Nonlinearity **14** (2001) 1141–1195 PII: S0951-7715(01)19371-0

Homoclinic billiard orbits inside symmetrically perturbed ellipsoids

Amadeu Delshams[1], **Yuri Fedorov**[2] **and Rafael Ramírez-Ros**[1]

[1] Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain
[2] Department of Mathematics and Mechanics, Moscow State University, Moscow 119899, Vorob'eby Gory, Russia

E-mail: amadeu@ma1.upc.es, fedorov@mech.math.msu.su and rafael@vilma.upc.es

Received 27 November 2000, in final form 18 May 2001
Published 30 July 2001
Online at stacks.iop.org/Non/14/1141

Recommended by L Bunimovich

Abstract
The billiard motion inside an ellipsoid of $\mathbb{R}^3$ is completely integrable. If the ellipsoid is not of revolution, there are many orbits bi-asymptotic to its major axis. The set of bi-asymptotic orbits is described from a geometrical, dynamical and topological point of view. It contains eight surfaces, called separatrices.

The splitting of the separatrices under symmetric perturbations of the ellipsoid is studied using a symplectic discrete version of the Poincaré–Melnikov method, with special emphasis on the following situations: close to the flat limit (when the minor axis of the ellipsoid is small enough), close to the oblate limit (when the ellipsoid is close to an ellipsoid of revolution around its minor axis) and close to the prolate limit (when the ellipsoid is close to an ellipsoid of revolution around its major axis).

It is proved that any non-quadratic entire symmetric perturbation breaks the integrability and splits the separatrices, although (at least) 16 symmetric homoclinic orbits persist. Close to the flat limit, these orbits become transverse under very general polynomial perturbations of the ellipsoid.

Finally, a particular quartic symmetric perturbation is analysed in great detail. Close to the flat and to the oblate limits, the 16 symmetric homoclinic orbits are the unique primary homoclinic orbits. Close to the prolate limit, the number of primary homoclinic orbits undergoes infinitely many bifurcations. The first bifurcation curves are computed numerically.

The planar and high-dimensional cases are also discussed.

Mathematics Subject Classification: 37J15, 37J20, 37J30, 37J35, 37J40, 37J45, 37N05

0951-7715/01/051141+55$30.00

Contents

1. Introduction and main results

1.1. Billiard maps inside ellipsoids

Billiard maps inside convex sets were introduced by Birkhoff [Bir27], and are possibly the most natural and genuine twist maps. Among them, the billiard map inside an ellipsoid (also called the elliptic billiard map) is the most famous integrable map. Its integrability is closely related to the remarkable fact that a billiard trajectory inside an ellipsoid Q in $\mathbb{R}^{n+1}$ is tangent to n fixed confocal quadrics (see, e.g., [KT91, Tab95]).

However, action-angle coordinates cannot be introduced over the whole phase space of the billiard map f inside an ellipsoid Q, due to the existence of several classes of bi-asymptotic motions. In the present paper we study in detail the bi-asymptotic motions to the diameter, i.e. the major axis.

For generic ellipsoids, those with three different axes, the trajectory along the diameter is generated by a hyperbolic two-periodic orbit $\{m^{\mathrm{h}}_{+}, m^{\mathrm{h}}_{-}\}$. Using geometrical considerations (propositions 4.1, 5.1 and 6.1), it is not difficult to see that this hyperbolic two-periodic orbit possesses an n-dimensional set $\mathcal{W}$ of bi-asymptotic motions, formed both by its unstable and stable invariant manifolds, which are doubled, i.e. they coincide: $\mathcal{W} = \mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}}$.

To provide a complete description of the dynamics on the *bi-asymptotic set* $\mathcal{W}$, we present in this paper *natural parametrizations* for the invariant manifolds $\mathcal{W}^{\mathrm{u,s}}_{\pm}$ associated with the fixed points $m^{\mathrm{h}}_{\pm}$ of f^2, the square of the billiard map f inside the ellipsoid. These are analytic diffeomorphisms $m^{\mathrm{u,s}}_{\pm} : \mathbb{R}^n \to \mathcal{W}^{\mathrm{u,s}}_{\pm}$ that conjugate f^2 on $\mathcal{W}^{\mathrm{u,s}}_{\pm}$ to a diagonal linear map with the (real) characteristic multipliers of $m^{\mathrm{h}}_{\pm}$ at their diagonal entries (lemmas 4.1, 5.1 and 6.1). Although the generic billiard trajectories inside an ellipsoid can be obtained in terms of theta functions, it is worth noting that the natural parametrizations $m^{\mathrm{u,s}}_{\pm}$ presented here are expressed as quotients of tau functions which are simply polynomial functions in the parameters we use.

It is very important to note that the dimension of the vectorial subspace of $\mathbb{R}^{n+1}$ generated by each billiard trajectory is *not* the same for all billiard trajectories on this bi-asymptotic set $\mathcal{W}$, but follows a hierarchy. Moreover, the asymptotic behaviour of a bi-asymptotic trajectory depends on the parity of its dimension. Thus, for even l, $1 \leqslant l \leqslant n$, the set of bi-asymptotic trajectories of dimension $l + 1$ consists of homoclinic orbits for f^2 to the fixed points $m^{\mathrm{h}}_{\pm}$ of f^2, whereas for odd l, it consists of heteroclinic orbits for f^2 between m^{h}_{+} and m^{h}_{-}.

In this paper, we call the *separatrix* the set $\mathcal{S}$ formed by the bi-asymptotic trajectories of the largest dimension $n + 1$, and the *bifurcation set* its complementary set $\mathcal{B} = \mathcal{W} \setminus \mathcal{S}$ which is formed by the bi-asymptotic trajectories of dimension $l + 1 < n + 1$, $1 \leqslant l \leqslant n - 1$.

In terms of the above-mentioned natural parametrizations, it turns out that any point m in the separatrix $\mathcal{S}$ is parametrized by values $r \in \mathbb{R}^n$ outside the n coordinate hyperplanes $r_j = 0$ of $\mathbb{R}^n$, $1 \leqslant j \leqslant n$ and, consequently (propositions 4.2, 5.2 and 6.2), the separatrix $\mathcal{S}$ has 2^{n+1} diffeomorphic connected components. In contrast, $m = m^{\mathrm{u}}_{\pm}(r)$, with $r_j = 0$ for some $j = 1, \ldots, n$, for any point m in the bifurcation set $\mathcal{B}$.

The topology of both the separatrix and the bifurcation set is described in sections 4.1, 5.3 and 6, and is depicted in figure 5 for the planar case, and in figure 8 for the spatial case. It is worth noting that, in the spatial case, the bi-asymptotic set falls into the last case of the classification of bi-asymptotic sets of saddle points of four-dimensional integrable Hamiltonian systems carried out by Lerman and Umanskiĭ [LU94]. (Incidentally, our 'bifurcation set' is denoted as 'garland' in [LU94].)

1.2. Symmetrically perturbed ellipsoids

Integrable billiard maps seem to be very rare. Indeed, there is a famous conjecture (due to Birkhoff, at least for $n = 1$) that states that among all the billiard maps inside convex hypersurfaces $\mathcal{Q}$ in $\mathbb{R}^{n+1}$, the only ones that are integrable occur when $\mathcal{Q}$ is an ellipsoid.

The main goal of the present paper is to study the break-up of the bi-asymptotic set $\mathcal{W}$, and more precisely of the separatrix $\mathcal{S}$, for the billiard map inside a symmetric perturbation $\mathcal{Q}_\varepsilon$ of an ellipsoid $\mathcal{Q} \in \mathbb{R}^{n+1}$.

As a general rule, only some of the bi-asymptotic motions in $\mathcal{W}$ persist under perturbations. In the present paper, we restrict ourselves to *symmetric* perturbations, that is, to hypersurfaces $\mathcal{Q}_\varepsilon \in \mathbb{R}^{n+1}$ that are symmetric with regard to all the coordinate axes of the Euclidean space $\mathbb{R}^{n+1}$, to ensure the preservation of the symmetric bi-asymptotic billiard orbits. By a symmetric billiard orbit we mean an orbit such that its billiard configuration is symmetric with regard to some coordinate subspace of $\mathbb{R}^{n+1}$.

Using symmetry arguments, we see that there are $\binom{n}{l}2^{l+2}$ symmetric bi-asymptotic billiard orbits of dimension $l+1$, $l = 1, \ldots, n$, in the billiard map inside a generic ellipsoid $\mathcal{Q}$ in $\mathbb{R}^{n+1}$, and that all of them persist under symmetric perturbations (theorems 4.1, 5.1 and 6.1). This gives a total number of $4(3^n - 1)$ symmetric bi-asymptotic billiard orbits.

To check whether these symmetric bi-asymptotic orbits are the only ones preserved under the symmetric perturbation, we use the *Melnikov potential* $L : \mathcal{S} \to \mathbb{R}$ which is a well defined and smooth function on the unperturbed separatrix (see section 2.2). It is at this moment that the existence of 'nice' natural parametrizations for each connected component of the separatrix becomes a crucial tool for the explicit representation of the Melnikov potential as an absolutely convergent series (lemmas 4.2, 5.2 and 6.2).

As a first application of the Melnikov potential, it is proved (theorems 5.2 and 6.2) that non-trivial entire symmetric perturbed billiards are non-uniformly integrable. More precisely, we first note that every symmetric perturbation $\mathcal{Q}_\varepsilon$ of the ellipsoid $\mathcal{Q} = \{q \in \mathbb{R}^{n+1} : \langle q, D^{-2}q\rangle = 1\}$, where $D = \mathrm{diag}(d_0, \ldots, d_n)$, $d_0 > \cdots > d_n > 0$, can be written in the form

$$\mathcal{Q}_\varepsilon = \{q \in \mathbb{R}^{n+1} : \langle q, D^{-2}q\rangle = 1 + \varepsilon P(q_1^2/d_1^2, \ldots, q_n^2/d_n^2)\}$$

for some function $P : \mathbb{R}^n \to \mathbb{R}$ such that $P(0, \ldots, 0) = 0$. It is clear that if P is a linear function (the case of a *quadratic* perturbation), $\mathcal{Q}_\varepsilon$ remains a generic ellipsoid for $|\varepsilon|$ small enough, and therefore integrable, as a matter of fact with n independent (and explicit, see equation (6.2)) first integrals in involution $I_1, \ldots, I_n$. For an arbitrary *entire* function P, we see that the perturbed billiard map inside $\mathcal{Q}_\varepsilon$ is not uniformly integrable, unless P is linear. By uniformly integrable we mean that the first integrals $I_1, \ldots, I_n$ can be smoothly extended as a function of ε. This result can be considered as a local weak version of the Birkhoff conjecture. Its proof consists simply in showing that for nonlinear functions P, the Melnikov potential L has a singularity on $\mathbb{C}^n$, and therefore is non-constant, forbidding the persistence of the whole separatrix $\mathcal{S}$, and consequently, preventing the existence of first integrals regular in ε.

Stronger results can be obtained for concrete values of the dimension $n + 1$ where the ellipsoid lives. Thus, for planar billiards ($n = 1$, symmetrically perturbed elliptic billiard tables), one has non-integrability, i.e. the absence of any analytical non-constant first integral, for any nonlinear entire function P (theorem 4.2).

In particular, specific computations are carried out on two cases for the planar billiard. One is the close to flat limit that takes place when the minor semi-axis goes to zero. In this case, it is proved that if the ellipse is narrow enough, under *any* non-quadratic analytic symmetric small enough perturbation, *all* the symmetric bi-asymptotic orbits become transverse (theorem 4.3). The other one is the case of the simplest non-quadratic symmetric perturbation, a quartic perturbation of the form

$$\mathcal{Q}_\varepsilon = \left\{q = (x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 + \varepsilon\frac{y^4}{b^4}\right\} \qquad a > b > 0.$$

In this case, one can compute the Melnikov potential explicitly, as well as its critical points, and consequently, conclude that there exist only eight primary (that is, depending smoothly on ε) bi-asymptotic orbits, which are precisely the symmetric ones and which are transverse (theorem 4.4).

The spatial billiard ($n = 2$, symmetrically perturbed ellipsoids in $\mathbb{R}^3$) of the form

$$\mathcal{Q}_\varepsilon = \left\{q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \varepsilon P(y^2/b^2, z^2/c^2)\right\} \qquad (1.1)$$

where $a > b > c > 0$, is the 'star' case of this paper, for at least two reasons. First, thanks to careful study, we have been able to generalize the planar results to the higher-dimensional case. Second, for a particular perturbation, we present a pretty complete description of the

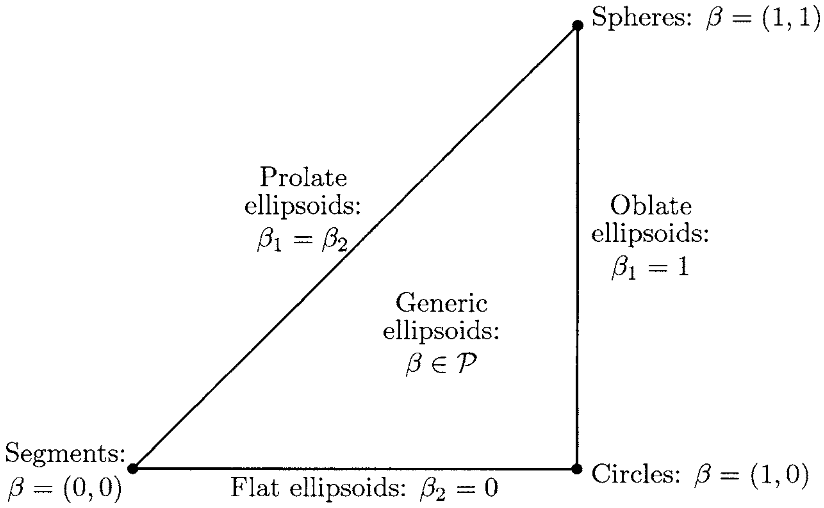

Spheres: $\beta = (1,1)$
Prolate ellipsoids: $\beta_1 = \beta_2$
Oblate ellipsoids: $\beta_1 = 1$
Generic ellipsoids: $\beta \in \mathcal{P}$
Segments: $\beta = (0,0)$
Flat ellipsoids: $\beta_2 = 0$
Circles: $\beta = (1,0)$

Figure 1. The parameter space $\mathcal{P}$ of generic spatial ellipsoidal billiards, where $\beta_1 = b^2/a^2$, $\beta_2 = c^2/a^2$ and $a \geqslant b \geqslant c$ are the semi-axes of the ellipsoid. Its border corresponds to degenerate ellipsoids.

spatial primary homoclinic orbits scenario in terms of the two intrinsic parameters of the system $\beta_1 = b^2/a^2$, $\beta_2 = c^2/a^2$.

The parameter space of generic spatial ellipsoidal billiards is the triangle

$$\mathcal{P} = \left\{\beta = (\beta_1, \beta_2) \in \mathbb{R}^2 : 0 < \beta_2 < \beta_1 < 1\right\}$$

whose edges consist of degenerate spatial ellipsoids (the oblate and prolate case) or flat ellipsoids, as illustrated in figure 1.

Two different kinds of specific computations are carried out for the spatial billiard. One is the close to flat limit that takes place when the minor semi-axis goes to zero. In terms of the parameter β, this is equivalent to saying that $\beta_2 \to 0^+$ (we approach the lower edge of figure 1).

To avoid cumbersome computations, we have restricted ourselves to the symmetric polynomial perturbations $\mathcal{Q}_\varepsilon$ preserving the horizontal section of $\mathcal{Q} = \mathcal{Q}_0$, that is, the function P has the form $P(s_1, s_2) = s_2 R(s_1, s_2)$, and non-quadratic perturbations correspond to nonlinear P, that is, to non-constant R. In this case, explicit generic inequalities (corollary 5.3) on the 1-jet of R imply the transversality of each one of the spatial symmetric bi-asymptotic orbits. We believe that such a result also holds for any non-quadratic analytic symmetric perturbation (as in theorem 4.3 for the planar case), but the much more complicated computations of the Melnikov potential require a considerable extra effort that has not been carried out by the authors.

To realize the complexity of the computations in the spatial case, it is worth noting one of the main differences between the spatial and the planar case, at least for what concerns the Melnikov potential. One of the main features of the Melnikov potential $L(r)$ for the planar case is that the function $t \mapsto L(\mathrm{e}^t)$ is an *elliptic* function for any polynomial perturbation, a fact that makes the computation of the series defining $L(r)$ easier. Unfortunately, in the spatial case, the function $t = (t_1, t_2) \mapsto L(\mathrm{e}^{t_1}, \mathrm{e}^{t_2})$ only possesses three independent periods (instead of four) and therefore is not a *hyperelliptic* or an *Abelian* function, and this means that in general we are not able to obtain closed formulae for the Melnikov potential $L(r)$.

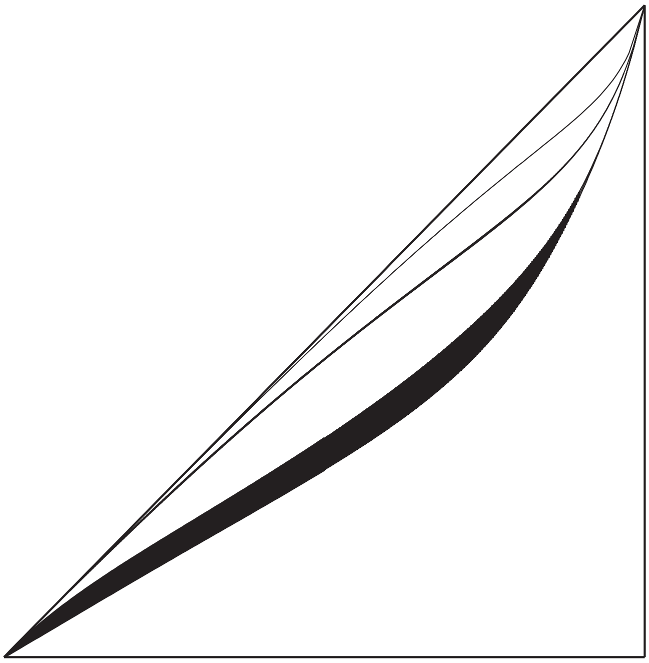


Figure 2. Partition of the parameter space in white (16 primary homoclinic orbits) and black strips (32 primary homoclinic orbits) for a perturbed ellipsoid $\mathcal{Q}_\varepsilon$ of the form (1.1) with $P(s_1, s_2) = s_1 s_2$. (Only a finite number of strips are visible in this figure.)

The second kind of specific computation has been focused on the number $H = H(\beta)$ of primary homoclinic orbits for the perturbation $\mathcal{Q}_\varepsilon$ in the concrete case $P(s_1, s_2) = s_1 s_2$. Briefly, we see that $H(\beta) = 16$ close to the flat, circular and oblate limits, but that $H(\beta)$ undergoes infinitely many bifurcations when β approaches the prolate limit, oscillating between 16 and 32. The statements $H(\beta) \geqslant 16$ close to the flat or circular limit, and $H(\beta) = 16$ close to the oblate limit, are proved analytically, whereas the rest are contained in the numerical results 5.4 and 5.5. In particular, the result concerning the oscillation of the number of homoclinic orbits close to the prolate case is completely new and seems amazing to us.

An illustration of these results is provided by figure 2, where the points of the set $\mathcal{P}_- = \{\beta \in \mathcal{P} : H(\beta) = 16\}$ are drawn in white, whereas the points of the set $\mathcal{P}_+ = \{\beta \in \mathcal{P} : H(\beta) = 32\}$ are drawn in black. These sets are formed by *infinitely many strips* connecting $\beta = (0, 0)$ and $\beta = (1, 1)$; that is, their extrema correspond to segments and spheres. The reader will be able to find only the first three black strips, since the other black strips are too thin to be seen on the scale of the picture.

1.3. Related open problems

The tools presented in this paper, which rely on the study of the Melnikov potential, can also be applied to perturbations of any integrable map, whose bi-asymptotic motions admit good natural parametrizations. We mention some such problems.

