$$Q(\kappa) = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2 - \kappa^2} + \frac{y^2}{b^2 - \kappa^2} + \frac{z^2}{c^2 - \kappa^2} = 1 \right\}$$

be the family of quadrics *confocal to* the ellipsoid Q. It is clear that $Q(\kappa)$ is an ellipsoid for $0 < \kappa < c$, an one-sheet hyperboloid when $c < \kappa < b$, and a two-sheet hyperboloid if $b < \kappa < a$. No real quadric exists for $\kappa > a$.

For $\kappa \to c^-$ (respectively, $\kappa \to c^+$), the quadric $Q(\kappa)$ flattens into the region of the xy-plane enclosed by (respectively, outside) the *focal ellipse*

$$E = \left\{ q = (x, y, 0) \in \mathbb{R}^3 : \frac{x^2}{a^2 - c^2} + \frac{y^2}{b^2 - c^2} = 1 \right\}.$$

For $\kappa \to b^-$ (respectively, $\kappa \to b^+$), the quadric $Q(\kappa)$ flattens into the region of the xz-plane between (respectively, outside) the branches of the *focal hyperbola*

$$H = \left\{ q = (x, 0, z) \in \mathbb{R}^3 : \frac{x^2}{a^2 - b^2} - \frac{z^2}{b^2 - c^2} = 1 \right\}.$$

Finally, for $\kappa \to a^-$, the quadric flattens into the yz-plane.

We shall use the term *focal conics* when we refer to both E and H. They are represented in figure 1.

The integrability of the billiard map f is closely related to the following property: *any segment (or its prolongation) of a billiard trajectory inside $Q = Q(0)$ is tangent*[2] *to two fixed confocal quadrics $Q(\kappa_1)$ and $Q(\kappa_2)$*, see [17, §2.3]. The quantities κ_1 and κ_2, regarded as functions defined on the phase space M, are first integrals of f.

There is a simpler family of first integrals in involution, namely

$$\begin{aligned}
I_{\mathrm{x}}(m) &= \frac{(\kappa_1^2(m) - a^2)(\kappa_2^2(m) - a^2)}{(a^2 - b^2)(a^2 - c^2)} = u^2 + \frac{(xv - yu)^2}{a^2 - b^2} + \frac{(xw - zu)^2}{a^2 - c^2} \\
I_{\mathrm{y}}(m) &= \frac{(\kappa_1^2(m) - b^2)(\kappa_2^2(m) - b^2)}{(b^2 - a^2)(b^2 - c^2)} = v^2 - \frac{(yu - xv)^2}{a^2 - b^2} + \frac{(yw - zv)^2}{b^2 - c^2} \\
I_{\mathrm{z}}(m) &= \frac{(\kappa_1^2(m) - c^2)(\kappa_2^2(m) - c^2)}{(a^2 - c^2)(b^2 - c^2)} = w^2 - \frac{(zu - xw)^2}{a^2 - c^2} - \frac{(zv - yw)^2}{b^2 - c^2}
\end{aligned}$$

[2] Tangent in a projective sense; that is, the points of tangency can be proper or improper.

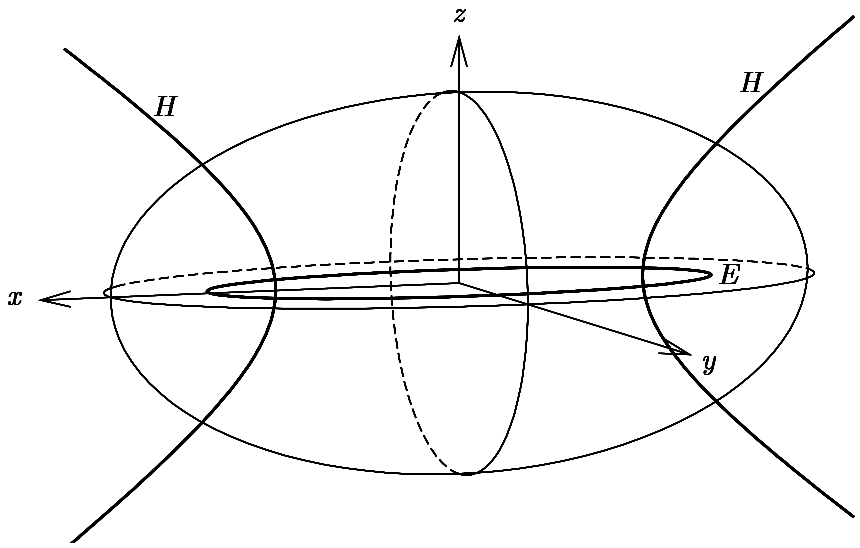

Figure 1: The focal conics of a generic ellipsoid.

where $m = (q, p) \in M$, with $q = (x, y, z) \in Q$ and $p = (u, v, w) \in \mathbb{S}^2$.

These first integrals are dependent: $I_{\rm x}(m)+I_{\rm y}(m)+I_{\rm z}(m) = u^2+v^2+w^2 \equiv 1$, but skipping one of them the rest are independent almost everywhere. Therefore, billiards inside generic ellipsoids are completely integrable. The above formulae can also be used to compute the value of κ_1 and κ_2 at any point of the phase space.