- Billiard maps inside convex closed surfaces can be perturbed in many ways, and not only by deforming the initial surface. For instance, one can consider that a constant weak magnetic field acts on the particle between consecutive impacts [RB85, BK96]. In a first approximation, this perturbation is equivalent to a rotation of the surface around the direction of the magnetic field at a slow constant velocity. The elliptic planar case (a particle inside a slowly rotating ellipse) was studied in [Koz98]. It seems reasonable to deal with the spatial case (a particle inside a slowly rotating ellipsoid).

- Another related problem is the billiard with an oscillating ellipsoidal boundary, which is a rather popular problem with several applications in physics [KMOC95, KMOC96].
- It is known that the ellipsoidal billiard remains integrable in the presence of certain (separable) polynomial potentials of even degree, in particular, the quadratic Hooke potential. Such kinds of potentials were described by Jacobi (see, for instance, [WT85]). One can investigate ellipsoidal billiards with perturbations of the above potentials.
- Nowadays various non-trivial integrable discretizations of the classical Euler top and its multidimensional generalizations are known (see [MV91, BLS98, Fed00]). Our approach can equally be applied to perturbations of the discrete Euler top.

Going back to the billiard inside a perturbed ellipsoid, we list other open problems.

First, the bifurcations that appear in infinite cycles close to the prolate limit look amazing. At the present time, we do not know of an ultimate reason for their appearance. Since the homoclinic orbits inside a generic ellipsoid tend to a concatenation of two heteroclinic orbits inside a prolate ellipsoid when the generic ellipsoid tends to the prolate one, it seems necessary to develop some kind of 'secondary' Poincaré–Melnikov method to study this phenomenon (classical methods only detect primary orbits). One can use the variational approach introduced in [BM00] to detect these secondary (or two-bump) homoclinic orbits.

Second, in this paper, we have restricted ourselves to bi-asymptotic motions of maximal dimension (for instance, spatial bi-asymptotic motions in the spatial billiard), since they lie on the separatrix. It would be very interesting to investigate the persistence of homoclinic or heteroclinic orbits of lower dimension (those originating from the bifurcation set).

We finish this introduction with the organization of this paper. The logical development of its content gives rise to a structure that is very different from this introduction. Thus, we first need to introduce (in section 2) the concepts of twist maps, and for them the notion of doubled invariant manifolds and the Melnikov potential. Next, we introduce convex billiards in section 3. Afterwards, we deal in section 4 with the planar case, to make the reader familiar with the concept of the separatrix. The Melnikov potential is computed explicitly for a flat case and for a quartic perturbation (some of its computations are postponed to appendix B).

Subsequently, in section 5 we introduce the spatial billiard and the geometry, dynamics and topology of the bi-asymptotic set are carefully explained, as well as the role of the separatrix. Several kinds of symmetries are introduced. We are confronted with non-uniform integrability instead of non-integrability as another difference with the planar case. The situation close to the flat limit is studied analytically (some of these computations are postponed to appendix A). In the case of a particular quartic perturbation, both analytical and numerical results are presented. Finally, all the results that do not involve specific computations are generalized to higher dimensions in section 6.

2. The Poincaré–Melnikov method for twist maps

In this section the Poincaré–Melnikov method for twist maps from [DR97] is reviewed. Related ideas are contained in [Lom97].

For the sake of simplicity, we will assume that the objects considered here are smooth. For a general background on the symplectic geometry we refer to [AM78]. The review [Mei92] is a good reference for twist maps.

2.1. Twist maps

An *exact symplectic manifold* is an even-dimensional manifold $\mathcal{M}$ endowed with a *symplectic form* ω which is exact: $\omega = -\mathrm{d}\phi$. An *exact symplectic map* is a map $f : \mathcal{M} \to \mathcal{M}$ such that $f^*\phi - \phi = \mathrm{d}S$ for some function $S : \mathcal{M} \to \mathbb{R}$, called a *generating function* of f.

Typical examples of exact symplectic manifolds are cotangent bundles endowed with their canonical forms ϕ_0 and $\omega_0 = -\mathrm{d}\phi_0$. Typical examples of exact symplectic maps are the so-called *twist maps*. Although there exist several almost-equivalent definitions of twist maps, the following one suffices for the study of convex billiards.

Let $\mathcal{M}$ be an open subset of a cotangent bundle $T^*\mathcal{Q}$. A map $f : \mathcal{M} \to \mathcal{M}$ will be called *twist* if it is exact symplectic and there exists an open $\mathcal{U} \subset \mathcal{Q} \times \mathcal{Q}$ such that $\pi \times \pi \circ f : \mathcal{M} \to \mathcal{U}$ is a diffeomorphism. Here, π denotes the canonical projection of the phase space $\mathcal{M} \subset T^*\mathcal{Q}$ onto the configuration space $\mathcal{Q}$. The quantity $n = \dim \mathcal{Q}$ is called the *number of degrees of freedom* of f, the map $l = (\pi \times \pi \circ f)^{-1} : \mathcal{U} \to \mathcal{M}$ is called the *Legendre transformation* of f, and the function $\mathcal{L} = S \circ l : \mathcal{U} \to \mathbb{R}$ is called the *twist generating function* or *Lagrangian* of f.

The term Lagrangian is motivated by the following variational principle.

Let $f : \mathcal{M} \to \mathcal{M}$ be a twist map. Its *orbits* are the sequences $\mathcal{O} = (m_k)_{k\in\mathbb{Z}} \in \mathcal{M}^{\mathbb{Z}}$ such that $f(m_k) = m_{k+1}$. Its *configurations* are the sequences $\mathcal{C} = (q_k)_{k\in\mathbb{Z}} \in \mathcal{Q}^{\mathbb{Z}}$ such that $f(m_k) = m_{k+1}$ when $m_k = l(q_k, q_{k+1})$. (Configurations are in a one-to-one correspondence with orbits via the Legendre transformation.) Then the configurations are just the critical points of the formal series (called the *action functional*)

$$\mathcal{Q}^{\mathbb{Z}} \ni \mathcal{C} \mapsto \sum_{k\in\mathbb{Z}} \mathcal{L}(q_k, q_{k+1}) \in \mathbb{R}. \tag{2.1}$$

Of course, this series can be divergent, but there are many special cases in which it makes sense. For instance, let $\mathcal{O} = (m_k)_{k\in\mathbb{Z}}$ be a homoclinic orbit to a hyperbolic fixed point m^{h} and let $\mathcal{C} = (q_k)_{k\in\mathbb{Z}}$ be the corresponding configuration. Then the series

$$W[\mathcal{O}] := W[\mathcal{C}] := \sum_{k\in\mathbb{Z}} [\mathcal{L}(q_k, q_{k+1}) - \mathcal{L}(q^{\mathrm{h}}, q^{\mathrm{h}})] \qquad q^{\mathrm{h}} = \pi(m^{\mathrm{h}})$$

converges to a quantity called the *homoclinic action* of the orbit $\mathcal{O}$ (or the configuration $\mathcal{C}$).

In cotangent coordinates (q, p) (q positions, p momenta) the canonical forms read as $\phi_0 = p\,\mathrm{d}q$ and $\omega_0 = \mathrm{d}q \wedge \mathrm{d}p$, whereas the canonical projection is $\pi(q, p) = q$. Writing $f(q, p) = (q', p')$, the exactness property $f^*\phi_0 - \phi_0 = \mathrm{d}S$ reads as $p'\,\mathrm{d}q' - p\,\mathrm{d}q = \mathrm{d}S(q, p)$, whereas the Legendre transformation l is simply given by $(q, q') \mapsto (q, p)$ and the twist generating function is $\mathcal{L}(q, q') = S(q, p)$. Hence, $p'\,\mathrm{d}q' - p\,\mathrm{d}q = \mathrm{d}\mathcal{L}(q, q')$, so that one can retrieve the map f implicitly from

$$p = -\partial_1\mathcal{L}(q, q') \qquad p' = \partial_2\mathcal{L}(q, q').$$

This can be done over the whole phase space $\mathcal{M}$, because in the above definition of twist maps it is assumed that the momenta can be expressed globally in terms of old and new positions, via the Legendre transformation l.

2.2. Doubled invariant manifolds

Let $f : \mathcal{M} \to \mathcal{M}$ be a twist diffeomorphism on an open subset of a cotangent bundle $T^*\mathcal{Q}$ with Lagrangian $\mathcal{L} : \mathcal{U} \to \mathbb{R}$, and assume that f has a hyperbolic fixed point m^{h}. We will say that its n-dimensional *unstable and stable invariant manifolds*

$$\mathcal{W}^{\mathrm{u}} := \left\{ m \in \mathcal{M} : \lim_{k\to-\infty} f^k(m) = m^{\mathrm{h}} \right\} \qquad \mathcal{W}^{\mathrm{s}} := \left\{ m \in \mathcal{M} : \lim_{k\to+\infty} f^k(m) = m^{\mathrm{h}} \right\}$$

are *doubled* when they coincide: $\mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}}$. Nevertheless, the invariant manifolds could not coincide as smooth manifolds. We pause here to explain this subtle, although crucial, fact. For more details, the reader can consult [DR97].

As a matter of fact, the invariant manifolds $\mathcal{W}^{\mathrm{u,s}}$ need not be submanifolds of $\mathcal{M} \subset T^*\mathcal{Q}$, but just connected *immersed submanifolds*. More precisely, $\mathcal{W}^{\mathrm{u,s}} = m^{\mathrm{u,s}}(\mathbb{R}^n)$ for some one-to-one immersions $m^{\mathrm{u,s}} : \mathbb{R}^n \to \mathcal{M}$, such that $m^{\mathrm{u,s}}(0) = m^{\mathrm{h}}$ and $\mathrm{d}m^{\mathrm{u,s}}(0)[\mathbb{R}^n]$ is the tangent space to $\mathcal{W}^{\mathrm{u,s}}$ at m^{h}. Hence, it is natural to endow $\mathcal{W}^{\mathrm{u}}$ (respectively, $\mathcal{W}^{\mathrm{s}}$) with the smooth structure induced by the immersion m^{u} (respectively, m^{s}).

From now on, in the case of doubled manifolds, we will reserve the letter $\mathcal{W}$ (without a superscript) for the *bi-asymptotic set* $\mathcal{W} := \{m \in \mathcal{M} : \lim_{|k|\to\infty} f^k(m) = m^{\mathrm{h}}\}$, whereas $\mathcal{W}^{\mathrm{u}}$ and $\mathcal{W}^{\mathrm{s}}$ will denote the invariant manifolds equipped with the above-mentioned smooth structure.

We define the *separatrix* $\mathcal{S}$ (respectively, the *bifurcation set* $\mathcal{B}$) as the subset of $\mathcal{W}$ of the points where the invariant manifolds have (respectively, do not have) the same smooth structure. In particular, the bifurcation set contains all the points where the tangent spaces of the invariant manifolds differ.

The separatrix is an *exact Lagrangian submanifold* of the phase space invariant by the twist diffeomorphism. Obviously, it does not contain the hyperbolic fixed point, since the invariant manifolds are transverse at it. Moreover, it is easy to see [DR97] that all the orbits on a connected component of the separatrix have the same homoclinic action.

In the planar case, that is, twist maps with one degree of freedom, the bifurcation set becomes just the hyperbolic fixed point. For more degrees of freedom, the determination of the bifurcation set (and as a consequence of the separatrix) is itself an interesting problem.

The following characterization of the bifurcation set turns out to be very useful in determining it. Given a point m in the bi-asymptotic set $\mathcal{W}$, it is parametrized by means of the above-mentioned immersions $m^{\mathrm{u,s}} : \mathbb{R}^n \to \mathcal{M}$ as $m = m^{\mathrm{u}}(r^{\mathrm{u}}) = m^{\mathrm{s}}(r^{\mathrm{s}})$ for *unique* parameter values r^{u} and r^{s}. The possibility that m belongs to the bifurcation set $\mathcal{B}$ can then only take place when $m = \lim_{j\to\infty} m^{\mathrm{u}}(r_j)$ or $m = \lim_{j\to\infty} m^{\mathrm{s}}(r_j)$, for some unbounded sequence $(r_j)_{j\geqslant 0} \subset \mathbb{R}^n$. Roughly speaking, this means that the bifurcation set is formed by the self-intersections of the bi-asymptotic set in the phase space.

Our next goal is to investigate the effect of small twist perturbations on this structure. Typically, the separatrix splits (does not persist), and breaks down into isolated homoclinic orbits, some of them transverse. The standard tool for measuring this splitting of the separatrix is the Melnikov potential.

2.3. The Melnikov potential

Let $f_\varepsilon : \mathcal{M} \to \mathcal{M}$ be a twist perturbation of f. Let $\mathcal{L}_\varepsilon = \mathcal{L} + \varepsilon\mathcal{L}_1 + \mathrm{O}(\varepsilon^2) : \mathcal{U} \to \mathbb{R}$ be its Lagrangian. It is not restrictive to normalize the problem in such a way that the hyperbolic fixed point does not change: $f_\varepsilon(m^{\mathrm{h}}) = m^{\mathrm{h}}$. Set $q^{\mathrm{h}} = \pi(m^{\mathrm{h}})$. Then we introduce the *Melnikov potential* as the function $L : \mathcal{S} \longrightarrow \mathbb{R}$ given by

$$L(m) = \sum_{k\in\mathbb{Z}} \left[\mathcal{L}_1(q_k, q_{k+1}) - \mathcal{L}_1(q^{\mathrm{h}}, q^{\mathrm{h}})\right] \qquad q_k = \pi(m_k) \quad m_k = f^k(m). \tag{2.2}$$

This function is well defined, smooth and invariant under the unperturbed map: $L = L \circ f$. Due to the hyperbolic character of the fixed point, the series in (2.2) is absolutely convergent. The fact that the perturbed invariant manifolds are exact Lagrangian immersed submanifolds of $\mathcal{M}$ plays an essential role in its derivation. In particular, if (x, y) are coordinates *symplectically adapted* to $\mathcal{S}$—that is, in these coordinates $\mathcal{S}$ is given locally by $\{y = 0\}$ and the Liouville form ϕ_0 reads as $y\,\mathrm{d}x$—it turns out [DR97] that the perturbed invariant manifolds can be expressed as $\mathcal{W}^{\mathrm{u,s}}_{\varepsilon} = \{y = \varepsilon\nabla L^{\mathrm{u,s}}(x) + \mathrm{O}(\varepsilon^2)\}$ for some smooth functions $L^{\mathrm{u,s}} : \mathcal{W}^{\mathrm{u,s}} \to \mathbb{R}$. Restricting the base points of the invariant manifolds to the separatrix where their smooth structures coincide, we arrive at the Melnikov potential $L = L^{\mathrm{u}} - L^{\mathrm{s}} : \mathcal{S} \to \mathbb{R}$. It does not depend on the coordinates (x, y), as expression (2.2) shows. Therefore, the differential $M = \mathrm{d}L$ (called the *Melnikov function*) gives the first-order distance, along the coordinate y in any coordinates (x, y) symplectically adapted to the separatrix, between the perturbed invariant manifolds.

The following properties, which hold for $0 < |\varepsilon|$ small enough, are obtained using this geometrical construction:

L1 If L is not locally constant, the separatrix splits.

L2 If L is not locally constant, and the unperturbed map f is (completely) integrable, then the perturbed map is not uniformly integrable.

L3 The non-degenerate critical points of L give rise to transverse primary homoclinic orbits of the perturbed map. If all the critical points of L are non-degenerate, any primary homoclinic orbit arising from the separatrix can be associated with some critical point.

L4 If $\tilde{\mathcal{O}}(\varepsilon)$ and $\hat{\mathcal{O}}(\varepsilon)$ are two primary homoclinic orbits associated with the critical points $\tilde{m}$ and $\hat{m}$, then $W[\tilde{\mathcal{O}}(\varepsilon)] - W[\hat{\mathcal{O}}(\varepsilon)] = \varepsilon[L(\tilde{m}) - L(\hat{m})] + \mathrm{O}(\varepsilon^2)$, provided that $\tilde{m}$ and $\hat{m}$ belong to the same connected component of the separatrix.

The proofs of L1, L3 and L4 can be found in [DR97], whereas L2 is established in [DLR01]. We will restrict ourselves to pointing out some comments about them.

C1 If the Melnikov potential is not locally constant, the Melnikov function is not identically zero and the perturbed invariant manifolds do not coincide. This is the phenomenon of separatrix splitting.

C2 A twist map with n degrees of freedom is called *(completely) integrable* if it has n functionally independent almost-everywhere first integrals in involution. A family of twist maps is called *uniformly integrable* if each map of the family is completely integrable and the first integrals depend smoothly on the parameter of the family. Hence, in L2 it is stated that the unperturbed first integrals cannot be smoothly continued to perturbed ones when the Melnikov potential is not locally constant.

C3 In the planar case, the invariant manifolds become invariant curves and the non-integrability criterion L2 can be strengthened. The result is the following: if the Melnikov potential is not locally constant, then the perturbed invariant curves cross topologically, and so the perturbed map becomes non-integrable [Cus78]. That is, we can deal with just integrability, instead of uniform integrability.

C4 Let $\mathcal{O}(\varepsilon)$ be a perturbed homoclinic orbit; that is, $\mathcal{O}(\varepsilon) \subset \left(\mathcal{W}^{\mathrm{u}}_{\varepsilon} \bigcap \mathcal{W}^{\mathrm{s}}_{\varepsilon}\right) \setminus \{m^{\mathrm{h}}\}$. It is called *transverse* when the intersections of the invariant manifolds at its points are transverse. It is called *primary* when it depends smoothly on $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ for some $\varepsilon_0 > 0$, and $\mathcal{O}(\varepsilon) = \mathcal{O} + \mathrm{O}(\varepsilon)$ for some unperturbed homoclinic orbit $\mathcal{O} = (m_k)_{k\in\mathbb{Z}}$, $m_k = f^k(m)$. These are the kind of orbits that can be detected by a perturbative theory based on the Melnikov potential.

C5 According to the *Birkhoff–Smale homoclinic theorem* [Sma65], maps with transverse homoclinic orbits are *chaotic*: the restriction of some power of the map to some Cantor set close to the homoclinic orbit is conjugated to a *transitive topological Markov chain*. This existence of chaotic behaviour explains the importance of transverse homoclinic orbits.

C6 To motivate L4, let us first focus on the planar case. Let $\tilde{\mathcal{O}}(\varepsilon)$ and $\hat{\mathcal{O}}(\varepsilon)$ be two primary homoclinic orbits of f_ε, associated with some critical points $\tilde{m}$ and $\hat{m}$ of the Melnikov potential L, lying on the same connected component of the separatrix. Let $A(\varepsilon)$ be the area of the region enclosed by the perturbed invariant curves with endpoints on these orbits. (Such regions are called *lobes*.) In the *MacKay–Meiss–Percival action principle* [MMP84] this lobe area is interpreted as a difference of homoclinic actions: $A(\varepsilon) = W[\tilde{\mathcal{O}}(\varepsilon)] - W[\hat{\mathcal{O}}(\varepsilon)]$. Therefore, the difference $L(\tilde{m}) - L(\hat{m})$ measures this lobe area in first order in ε. Moreover, when this difference is zero and L is not constant, L has other critical points and there exist more primary homoclinic orbits.

This is due to the fact that, since the Melnikov potential L is invariant under the unperturbed map f: $L = L \circ f$, one can define it on the *reduced separatrix* $\mathcal{S}^* = \mathcal{S}/\{f\}$, which is the quotient of the separatrix under the action of the unperturbed map. In the planar case, the separatrix is one dimensional and therefore every connected component of the reduced separatrix is diffeomorphic to a one-dimensional torus $\mathbb{T}$. Consequently, the Melnikov potential L is a periodic function in suitable coordinates, and therefore it has more than one critical point if it is non-constant.

In the general case of n degrees of freedom, if the reduced separatrix $\mathcal{S}^*$ is compact, this latter property also holds: if $L(\tilde{m}) - L(\hat{m}) = 0$ for two critical points $\tilde{m}$ and $\hat{m}$ of the Melnikov potential L on the same connected component of the separatrix, there exist more critical points of L and therefore more primary homoclinic orbits. Sometimes compactness of $\mathcal{S}^*$ is not necessary. In those cases, it is important to check the difference $L(\tilde{m}) - L(\hat{m})$ for each pair of critical points $\tilde{m}$, $\hat{m}$ of L on the same connected component of the separatrix.

We end this section by noting that heteroclinic orbits (instead of homoclinic ones) and periodic points (instead of fixed ones) can also be analysed with minor changes.

3. Convex billiards

We consider the convex billiard problem [Bir27, KT91, Tab95]. Let $\mathcal{Q}$ be a closed convex hypersurface of $\mathbb{R}^{n+1}$. A material point moves inside $\mathcal{Q}$ and collides elastically with $\mathcal{Q}$; that is, at the impact points the velocity is reflected so that its tangential component remains the same, while the sign of its normal component is changed.

This billiard motion can be modelled by means of a diffeomorphism defined on a phase space $\mathcal{M}$ consisting of positions q on the hypersurface $\mathcal{Q}$ and unitary velocities p directed outward $\mathcal{Q}$ at q:

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

Indeed, we define the *billiard map* $f : \mathcal{M} \to \mathcal{M}$, $f(q, p) = (q', p')$, in the following way (see figure 3): p' is the reflection of the velocity p described above, and $q' \in \mathcal{Q}$ is determined by $p' = (q' - q)/\left|q' - q\right|$.

Let $\mathcal{U} = \{(q, q') \in \mathcal{Q} \times \mathcal{Q} : q \neq q'\}$. It is very well known (see, for instance, [Tab95, section 2.9]), that the convexity of $\mathcal{Q}$ implies that f is a twist map with Lagrangian

$$\mathcal{L} : \mathcal{U} \to \mathbb{R} \qquad \mathcal{L}(q, q') = \left|q - q'\right|. \tag{3.1}$$

A *billiard orbit* is a sequence $\mathcal{O} = (m_k)_{k\in\mathbb{Z}} \in \mathcal{M}^{\mathbb{Z}}$ such that $f(m_k) = m_{k+1}$. A *billiard configuration* is a sequence of impact points $\mathcal{C} = (q_k)_{k\in\mathbb{Z}} \in \mathcal{Q}^{\mathbb{Z}}$ such that $f(q_k, p_k) = (q_{k+1}, p_{k+1})$ for $p_{k+1} = (q_{k+1} - q_k)/\left|q_{k+1} - q_k\right|$. A *billiard trajectory* is a sequence of oriented segments $\mathcal{T} = (s_k)_{k\in\mathbb{Z}}$ such that $s_k = [q_k, q_{k+1}]$ for some billiard configuration $(q_k)_{k\in\mathbb{Z}}$, where

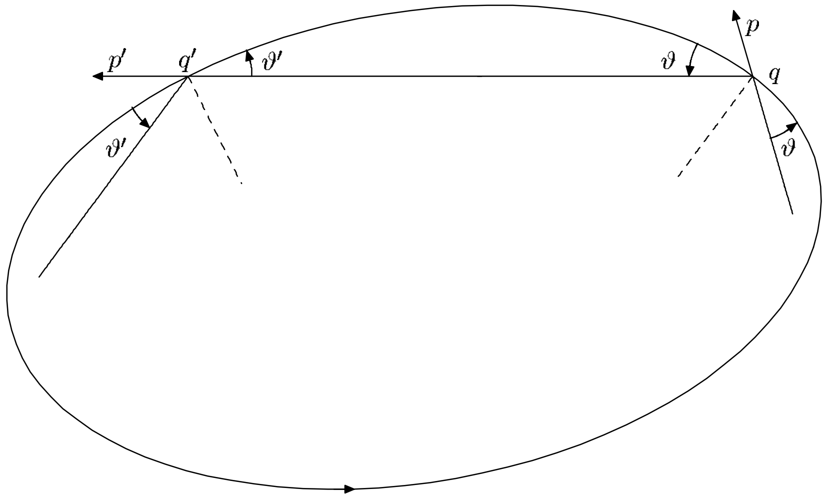


Figure 3. The billiard map $f(q, p) = (q', p')$.

$s = [q, q']$ denotes the segment from q to q'. It is clear that orbits, configurations and trajectories are in one-to-one correspondence. Hence, we can use them indistinctly.

The *chords* of the hypersurface $\mathcal{Q}$ are the segments perpendicular to $\mathcal{Q}$ at their ends. The greatest chords are called *diameters*. A chord gives rise to a couple of two-periodic points. Generically, the two-periodic points are hyperbolic when the chord is a *diameter*. Such a diameter will be called *hyperbolic*.

Let $\mathcal{T} = (s_k)_{k\in\mathbb{Z}}$ be a homoclinic trajectory to a hyperbolic diameter of a hypersurface $\mathcal{Q}$. Let $\mathcal{O}$ and $\mathcal{C}$ be its associated orbit and configuration, respectively. We denote by diam $\mathcal{Q}$ the length of the diameter. Then the series

$$\text{Length}\,\mathcal{O} := \text{Length}\,\mathcal{C} := \text{Length}\,\mathcal{T} := \sum_{k\in\mathbb{Z}}(\text{length}\, s_k - \text{diam}\,\mathcal{Q})$$

converges to a negative quantity called the *(homoclinic) length* of $\mathcal{O}$, $\mathcal{C}$ or $\mathcal{T}$. Clearly, it coincides with the (homoclinic) action defined in section 2.1: Length $\mathcal{O} = W[\mathcal{O}]$.

This leads to the following interpretation of the lobe area for planar billiards (see comment C6 of section 2.3): *the lobe area enclosed between two primary homoclinic billiard orbits to a hyperbolic diameter is equal to the difference of lengths*. For spatial and higher-dimensional billiards, these differences of lengths are symplectic invariants which are useful in estimating the splitting size [DR97].

4. The planar case

We collect here several results on billiards inside perturbed ellipses, adapted from the studies contained in [LT93, Tab94, DR96, Lom96, Lev97]. They are intended to prepare the scenario for the spatial case, which is conceptually similar but technically harder. So, in this section we will lay the foundations for the next one.

We will consider a *non-circular ellipse*

$$\mathcal{Q} = \left\{ q = (x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \right\} \qquad a > b > 0 \tag{4.1}$$

whose diameter is given by the chord joining the vertices $(-a, 0)$ and $(a, 0)$. (We avoid dealing with the circumference $a = b$, since it has a continuous family of diameters.) We will denote the set formed by the two-periodic points associated with the diameter by

$$\mathcal{M}^{\mathrm{h}} = \{m_+^{\mathrm{h}}, m_-^{\mathrm{h}}\} \qquad m_\pm^{\mathrm{h}} = (q_\pm^{\mathrm{h}}, p_\pm^{\mathrm{h}}) \qquad q_\pm^{\mathrm{h}} = (\pm a, 0) \qquad p_\pm^{\mathrm{h}} = (\pm 1, 0).$$

4.1. The bi-asymptotic set

We are going to see that m_+^{h} and m_-^{h} are hyperbolic two-periodic points of the planar elliptic billiard map f whose unstable and stable invariant manifolds are doubled. Our first step will be to give a geometric characterization of the invariant sets

$$\begin{aligned} \mathcal{W} &:= \left\{ m \in \mathcal{M} : \lim_{|k|\to\infty} \operatorname{dist}\left(f^k(m), \mathcal{M}^{\mathrm{h}}\right) = 0 \right\} \\ \mathcal{W}^{\mathrm{u}} &:= \left\{ m \in \mathcal{M} : \lim_{k\to-\infty} \operatorname{dist}\left(f^k(m), \mathcal{M}^{\mathrm{h}}\right) = 0 \right\} \\ \mathcal{W}^{\mathrm{s}} &:= \left\{ m \in \mathcal{M} : \lim_{k\to+\infty} \operatorname{dist}\left(f^k(m), \mathcal{M}^{\mathrm{h}}\right) = 0 \right\}. \end{aligned} \tag{4.2}$$

As a by-product of this characterization, we will prove that the unstable and stable invariant manifolds are doubled and coincide with the bi-asymptotic set:

$$\mathcal{W} = \mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}}.$$

To begin with, let us recall a geometric property of the ellipses [Tab95, section 2.1]. Let

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

be the family of *confocal conics* to the ellipse Q. It is clear that $Q(\kappa)$ is an ellipse for $0 < \kappa < b$, and a hyperbola for $b < \kappa < a$. No real conic exists for $\kappa > a$.

Concerning the degenerate cases $\kappa = a, b$, we first note that for $\kappa \to b^-$ (respectively, $\kappa \to b^+$) the conic $Q(\kappa)$ flattens into the region of the x-axis enclosed by (respectively, outside) the foci of the ellipse Q. On the other hand, for $\kappa \to a^-$, the hyperbola flattens into the whole y-axis.

The fundamental property of planar elliptic billiards is that *any segment (or its prolongation) of a billiard trajectory inside the ellipse $Q = Q(0)$ is tangent*[3] *to one fixed confocal conic $Q(\kappa)$*. Thus, the confocal conics are *caustics* of the elliptic billiard (see figure 4).

The notion of tangency in the degenerate cases is the following: a line is *tangent to $Q(b)$* when it passes through the foci

$$\mathcal{F} = \{(-c, 0), (c, 0)\} \qquad c = \sqrt{a^2 - b^2}$$

and it is *tangent to $Q(a)$* when it coincides with the y-axis.

Therefore, the function $\kappa : \mathcal{M} \to \mathbb{R}$ is a first integral of the elliptic billiard map f, that is, $\kappa \circ f = \kappa$. A straightforward computation gives

$$\kappa(m) = ab(xu/a^2 + yv/b^2) \qquad m = (q, p) \quad q = (x, y) \quad p = (u, v).$$

Now, it is clear that the billiard orbits bi-asymptotic to the diameter are those ones that are tangent to $Q(b)$ or, equivalently, pass through the foci. Thus, if $q + \langle p \rangle$ denotes the line passing by q with direction p, the next proposition holds.

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

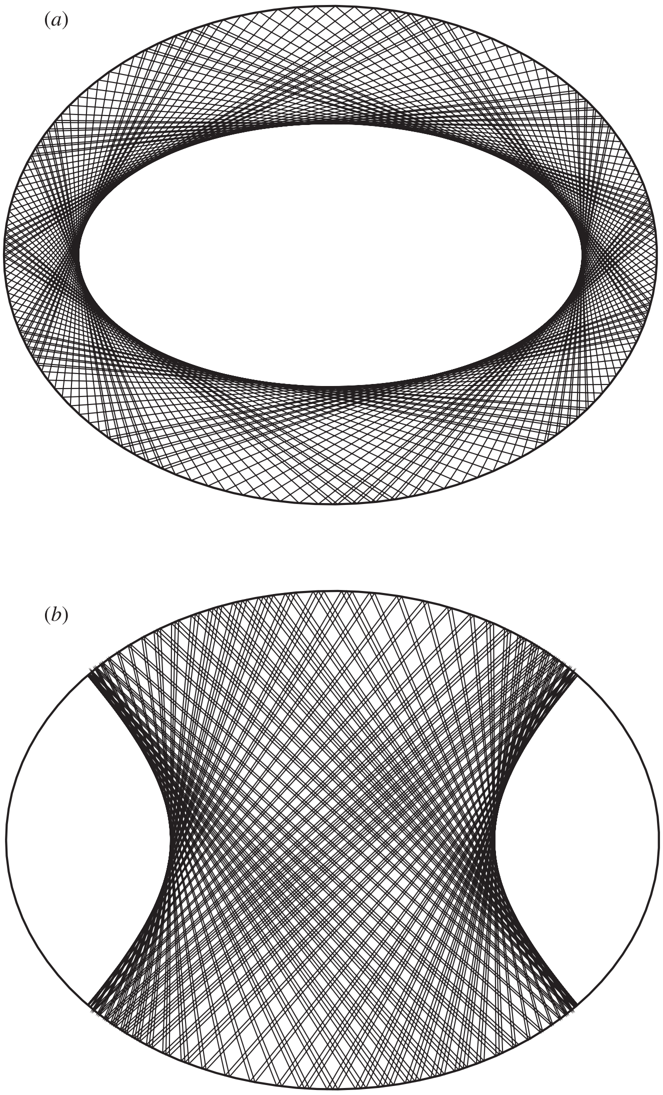


Figure 4. The two kinds of confocal caustics: ellipses (*a*) and hyperbolas (*b*).

Proposition 4.1. *$\mathcal{W} = \mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}} = \mathcal{M}_b$, where*

$$\mathcal{M}_b := \{m \in \mathcal{M} : \kappa(m) = b\} = \{m = (q, p) \in \mathcal{M} : q + \langle p \rangle \textit{ intersects } \mathcal{F}\}.$$

This discussion shows that elliptic planar billiards give a suitable framework with which to apply the Poincaré–Melnikov method described in section 2, because they are integrable twist diffeomorphisms with bi-asymptotic connections. Nevertheless, as was already stressed in [GPB89], the explicit implementation of this method requires a closed-form solution of the map on (and not only a parametrization of) the bi-asymptotic set. This closed form solution is given in the following lemma. Equivalent or related formulae can be found in [LT93, Tab94, DR96, Lom96, Lev97, Koz98].

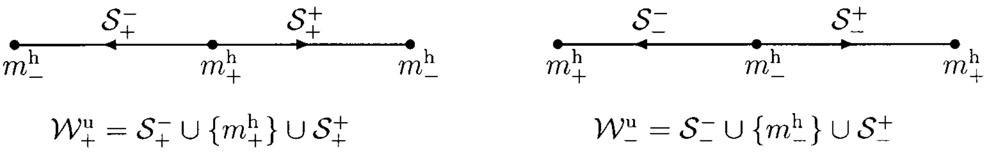

$\mathcal{S}^-_+$
$\mathcal{S}^+_+$
$m^{\rm h}_-$
$m^{\rm h}_+$
$m^{\rm h}_-$
$\mathcal{W}^{\rm u}_+ = \mathcal{S}^-_+ \cup \{m^{\rm h}_+\} \cup \mathcal{S}^+_+$
$\mathcal{S}^-_-$
$\mathcal{S}^+_-$
$m^{\rm h}_+$
$m^{\rm h}_-$
$m^{\rm h}_+$
$\mathcal{W}^{\rm u}_- = \mathcal{S}^-_- \cup \{m^{\rm h}_-\} \cup \mathcal{S}^+_-$

Figure 5. Topological representation of the bi-asymptotic set in the planar case.

Lemma 4.1. *Let τ, $\tau_{\rm x}$, $\tau_{\rm y}$ be the polynomials*

$$\tau(r) := 1 + r^2 \qquad \tau_{\rm x}(r) := \tau({\rm i}r) = 1 - r^2 \qquad \tau_{\rm y}(r) := 2r$$

and $\chi = (\tau_{\rm x}/\tau, \tau_{\rm y}/\tau)$. Let $q : \mathbb{R} \to \mathcal{Q}$ and $p : \mathbb{R} \to \mathbb{S}$ be the maps

$$q(r) := D\chi(r) = \left(a\frac{1-r^2}{1+r^2}, \frac{2br}{1+r^2} \right) \qquad p(r) := \chi(\lambda^{-1/2}r) = \left(\frac{\lambda - r^2}{\lambda + r^2}, \frac{2\lambda^{1/2}r}{\lambda + r^2} \right)$$

where $D = {\rm diag}(a, b)$ and $\lambda = (1+e)/(1-e)$ with $e = \sqrt{1 - b^2/a^2}$, and let $m = (q, p)$.

Then the maps $m^{\rm u,s}_\pm : \mathbb{R} \to \mathcal{M}$ defined by $m^{\rm u}_\pm(r) = \pm m(r)$ and $m^{\rm s}_\pm(r) = \mp m(1/r)$, are natural parametrizations of the curves

$$\begin{aligned} \mathcal{W}^{\rm u}_\pm &:= \left\{ m \in \mathcal{M} : \lim_{k \to -\infty} {\rm dist}\left(f^k(m), f^k(m^{\rm h}_\pm)\right) = 0 \right\} \\ \mathcal{W}^{\rm s}_\pm &:= \left\{ m \in \mathcal{M} : \lim_{k \to +\infty} {\rm dist}\left(f^k(m), f^k(m^{\rm h}_\pm)\right) = 0 \right\} \end{aligned} \tag{4.3}$$

which are invariant under the square of the billiard map. That is, $m^{\rm u,s}_\pm : \mathbb{R} \to \mathcal{W}^{\rm u,s}_\pm$ are analytic diffeomorphisms such that

$$m^{\rm u,s}_\pm(0) = m^{\rm h}_\pm \qquad f(m^{\rm u}_\pm(r)) = m^{\rm u}_\mp(\lambda r) \qquad f(m^{\rm s}_\pm(r)) = m^{\rm s}_\mp(r/\lambda).$$

The quantity λ is the *characteristic multiplier* of the hyperbolic two-periodic points $m^{\rm h}_\pm$; that is, the modulus of λ is greater than one and the eigenvalues of the differential of the elliptic billiard map at the hyperbolic two-periodic points are λ and $1/\lambda$ (see [LT93]).

Remark 4.1. From the definitions (4.2) and (4.3), it is clear that the bi-asymptotic set can be written as a disjoint union in two ways: $\mathcal{W} = \mathcal{W}^{\rm u}_+ \cup \mathcal{W}^{\rm u}_- = \mathcal{W}^{\rm s}_+ \cup \mathcal{W}^{\rm s}_-$, and that $f\left(\mathcal{W}^{\rm u,s}_+\right) = \mathcal{W}^{\rm u,s}_-$. We shall classify the bi-asymptotic orbits through the foci as heteroclinic (in opposition to homoclinic), since they have that character for the square of the billiard map.

Let us recall (see section 2.2) that, for planar maps, all the heteroclinic points are contained in the separatrix. Hence, the separatrix $\mathcal{S}$ of the planar elliptic billiard map has four connected components, namely

Proposition 4.2. *$\mathcal{S} = \mathcal{W} \setminus \mathcal{M}^{\rm h} = \mathcal{S}^+_+ \cup \mathcal{S}^-_+ \cup \mathcal{S}^+_- \cup \mathcal{S}^-_-$, where $\mathcal{S}^\sigma_\varsigma = \{\varsigma m(\sigma r) : r > 0\}$.*

We have depicted a topological representation of the bi-asymptotic set in figure 5. (For a dynamical representation the reader can see figure 7.) Points with equal labels are identified, so $\mathcal{W} = \mathcal{W}^{\rm u}_+ \cup \mathcal{W}^{\rm u}_-$ is homeomorphic to two circumferences glued along the couple of points $m^{\rm h}_\pm = \pm m(0) = \mp m(\infty)$. The connected components of the separatrix are invariant under the square of the elliptic billiard map. The arrows in the figure show this dynamics.

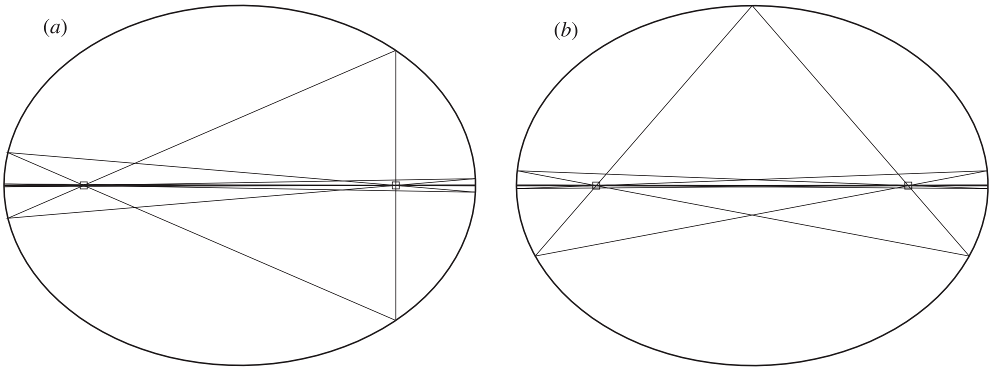


Figure 6. The two kinds of axial bi-asymptotic billiard trajectories inside an ellipse: x-axial (a) and y-axial (b). The foci are marked with squares.