There exist some restrictions in the place where the quantities κ_1 and κ_2 range. We can assume that $\kappa_1 \leq \kappa_2$. Then $\kappa_1 > 0$ and $\kappa_2 \leq a$. On the other hand, a line can not be tangent to two different ellipsoids or to two different hyperboloids of two sheets. Hence, $\kappa_1, \kappa_2 < c$ and $\kappa_1, \kappa_2 > b$ are impossible configurations. There are no more restrictions.

The *diameter* of the ellipsoid is the chord joining the vertices $(-a, 0, 0)$ and $(a, 0, 0)$. It gives rise to a couple of two-periodic points of f, since

$$f(m^{\rm h}_{\pm}) = m^{\rm h}_{\mp} \qquad m^{\rm h}_{\pm} = (q^{\rm h}_{\pm}, p^{\rm h}_{\pm}) = ((\pm a, 0, 0), (\pm 1, 0, 0)).$$

The two-periodic set $P^{\rm h} = \{m^{\rm h}_+, m^{\rm h}_-\}$ is *hyperbolic*: the spectrum of the differential of f at its points does not intersect the unit circumference. In fact, the spectrum has the form $\{\lambda_1, \lambda_2, 1/\lambda_1, 1/\lambda_2\}$ for some $\lambda_1, \lambda_2 > 1$ which are called the *characteristic multipliers* of $P^{\rm h}$, namely

$$\lambda_j = \frac{1+e_j}{1-e_j} \qquad e_1 = \sqrt{1-b^2/a^2} \qquad e_2 = \sqrt{1-c^2/a^2}.$$

Note that e_1 (respectively, e_2) is the *eccentricity* of the elliptic section of the ellipsoid Q with the horizontal plane $\{z = 0\}$ (respectively, the vertical plane $\{y = 0\}$).

Moreover, the *unstable and stable invariant surfaces*

$$\begin{aligned} W^{\mathrm{u}} &:= W^{\mathrm{u}}(P^{\mathrm{h}}) = \left\{ m \in M : \lim_{k \to -\infty} \operatorname{dist}\left(f^k(m), P^{\mathrm{h}} \right) = 0 \right\} \\ W^{\mathrm{s}} &:= W^{\mathrm{s}}(P^{\mathrm{h}}) = \left\{ m \in M : \lim_{k \to +\infty} \operatorname{dist}\left(f^k(m), P^{\mathrm{h}} \right) = 0 \right\} \end{aligned}$$

of P^{h} are *doubled*: $W^{\mathrm{u}} = W^{\mathrm{s}} = W$, where

$$W := \left\{ m \in M : \lim_{|k| \to \infty} \operatorname{dist}\left(f^k(m), P^{\mathrm{h}} \right) = 0 \right\}$$

is the *bi-asymptotic set*. This fact follows from a geometric characterization of these invariant surfaces contained in the following theorem.

Theorem 1 $W^{\mathrm{u}} = \{ m = (q,p) \in M : q + \langle p \rangle$ intersects E and $H \} = W^{\mathrm{s}}$, where $q + \langle p \rangle$ denotes the line passing by q with direction p.

We remark two consequences of the above geometric characterization.

On the one hand, when a segment (or its prolongation) of a billiard trajectory inside a generic ellipsoid intersects both focal conics of the ellipsoid, all the other segments (or their prolongations) also do the same. On the other hand, a billiard trajectory inside a generic ellipsoid is bi-asymptotic to the diameter if and only if all the segments (or their prolongations) of the trajectory intersect both focal conics.

Now, we describe the billiard dynamics on the bi-asymptotic set W. To be more precise, we shall linearize the billiard motion on the invariant surfaces

$$\begin{aligned} W^{\mathrm{u}}_{\pm} &:= W^{\mathrm{u}}(m^{\mathrm{h}}_{\pm}) = \left\{ m \in M : \lim_{k \to -\infty} \operatorname{dist}\left(f^k(m), f^k(m^{\mathrm{h}}_{\pm}) \right) = 0 \right\} \\ W^{\mathrm{s}}_{\pm} &:= W^{\mathrm{u}}(m^{\mathrm{h}}_{\pm}) = \left\{ m \in M : \lim_{k \to +\infty} \operatorname{dist}\left(f^k(m), f^k(m^{\mathrm{h}}_{\pm}) \right) = 0 \right\}. \end{aligned}$$

That is, we shall compute a conjugation between the restrictions $f \colon W^{\mathrm{u}}_{\pm} \to W^{\mathrm{u}}_{\mp}$ (respectively, $f \colon W^{\mathrm{s}}_{\pm} \to W^{\mathrm{s}}_{\mp}$) and the linear map $r \mapsto \Lambda r$ (respectively, $r \mapsto \Lambda^{-1} r$), where the entries of the matrix $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2)$ are the characteristic multipliers.