An ellipse is the geometric locus of the points whose sum of distances to two given points (the foci of the ellipse) is a fixed quantity (the diameter of the ellipse). Using this characterization and a straightforward telescopic argument, one finds that the length of *all* the orbits in the separatrix is equal to minus the focal distance; that is,

Proposition 4.3. Length $\mathcal{O} = -2\sqrt{a^2 - b^2}$, *for all* $\mathcal{O} \subset \mathcal{S}$.

4.2. Persistence of symmetric bi-asymptotic orbits

Next, our goal is to prove that some distinguished heteroclinic orbits persist under suitable perturbations. Let us introduce these perturbations and orbits. A curve in the plane will be called *symmetric* when it is symmetric with regard to *both* coordinate axes of the plane. A perturbation of the ellipse (4.1) will be called *symmetric* if the perturbed ellipse is symmetric. Finally, a billiard orbit inside a symmetric curve will be called *central* (respectively, *axial*) when its billiard configuration is symmetric with regard to the origin (respectively, to some axis of coordinates). Inside an ellipse there are no central bi-asymptotic orbits, but there are two kinds of axial bi-asymptotic orbits: x*-axial* and y*-axial* (see figure 6). Their axes of symmetry are the x-axis and the y-axis, respectively.

Theorem 4.1. *Inside a non-circular ellipse there are four* x*-axial (and four* y*-axial) billiard orbits bi-asymptotic to the diameter. They persist under symmetric perturbations.*

Proof. Any billiard orbit inside a symmetric curve with a point on the set of symmetry

$$\tilde{\mathcal{F}} = \{m \in \mathcal{M} : x = 0\}$$

is y-axial. The connected component $\mathcal{S}^\sigma_\varsigma$ of the separatrix intersects $\tilde{\mathcal{F}}$ at the point

$$\tilde{m}^\sigma_\varsigma = \varsigma m(\sigma \tilde{r}) = \varsigma\,(\tilde{q}^\sigma, \tilde{p}^\sigma) \qquad \tilde{q}^\sigma = (\tilde{x}, \sigma\tilde{y}) \qquad \tilde{p}^\sigma = (\tilde{u}, \sigma\tilde{v}) \qquad \varsigma, \sigma = \pm$$

where $\tilde{r} = 1$, $\tilde{x} = 0$, $\tilde{y} = b$, $\tilde{u} = \sqrt{a^2 - b^2}/a$ and $\tilde{v} = b/a$. The intersection is transverse, because the equations of the phase space ($x^2/a^2 + y^2/b^2 = 1$ and $u^2 + v^2 = 1$), the separatrix ($xu/a^2 + yv/b^2 = 1/a$), and the set of symmetry ($x = 0$), are functionally independent at the points $\tilde{m}^\sigma_\varsigma$. Hence, the stable invariant curve of the billiard map associated with any symmetric perturbation of the ellipse intersects $\tilde{\mathcal{F}}$ in (at least) four points. The orbits by these points are y-axial and forward asymptotic. Due to the

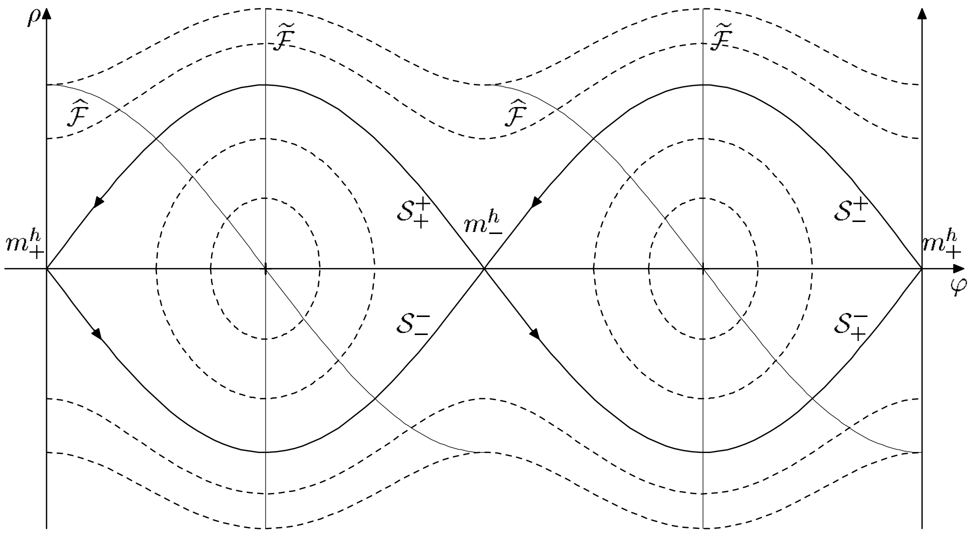


Figure 7. Phase portrait of the planar elliptic billiard map in (φ, ρ) coordinates.

axial symmetry, they are also backward asymptotic and, therefore, bi-asymptotic. This ends the proof concerning the existence of four y-axial persistent bi-asymptotic billiard orbits.

It turns out that there also exist four x-axial persistent bi-asymptotic billiard orbits. These x-axial orbits pass through the points

$$\hat{m}^\sigma_\varsigma = \varsigma m(\sigma\hat{r}) = \varsigma\left(\hat{q}^\sigma, \hat{p}^\sigma\right) \qquad \hat{q}^\sigma = \left(\hat{x}, \sigma\hat{y}\right) \qquad \hat{p}^\sigma = \left(\hat{u}, \sigma\hat{v}\right) \qquad \varsigma, \sigma = \pm$$

where $\hat{r} = \sqrt{\lambda}$, $\tilde{x} = -\sqrt{a^2 - b^2}$, $\tilde{y} = b^2/a$, $\tilde{u} = 0$ and $\tilde{v} = 1$. It suffices to check that $\mathcal{S}^\sigma_\varsigma$ intersects transversely at $\hat{m}^\sigma_\varsigma$ another set of symmetry: $\hat{\mathcal{F}} = \{m \in \mathcal{M} : u = 0\}$. □

Remark 4.2. We have used the axial symmetries to determine the natural parametrizations of lemma 4.1 in the following way. Natural parametrizations are unique except for linear changes of variables $r \mapsto \mu r$, for some $\mu \neq 0$. Between all the maps $m : \mathbb{R} \to \mathcal{M}$ verifying $m(0) = m^{\mathrm{h}}_+$ and $f(m(r)) = -m(\lambda r)$, we have chosen one in such a way that the 'natural parameter' of the y-axial bi-asymptotic points $\tilde{m}^\sigma_\varsigma$ (respectively, the x-axial ones $\hat{m}^\sigma_\varsigma$) is $r = \sigma\tilde{r}$ with $\tilde{r} = 1$ (respectively, $r = \sigma\hat{r}$ with $\hat{r} = \sqrt{\lambda}$).

For visualization purposes, it is useful to identify the phase space $\mathcal{M}$ with the annulus

$$\mathcal{A} = \{(\varphi, \rho) \in \mathbb{T} \times \mathbb{R} : |\rho| < |\dot{\gamma}(\varphi)|\} \qquad \gamma(\varphi) = (a\cos\varphi, b\sin\varphi)$$

by means of the relations $q = \gamma(\varphi)$ and $\rho = \langle\dot{\gamma}(\varphi), p\rangle = |\dot{\gamma}(\varphi)|\cos\vartheta$, where $\vartheta \in (0, \pi)$ is the angle between the tangent vector $\dot{\gamma}(\varphi)$ and the velocity p. In these coordinates, $\kappa^2 = b^2 + c^2\sin^2\varphi - \rho^2$. The partition of the annulus into invariant level curves of the billiard map f is shown in figure 7. The ∞-shaped curve is the *bi-asymptotic set* $\mathcal{W} = \{(\varphi, \rho) \in \mathcal{A} : \rho = \pm c\sin\varphi\}$. In particular, it becomes clear that the separatrix $\mathcal{S}$ has four connected components and that it intersects transversely the sets of symmetry $\tilde{\mathcal{F}}$ and $\hat{\mathcal{F}}$ just at eight points.

Once this persistence has been confirmed, several questions arise. Is the perturbed billiard integrable? Are the perturbed axial bi-asymptotic orbits transverse? Do all the perturbed axial bi-asymptotic orbits have the same length? The answers to these questions have been considered in several papers, by means of the Poincaré–Melnikov method. We summarize some results below.

4.3. The Melnikov potential

Let $\mathcal{Q}_\varepsilon$ be a symmetric perturbation of the ellipse (4.1). We can assume without loss of generality that the perturbation preserves the diameter of the ellipse; that is, diam $\mathcal{Q}_\varepsilon \equiv 2a$. Modulo $\mathrm{O}(\varepsilon^2)$ terms, which do not play any role in our first-order perturbative analysis, $\mathcal{Q}_\varepsilon$ can be put in the following *explicit form*:

$$\mathcal{Q}_\varepsilon = \phi_\varepsilon(\mathcal{Q}) \qquad \phi_\varepsilon(q) = [1 + \varepsilon\psi(q)]q \tag{4.4}$$

for some *symmetric* function $\psi : \mathcal{Q} \to \mathbb{R}$ such that $\psi(q^{\mathrm{h}}_\pm) = 0$. (Symmetric means that $\psi(x, y) = \psi(-x, y) = \psi(x, -y)$ for all $(x, y) \in \mathcal{Q}$.) If $P : \mathbb{R} \to \mathbb{R}$ is a function such that

$$P(y^2/b^2) = 2\psi(q) \qquad \forall q = (x, y) \in \mathcal{Q} \tag{4.5}$$

then the explicit form (4.4) is equivalent, modulo $\mathrm{O}(\varepsilon^2)$ terms, to the *implicit form*

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 + \varepsilon P(y^2/b^2) \right\}. \tag{4.6}$$

(Note that $P(0) = 2\psi(q^{\mathrm{h}}_\pm) = 0$.) We shall call this perturbation *polynomial*, *entire* or *analytic*, if the function P is polynomial, entire or analytic, respectively. In the polynomial case, we shall say that the *order of the perturbation* is twice the degree of the polynomial P. Thus, quadratic perturbations correspond to linear functions P.

Next, we look for an expression of the Melnikov potential (2.2) which is as simple as possible. The methodology for the spatial case is the same, and so it will not be repeated in the next section.

The Lagrangian $\mathcal{L}_\varepsilon = \mathcal{L}_0 + \varepsilon\mathcal{L}_1 + \mathrm{O}(\varepsilon^2)$ of the perturbed billiard inside (4.4) is

$$\mathcal{L}_\varepsilon(q, q') = \left|\phi_\varepsilon(q) - \phi_\varepsilon(q')\right| = \left|q - q'\right| + \varepsilon\langle p', \psi(q')q' - \psi(q)q\rangle + \mathrm{O}(\varepsilon^2)$$

where $p' = (q - q')/\left|q - q'\right|$. Hence, $\mathcal{L}_1(q, q') = \langle p', \psi(q')q' - \psi(q)q\rangle$ and $\mathcal{L}_1(q^{\mathrm{h}}_\pm, q^{\mathrm{h}}_\mp) = 0$. Given an unperturbed bi-asymptotic orbit $\mathcal{O} = (m_k)_{k\in\mathbb{Z}} \subset \mathcal{S}$, $m_k = (q_k, p_k) = f^k(m)$, we introduce the notation $\nu_k = \langle p_k - p_{k+1}, q_k\rangle$ and $\psi_k = \psi(q_k)$. Then the absolutely convergent series in (2.2) can be rearranged in the following way:

$$\begin{aligned} L(m) &= \sum_{k\in\mathbb{Z}} \mathcal{L}_1(q_k, q_{k+1}) = \sum_{k\in\mathbb{Z}} \langle p_{k+1}, \psi_{k+1}q_{k+1} - \psi_k q_k\rangle \\ &= \sum_{k\in\mathbb{Z}} \langle p_k - p_{k+1}, \psi_k q_k\rangle = \sum_{k\in\mathbb{Z}} \nu_k\psi_k \end{aligned}$$

which is the simple formula for the Melnikov potential we were looking for.

The separatrix has four connected components, but symmetric perturbations cause the same effect on any of them. Therefore, we can restrict our study to one component, namely $\mathcal{S}^+_+ = \{m(r) : r \in (0, +\infty)\}$. If we take the variable r as a natural coordinate over $\mathcal{S}^+_+$, the Melnikov potential can be written in the following form:

$$L : (0, +\infty) \to \mathbb{R} \qquad L(r) = L(m(r)) = \sum_{k\in\mathbb{Z}} \nu(\lambda^k r)\psi(\lambda^k r)$$

where $\nu, \psi : (0, +\infty) \to \mathbb{R}$ are the continuous versions of the sequences ν_k, ψ_k; that is,

$$\nu(r) := \langle p(r) + p(\lambda r), q(r) \rangle = \frac{2a\tau^2(r)}{\tau(\lambda^{-1/2}r)\tau(\lambda^{1/2}r)} = \frac{2a(1+r^2)^2}{(1+\lambda^{-1}r^2)(1+\lambda r^2)}$$

$$\psi(r) := \psi(q(r)) = \tfrac{1}{2}P\big(\tau_y^2(r)/\tau^2(r)\big) = \tfrac{1}{2}P\left(4r^2/(1+r^2)^2\right)$$

see lemma 4.1 and relation (4.5). The final result can be stated as follows.

Lemma 4.2. *The Melnikov potential associated with the billiard inside (4.6) consists of four copies of the function $L : (0, +\infty) \to \mathbb{R}$ defined by*

$$L(r) = a\sum_{k\in\mathbb{Z}} \ell(\lambda^k r) \qquad \ell(r) = \frac{\tau^2(r)}{\tau(\lambda^{-1/2}r)\tau(\lambda^{1/2}r)} P\left(\frac{\tau_y^2(r)}{\tau^2(r)}\right) \tag{4.7}$$

where the characteristic multiplier λ and the polynomials τ, τ_y are defined in lemma 4.1. Moreover, the function L is invariant: $L(\lambda r) = L(r)$, and symmetric: $L(r) = L(1/r)$. In particular, $\tilde{r} = 1$ and $\hat{r} = \sqrt{\lambda}$ are critical points of L: $L'(\tilde{r}) = L'(\hat{r}) = 0$.

Once we have a suitable expression for the Melnikov potential, we are ready to study the splitting of separatrices, both from a qualitative and quantitative point of view. In the qualitative part, we shall establish a non-integrability criterion. In the quantitative part, we shall prove that the perturbed axial bi-asymptotic orbits are transverse *close to the flat limit*; that is, when the unperturbed ellipse is close to a segment whose extrema are the foci: $\beta = b^2/a^2 \to 0^+$, so that $e = \sqrt{1-\beta} \to 1^-$ and $\lambda \to +\infty$.

4.4. Non-integrability and splitting

We begin with the non-integrability criterion. It is clear that the quadratic perturbations preserve the elliptic (and consequently, integrable) character of the billiard inside an ellipse. It is remarkable that, for entire symmetric perturbations, the converse also holds.

Theorem 4.2 (see [DR96]). *An entire symmetric perturbation of a non-circular ellipse gives rise to an integrable billiard if and only if it is quadratic.*

Proof. Let P be an entire function. It suffices to check that if P is not linear, then the Melnikov potential (4.7) is non-constant, see comment C3 in section 2.3. (Indeed, one can check that $L(r) \equiv p_1 b^2/c$ when $P(s) = p_1 s$.)

The key point is to observe that the Melnikov potential can be analytically extended to the complex plane and to study its complex singularities, since it can be constant only when it has no complex singularities.

Let $r^* = \mathrm{i}$ be the imaginary unit. It is clear that $L(r) - a\ell(r) = a\sum_{0\neq k\in\mathbb{Z}} \ell(\lambda^k r)$ is analytic at r^*. On the other hand, $\ell(r)$ is analytic at r^* if and only if P is linear. Consequently, if P is not linear, then $L(r)$ is not analytic at r^* and therefore is non-constant. □

Birkhoff conjectured that elliptic billiards are the unique integrable smooth convex billiards. Theorem 4.2 can be considered as a local version of this conjecture around non-circular ellipses in the set of entire symmetric curves.

From the proof and property L1 in section 2.3, we deduce that the separatrix splits under any non-quadratic entirely symmetric perturbation.

At this point, we have finished the review of known results. Next, we present some new results concerning transversality close to the flat limit.

4.5. Transversality close to the flat limit

The transversality conditions are obtained from the asymptotic behaviour of the Melnikov potential at its critical points $\tilde{r} = 1$ and $\hat{r} = \sqrt{\lambda}$ when

$$\beta := \frac{b^2}{a^2} = \frac{4\lambda}{(1+\lambda)^2} \to 0^+.$$

This behaviour is contained in the following lemma. In the spatial case a similar study will be carried out, although with more cumbersome computations (see appendix A). Here, we shall only sketch the proof to avoid tedious repetitions.

Lemma 4.3. *Let* $d, \tilde{d}, \hat{d} : (0,1) \to \mathbb{R}$ *be the functions*

$$d(\beta) = a^{-1}[L(\tilde{r}) - L(\hat{r})] \qquad \tilde{d}(\beta) = a^{-1}L''(\tilde{r}) \qquad \hat{d}(\beta) = a^{-1}L''(\hat{r}).$$

If the perturbation (4.6) is analytic, then

$$d(\beta) = [P(1) - P'(0)]\beta + \mathrm{O}(\beta^2)$$

$$\tilde{d}(\beta) = 2[P(1) - P'(1)]\beta + 2[P'(0) - P(1)]\beta^2 + \mathrm{O}(\beta^4)$$

$$\hat{d}(\beta) = P''(0)\beta^3 + \mathrm{O}(\beta^4).$$

Proof. Let $P(s) = \sum_{j\geqslant 1} p_j s^j$ be the Taylor expansion of the perturbation around zero. Then the function $\ell(r)$ defined in (4.7) can be written as follows:

$$\ell(r) = \sum_{j\geqslant 1} p_j \ell_j(r) \qquad \ell_j(r) = \frac{(2r)^{2j}}{(1+\lambda^{-1}r^2)(1+r^2)^{2(j-1)}(1+\lambda r^2)}.$$

We note that $0 < \ell_j(\lambda^k) = \ell_j(\lambda^{-k}) \leqslant (2\lambda^{-|k|})^{2j}$, for all integers $k, j \geqslant 1$. On the other hand, $\ell(\lambda^0) = \ell(1) = 4\lambda P(1)/(1+\lambda)^2 = P(1)\beta$. Therefore,

$$\begin{aligned} \left|a^{-1}L(\tilde{r}) - P(1)\beta\right| &\leqslant \sum_{k\neq 0} \left|\ell(\lambda^k)\right| \leqslant \sum_{j\geqslant 1}\sum_{k\neq 0} \left|p_j \ell_j(\lambda^k)\right| \\ &\leqslant 2\sum_{j\geqslant 1} \frac{2^{2j}\left|p_j\right|}{\lambda^{2j}-1} = \mathrm{O}(\lambda^{-2}) = \mathrm{O}(\beta^2) \end{aligned}$$

using that $\lambda^{-1} \sim \beta/4$ when $\beta \to 0$.

The formulae for $L(\hat{r})$, $L''(\tilde{r})$, and $L''(\hat{r})$ can be obtained in the same way. □

We recall that the Melnikov potential is constant for quadratic perturbations or, equivalently, for a linear P: $P(s) = p_1 s$, which satisfies $P(1) = P'(0) = P'(1)$ and $P''(0) = 0$.