Such conjugations are called *natural parameterizations*. In our setting, they have a rational character due to the algebraic integrability of billiards inside ellipsoids. In fact, they can be expressed as quotients of the tau-polynomials

$$\begin{aligned} \tau(r_1, r_2) &= 1 + r_1^2 r_2^2 + \alpha^2 (r_1^2 + r_2^2) \\ \tau_{\mathrm{x}}(r_1, r_2) &= 1 + r_1^2 r_2^2 - \alpha^2 (r_1^2 + r_2^2) \\ \tau_{\mathrm{y}}(r_1, r_2) &= 2\alpha r_1 (1 - r_2^2) \\ \tau_{\mathrm{z}}(r_1, r_2) &= 2\alpha r_2 (1 + r_1^2) \end{aligned}$$

where $\alpha^2 = (e_2 + e_1)/(e_2 - e_1)$ with $\alpha > 1$.

To define these natural parameterizations, we introduce the following notations.

- The rational map $\chi : \mathbb{R}^2 \to \mathbb{S}^2$ defined by $\chi = (\tau_{\rm x}/\tau, \tau_{\rm y}/\tau, \tau_z/\tau)$.
- The diagonal matrix $D = \mathrm{diag}(a, b, c)$.
- Given any real s and any map g defined on $\mathbb{R}^2$, we denote by $g \circ \Lambda^s$ the map $r = (r_1, r_2) \mapsto g(\Lambda^s r) = g({\lambda_1}^s r_1, {\lambda_2}^s r_2)$.
- The maps $q : \mathbb{R}^2 \to Q$, $p : \mathbb{R}^2 \to \mathbb{S}^2$, and $m : \mathbb{R}^2 \to M$ defined by $q = D\chi$, $p = \chi \circ \Lambda^{-1/2}$, and $m = (q, p)$.
- The involution $\mathrm{I}(r_1, r_2) = (1/r_1, 1/r_2)$, where $1/0 = \infty$ and $1/\infty = 0$.
- The maps $m^{\rm u,s}_\pm : \mathbb{R}^2 \to M$ defined by $m^{\rm u}_\pm = \pm m$ and $m^{\rm s}_\pm = \pm m \circ \mathrm{I}$.

Theorem 2 The maps $m^{\rm u,s}_\pm : \mathbb{R}^2 \to M$ are natural parameterizations of the invariant surfaces $W^{\rm u,s}_\pm$. That is, $m^{\rm u,s}_\pm$ are analytic diffeomorphisms onto $W^{\rm u,s}_\pm$ such that

$$m^{\rm u,s}_\pm(0) = m^{\rm h}_\pm \qquad f \circ m^{\rm u}_\pm = m^{\rm u}_\mp \circ \Lambda \qquad f \circ m^{\rm s}_\pm = m^{\rm s}_\mp \circ \Lambda^{-1}.$$

Finally, the topology of the bi-asymptotic set is described below.

Theorem 3 The bi-asymptotic set is a two-dimensional CW-complex with two zero-dimensional cells (the periodic points), eight one-dimensional cells (the *loops*) and eight two-dimensional cells (the *squares*).

To describe the cells, we note that the bi-asymptotic set is the disjoint union $W = W^{\rm u}_+ \cup W^{\rm u}_-$, where $W^{\rm u}_\pm = m^{\rm u}_\pm(\mathbb{R}^2)$ and $m^{\rm u}_\pm = \pm m$. Set $\mathbb{R}_- = (-\infty, 0)$, $\mathbb{R}_0 = \{0\}$, and $\mathbb{R}_+ = (0, +\infty)$. Then the eighteen cells are

$$C^{\sigma_1,\sigma_2}_\varsigma = \varsigma m(\mathbb{R}_{\sigma_1} \times \mathbb{R}_{\sigma_2}), \qquad \varsigma \in \{-, +\}, \qquad \sigma_1, \sigma_2 \in \{-, 0, +\}.$$

The eight cells with $\sigma_1, \sigma_2 \in \{-, +\}$ are the squares. The two cells with $\sigma_1 = \sigma_2 = 0$ are the periodic points. The others are the loops. The cells are represented in figure 2. Points and loops with equal labels are identified. All the cells are invariant under the square map f^2. The arrows show the dynamics in the loops. In the squares the dynamics is compatible with the dynamics in the loops. Hence, the points on the loops are *heteroclinic* points of f^2, whereas the points on the squares are *homoclinic*.

There is another difference between loops and squares. The billiard trajectories corresponding to points of the loops are *planar*, that is, they are contained in a plane: the vertical plane $\{y = 0\}$ for the loops with $\sigma_1 = 0$ and the horizontal plane $\{z = 0\}$ for the loops with $\sigma_2 = 0$. This has to do with the fact that the tau-polynomials $\tau_{\rm y}(r)$ and $\tau_{\rm z}(r)$ vanish for $r_1 = 0$ and $r_2 = 0$, respectively.

For further reference, let us introduce the sets S, N, N_1, and N_2 formed by the 8 squares, the 8 loops, the 4 loops with $\sigma_2 = 0$, and the 4 loops with $\sigma_1 = 0$, respectively. Obviously, $N = N_1 \cup N_2$. Following [7], we say that S is the *separatrix* of the unperturbed system, whereas $B := W \setminus S = N \cup P^{\rm h}$ is the *bifurcation set*.

Once presented the unperturbed setup, we can begin the study of billiards inside perturbed ellipsoids. Before exposing the main results, some comments are in order.