Corollary 4.1. *If P is an analytic function such that $P(1) \neq P'(1)$ or $P(1) \neq P'(0)$ (respectively, $P''(0) \neq 0$) (respectively, $P(1) \neq P'(0)$) and the ellipse is narrow enough, then for a small enough perturbation the y-axial bi-asymptotic orbits become transverse (respectively, the x-axial bi-asymptotic orbits become transverse) (respectively, the length of the x-axial bi-asymptotic orbits is different from the length of the y-axial ones).*

Proof. Let P be an analytic function such that $P(1) \neq P'(1)$ or $P(1) \neq P'(0)$. Then there exists a constant $\beta_0 > 0$ such that $a^{-1}L''(\tilde{r}) = \tilde{d}(\beta) \neq 0$ for all $\beta \in (0, \beta_0)$, see the expansion of $\tilde{d}(\beta)$ in lemma 4.3. This is equivalent to saying that the points $\tilde{m}^{\sigma}_{\varsigma} = \varsigma m(\sigma\tilde{r})$ are non-degenerate critical points of the Melnikov potential if the ellipse is narrow enough for all $\varsigma, \sigma \in \{-, +\}$. Thus, taking into account that the billiard orbits passing by the points $\tilde{m}^{\sigma}_{\varsigma}$ are the y-axial bi-asymptotic ones, the first claim of the corollary follows from property L3 in section 2.3. The other claims are established using similar arguments. □

In some degenerate cases, this corollary cannot be applied and further computations are necessary to obtain the following general theorem.

Theorem 4.3. *If the ellipse is narrow enough, under any non-quadratic analytic symmetric small enough perturbation, all the axial bi-asymptotic orbits become transverse, and the length of the x-axial bi-asymptotic orbits is different from the length of the y-axial ones.*

Proof. Let P be a nonlinear analytic function such that $P(0) = 0$, and let $j \geqslant 2$ be the smallest integer such that $p_j = P^{(j)}(0)/j! \neq 0$. After some rather tedious, but simple, manipulations, it turns out that the following estimates hold:

$$d(\beta) = [P(1) - P'(0)]\beta - p_j\beta^j + \mathrm{O}(\beta^{j+1})$$

$$\tilde{d}(\beta) = 2[P(1) - P'(1)]\beta + 2[P'(0) - P(1)]\beta^2 + 2^{3-2j} j^2 p_j \beta^{2j} + \mathrm{O}(\beta^{2j+1})$$

$$\hat{d}(\beta) = (j^2 - j)p_j\beta^{j+1} + \mathrm{O}(\beta^{j+2}).$$

Hence, the functions d, $\tilde{d}$ and $\hat{d}$ are non-zero for $0 < \beta \ll 1$, and the theorem follows. □

This theorem is rather powerful, in the sense that it establishes that, close to the flat limit, an analytic perturbation is either quadratic and preserves the separatrix, or breaks up the separatrix in a transverse way. Behind the chunk of computations used to prove the above theorem, the main idea relies on the fact that the functions $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ can be extended analytically to the flat limit $\beta = 0$. Hence, it suffices to prove that the Taylor expansions of $d(\beta)$, $\tilde{d}(\beta)$ and $\hat{d}(\beta)$ around $\beta = 0$ have some non-zero Taylor coefficient in order to find that these functions do not vanish for small enough, but positive, values of β. (The zeros of analytic functions are isolated.)

On the other hand, this discussion motivates the following question: can $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ be analytically extended to the circular limit $\beta = 1$; that is, when the unperturbed ellipse is close to a circumference? The answer is 'no'. We do not pursue a detailed explanation of this claim, but only sketch a counter-example. Similar results hold for any polynomial perturbation.

4.6. A quartic perturbation

The simplest non-quadratic symmetric perturbation is, of course, a quartic one. So, let us consider the quartic perturbation

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 + \varepsilon \frac{y^4}{b^4} \right\}. \tag{4.8}$$

Then $P(s) = s^2$ and the Melnikov potential (4.7) becomes

$$L(r) = a \sum_{k \in \mathbb{Z}} \ell(\lambda^k r) \qquad \ell(r) = \frac{16r^4}{(1 + \lambda^{-1}r^2)(1 + r^2)^2(1 + \lambda r^2)}. \tag{4.9}$$

It is clear that the function $t \mapsto L(\mathrm{e}^t)$ is *elliptic*; that is, it is meromorphic in the whole complex plane and has two complex periods independent over the reals:

$$\omega_1 := \ln \lambda \qquad \omega_2 := \pi \mathrm{i}.$$

This crucial observation goes back to the work of Levallois [Lev97], and allows us to apply the powerful theory of elliptic functions to our problem, and to compute the Melnikov potential (4.9) explicitly in terms of the classical *Jacobian elliptic functions*. Concretely, in appendix B it is shown that

$$a^{-1}L(r) = \text{constant} + \frac{4\lambda}{(\lambda-1)^2}\left(\frac{2K}{\ln\lambda}\right)^2 \mathrm{dn}^2\left(\frac{2K\log r}{\ln\lambda}, k\right) \tag{4.10}$$

where, if $K = K(k)$ is the *complete elliptic integral of the first kind*, the *modulus* $k \in (0, 1)$ of the Jacobian elliptic function $\mathrm{dn}(u) = \mathrm{dn}(u, k)$ is determined by imposing the condition $K(k') = K(k)\pi/\ln\lambda$ with $k^2 + k'^2 = 1$. As a corollary, one can obtain the following lemma.

Lemma 4.4. *The critical points of the Melnikov potential (4.9) are the points in the set* $\lambda^{\mathbb{Z}/2}$, *all of which are non-degenerate. The functions* $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ *associated with the quartic perturbation (4.8) are*

$$d(\beta) = \frac{4\pi^2\lambda}{(\lambda-1)^2\ln^2\lambda}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4$$

$$\tilde{d}(\beta) = \frac{-8\pi^4\lambda}{(\lambda-1)^2\ln^4\lambda}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4\left(\sum_{n\in\mathbb{Z}} q^{n^2}\right)^4$$

$$\hat{d}(\beta) = \frac{8\pi^4}{(\lambda-1)^2\ln^4\lambda}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4\left(\sum_{n\in\mathbb{Z}} (-q)^{n^2}\right)^4$$

where $q = \mathrm{e}^{-\pi^2/\ln\lambda}$. *In particular, the functions* $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ *never vanish.*

For the sake of brevity, we skip the proof. (We shall describe a similar proof with full details in the spatial case, see lemma 5.4.)

Several interesting results can easily be deduced from this lemma. Let us mention only a couple of them. First, we present a result on the number of primary bi-asymptotic orbits of symmetric quartic perturbations of non-circular ellipses, in which no ‘flat’ hypothesis is required.

Theorem 4.4. *The billiard inside a small enough quartic symmetric perturbation of a non-circular ellipse has only eight primary bi-asymptotic orbits (the axial ones), which are transverse. Moreover, the x-axial and y-axial bi-asymptotic orbits have different lengths.*

Proof. It is a consequence of properties L3 and L4 in section 2.3 and the previous lemma. □

Next, we address the non-analyticity of the functions $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ at the circular limit $\beta = 1$. The formulae in lemma 4.4 imply that the quantities $d(\beta)$, $\tilde{d}(\beta)$ and $\hat{d}(\beta)$ are exponentially small in $\ln\lambda \sim 2\sqrt{1-\beta}$ when $\beta \to 1^-$. For instance,

$$d(\beta) \sim \frac{4\pi^2}{(1-\beta)^2}\exp\left(-\frac{\pi^2}{2\sqrt{1-\beta}}\right) \qquad (\beta \to 1^-).$$

Similar formulae hold for $\tilde{d}(\beta)$ and $\hat{d}(\beta)$. Hence, the functions $d, \tilde{d}, \hat{d} : (0, 1) \to \mathbb{R}$ cannot be analytic at $\beta = 1$.

To end this section, let us introduce a problem arising from the exponential smallness of the function $d(\beta)$ near $\beta = 1$. For *regular* perturbations (β remains fixed, whereas $\varepsilon \to 0$), the Melnikov term $\varepsilon a d(\beta)$ is the dominant term of the lobe area between the y-axial and x-axial bi-asymptotic orbits; that is, of the difference of the lengths of these two kinds of orbits. In contrast, for *singular* perturbations ($\beta \to 1^-$ and $\varepsilon \to 0$), one is confronted with the difficult problem of justifying the following exponentially small asymptotic expression provided by the Poincaré–Melnikov method:

$$\text{lobe area} = \text{difference of lengths} \sim \frac{4\pi^2 a\varepsilon}{(1-\beta)^2} \exp\left(-\frac{\pi^2}{2\sqrt{1-\beta}}\right) \qquad (\varepsilon \to 0, \beta \to 1^-).$$

We refer to [DR99] for a brief account of results on singular splittings for analytic area-preserving maps. The perturbations of elliptic planar billiards close to the circular limit are still an open problem in that subject.

5. The spatial case

In this section, we extend to the spatial case, with the appropriate modifications, the lemmas, propositions and theorems concerning the planar case presented in the previous section. There are two important exceptions: theorem 4.3 and the final comments concerning the quartic perturbation.

Let $a \geqslant b \geqslant c$ be the semi-axes of a given ellipsoid. The ellipsoid will be called *generic*, *prolate*, *oblate* or *spherical* when $a > b > c$, $a > b = c$, $a = b > c$ or $a = b = c$, respectively. Oblate ellipsoids and spheres do not fall into our set-up, because they have a continuous family of diameters. Prolate ellipsoids have already been considered in [DR98] as a first step in gaining insight into the spatial case, since they are much simpler than generic ellipsoids.

We will consider a *generic ellipsoid*

$$Q = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \right\} \qquad a > b > c > 0 \qquad (5.1)$$

whose diameter is given by the chord joining the vertices $(-a, 0, 0)$ and $(a, 0, 0)$. We will denote the set formed by the two-periodic points associated with the diameter by

$$\mathcal{M}^{\mathrm{h}} = \{m_+^{\mathrm{h}}, m_-^{\mathrm{h}}\} \qquad m_\pm^{\mathrm{h}} = (q_\pm^{\mathrm{h}}, p_\pm^{\mathrm{h}}) \qquad q_\pm^{\mathrm{h}} = (\pm a, 0, 0) \qquad p_\pm^{\mathrm{h}} = (\pm 1, 0, 0).$$

5.1. Geometry of the bi-asymptotic set

We are going to see that m_+^{h} and m_-^{h} are hyperbolic two-periodic points of the spatial elliptic billiard map f whose unstable and stable invariant manifolds are doubled.

The sets $\mathcal{W}$, $\mathcal{W}^{\mathrm{u}}$, $\mathcal{W}^{\mathrm{s}}$, $\mathcal{W}_\pm^{\mathrm{u}}$, $\mathcal{W}_\pm^{\mathrm{s}}$ have the same meaning, (4.2) and (4.3), as in the planar case. Let

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

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 $\kappa > a$. We now consider the degenerate cases $\kappa = a, b, c$.

On the one hand, 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*

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

On the other hand, 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*

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

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

We shall use the term *focal conics* when we refer to both the focal ellipse and the focal hyperbola. The four points on the intersection of the focal hyperbola with the ellipsoid are called *umbilical points* and their coordinates are $(\pm\tilde{x}, \tilde{y}, \pm\tilde{z})$, where

$$\tilde{x} = -a\frac{\sqrt{a^2 - b^2}}{\sqrt{a^2 - c^2}} \qquad \tilde{y} = 0 \qquad \tilde{z} = c\frac{\sqrt{b^2 - c^2}}{\sqrt{a^2 - c^2}}. \tag{5.2}$$

The four unitary velocities determined by the couple of asymptotes of the focal hyperbola will be called *asymptotic velocities* and their coordinates are $(\pm\hat{u}, \hat{v}, \pm\hat{w})$, where

$$\hat{u} = -\frac{\sqrt{a^2 - b^2}}{\sqrt{a^2 - c^2}} \qquad \hat{v} = 0 \qquad \hat{w} = \frac{\sqrt{b^2 - c^2}}{\sqrt{a^2 - c^2}}. \tag{5.3}$$

(One must beware of the terms ‘asymptotic’ and ‘bi-asymptotic’; the former has a purely geometric nature, whereas the latter describes a dynamical behaviour.)

The integrability of spatial elliptic billiards is closely related to the following nice property: *any segment (or its prolongation) of a billiard trajectory inside the ellipsoid $Q = Q(0)$ is tangent*[4] *to two fixed confocal quadrics $Q(\kappa_1)$ and $Q(\kappa_2)$*, see [Tab95, section 2.3]. The quantities κ_1 and κ_2, regarded as functions of the starting point of the billiard orbit, are first integrals of the elliptic billiard map. Hence, the level sets

$$\mathcal{M}_\kappa = \mathcal{M}_{(\kappa_1,\kappa_2)} = \{m = (q, p) \in \mathcal{M} : \ q + \langle p\rangle \text{ is tangent to } Q(\kappa_1) \text{ and } Q(\kappa_2)\}$$

are invariant by f, where $q + \langle p\rangle$ denotes the line passing by q with direction p.

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

$$I_{\mathrm{x}}(m) = \frac{(\kappa_1^2(m) - a^2)(\kappa_2^2(m) - a^2)}{(a^2 - b^2)(a^2 - c^2)} = u^2 + \frac{(xv - yu)^2}{a^2 - b^2} + \frac{(xw - zu)^2}{a^2 - c^2}$$

$$I_{\mathrm{y}}(m) = \frac{(\kappa_1^2(m) - b^2)(\kappa_2^2(m) - b^2)}{(b^2 - a^2)(b^2 - c^2)} = v^2 - \frac{(yu - xv)^2}{a^2 - b^2} + \frac{(yw - zv)^2}{b^2 - c^2}$$

$$I_{\mathrm{z}}(m) = \frac{(\kappa_1^2(m) - c^2)(\kappa_2^2(m) - c^2)}{(a^2 - c^2)(b^2 - c^2)} = w^2 - \frac{(zu - xw)^2}{a^2 - c^2} - \frac{(zv - yw)^2}{b^2 - c^2}$$

where $m = (q, p)$ with $q = (x, y, z) \in Q$ and $p = (u, v, w) \in \mathbb{S}^2$ (see, for instance, [Tab95, section 2.3]). These first integrals are dependent: $I_{\mathrm{x}}(m) + I_{\mathrm{y}}(m) + I_{\mathrm{z}}(m) = u^2 + v^2 + w^2 \equiv 1$, but skipping one of them the rest are independent almost-everywhere. This shows that spatial elliptic billiards 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 over where the quantities κ_1 and κ_2 range. We can assume that $\kappa_1 \leqslant \kappa_2$. Then $\kappa_1 > 0$ and $\kappa_2 \leqslant a$. On the other hand, a line cannot 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.

Next, we shall explain what we mean by tangency of a billiard trajectory to the confocal quadrics $Q(\kappa_1)$ and $Q(\kappa_2)$ in the degenerate situations. There are two kinds of degenerations.

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

First, the coincidence of the confocal quadrics: $\kappa_1 = \kappa_2$, where we shall say that the trajectory is *bi-tangent to* $\mathcal{Q}(\kappa_1) = \mathcal{Q}(\kappa_2)$, and second, the degeneration of a confocal quadric: $\kappa_j = a, b$ or c, for some $j = 1, 2$. All of these situations are covered by the following definitions, easily deduced by limit procedures passing from generic cases (in which the definition is clear) to degenerate ones. A line is:

- *bi-tangent to* $\mathcal{Q}(\kappa)$, $c < \kappa < b$, when it is a generatrix of the ruled quadric $\mathcal{Q}(\kappa)$;
- *tangent to* $\mathcal{Q}(c)$ when it is contained in the xy-plane or it intersects $\mathcal{E}_{\text{focal}}$;
- *bi-tangent to* $\mathcal{Q}(c)$ when it is contained in the xy-plane and it is tangent to $\mathcal{E}_{\text{focal}}$;
- *tangent to* $\mathcal{Q}(b)$ when it is contained in the xz-plane or it intersects $\mathcal{H}_{\text{focal}}$;
- *bi-tangent to* $\mathcal{Q}(b)$ when it is contained in the xz-plane and it is tangent to $\mathcal{H}_{\text{focal}}$;
- *tangent to* $\mathcal{Q}(a)$ when it is contained in the yz-plane.

Obviously, these intersections and tangencies are understood in a projective sense. In particular, it must be retained in what follows that *any line parallel to an asymptote of the focal hyperbola intersects the focal hyperbola* (at an improper point, of course).

Now, we are ready to prove that the invariant manifolds are doubled.

Proposition 5.1. *$\mathcal{W} = \mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}} = \mathcal{M}_{(c,b)}$, where*

$$\mathcal{M}_{(c,b)} = \{m = (q, p) \in \mathcal{M} : q + \langle p \rangle \textit{ intersects } \mathcal{E}_{\text{focal}} \textit{ and } \mathcal{H}_{\text{focal}}\}$$

$$= \{m \in \mathcal{M} : I_{\mathrm{y}}(m) = I_{\mathrm{z}}(m) = 0\}.$$

Proof. The first formula for the level set $\mathcal{M}_{(c,b)}$ is obtained simply by noting that a line is tangent to $\mathcal{Q}(c)$ and $\mathcal{Q}(b)$ if and only if it intersects both focal conics.

The second one follows from the relations between the two families of first integrals.

The inclusion $\mathcal{W} \subset \mathcal{W}^{\mathrm{u,s}}$ is obvious.

The inclusion $\mathcal{W}^{\mathrm{u,s}} \subset \mathcal{M}_{(c,b)}$ is a direct consequence of $\mathcal{M}^{\mathrm{h}} \subset \mathcal{M}_{(c,b)}$.

It remains to prove that $\mathcal{M}_{(c,b)} \subset \mathcal{W}$. Let $\mathcal{O} = (m_k)_{k\in\mathbb{Z}} \subset \mathcal{M}_{(c,b)}$.

If the lines $l_k = q_k + \langle p_k \rangle$, $m_k = (q_k, p_k)$, are contained in the xy-plane, then $\mathcal{O} \subset \mathcal{W}$ because $\mathcal{O}$ is a bi-asymptotic billiard orbit inside the ellipse $\mathcal{Q}_{\mathrm{xy}} := \mathcal{Q} \cap \{z = 0\}$. It suffices to note that the xy-plane cuts the focal hyperbola at the foci of the ellipse $\mathcal{Q}_{\mathrm{xy}}$.

If the lines l_k are contained in the xz-plane, a similar argument shows that $\mathcal{O} \subset \mathcal{W}$, since the xz-plane cuts the focal ellipse at the foci of the ellipse $\mathcal{Q}_{\mathrm{xz}} := \mathcal{Q} \cap \{y = 0\}$.

If the lines l_k are contained neither in the xy-plane nor in the xz-plane, they intersect each focal conic at a single point. Let

$$q_k^{\mathcal{E}} := \left(x_k^{\mathcal{E}}, y_k^{\mathcal{E}}, 0\right) := \mathcal{E}_{\text{focal}} \cap l_k \qquad q_k^{\mathcal{H}} := \left(x_k^{\mathcal{H}}, 0, z_k^{\mathcal{H}}\right) := \mathcal{H}_{\text{focal}} \cap l_k$$

be these points. They are mapped onto points $q_k^{\mathrm{xy}} \in \mathcal{Q}_{\mathrm{xy}}$ and $q_k^{\mathrm{xz}} \in \mathcal{Q}_{\mathrm{xz}}$ by means of the transformations (note that $x_k^{\mathcal{H}} \neq 0$ for all $k \in \mathbb{Z}$):

$$q_k^{\mathrm{xy}} := \left(x_k^{\mathrm{xy}}, y_k^{\mathrm{xy}}, 0\right) := \left(\frac{a x_k^{\mathcal{E}}}{\sqrt{a^2 - c^2}}, \frac{b y_k^{\mathcal{E}}}{\sqrt{b^2 - c^2}}, 0\right)$$

$$q_k^{\mathrm{xz}} := \left(x_k^{\mathrm{xz}}, 0, z_k^{\mathrm{xz}}\right) := \left(\frac{a\sqrt{a^2 - b^2}}{x_k^{\mathcal{H}}}, 0, (-1)^k c \frac{\sqrt{a^2 - b^2}}{\sqrt{b^2 - c^2}} \frac{z_k^{\mathcal{H}}}{x_k^{\mathcal{H}}}\right).$$

In [Fed99] it is stated that the sequences $(q_k^{\mathrm{xy}})_{k\in\mathbb{Z}}$ and $(q_k^{\mathrm{xz}})_{k\in\mathbb{Z}}$ are bi-asymptotic billiard configurations inside the ellipses $\mathcal{Q}_{\mathrm{xy}}$ and $\mathcal{Q}_{\mathrm{xz}}$, respectively. Hence, $\lim_{|k|\to\infty} y_k^{\mathrm{xy}} = \lim_{|k|\to\infty} z_k^{\mathrm{xz}} = 0$ or, equivalently, $\lim_{|k|\to\infty} y_k^{\mathcal{E}} = \lim_{|k|\to\infty} z_k^{\mathcal{H}} = 0$. This implies that the line l_k tends to the x-axis when $|k| \to \infty$, and so $\mathcal{O} \subset \mathcal{W}$. □

We can summarize the above results in the following remarks:

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

In order to better understand the structure of the bi-asymptotic set, we describe the set of lines passing through a given impact point (parallel to a given unitary velocity) that intersect both focal conics. Such lines will be called *bi-asymptotic*.