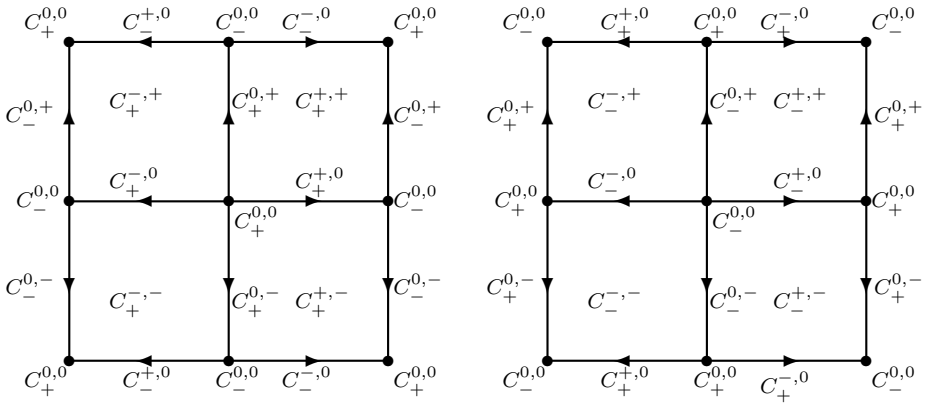

Figure 2: Topological representation of the bi-asymptotic set W.

Firstly, we suppose that the perturbed ellipsoids are at least C^3, although for some results it is required a greater regularity.

Secondly, any Melnikov method can only detect *primary* bi-asymptotic orbits. A perturbed bi-asymptotic orbit is called primary when it depends smoothly on the perturbative parameter ϵ and it is $O(\epsilon)$-close to an unperturbed bi-asymptotic orbit. The orbits obtained in this work are primary.

Finally, there are two kinds of unperturbed bi-asymptotic orbits: heteroclinic on the squares and homoclinic on the loops. The unstable and stable surfaces have a non-degenerate intersection along the loops, that is, at the loops the intersection of the planes tangent to the invariant surfaces coincides with the line tangent to the loop. On the contrary, the invariant surfaces have the same tangent planes at the squares. Thus, loops and squares must be studied separately.

4 The Melnikov potential for perturbed generic ellipsoids

Following [7] and [4], we introduce a function (the *Melnikov potential*) which gives information about the splitting of separatrices in problems like this one.

In the first paper [7], the Melnikov potential was defined only on the separatrix S, since it is generically discontinuous at the bifurcation set $B = W \setminus S$. Nevertheless, it became soon clear that it should be considered as a function defined over the whole bi-asymptotic set, although its restrictions to the set of loops and to the set of squares must be studied separately [4].

Henceforth, $L : W \to \mathbb{R}$ stands for the Melnikov potential (to be defined at the end of next page), whereas $L_S : S \to \mathbb{R}$, $L_N : N \to \mathbb{R}$, $L_{N_1} : N_1 \to \mathbb{R}$, and $L_{N_2} : N_2 \to \mathbb{R}$ denote its restrictions to the sets S, N, N_1, and N_2. Finally, given any cell C of the bi-asymptotic set, $L_C : C \to \mathbb{R}$ is the restriction of the Melnikov potential to C.

In the current frame of billiards inside perturbed generic ellipsoids, the main properties of the Melnikov potential are listed below.

- L1 It is invariant under the unperturbed billiard map: $L \circ f = L$.
- L2 Its restrictions inherit the regularity from the perturbation of the ellipsoid.
- L3 If L_S is not locally constant, then:
  - the separatrix S splits; that is, it cannot be continued smoothly for $\epsilon \neq 0$.
  - the perturbed billiard is not uniformly integrable; that is, the first integrals cannot be continued smoothly for $\epsilon \neq 0$.
- L4 If L_N is not locally constant, then the perturbed unstable and stable invariant surfaces cross topologically.
- L5 Let $C \subset W$ be a cell. If L_C is not constant, C breaks out; that is, it cannot be continued smoothly for $\epsilon \neq 0$.
- L6 The non-degenerate critical points of the restriction L_S give rise to (primary) transverse homoclinic orbits, and hence, to chaotic behavior.
- L7 The non-degenerate critical points of the restriction L_N give rise to (primary) transverse heteroclinic orbits.

In order to find an explicit expression for the Melnikov potential, we need an explicit expression for the perturbations. There are many ways to do that. When one is confronted to the choice, it must be taken into account that the more general is the perturbation, the harder will be the study.

In this section, we have restricted ourselves to global perturbations of the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \epsilon P(y/b, z/c)$$

for some function $P : \mathbb{R}^2 \to \mathbb{R}$ such that $P(0,0) = \partial_1 P(0,0) = \partial_2 P(0,0) = 0$. This means that the diameter of the ellipsoid (and so, the two-periodic set P^{h}) is preserved by the perturbation.

These perturbations are symmetric with regard to the change $x \mapsto -x$, so that not all of the perturbations fit into this frame, but they suffice for our purposes.

We said that the bi-asymptotic set is the disjoint union $W = W^{\mathrm{u}}_- \cup W^{\mathrm{u}}_+$, where $W^{\mathrm{u}}_\pm = m^{\mathrm{u}}_\pm(\mathbb{R}^2)$ and $m^{\mathrm{u}}_\pm = \pm m$ are the parameterizations of theorem 2. Hence, it is natural to define the Melnikov potential $L : W \to \mathbb{R}$ by means of some functions $L_\pm : \mathbb{R}^2 \to \mathbb{R}$ and the relations $L(\pm m(r)) = L_\pm(r)$.