Given $q \in \mathcal{Q}$, four different cases arise:

(1) if q is an umbilical point, there is *a continuous family* of bi-asymptotic lines through q: the generatrices of the cone with vertex at q and containing the focal ellipse;
(2) if $q = q^{\mathrm{h}}_{\pm}$, there is just *one*: the x-axis;
(3) if q is contained in the xy-plane or in the xz-plane, but q is not umbilical and $q \neq q^{\mathrm{h}}_{\pm}$, there are *two* such lines: those passing through the foci of the ellipse $\mathcal{Q}_{\mathrm{xy}}$ or $\mathcal{Q}_{\mathrm{xz}}$, respectively; and
(4) if q is contained neither in the xy-plane nor in the xz-plane, there are *four* bi-asymptotic lines through q.

Given $p \in \mathbb{S}^2$, there arise also four different cases:

(1′) if p is an asymptotic velocity, there is *a continuous family* of bi-asymptotic lines parallel to p: the generatrices of the cylinder with direction p and containing the focal ellipse;
(2′) if $p = p^{\mathrm{h}}_{\pm}$, there is just *one*: the x-axis;
(3′) if p is parallel to the xy-plane or to the xz-plane, but p is not asymptotic and $p \neq p^{\mathrm{h}}_{\pm}$, there are *two* such lines: those passing through the foci of the ellipse $\mathcal{Q}_{\mathrm{xy}}$ or $\mathcal{Q}_{\mathrm{xz}}$, respectively; and
(4′) if p is not parallel to the xy-plane nor in the xz-plane, there are *four* bi-asymptotic lines parallel to p.

The pictures displayed in figure 9 help to visualize cases (4) and (4′). It suffices to realize that the projections of the focal conics from a point (or with a direction) onto the plane of the paper have four points of intersection in those cases.

5.2. Dynamics of the bi-asymptotic set

Here, we shall describe the billiard dynamics on the bi-asymptotic set. To be more precise, we shall linearize the billiard motion on the invariant manifolds $\mathcal{W}^{\mathrm{u}}_{\pm}$ and $\mathcal{W}^{\mathrm{s}}_{\pm}$ introduced in (4.3); that is, we shall compute an analytic conjugation between the restrictions $f\colon \mathcal{W}^{\mathrm{u}}_{\pm} \to \mathcal{W}^{\mathrm{u}}_{\mp}$ (respectively, $f\colon \mathcal{W}^{\mathrm{s}}_{\pm} \to \mathcal{W}^{\mathrm{s}}_{\mp}$) and the linear map $r \mapsto \Lambda r$ (respectively, $r \mapsto \Lambda^{-1} r$), where the entries of the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ are the *characteristic multipliers* of the hyperbolic periodic set $\mathcal{M}^{\mathrm{h}}$: $\mathrm{Spec}[\,\mathrm{d}f(\mathcal{M}^{\mathrm{h}})] = \{\lambda_1, \lambda_2, 1/\lambda_1, 1/\lambda_2\}$ and $|\lambda_1|, |\lambda_2| > 1$. Such conjugations will be called *natural parametrizations*.

Remark 5.1. In general, resonances between characteristic multipliers are an obstruction for the analytic linearization of the dynamics on the unstable and stable invariant manifolds. In our setting, the algebraic integrability of elliptic billiard maps implies not only the existence of the natural parametrizations, but also their rational character.

To parametrize the bi-asymptotic motions, let us first recall that the real invariant tori of algebraic completely integrable systems can be extended to complex algebraic tori (Abelian varieties) related to regular algebraic curves, and the complex parametrizations of the tori are given by theta-functions of the above-mentioned curves (see, e.g., [AvM89]). In the case of algebraic completely integrable maps the dynamics on the complex tori becomes just a translation by a constant vector (see [Ves88]). On the other hand, some complex invariant manifolds turn out to be non-compact Abelian varieties related to singular Riemann surfaces. These invariant manifolds also have a shift dynamics, and their parametrizations are given by generalized theta-functions, which reduce to tau-functions, finite sums of exponents, in the most degenerate cases. A description of tau-functions can be found in [Mum84].

Presumably, the most celebrated examples of algebraic completely integrable maps are the elliptic billiard maps, which have been studied in several papers using quite different approaches. The generic and bi-asymptotic elliptic billiard motions were first integrated in [Ves88] and [Fed99], respectively. The following results are contained in [Fed99].

If $\mathcal{M}_\kappa$ is a (real) generic level set with complex extension $\mathcal{M}_\kappa^{\mathbb{C}}$, there exists a map $m_\kappa = (q_\kappa, p_\kappa) : \mathbb{C}^2 \to \mathcal{M}_\kappa^{\mathbb{C}}$ and a constant shift $h_\kappa \in \mathbb{R}^2$ such that $f(m_\kappa(t)) = m_\kappa(t + h_\kappa)$. The map $m_\kappa = (q_\kappa, p_\kappa)$ has the form $q_\kappa(t) = D\chi_\kappa(t)$, $p_\kappa(t) = \chi_\kappa(t - h_\kappa/2)$, where

$$\chi_\kappa = (\alpha_{\mathrm{x}}\theta_{\mathrm{x}}/\theta, \alpha_{\mathrm{y}}\theta_{\mathrm{y}}/\theta, \alpha_{\mathrm{z}}\theta_{\mathrm{z}}/\theta) \qquad D = \mathrm{diag}(a, b, c)$$

and $\theta, \theta_{\mathrm{x}}, \theta_{\mathrm{y}}, \theta_{\mathrm{z}}$ are some theta-functions with half-integer theta-characteristics related to the hyperelliptic curve of genus two,

$$\Gamma_\kappa = \Gamma_{(\kappa_1,\kappa_2)} = \left\{\omega^2 = -(\mu - a^2)\cdot(\mu - b^2)\cdot(\mu - c^2)\cdot(\mu - \kappa_1^2)\cdot(\mu - \kappa_2^2)\right\}$$

and $\alpha_{\mathrm{x}}, \alpha_{\mathrm{y}}, \alpha_{\mathrm{z}}$ are constants depending on the moduli of Γ_κ only.

The generic cases correspond to regular hyperelliptic curves; that is, when the numbers $\kappa_1, \kappa_2, a, b, c$ are all different. When $\kappa = (c, b)$, the above theta-functions degenerate into some tau-functions $\tau, \tau_{\mathrm{x}}, \tau_{\mathrm{y}}, \tau_{\mathrm{z}}$. Their exact expressions can be found in [Fed99].

From a dynamical point of view, it is better to adopt a multiplicative notation. Thus, we have substituted the additive variable $t = (t_1, t_2)$ by the multiplicative one $r = (r_1, r_2)$ defined by $r_1 = \exp(t_1)$ and $r_2 = \exp(t_2)$. Then the tau-functions $\tau, \tau_{\mathrm{x}}, \tau_{\mathrm{y}}, \tau_{\mathrm{z}}$ become the tau-polynomials to be introduced in lemma 5.1, and the shifts $t_j \mapsto t_j + h_j$ read as $r_j \mapsto \lambda_j r_j$ with $\lambda_j = \exp(h_j)$, $j = 1, 2$.

All of these comments are intended to clarify the origin of the polynomials $\tau, \tau_{\mathrm{x}}, \tau_{\mathrm{y}}, \tau_{\mathrm{z}}$. The proof of the lemma, although self-contained and short, does not illuminate it at all.

The following notation is used in the statement of lemma 5.1.

Let $\hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ be the extended real line. Let i be the imaginary unit. Let $\mathrm{I} : \hat{\mathbb{R}}^2 \to \hat{\mathbb{R}}^2$ be the involution $\mathrm{I}(r_1, r_2) = (r_1^{-1}, r_2^{-1})$, where $0^{-1} = \infty$ and $\infty^{-1} = 0$. Henceforth, if $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ is a diagonal matrix, s is a real number and m is a map defined on $\hat{\mathbb{R}}^2$ or $\mathbb{R}^2$, we shall denote by $m \circ \Lambda^s$ the map $r \mapsto m(\Lambda^s r) = m(\lambda_1^s r_1, \lambda_2^s r_2)$.

Lemma 5.1. *Let* $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ *be the diagonal matrix whose entries are the characteristic multipliers*

$$\lambda_1 = \frac{1 + e_1}{1 - e_1} \qquad \lambda_2 = \frac{1 + e_2}{1 - e_2}$$

$$e_1 = \sqrt{1 - \beta_1} \qquad e_2 = \sqrt{1 - \beta_2}$$

$$\beta_1 = \frac{b^2}{a^2} \qquad \beta_2 = \frac{c^2}{a^2}.$$

Let $\alpha^2 = (e_2 + e_1)/(e_2 - e_1)$ *with* $\alpha > 1$. *Let* $\tau, \tau_{\rm x}, \tau_{\rm y}, \tau_{\rm z}$ *be the tau-polynomials*

$$\tau(r) = 1 + r_1^2 r_2^2 + \alpha^2(r_1^2 + r_2^2) \qquad \tau_{\rm x}(r) = \tau({\rm i}r)$$
$$\tau_{\rm y}(r) = 2\alpha r_1(1 - r_2^2) \qquad \tau_{\rm z}(r) = 2\alpha r_2(1 + r_1^2).$$

Let $\chi = (\tau_{\rm x}/\tau, \tau_{\rm y}/\tau, \tau_z/\tau) : \hat{\mathbb{R}}^2 \to \hat{\mathbb{R}}^3$. *Let* $D = {\rm diag}(a, b, c)$. *Let* $q = D\chi : \mathbb{R}^2 \to \mathcal{Q}$, $p = \chi \circ \Lambda^{-1/2} : \mathbb{R}^2 \to \mathbb{S}^2$ *and* $m = (q, p) : \mathbb{R}^2 \to \mathcal{M}$.

Then the maps $m_\pm^{\rm u,s} : \mathbb{R}^2 \to \mathcal{M}$ *defined by* $m_\pm^{\rm u} = \pm m$ *and* $m_\pm^{\rm s} = \pm m \circ {\rm I}$ *are natural parametrizations of the invariant manifolds* $\mathcal{W}_\pm^{\rm u}$ *and* $\mathcal{W}_\pm^{\rm s}$*; that is,* $m_\pm^{\rm u,s} : \mathbb{R}^2 \to \mathcal{W}_\pm^{\rm u,s}$ *are analytic diffeomorphisms such that*

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

Proof. First, we note that the rational map χ has the following fundamental properties:

$$\chi(0,0) = (1,0,0) \qquad {\rm rank}[\,{\rm d}\chi(0,0)] = 2$$
$$|\chi| \equiv 1 \qquad \chi \circ \Lambda^{-1/2} + \chi \circ \Lambda^{1/2} = \nu D^{-1}\chi$$

where $\nu : \hat{\mathbb{R}} \to (0, +\infty)$ is the rational function $\nu = 2a\tau^2/(\tau \circ \Lambda^{-1/2} \cdot \tau \circ \Lambda^{1/2})$. These properties can be checked by means of straightforward computations. They will be used repeatedly throughout the proof.

Let $(m_k)_{k\in\mathbb{Z}} \subset \mathcal{M}^{\mathbb{Z}}$, $m_k = (q_k, p_k)$, be a billiard orbit associated with the ellipsoid $\mathcal{Q} = \{q \in \mathbb{R}^3 : |D^{-1}q| = 1\}$. The difference $q_{k+1} - q_k$ of two consecutive impact points has the same direction and sense as the unitary outward velocity p_{k+1}, whereas the difference $p_k - p_{k+1}$ of two consecutive unitary outward velocities is an outward normal vector to the ellipsoid at the impact point q_k. Thus, taking into account that $D^{-2}q_k$ is an outward normal vector to the ellipsoid at q_k, we deduce that there exist a couple of positive numbers μ_k and ν_k such that

$$q_{k+1} - q_k = \mu_k p_{k+1} \qquad p_k - p_{k+1} = \nu_k D^{-2} q_k. \tag{5.4}$$

The converse also holds. A sequence $(m_k)_k \in \mathcal{M}^{\mathbb{Z}}$ is a billiard orbit if and only if it verifies (5.4) for some sequences of positive numbers $(\mu_k)_k$ and $(\nu_k)_k$, which are determined by the conditions $q_k \in \mathcal{Q}$ and $p_k \in \mathbb{S}^2$, or equivalently, by the condition $|D^{-1}q_k| = |p_k| = 1$. Concretely, $\nu_k = \langle p_k - p_{k+1}, q_k\rangle$ and $\mu_k = \langle q_{k+1} - q_k, p_{k+1}\rangle$.

From $\chi(0,0) = (1,0,0)$ and $|\chi| \equiv 1$ we deduce that $q(0,0) = q_+^{\rm h}$, $p(0,0) = p_+^{\rm h}$, and $|D^{-1}q| \equiv 1$, $|p| \equiv 1$. Hence, $m(0,0) = m_+^{\rm h}$ and $m(\mathbb{R}^2) \subset \mathcal{M}$. Besides, the sequences of maps

$$q_k = (-1)^k q \circ \Lambda^k \qquad p_k = (-1)^k p \circ \Lambda^k \qquad \mu_k = \nu \circ \Lambda^{k+1/2} \qquad \nu_k = \nu \circ \Lambda^k$$

verify the billiard equations (5.4). This follows from $\chi \circ \Lambda^{-1/2} + \chi \circ \Lambda^{1/2} = \nu D^{-1}\chi$.

Therefore, $m : \mathbb{R}^2 \to \mathcal{M}$ is a well defined rational map such that $m(0,0) = m_+^{\rm h}$ and $f \circ m = -m \circ \Lambda$. It remains to see that it is a diffeomorphism onto $\mathcal{W}_+^{\rm u}$.

To begin with, it is clear that m maps $\mathbb{R}^2$ onto $\mathcal{W}_+^{\rm u}$:

$$\lim_{k\to-\infty} {\rm dist}\left(f^k(m(r)), f^k(m_+^{\rm h})\right) = \lim_{k\to-\infty} {\rm dist}\left(m(\Lambda^k r), m_+^{\rm h}\right) = {\rm dist}\left(m_+^{\rm h}, m_+^{\rm h}\right) = 0.$$

Besides, $m : \mathbb{R}^2 \to \mathcal{W}_+^{\rm u}$ is a local diffeomorphism at $r = (0,0)$, since ${\rm rank}[\,{\rm d}\chi(0,0)] = 2$. This implies that $m : \mathbb{R}^2 \to \mathcal{W}_+^{\rm u}$ is a global diffeomorphism, because the whole manifold $\mathcal{W}_+^{\rm u}$ is obtained from its local part by iterating the square of the map f.

This proves that $m_+^{\rm u} = m$ is a natural parametrization of $\mathcal{W}_+^{\rm u}$. The other cases can be analysed in a similar way. □

Remark 5.2. Although the map $m = (q, p) : \mathbb{R}^2 \to \mathcal{M}$ is a diffeomorphism onto the invariant manifold $\mathcal{W}_+^{\mathrm{u}} \subset \mathcal{M}$, its components $q : \mathbb{R}^2 \to \mathcal{Q}$ and $p : \mathbb{R}^2 \to \mathbb{S}^2$ are not injective. For instance, q maps all the points $r = (r_1, r_2) \in \mathbb{R}^2$ such that $r_2^2 = 1$ onto umbilical points:

$$q(r_1, \pm 1) = \left(a\frac{1-\alpha^2}{1+\alpha^2}, 0, \frac{\pm 2\alpha c}{1+\alpha^2} \right) = (\tilde{x}, \tilde{y}, \pm\tilde{z})$$

see equation (5.2). Analogously, but using equation (5.3), p maps all the points $r = (r_1, r_2) \in \mathbb{R}^2$ such that $r_2^2 = \lambda_2$ onto asymptotic velocities:

$$p(r_1, \pm\lambda_2^{1/2}) = \left(\frac{1-\alpha^2}{1+\alpha^2}, 0, \frac{\pm 2\alpha}{1+\alpha^2} \right) = (\hat{u}, \hat{v}, \pm\hat{w}).$$

Remark 5.3. The natural parametrizations of the spatial case (lemma 5.1) are related to those of the planar case (lemma 4.1), since

$$q(r_1/\alpha, 0) = \left(a\frac{1-r_1^2}{1+r_1^2}, \frac{2br_1}{1+r_1^2}, 0 \right) \qquad q(0, r_2/\alpha) = \left(a\frac{1-r_2^2}{1+r_2^2}, 0, \frac{2cr_2}{1+r_2^2} \right). \tag{5.5}$$

To explain these relations, let us consider the sections of the ellipsoid

$$\mathcal{Q}_{\mathrm{xy}} = \mathcal{Q} \cap \{z = 0\} \qquad \mathcal{Q}_{\mathrm{xz}} = \mathcal{Q} \cap \{y = 0\} \qquad \mathcal{Q}_{\mathrm{yz}} = \mathcal{Q} \cap \{x = 0\}.$$

If two consecutive impact points are on one of them, all the rest also are. Therefore, the spatial elliptic billiard has three invariant sub-systems with the same properties as a planar elliptic billiard. The section $\mathcal{Q}_{\mathrm{yz}}$ has no interest here, because there are no orbits in the bi-asymptotic set $\mathcal{W}$ whose impact points are entirely contained in the ellipse $\mathcal{Q}_{\mathrm{yz}}$. Returning to the relations (5.5), the bi-asymptotic billiard orbits inside the ellipsoid with $r_2 = 0$ (respectively, $r_1 = 0$) can be viewed as bi-asymptotic orbits of the planar elliptic billiard inside the ellipse $\mathcal{Q}_{\mathrm{xy}}$ (respectively, $\mathcal{Q}_{\mathrm{xz}}$).

5.3. Topology of the bi-asymptotic set

In this subsection, we describe the bifurcation set $\mathcal{B}$ and the separatrix $\mathcal{S} = \mathcal{W} \setminus \mathcal{B}$ of the spatial elliptic billiard. According to the characterization of the bifurcation set of section 2.2, this is accomplished through the study of the map $m : \mathbb{R}^2 \to \mathcal{M}$ at infinity. (Let us recall that, roughly speaking, the bifurcation set is formed by the self-intersections of the bi-asymptotic set.)