Using results of [5], it can be shown that $L_\pm(r) = a \sum_{k\in\mathbb{Z}} \ell_\pm(\Lambda^k r)$ where

$$\ell_\pm(r) = \frac{\tau^2(r)}{\tau(\Lambda^{-1/2} r) \cdot \tau(\Lambda^{1/2} r)} P\left(\pm \frac{\tau_{\mathrm{y}}(r)}{\tau(r)}, \pm \frac{\tau_{\mathrm{z}}(r)}{\tau(r)} \right).$$

The series $\sum_{k\in\mathbb{Z}} \ell_\pm(\Lambda^k r)$ is absolutely convergent for any $r \in \mathbb{R}^2$, so the Melnikov potential is well defined on the whole bi-asymptotic set.

Nevertheless, as we have already mentioned, the Melnikov potential L can be discontinuous at the bifurcation set $B = W \setminus S = N \cup P^{\mathrm{h}} = N_1 \cup N_2 \cup \{m^{\mathrm{h}}_+, m^{\mathrm{h}}_-\}$. For instance,

$$L(m^{\mathrm{h}}_{\pm}) = 0 \qquad L_{N_j} \equiv a(1/e_j - e_j) \qquad L_S \equiv a(1/e_1 + 1/e_2 - e_1 - e_2).$$

for the quadratic perturbation $P(y/b, z/c) = y^2/b^2 + z^2/c^2$.

For this example, the restrictions L_C are constant for any cell $C \subset W$. Taking into account properties L3–L5, the first step in our study should be to characterize for which perturbations some of these restrictions is not locally constant. The result is summarized in the following theorem.

Theorem 4 Assume that $P : \mathbb{C}^2 \to \mathbb{C}$ is an entire function and let $C \subset W$ be a cell.

- If $C \subset S$, L_C is constant iff $P(y/b, z/c)$ is a polynomial of degree two.
- If $C \subset N_1$, L_C is constant iff $P(y/b, 0)$ is a polynomial of degree two in y.
- If $C \subset N_2$, L_C is constant iff $P(0, z/c)$ is a polynomial of degree two in z.

Proof. We explain briefly the proof of the first claim. The others are simpler.

If $C \subset S$, then $C = C^{\sigma_1,\sigma_2}_{\varsigma}$ for some $\varsigma, \sigma_1, \sigma_2 \in \{-, +\}$. The key point is to realize that the restrictions

$$L_{\varsigma}|(\mathbb{R}_{\sigma_1} \times \mathbb{R}_{\sigma_2}) : \mathbb{R}_{\sigma_1} \times \mathbb{R}_{\sigma_2} \to \mathbb{R}$$

can be analytically extended to the whole complex bi-plane $\mathbb{C}^2$ and to study their complex singularities. Constant functions do not have singularities.

Set $r_* := (1, \mathrm{i})$, where i is the imaginary unit. It can be shown that the complex extensions are analytic at r_* if and only if P is a polynomial of degree two. □

This characterization has important consequences. For instance, the separatrix splits and the perturbed billiard is not uniformly integrable under any non-quadratic entire perturbation of the form here considered, see L3. (For symmetric perturbations, this result was obtained in [5, §5.6].) Similar results can be obtained using L4 and L5. We skip the details.

In [5] it is also analyzed the quartic symmetric perturbation

$$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1 + \epsilon(y^2/b^2)(z^2/c^2).$$

For that quartic perturbation, L7 gives no information since $L_N \equiv 0$. Nevertheless, L6 can be applied and the perturbed billiard turns out to be chaotic for almost all the triples of semi-lengths (a, b, c). It suffices to note that L_S has non-degenerate critical points for an open set of the parameter space $\{(a, b, c) : a > b > c > 0\}$ whose complementary has zero measure. Finally, the number of non-degenerate critical points undergoes infinitely many ciclic bifurcations when $c \to b^-$. See [5, §5.9] for the details.

5 Persistence of bi-asymptotic orbits inside perturbed generic ellipsoids

The following theorem is contained in [4].

Theorem 5 Inside any small enough perturbation of a generic ellipsoid there exist at least sixteen (primary) heteroclinic billiard orbits close to the unperturbed loops.

Proof. The eight loops of the unperturbed billiard map are non-degenerate and the perturbed hyperbolic two-periodic sets have the same action, see the appendix. □

This lower bound is optimal. There are exactly sixteen (primary) heteroclinic billiard orbits for the quartic perturbation

$$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1 + \epsilon(y^4/b^4 + z^4/c^4).$$

We note that, generically, the perturbed heteroclinic trajectories are not contained in a plane, although the unperturbed ones are.

The rest of the section is taken from [5]. It deals with the persistence of some bi-asymptotic orbits when the perturbation preserves the symmetries of the ellipsoid.

A surface in $\mathbb{R}^3$ will be called *symmetric* when it is symmetric with regard to the three coordinate axis of $\mathbb{R}^3$. A billiard orbit inside a symmetric surface will be called *central* (respectively, *axial*) (respectively, *specular*) when its billiard configuration is symmetric with regard to the origin (respectively, to some axis of coordinates) (respectively, to some plane of coordinates). We shall say that an orbit is *symmetric* when it is central, axial or specular.