The map $m : \mathbb{R}^2 \to \mathcal{M}$ can be evaluated at the *infinity points* (i.e. the points $r = (r_1, r_2) \in \hat{\mathbb{R}}^2$ such that $r_1 = \infty$ or $r_2 = \infty$) by direct substitution, because $m = (q, p) = (D\chi, \chi \circ \Lambda^{-1/2})$ and the components of the map $\chi = (\tau_{\mathrm{x}}/\tau, \tau_y/\tau, \tau_z/\tau)$ are rational fractions. After some trivial computations, we find that the values of these rational fractions $\chi_{\mathrm{x}}(r) := \tau_{\mathrm{x}}(r)/\tau(r)$, $\chi_{\mathrm{y}}(r) := \tau_{\mathrm{y}}(r)/\tau(r)$ and $\chi_{\mathrm{z}}(r) := \tau_{\mathrm{z}}(r)/\tau(r)$ at the infinity points are related to their values at the *zero points* (i.e. the points $r = (r_1, r_2)$ such that $r_1 = 0$ or $r_2 = 0$). We summarize the relations in the following list:

$$\begin{array}{lll} \chi_{\mathrm{x}}(r_1, \infty) = -\chi_{\mathrm{x}}(\bar{r}_1, 0) & \chi_{\mathrm{y}}(r_1, \infty) = -\chi_{\mathrm{y}}(\bar{r}_1, 0) & \chi_{\mathrm{z}}(\cdot, \infty) = \chi_{\mathrm{z}}(\cdot, 0) = 0 \\ \chi_{\mathrm{x}}(\infty, r_2) = -\chi_{\mathrm{x}}(0, \bar{r}_2) & \chi_{\mathrm{y}}(\infty, \cdot) = \chi_{\mathrm{y}}(0, \cdot) = 0 & \chi_{\mathrm{z}}(\infty, r_2) = \chi_{\mathrm{z}}(0, \bar{r}_2) \\ \chi_{\mathrm{x}}(\infty, \infty) = \chi_{\mathrm{x}}(0, 0) & \chi_{\mathrm{y}}(\infty, \infty) = \chi_{\mathrm{y}}(0, 0) & \chi_{\mathrm{z}}(\infty, \infty) = \chi_{\mathrm{z}}(0, 0) \end{array}$$

where $\bar{r}_1 = r_1/\alpha^2$ and $\bar{r}_2 = r_2/\alpha^2$. These relations imply that

$$\{\pm m(\sigma_1 r_1, 0) : r_1 > 0\} =: \mathcal{Y}_\pm^{\sigma_1} := \{\mp m(\sigma_1 r_1, \infty) : r_1 < 0\}$$

$$\{\pm m(0, \sigma_2 r_2) : r_2 > 0\} =: \mathcal{Z}_\pm^{\sigma_2} := \{\mp m(\infty, \sigma_2 r_2) : r_2 > 0\}$$

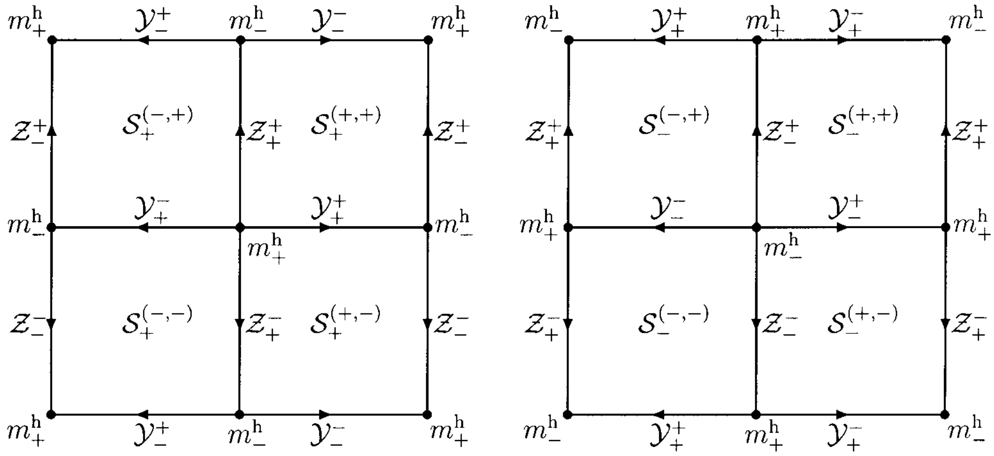


Figure 8. Topological representation of the bi-asymptotic set in the spatial case.

for $\sigma_1, \sigma_2 \in \{-, +\}$. Besides, $m^{\mathrm{h}}_{\pm} = \pm m(0, 0) = \mp m(\infty, 0) = \mp m(0, \infty) = \pm m(\infty, \infty)$.

Therefore, the self-intersections of the bi-asymptotic set take place along the curves

$$\mathcal{Y} := \mathcal{Y}^{+}_{+} \cup \mathcal{Y}^{-}_{+} \cup \mathcal{Y}^{+}_{-} \cup \mathcal{Y}^{-}_{-} \qquad \mathcal{Z} := \mathcal{Z}^{+}_{+} \cup \mathcal{Z}^{-}_{+} \cup \mathcal{Z}^{+}_{-} \cup \mathcal{Z}^{-}_{-}$$

and, of course, at the hyperbolic points $m^{\mathrm{h}}_{\pm}$. After excluding all of them, the separatrix has eight connected components, namely

Proposition 5.2. $\mathcal{S} = \mathcal{S}^{(+,+)}_{+} \cup \mathcal{S}^{(-,+)}_{+} \cup \mathcal{S}^{(-,-)}_{+} \cup \mathcal{S}^{(+,-)}_{+} \cup \mathcal{S}^{(+,+)}_{-} \cup \mathcal{S}^{(-,+)}_{-} \cup \mathcal{S}^{(-,-)}_{-} \cup \mathcal{S}^{(+,-)}_{-}$, *where* $\mathcal{S}^{\sigma}_{\varsigma} := \{\varsigma m(\sigma_1 r_1, \sigma_2 r_2) : r_1, r_2 > 0\}$.

We have depicted a topological representation of the bi-asymptotic set in figure 8. Points and segments with equal labels are identified. All the elements that appear in the figure (points, segments and squares) are invariant under the square of the billiard map; that is, under f^2. The arrows show the sense of the dynamics in the segments. In the squares the dynamics must be compatible with the arrows of the segments. This implies that the points on the segments (that is, on the curves $\mathcal{Y}$ and $\mathcal{Z}$) are heteroclinic points of the map f^2, whereas those in the squares (that is, on the separatrix) are homoclinic.

Remark 5.4. It is interesting to compare our figure with those that appear in the topological classification of the energy levels of saddle points of four-dimensional integrable Hamiltonians obtained by Lerman and Umanskiĭ [LU94]. They prove that, under some mild hypotheses, the bi-asymptotic set of a saddle point with doubled invariant manifolds is a CW-complex with one zero-dimensional cell (the saddle point), four one-dimensional cells (called *loops*) and four two-dimensional cells (called, say, *squares*). Each loop is either *orientable* or *non-orientable* (see definition 4.1 in [LU94]), but some combinations can never take place. Lerman and Umanskiĭ list all the feasible cases. In the last one, exactly one half of the loops are orientable.

Since we are not dealing with fixed points, but with two-periodic points, all the cells appear repeated in figure 8. There are two one-dimensional cells (the periodic points), eight loops and eight squares. Moreover, the four loops $\mathcal{Y}^{\sigma_1}_{\pm}$ are non-orientable and the other four loops $\mathcal{Z}^{\sigma_2}_{\pm}$ are orientable. Therefore, the spatial elliptic billiard is a realization of the last case in the above-mentioned Lerman–Umanskiĭ classification, but for two-periodic points of discrete systems, instead of fixed points of continuous systems.

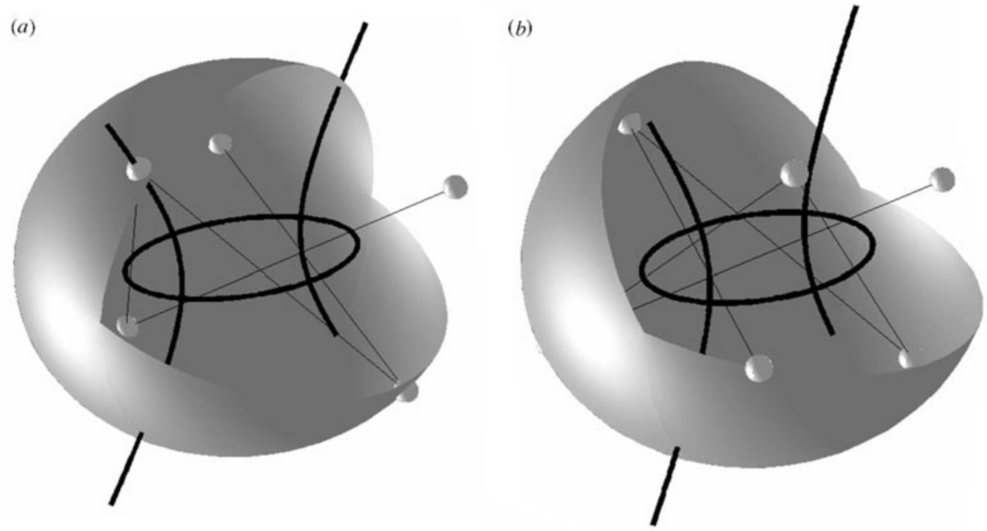


Figure 9. The two kinds of spatial symmetric bi-asymptotic trajectories inside a generic ellipsoid: xz-specular with an umbilical impact point (a) and y-axial with an asymptotic unitary velocity (b). The focal conics are represented by thick curves and a section of the ellipsoid is skipped for visualization purposes.

To end this section, we give a geometric characterization of the separatrix. We define the *dimension* of a billiard orbit as the dimension of the vectorial subspace generated by its velocities. A billiard orbit will be called *linear*, *planar* or *spatial* if its dimension is one, two or three, respectively. Billiard orbits inside a hypersurface of $\mathbb{R}^3$ are generically spatial. The only linear billiard orbits are the two-periodic orbits.

Coming back to the bi-asymptotic orbits inside the ellipsoid $\mathcal{Q}$, the planar ones are those on the curves $\mathcal{Y}$ and $\mathcal{Z}$, which can be interpreted as the separatrices of the planar billiards inside the ellipses $\mathcal{Q}_{\mathrm{xy}}$ and $\mathcal{Q}_{\mathrm{xz}}$, respectively. (See the end of subsection 5.2.) Thus, the separatrix is characterized as the set of *spatial bi-asymptotic billiard orbits*.

As in the planar case, all the orbits in the separatrix have the same (homoclinic) length, namely

Proposition 5.3. Length $\mathcal{O} = -2\left(\sqrt{a^2-b^2}+\sqrt{a^2-c^2}\right)$, *for all* $\mathcal{O} \subset \mathcal{S}$.

This result is stated without proof, since it will not be used.

5.4. Persistence of symmetric bi-asymptotic orbits

A surface in $\mathbb{R}^3$ will be called *symmetric* when it is symmetric with regard to the three coordinate axes 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.

It turns out that there are two kinds of spatial symmetric bi-asymptotic orbits inside a generic ellipsoid (see figure 9). They are the xz*-specular* ones and the y*-axial* ones, which are symmetric with regard to the xz-plane and the y-axis, respectively. There are other symmetric bi-asymptotic orbits, but they live on coordinate planes; that is, they are planar.

Theorem 5.1. *Inside a generic ellipsoid there are eight xz-specular (and eight y-axial) spatial symmetric billiard orbits bi-asymptotic to the diameter. They persist under symmetric perturbations.*

Proof. Let $\mathcal{Q}$ be any symmetric convex surface, not necessarily an ellipsoid. Denote its associated billiard map by $f : \mathcal{M} \to \mathcal{M}$ and its section by the xz-plane is $\mathcal{Q}_{\mathrm{xz}}$. Let us consider the set of symmetry

$$\tilde{\mathcal{F}} = \{m = (q, p) \in \mathcal{M} : q \in \mathcal{Q}_{\mathrm{xz}} \text{ and } p \text{ is perpendicular to } \mathcal{Q}_{\mathrm{xz}} \text{ at } q\}.$$

The importance of this set relies on the fact that the orbits inside $\mathcal{Q}$ with a point on it are xz-specular: if $(q, p) \in \tilde{\mathcal{F}}$ and $(q', p') = f(q, p)$, the line $l' = q' + \langle p' \rangle = q + \langle p' \rangle$ is the xz-specular reflection of the line $l = q + \langle p \rangle$.

Next, following the proof of the planar case, we shall study the intersections of the separatrix and this set of symmetry. Concretely, we shall prove that the connected component $\mathcal{S}^\sigma_\varsigma$ of the separatrix intersects the set of symmetry $\tilde{\mathcal{F}}$ at the point

$$\tilde{m}^\sigma_\varsigma = \varsigma m(\sigma_1 \tilde{r}_1, \sigma_2 \tilde{r}_2) = \varsigma(\tilde{q}^\sigma, \tilde{p}^\sigma) \qquad \tilde{q}^\sigma = (\tilde{x}, \sigma_1 \tilde{y}, \sigma_2 \tilde{z}) \qquad \tilde{p}^\sigma = (\tilde{u}, \sigma_1 \tilde{v}, \sigma_2 \tilde{w})$$

where $\tilde{r} = (\tilde{r}_1, \tilde{r}_2) = (1, 1)$, $\sigma = (\sigma_1, \sigma_2) \in \{-, +\}^2$, $\varsigma = \pm$ and

$$\tilde{x} = -a\frac{\sqrt{a^2 - b^2}}{\sqrt{a^2 - c^2}} \qquad \tilde{y} = 0 \qquad \tilde{z} = c\frac{\sqrt{b^2 - c^2}}{\sqrt{a^2 - c^2}}$$

$$\tilde{u} = -\frac{c^2\sqrt{a^2 - b^2}}{b^2\sqrt{a^2 - c^2}} \qquad \tilde{v} = \frac{\sqrt{b^2 - c^2}}{b} \qquad \tilde{w} = \frac{ac\sqrt{b^2 - c^2}}{b^2\sqrt{a^2 - c^2}}.$$

The inclusion $\tilde{m}^\sigma_\varsigma \in \mathcal{S}^\sigma_\varsigma$ is direct from the definition of $\mathcal{S}^\sigma_\varsigma$. On the other hand, if the surface $\mathcal{Q}$ is the ellipsoid (5.1), the equations of the set of symmetry $\tilde{\mathcal{F}}$ in the coordinates $q = (x, y, z) \in \mathcal{Q}$ and $p = (u, v, w) \in \mathbb{S}^2$ are $y = 0$ and $a^2uz = c^2xw$, so $\tilde{m}^\sigma_\varsigma \in \tilde{\mathcal{F}}$.

We also note that all of these intersections are transverse: the equations of the phase space ($x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ and $u^2 + v^2 + w^2 = 1$), the separatrix ($I_{\mathrm{y}}(m) = 0$ and $I_{\mathrm{z}}(m) = 0$, see proposition 5.1), and the set of symmetry ($y = 0$ and $a^2uz = c^2xw$), are functionally independent at the points $\tilde{m}^\sigma_\varsigma$.

Therefore, inside a generic ellipsoid there are eight xz-specular symmetric billiard orbits bi-asymptotic to the diameter: the orbits by the points $\tilde{m}^\sigma_\varsigma$, $\sigma \in \{-, +\}^2$, $\varsigma = \pm$. Their persistence is a consequence of the above established transversality (see the argument explained in the planar case).

It turns out that there exist eight y-axial persistent bi-asymptotic billiard orbits, too. These y-axial orbits pass through the points

$$\hat{m}^\sigma_\varsigma = \varsigma m(\sigma_1 \hat{r}_1, \sigma_2 \hat{r}_2) = \varsigma\left(\hat{q}^\sigma, \hat{p}^\sigma\right) \qquad \hat{q}^\sigma = (\hat{x}, \sigma_1 \hat{y}, \sigma_2 \hat{z}) \qquad \hat{p}^\sigma = (\hat{u}, \sigma_1 \hat{v}, \sigma_2 \hat{w})$$

where $\hat{r} = (\hat{r}_1, \hat{r}_2) = (\lambda_1^{1/2}, \lambda_2^{1/2})$, $\sigma = (\sigma_1, \sigma_2) \in \{-, +\}^2$, $\varsigma = \pm$ and

$$\hat{x} = -\frac{ac^2\sqrt{a^2 - b^2}}{b^2\sqrt{a^2 - c^2}} \qquad \hat{y} = -\sqrt{b^2 - c^2} \qquad \hat{z} = \frac{ac^2\sqrt{b^2 - c^2}}{b^2\sqrt{a^2 - c^2}}$$

$$\hat{u} = -\frac{\sqrt{a^2 - b^2}}{\sqrt{a^2 - c^2}} \qquad \hat{v} = 0 \qquad \hat{w} = \frac{\sqrt{b^2 - c^2}}{\sqrt{a^2 - c^2}}.$$

In order to prove it, let us consider another set of symmetry, namely

$$\hat{\mathcal{F}} = \{m = (q, p) \in \mathcal{M} : q + \langle p \rangle \text{ cuts the } y\text{-axis perpendicularly}\}.$$

Obviously, any billiard orbit inside a symmetric surface $\mathcal{Q}$ with a point on $\hat{\mathcal{F}}$ is y-axial. Moreover, the equations of $\hat{\mathcal{F}}$ in the coordinates $q = (x, y, z) \in \mathcal{Q}$ and $p = (u, v, w) \in \mathbb{S}^2$ are $v = 0$ and $uz = xw$. To end, it suffices to check that $\mathcal{S}^\sigma_\varsigma$ intersects $\hat{\mathcal{F}}$ transversely at the point $\hat{m}^\sigma_\varsigma$. □

It is interesting to note that imposing some symmetries is not necessary; to prove the persistence of the xz-specular (respectively, y-axial) bi-asymptotic orbits we have only used the symmetry with regard to the xz-plane (respectively, y-axis). Therefore, the next corollary follows.

Corollary 5.1. *The eight xz-specular (respectively, y-axial) symmetric bi-asymptotic orbits persist under any small perturbation preserving the symmetry with regard to the xz-plane (respectively, y-axis).*

Remark 5.5. The xz-specular bi-asymptotic trajectories inside a generic ellipsoid have an umbilical impact point, whereas the y-axial ones have an asymptotic unitary velocity, see (5.2) and (5.3). This gives rise to some nice geometric characterizations of these orbits. For instance, the eight y-axial spatial bi-asymptotic trajectories inside a generic ellipsoid 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 eight xz-specular trajectories admit a similar characterization. In the perturbed case, the symmetric bi-asymptotic orbits do not admit any characterization of this type, because the perturbed surface is not a quadric and concepts such as focal conics are missing.

Another interesting observation is that the persistence of the 16 symmetric homoclinic orbits can be obtained from purely dynamic arguments and, to be more precise, from a well known property of reversible maps. This dynamic approach is less intuitive than the geometric one, but its generalization to higher-dimensional cases is easier. We limit ourselves to describing the main ideas of the proof.

A map f is *reversible* when $f \circ h = h \circ f^{-1}$ for some involution h, which is called a *reversor* of f. If m^{h} is a hyperbolic fixed point of f, $h(m^{\mathrm{h}}) = m^{\mathrm{h}}$ and $m \in \mathcal{W}^{\mathrm{u}}$, then

$$\lim_{k\to+\infty} f^k(h(m)) = \lim_{k\to+\infty} h(f^{-k}(m)) = h\left(\lim_{k\to-\infty} f^k(m)\right) = h(m^{\mathrm{h}}) = m^{\mathrm{h}}.$$

Thus, reversors interchange unstable and stable invariant manifolds: $h(\mathcal{W}^{\mathrm{u,s}}) = \mathcal{W}^{\mathrm{s,u}}$, so the points in the intersection of $\mathcal{W}^{\mathrm{u}}$ (or $\mathcal{W}^{\mathrm{s}}$) with the set $\mathcal{F} = \{m : h(m) = m\}$ are homoclinic to m^{h}. In particular, the homoclinic points associated with transverse intersections of $\mathcal{W}^{\mathrm{u}}$ (or $\mathcal{W}^{\mathrm{s}}$) with $\mathcal{F}$ persist under reversible perturbations of the map.