Inside an ellipsoid there are several kinds of symmetric trajectories bi-asymptotic to the diameter. On the one hand, in the squares of the bi-asymptotic set there are eight *xz-specular* ones and eight *y-axial* ones, which are symmetric with regard to the xz-plane and the y-axis, respectively. The y-axial trajectories are characterized as follows: the prolongation of some of their segments intersects the focal hyperbola at an improper point and the focal ellipse at a vertex of its minor axis. The xz-specular trajectories have an umbilical impact point. In [5] there are some figures to visualize them better. On the other hand, there are sixteen symmetric bi-asymptotic orbits more in the loops. All of them are preserved under symmetric perturbations.

Theorem 6 Inside any small enough symmetric perturbation of a generic ellipsoid there exist at least thirty-two (primary) bi-asymptotic billiard orbits close to the bi-asymptotic set. Sixteen are homoclinic and arise from the squares, and sixteen are heteroclinic and arise from the loops.

We end this section with a couple of remarks. First, the perturbed heteroclinic orbits are always planar under symmetric perturbations (compare with the general case). Second, to obtain the persistence of the 8 xz-specular (respectively, y-axial) homoclinic orbits it would suffice to assume that the perturbed ellipsoid is symmetric with regard to the xz-plane (respectively, y-axis).

6 Persistence of bi-asymptotic orbits inside perturbed prolate ellipsoids

In this section we present some results concerning billiards inside perturbed prolate ellipsoids that can be found in [4]. An ellipsoid is called *prolate* when it is an ellipsoid of revolution around its major axis. It is instructive to compare the prolate case with the generic one already considered.

Let $f : M \to M$ be the billiard map associated to the prolate ellipsoid

$$Q = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2 + z^2}{b^2} = 1 \right\} \qquad a > b > 0.$$

The diameter of a prolate ellipsoid gives also rise to a two-periodic hyperbolic set. The invariant surfaces $W^{\mathrm{u,s}} = W^{\mathrm{u,s}}(P^{\mathrm{h}})$ are defined in the same way that for generic ellipsoids. Since we are confronted with the degeneration $c = b$, the characteristic multipliers of the hyperbolic two-periodic set coincide: $\lambda_1 = \lambda_2 = \lambda$ where

$$\lambda = \frac{1+e}{1-e} \qquad e = \sqrt{1 - b^2/a^2}.$$

Here, e is the eccentricity of the elliptic sections of the prolate ellipsoid with any plane containing the x-axis. All of these elliptic sections have the same set of foci:

$$F = \{(+\gamma, 0, 0), (-\gamma, 0, 0)\} \qquad \gamma = \sqrt{a^2 - b^2}.$$

The set F plays here the same rôle that focal conics played in generic ellipsoids (see Theorem 1), namely $W^{\mathrm{u}} = \{m = (q, p) \in M : q + \langle p \rangle \text{ intersects } F\} = W^{\mathrm{s}}$. Therefore, in this degenerate case the invariant manifolds case are also doubled.

The billiard dynamics on the bi-asymptotic set W can be written as follows. Let $\mathbb{A} = (0, +\infty) \times \mathbb{T}$ be a cilynder. If $q : \mathbb{A} \to Q$ and $p : \mathbb{A} \to \mathbb{S}^2$ are the maps

$$\begin{aligned} q(r, \theta) &= \left(a\frac{1 - r^2}{1 + r^2}, \frac{2br}{1 + r^2} \cos\theta, \frac{2br}{1 + r^2} \sin\theta \right) \\ p(r, \theta) &= \left(\frac{\lambda - r^2}{\lambda + r^2}, \frac{2\lambda^{1/2} r}{\lambda + r^2} \cos\theta, \frac{2\lambda^{1/2} r}{\lambda + r^2} \sin\theta \right) \end{aligned}$$

then $m_\pm = \pm(q, p) : \mathbb{A} \to W^{\mathrm{u}}_\pm \setminus \{m^{\mathrm{h}}_\pm\}$ are analytic diffeomorphisms such that

$$m_\pm(0, \theta) = m^{\mathrm{h}}_\pm = m_\mp(\infty, \theta) \qquad f(m_\pm(r, \theta)) = m_\mp(\lambda r, \theta).$$

The angular variable $\theta \in \mathbb{T}$ does not change under the billiard dynamics due to the continuous symmetry of the prolate ellipsoid around its diameter. This is related to the fact that all of these bi-asymptotic billiard trajectories are *planar*. In fact, they are contained in planes passing by the x-axis. The variable θ labels this pencil of planes.

There is a fundamental difference in the topology of the bi-asymptotic set W. It is a CW-complex with two zero-dimensional cells (the periodic points), and two

two-dimensional cells given by $C_\varsigma = m_\varsigma(\mathbb{A})$ for $\varsigma \in \{-,+\}$. The points on the cells C_- and C_+ are heteroclinic points of f^2, since $m_\pm(0,\theta) \neq m_\pm(\infty,\theta)$.

The invariant surfaces $W^{\rm u}$ and $W^{\rm s}$ are not only doubled, but *completely doubled* (see the appendix for the definition), so that corollary 9 (see again the appendix) can be applied to obtain the following theorem.

Theorem 7 Inside any small enough perturbation of a prolate ellipsoid there exist at least six (primary) heteroclinic billiard orbits close to the separatrix.

We note that, generically, the perturbed heteroclinic trajectories are not contained in a plane, although the unperturbed ones are.

Concerning optimality, there are exactly 8 (primary) heteroclinic billiard orbits (which moreover are transverse) for the quartic perturbation

$$x^2/a^2 + (y^2+z^2)/b^2 = 1 + \epsilon x(y+z)(y^2+z^2).$$

Appendix. A theorem for twist maps

We are going to present a theorem about the persistence of heteroclinic orbits for twist maps. This theorem is the key point of theorems 5 and 7. It can be found in [4], although its statement has been slightly modified to avoid unnecessary technicalities. Before stating this theorem, a description of the related framework is necessary.

Let $f : M \to M$ be a twist[3] diffeomorphism defined on an open set of a cotangent bundle T^*Q. Let us assume that f has two hyperbolic s-periodic sets

$$P^{\rm h}_\pm = \left\{m^{\rm h}_\pm, f(m^{\rm h}_\pm), \ldots, f^{s-1}(m^{\rm h}_\pm)\right\} \qquad f^s(m^{\rm h}_\pm) = m^{\rm h}_\pm$$

whose unstable and stable invariant manifolds

$$\begin{aligned} W_- &:= W^{\rm u}(P^{\rm h}_-) = \left\{m \in M : \lim_{k\to-\infty} \operatorname{dist}\left(f^k(m), P^{\rm h}_-\right) = 0\right\} \\ W_+ &:= W^{\rm s}(P^{\rm h}_+) = \left\{m \in M : \lim_{k\to+\infty} \operatorname{dist}\left(f^k(m), P^{\rm h}_+\right) = 0\right\} \end{aligned}$$

have a *non-degenerate* intersection along an invariant submanifold $N \subset M$, that is,

$$N \subset W_- \cap W_+ \qquad f(N) = N \qquad T_N W_- \cap T_N W_+ = TN.$$

Then N consists of heteroclinic orbits from $P^{\rm h}_-$ to $P^{\rm h}_+$. We look for sufficient conditions for the persistence of some of these heteroclinic orbits under perturbations.

The perturbation must be exact. On the contrary, one can construct very simple perturbations without heteroclinic orbits. For simplicity, we have restricted our study to the frame of twist maps instead of exact maps. Therefore, let $f_\epsilon : M \to M$ be a twist perturbation of f.

[3] There are many almost equivalent definitions of twist maps. See [4] for the one used here.

From the Implicit Function Theorem, we know that f_ϵ has also two perturbed hyperbolic s-periodic sets $P^{\rm h}_{\pm,\epsilon} = P^{\rm h}_{\pm} + {\rm O}(\epsilon)$, for small enough ϵ. We will assume that they have the same action[4]: $A_{f_\epsilon}[P^{\rm h}_{-,\epsilon}] = A_{f_\epsilon}[P^{\rm h}_{+,\epsilon}]$. (This hypothesis always holds in the homoclinic case.)

Now we are ready to state the theorem about the persistence of heteroclinic orbits close to the unperturbed heteroclinic connection N.

Theorem 8 For small enough ϵ, f_ϵ has at least ${\rm cat}(N/f^s)$ primary heteroclinic orbits from $P^{\rm h}_{-,\epsilon}$ to $P^{\rm h}_{+,\epsilon}$ close to N, provided that:

C1 $N \cup P^{\rm h}_- \cup P^{\rm h}_+$ is compact.

C2 Given any $\delta > 0$, there exists $j = j(\delta) > 0$ such that for all $k \in \mathbb{Z}$ such that $|k| > j$ and and for all $m \in N$ such that ${\rm dist}(m, P^{\rm h}_- \cup P^{\rm h}_+) > \delta$, it follows that

$$ {\rm dist}\left(f^k(m), P^{\rm h}_+ \cup P^{\rm h}_-\right) < \delta. $$

Conditions C1 and C2 deserve some remarks. For instance, one can ask whether they are necessary or can be weakened. We do not know that, but they are essential in our proof, which, roughly speaking, goes as follows.

The heteroclinic orbits of f_ϵ are the critical points of a functional S_ϵ defined on a Hilbert manifold X. Using that the invariant manifolds W_- and W_+ have a non-degenerate intersection along N, it can be deduced that S_0 has a finite-dimensional non-degenerate critical manifold[5] $Z \subset X$. The conditions of theorem 8 are used to check that certain quotient manifold $K = Z/t$ of this critical manifold is *compact*. The proof is finished with a standard Lyapunov-Schmidt reduction over K and the Lusternik-Schnirelmann category of $K \simeq N/f^s$. Thus, compactness of K—and so, conditions C1 and C2—seems unavoidable in our scheme of proof.

Once accepted this fact, it is useful to find some cases in which the conditions of theorem 8 always hold. A couple of simple cases is presented below.

As a first example, we consider the completely doubled case. The manifolds $W_- = W^{\rm u}(P^{\rm h}_-)$ and $W_+ = W^{\rm s}(P^{\rm h}_+)$ are said *doubled* if $W_- \setminus P^{\rm h}_- = W_+ \setminus P^{\rm h}_+$. They are said *completely doubled* if they are doubled and, in addition, $T_N W_- = T_N W_+$, for $N := W_- \setminus P^{\rm h}_- = W_+ \setminus P^{\rm h}_+$.

In the completely doubled case, it can be checked that N verifies the conditions of theorem 8 with ${\rm cat}(N/f^s) = 3s$. Hence, we have obtained the following corollary.

Corollary 9 (Xia) *If the unperturbed invariant manifolds are completely doubled and the perturbed s-periodic sets have the same action, the perturbed map has at least $3s$ primary heteroclinic orbits close to N.*

[4] The *action* of a s-periodic set P of a twist map f with Lagrangian l is $A_f[P] = \sum_{k=0}^{s-1} l(q_k, q_{k+1})$, where q_k is the canonical projection of $f^k(m)$ onto the configuration space Q and m is any point in P.

[5] Z is a *non-degenerate critical manifold* of a function $S : X \to \mathbb{R}$ if, for any $x \in Z$, $S'(x) = 0$, the operator $S'' : T_x X \to T_x^* X$ has a closed range and $T_x Z = \ker S''$.

Xia [18] studied this problem in the frame of exact maps, but he missed the condition $T_N W_- = T_N W_+$. To be more precise, he stated the persistence result for the (general) doubled case, but his proof works only in the completely doubled case.

As a second example, we consider heteroclinic orbits coming from unperturbed loops. A curve $C \subset (W_- \setminus P_-^{\mathrm{h}}) \cap (W_+ \setminus P_+^{\mathrm{h}})$ from a point in P_-^{h} to another point in P_+^{h} is a *non-degenerate loop* when

$$\dim(T_m W_- \cap T_m W_+) = 1 \qquad \forall m \in C.$$

It turns out that the set N of non-degenerate loops verifies the conditions of theorem 8 with $\mathrm{cat}(N/f^s) = 2n$, where n is the number of non-degenerate loops.

Corollary 10 If the unperturbed map has n non-degenerate loops and the perturbed periodic sets have the same action, then the perturbed map has at least $2n$ primary heteroclinic orbits close to the loops.

On the other hand, the conditions of theorem 8 fail when N is, for instance, a set formed by *squares* similar to the ones of the billiard map inside a generic ellipsoid. Both conditions fail at the borders of these squares.

Acknowledgments. This work has been partially supported by the INTAS grant 00-221. One of the authors (S. B.) has also been partially supported by the RFBR grant 99-01-00953. Two of the authors (A. D. and R. R.-R.) have also been partially supported by the Spanish grant DGICYT BFM2000-0805 and the Catalan grant CIRIT 2000SGR–00027. This paper was finished while A. D. was a visitor of the *Centre de Recerca Matemàtica*, for whose hospitality he is very grateful.

References

[1] G. Birkhoff, *Dynamical Systems*, Amer. Math. Soc., Providence, 1927. 1
[2] S. Bolotin, *Regul. Chaotic Dyn.*, 2000, **5**(2), 139–156. 2
[3] S. Bolotin, *Nonlinearity*, 2001, **14**(5), 1123–1140. 2
[4] S. Bolotin, A. Delshams, and R. Ramírez-Ros, in progress. 2, 8, 11, 12, 13
[5] A. Delshams, Yu. Fedorov, and R. Ramírez-Ros, *Nonlinearity*, 2001, **14**(5), 1141–1195. 2, 4, 9, 10, 11
[6] A. Delshams, and R. Ramírez-Ros, *Nonlinearity*, 1996, **9**(1), 1–26. 1
[7] A. Delshams, and R. Ramírez-Ros, *Comm. Math. Phys.*, 1997, **190**(1), 213–245. 2, 7, 8
[8] Yu. Fedorov, *Acta Appl. Math.*, 1999, **55**(3), 251–301. 2
[9] V. Kozlov, and D. Treshchëv, *Billiards: a Genetic Introduction to the Dynamics of Systems with Impacts*, Amer. Math. Soc., Providence, 1991. 1
[10] P. Levallois, *Ergodic Theory Dynam. Systems*, 1997, **17**(2), 435–444. 1
[11] H. Lomelí, *Physica D*, 1996, **99**(1), 59–80. 1
[12] H. Lomelí, *Ergodic Theory Dynam. Systems*, 1997 **17**(2), 445–462. 2
[13] L. Lerman, and Ya. Umanskiĭ, *Russian Acad. Sci. Sb. Math.*, 1994, **78**(2), 479–506. 2
[14] H. Poincaré, *Acta Math.*, 1890, **13**, 1–271. 1
[15] S. Smale, *Differential and Combinatorial Topology*, S.S. Cairns (Ed.), Princeton, 1965, 63–80. 1
[16] M. Tabanov, *Chaos*, 1994, **4**(4), 595–606. 1
[17] S. Tabachnikov, *Billiards*, Panor. Synth., Paris, 1995. 1, 4
[18] Z. Xia, *Discrete Contin. Dynam. Systems*, 2000, **6**(1), 243–253. 3, 15