These concepts are relevant because billiard maps inside symmetric surfaces are reversible. Indeed, among the infinitely many reversors of such billiards, there exists a distinguished couple $\tilde{h}, \hat{h} : \mathcal{M} \to \mathcal{M}$ such that $\tilde{h}(m^{\mathrm{h}}_\pm) = \hat{h}(m^{\mathrm{h}}_\pm) = m^{\mathrm{h}}_\pm$ and

$$\tilde{\mathcal{F}} = \{m \in \mathcal{M} : \tilde{h}(m) = m\} \qquad \hat{\mathcal{F}} = \{m \in \mathcal{M} : \hat{h}(m) = m\}$$

are the couple of sets of symmetry introduced in the proof of theorem 5.1. Therefore, it is clear that each transverse intersection of the separatrix with these sets gives rise to a persistent homoclinic point under symmetric (and hence, reversible) perturbations of the ellipsoid. Let us study those intersections. It turns out that

$$\tilde{h}(m(r_1, r_2)) = m(1/r_1, 1/r_2) \qquad \hat{h}(m(r_1, r_2)) = m(\lambda_1/r_1, \lambda_2/r_2)$$

where $m : \mathbb{R}^2 \to \mathcal{M}$ is the map defined in lemma 5.1. Then

$$\tilde{\mathcal{F}} \cap S^\sigma_\varsigma = \{\varsigma m(\sigma_1 r_1, \sigma_2 r_2) : m(r_1, r_2) = m(1/r_1, 1/r_2), r_1, r_2 > 0\} = \left\{\tilde{m}^\sigma_\varsigma\right\}$$

$$\hat{\mathcal{F}} \cap S^\sigma_\varsigma = \{\varsigma m(\sigma_1 r_1, \sigma_2 r_2) : m(r_1, r_2) = m(\lambda_1/r_1, \lambda_2/r_2), r_1, r_2 > 0\} = \left\{\hat{m}^\sigma_\varsigma\right\}.$$

Once we have computed these intersections, we must check their transversality to prove theorem 5.1 (again). We focus on $\tilde{\mathcal{F}} \cap S^\sigma_\varsigma$—the study of $\hat{\mathcal{F}} \cap S^\sigma_\varsigma$ is analogous.

Clearly, it suffices to check that the tangent planes of $\tilde{\mathcal{F}}$ and $\mathcal{S}^\sigma_\varsigma$ at the point $m = \tilde{m}^\sigma_\varsigma$ have zero intersection. This follows from the fact that any tangent vector to $\tilde{\mathcal{F}}$ (respectively, $\mathcal{S}^\sigma_\varsigma$) at the point $m = \tilde{m}^\sigma_\varsigma$ is an eigenvector of eigenvalue *one* (respectively, *minus one*) of the differential linear map $\mathrm{d}\tilde{h}(m^\sigma_\varsigma)$. The first claim follows from the equality $\tilde{h}(m) = m$ for $m \in \tilde{\mathcal{F}}$. The second one is obtained by differentiating the equality $\tilde{h}(m(r_1, r_2)) = m(1/r_1, 1/r_2)$ at $r = (\sigma_1 \tilde{r}_1, \sigma_2 \tilde{r}_2)$, where $\tilde{r} = (\tilde{r}_1, \tilde{r}_2) = (1, 1)$.

To end the results about persistent symmetric bi-asymptotic orbits, we note that some of them are planar, instead of spatial.

Corollary 5.2. *Inside a generic ellipsoid there are 16 planar (and 16 spatial) symmetric billiard orbits bi-asymptotic to the diameter. They persist under symmetric perturbations.*

Proof. The 16 planar ones are obtained by applying theorem 4.1 to the horizontal section $\mathcal{Q}_{\mathrm{xy}} = \mathcal{Q} \cap \{z = 0\}$ and the vertical section $\mathcal{Q}_{\mathrm{xz}} = \mathcal{Q} \cap \{y = 0\}$ of the ellipsoid.

The 16 spatial ones have been obtained in theorem 5.1. □

In the rest of the paper, we restrict our study to the spatial symmetric bi-asymptotic orbits; that is, to the eight xz-specular ones and the eight y-axial ones stated in theorem 5.1. The planar ones can be analysed as in the planar case.

5.5. The Melnikov potential

We consider the symmetric perturbations of the ellipsoid (5.1) defined by means of an implicit equation like

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \varepsilon P(y^2/b^2, z^2/c^2) \right\} \tag{5.6}$$

for some function $P : \mathbb{R}^2 \to \mathbb{R}$ such that $P(0, 0) = 0$. We shall call this perturbation *polynomial*, *entire* or *analytic*, if the function P is polynomial, entire or analytic, respectively. In the polynomial case, we shall say that the *order of the perturbation* is twice the total degree of the polynomial P. Thus, quadratic perturbations correspond to linear functions P.

The Melnikov potential associated with this kind of implicit symmetric perturbation can be computed in the same way as in the planar case. There are no substantial differences, and the result is summarized in the following lemma.

Lemma 5.2. *The Melnikov potential associated with the billiard inside (5.6) consists of eight copies of the function* $L : (0, +\infty)^2 \to \mathbb{R}$ *defined by*

$$L(r) = a \sum_{k \in \mathbb{Z}} \ell(\Lambda^k r) \qquad \ell(r) = \frac{\tau^2(r)}{\tau(\Lambda^{-1/2} r) \cdot \tau(\Lambda^{1/2} r)} P\left(\frac{\tau_{\mathrm{y}}^2(r)}{\tau^2(r)}, \frac{\tau_{\mathrm{z}}^2(r)}{\tau^2(r)} \right) \tag{5.7}$$

where the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ *and the tau-polynomials* τ, τ_{y}, τ_{z} *are defined in lemma 5.1. Moreover,* $L \circ \Lambda = L = L \circ \mathrm{I}$*. In particular,* $\tilde{r} = (1, 1)$ *and* $\hat{r} = (\lambda_1^{1/2}, \lambda_2^{1/2})$ *are critical points of* L*.*

5.6. Uniform non-integrability and splitting

Following Birkhoff, it is natural to conjecture that the elliptic billiards are the unique smooth integrable convex billiards, not only in the plane but also in any dimension. We are not ready to tackle this conjecture, not even a local version of it around generic ellipsoids in the set of entire convex hypersurfaces. The tools at our disposal only allow us to establish a local version in the frame of the uniform integrability introduced in section 2.3.

Theorem 5.2. *An entirely symmetric perturbation of a generic ellipsoid gives rise to a uniformly integrable billiard if and only if it is quadratic.*

Proof. Let P be an entire function. As in the planar case, it suffices to prove that the Melnikov potential (5.7) is non-constant if the perturbation is not quadratic; that is, P is not linear. This can be accomplished through the study of the complex singularities of the Melnikov potential, since it is non-constant if its complex extension has singularities. More precisely, the theorem follows from the following fact: *$L(r)$ is analytic at $r^* := (1, \mathrm{i})$ if and only if P is linear.*

The proof of this equivalence is performed in two steps.

Step 1. $L(r) - a\ell(r)$ is analytic at $r^* = (1, \mathrm{i})$. Only the points r for which some denominator $\tau(\Lambda^{j/2}r)$ vanishes for $0 \neq j \in \mathbb{Z}$ can be singularities of $L(r) - a\ell(r) = a\sum_{0\neq k\in\mathbb{Z}} \ell(\Lambda^k r)$. We must check that these denominators are uniformly far away from zero in some neighbourhood of r^*.

Given $\delta \in (0, 1)$, let V_δ be the neighbourhood of r^* defined by

$$V_\delta = V_\delta^+ \cap V_\delta^- \qquad V_\delta^\pm = \left\{ r \in \mathbb{C}^2 : \left|r_1^{\pm 2} + r_2^{\pm 2}\right| < \delta, \left|1 + r_1^{\pm 2} r_2^{\pm 2}\right| < \delta \right\}.$$

Using the relation $\tau(r_1, r_2) = r_1^2 r_2^2 \tau(1/r_1, 1/r_2)$ and the inequalities $1 < \lambda_1 < \lambda_2$, we obtain

$$\left|\tau(\Lambda^{-j/2}r)\right| \geqslant 1 - \frac{1+\delta}{\lambda_1^j \lambda_2^j} - \frac{\alpha^2\delta}{\lambda_1^j} \geqslant 1 - \frac{1+\delta}{\lambda_1\lambda_2} - \frac{\alpha^2\delta}{\lambda_1}$$

$$\left|\tau(\Lambda^{j/2}r)\right| \geqslant (1-\delta)\lambda_1^j\lambda_2^j \left(1 - \frac{1+\delta}{\lambda_1^j\lambda_2^j} - \frac{\alpha^2\delta}{\lambda_1^j}\right) \geqslant (1-\delta)(\lambda_1\lambda_2 - 1 - \delta - \alpha^2\lambda_2\delta)$$

for any $j \geqslant 1$ and $r \in V_\delta$. Therefore, the denominators $\tau(\Lambda^{j/2}r)$, for $r \in V_\delta$ and $0 \neq j \in \mathbb{Z}$, are uniformly far away from zero if δ is small enough.

Step 2. $\ell(r)$ is analytic at $r^* = (1, \mathrm{i})$ if and only if P is linear. Let $\eta, \zeta, \xi : \mathbb{C}^2 \to \mathbb{C}$ be the functions

$$\eta(r) = \frac{\tau_z^2(r)}{\tau(\Lambda^{-1/2}r)\tau(\Lambda^{1/2}r)}$$

$$\zeta(r) = \frac{\tau^2(r)}{\tau_z^2(r)} P\left(\frac{\tau_y^2(r)}{\tau^2(r)}, \frac{\tau_z^2(r)}{\tau^2(r)}\right)$$

$$\xi(t) = t_1 P\left(\frac{t_2}{t_1}, \frac{1}{t_1}\right).$$

We note that $\ell(r) = \eta(r)\zeta(r)$, $\eta(r)$ is analytic at r^*, $\eta(r^*) \neq 0$, and the rational map

$$r = (r_1, r_2) \mapsto t = (t_1, t_2) = \left(\tau^2(r)/\tau_z^2(r), \tau_y^2(r)/\tau_z^2(r)\right)$$

is an analytic change of variables from a small neighbourhood of $r^* = (1, \mathrm{i})$ onto a small neighbourhood of $t^* = (0, -1)$. Hence, the following three statements are equivalent: $\ell(r)$ is analytic at r^*, $\zeta(r)$ is analytic at r^* and $\xi(t)$ is analytic at t^*.

It remains to prove that $\xi(t)$ is analytic at t^* if and only if P is linear.

Let $P = \sum_{n\geqslant 1} P_n$ be the decomposition of the perturbation as a convergent series of homogeneous polynomials; that is, $P_n(\mu s) = \mu^n P_n(s)$ for all $s \in \mathbb{C}^2$ and $\mu \in \mathbb{C}$. Then $\xi(t) = \sum_{j\geqslant 0} P_{j+1}(t_2, 1)t_1^{-j}$. Using this equality, we find that $\xi(t)$ is analytic at $t^* = (0, -1)$ if and only if the Laurent coefficients $P_n(t_2, 1)$, $n \geqslant 2$, are identically zero for t_2 close to -1, or equivalently, if and only if P is linear. □

As a by-product, the separatrix splits under any non-quadratic entire symmetric perturbation of a generic ellipsoid.

5.7. The parameter space

The closed convex surface (5.6) is determined by some function $P : \mathbb{R}^2 \to \mathbb{R}$ such that $P(0, 0) = 0$, some lengths (a, b, c) such that $a > b > c > 0$, and the perturbative parameter $\varepsilon \neq 0$, which is supposed to be as small as necessary. The billiard dynamics inside two different surfaces of that form, determined by the same function P and the same perturbative parameter ε but with different lengths (a, b, c) and (a', b', c') such that $a'/a = b'/b = c'/c$, are clearly conjugated. Therefore, we can normalize these lengths.

Concretely, we shall work with the normalized parameter

$$\beta = (\beta_1, \beta_2) \qquad \beta_1 = b^2/a^2 \qquad \beta_2 = c^2/a^2$$

already introduced in lemma 5.1. Then the parameter space is the triangle

$$\mathcal{P} = \left\{\beta = (\beta_1, \beta_2) \in \mathbb{R}^2 : 0 < \beta_2 < \beta_1 < 1\right\}.$$

When some inequality in the expression $a > b > c > 0$ becomes an equality, the unperturbed ellipsoid (5.1) is degenerate. As we can observe in figure 1, there are six types of degenerate ellipsoids:

- *Flat ellipsoids*: $\mathcal{Q} = \left\{(x, y, 0) \in \mathbb{R}^3 : x^2/a^2 + y^2/b^2 \leqslant 1\right\}$, with $a > b > 0$. They correspond to the single degeneration $c = 0$; that is, to $\beta_2 = 0$.
- *Oblate ellipsoids*: $\mathcal{Q} = \left\{(x, y, z) : x^2 + y^2 + \eta z^2 = a^2\right\}$, with $a > 0$ and $\eta \in (0, 1)$. They correspond to the single degeneration $b = a$; that is, to $\beta_1 = 1$.
- *Prolate ellipsoids*: $\mathcal{Q} = \left\{(x, y, z) : x^2 + \eta(y^2 + z^2) = a^2\right\}$, with $a > 0$ and $\eta \in (0, 1)$. They correspond to the single degeneration $c = b$; that is, to $\beta_1 = \beta_2$.
- *Circles*: $\mathcal{Q} = \left\{(x, y, 0) : x^2 + y^2 \leqslant a^2\right\}$, with $a > 0$. They correspond to the double degeneration $c = 0$ and $a = b$; that is, to the point $\beta = (1, 0)$.
- *Segments*: $\mathcal{Q} = \{(x, 0, 0) : -a \leqslant x \leqslant a\}$, with $a > 0$. They correspond to the double degeneration $b = c = 0$; that is, to the point $\beta = (0, 0)$.
- *Spheres*: $\mathcal{Q} = \left\{(x, y, z) : x^2 + y^2 + z^2 = a^2\right\}$, with $a > 0$. They correspond to the double degeneration $a = b = c$; that is, to the point $\beta = (1, 1)$.

The flat, oblate and prolate ellipsoids correspond to degenerations of co-dimension one, whereas the circles, segments and spheres are associated with degenerations of co-dimension two. We have restricted our considerations to the first kind of degenerate ellipsoids.

For further reference, it should be recalled that we say that a property holds:

- *close to the flat limit* when for any $\beta^{\mathrm{f}} = (\beta_1^{\mathrm{f}}, 0)$, $0 < \beta_1^{\mathrm{f}} < 1$, there exists a positive constant δ such that the property holds for all $\beta \in \mathcal{P}$, $\left|\beta^{\mathrm{f}} - \beta\right| < \delta$.

- *close to the oblate limit* when for any $\beta^{\mathrm{o}} = (1, \beta_2^{\mathrm{o}})$, $0 < \beta_2^{\mathrm{o}} < 1$, there exists a positive constant δ such that the property holds for all $\beta \in \mathcal{P}$, $|\beta^{\mathrm{o}} - \beta| < \delta$.
- *close to the prolate limit* when for any $\beta^{\mathrm{p}} = (\beta_1^{\mathrm{p}}, \beta_2^{\mathrm{p}})$, $0 < \beta_1^{\mathrm{p}} = \beta_2^{\mathrm{p}} < 1$, there exists a positive constant δ such that the property holds for all $\beta \in \mathcal{P}$, $|\beta^{\mathrm{p}} - \beta| < \delta$.

Of course, the size of the constant δ depends on the points $\beta^{\mathrm{f}}, \beta^{\mathrm{o}}, \beta^{\mathrm{p}} \in \partial\mathcal{P}$.

Henceforth, we shall put the superscripts ‘f’, ‘o’ and ‘p’ on objects associated with flat, oblate and prolate ellipsoids, respectively.

In the remainder of this section, we present some partial answers to the following questions: Are the 16 spatial symmetric bi-asymptotic orbits transverse? Are there other primary spatial bi-asymptotic orbits? The answers are just partial for several reasons. First, a restrictive hypothesis is placed on the perturbation. Second, analytic results have been obtained only close to the flat and oblate limits.

At the end of the section, we have performed an accurate numerical study of a particular quartic perturbation (the simplest non-trivial one), in which infinitely many bifurcations are observed numerically close to the prolate limit.

5.8. Transversality close to the flat limit

For technical reasons, which are explained in the remark below, we restrict ourselves to the symmetric polynomial perturbations of the ellipsoid (5.1) preserving its horizontal section; that is,

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \varepsilon(z^2/c^2) \cdot R(y^2/b^2, z^2/c^2) \right\} \tag{5.8}$$

for some polynomial $R : \mathbb{R}^2 \to \mathbb{R}$. This perturbation is linked to the perturbation (5.6) by means of the relation $P(s_1, s_2) = s_2 R(s_1, s_2)$.

Following the ideas presented in the planar case, we shall study the asymptotic behaviour of the Melnikov potential and the determinant of its Hessian ($\operatorname{Hess} L(r) := \mathrm{d}^2 L(r)$) at its critical points $\tilde{r} = (1, 1)$ and $\hat{r} = (\lambda_1^{1/2}, \lambda_2^{1/2})$ when $\beta_2 = c^2/a^2 \to 0^+$. This behaviour is contained in the following lemma.

Lemma 5.3. *Let $d, \tilde{d}, \hat{d} : \mathcal{P} \to \mathbb{R}$ be the functions*

$$d(\beta) = a^{-1}[L(\tilde{r}) - L(\hat{r})]$$
$$\tilde{d}(\beta) = a^{-2} \det[\operatorname{Hess} L(\tilde{r})]$$
$$\hat{d}(\beta) = a^{-2} \det[\operatorname{Hess} L(\hat{r})].$$

If the perturbation (5.8) is polynomial, then

$$d(\beta) = [R(0, \beta_1) - R(1, 0)]\beta_2 + \mathrm{O}(\beta_2^2)$$
$$\tilde{d}(\beta) = -4\beta_1^2(1 - \beta_1)\,[\partial_2 R(0, \beta_1)]^2\,\beta_2^2 + \mathrm{O}(\beta_2^3)$$
$$\hat{d}(\beta) = 4\beta_1^{-1}\lambda_1^{-1}\partial_1 R(1, 0) \cdot [\partial_1 R(1, 0) - \partial_2 R(0, 0)]\,\beta_2^4 + \mathrm{O}(\beta_2^5).$$

Proof. The lemma follows directly from

$$a^{-1}L(\tilde{r}) = R(0, \beta_1)\beta_2 + \mathrm{O}(\beta_2^2)$$
$$a^{-1}L(\hat{r}) = R(1, 0)\beta_2 + \mathrm{O}(\beta_2^2)$$

$$a^{-1}\,\mathrm{Hess}\,L(\tilde{r}) = 2\beta_1\sqrt{1-\beta_1}\begin{pmatrix} \mathrm{O}(\beta_2^2) & \partial_2 R(0,\beta_1)\beta_2 + \mathrm{O}(\beta_2^2) \\ \partial_2 R(0,\beta_1)\beta_2 + \mathrm{O}(\beta_2^2) & \mathrm{O}(\beta_2) \end{pmatrix}$$

$$a^{-1}\,\mathrm{Hess}\,L(\hat{r}) = \begin{pmatrix} -\dfrac{2}{\lambda_1}\partial_1 R(1,0)\beta_2 + \mathrm{O}(\beta_2^2) & \mathrm{O}(\beta_2^{5/2}) \\ \mathrm{O}(\beta_2^{5/2}) & \dfrac{2}{\beta_1}[\partial_2 R(0,0) - \partial_1 R(1,0)]\beta_2^3 + \mathrm{O}(\beta_2^4) \end{pmatrix}.$$

The proof of these formulae has been deferred to appendix A. □

Remark 5.6. The preservation of the horizontal section of the ellipsoid plays an essential role in the computations. This implies that the function $\ell(r)$ in (5.7) has a common factor $\tau_z^2(r)$. Then only the central term of the series

$$\sum_{k\in\mathbb{Z}}(-1)^k \ell\left(\lambda_1^{k/2}, \lambda_2^{k/2}\right) = a^{-1}[L(\tilde{r}) - L(\hat{r})] = d(\beta)$$

contributes to the lowest-order coefficient of the Taylor expansion of the function $d(\beta)$ in powers of the small parameter β_2. The same behaviour is observed for $\tilde{d}(\beta)$ and $\hat{d}(\beta)$. This makes it possible to prove the lemma with a reasonable amount of work (see appendix A). (Otherwise, it would be necessary to consider *all* the terms of the series.)

Corollary 5.3. *If* $\partial_1 R(1,0) \neq 0, \partial_2 R(0,0)$ *(respectively,* $\partial_2 R(0,\beta_1) \neq 0$*) (respectively,* $R(0,\beta_1) \neq R(1,0)$*), the y-axial spatial bi-asymptotic orbits become transverse (respectively, the xz-specular ones become transverse) (respectively, the xz-specular and y-axial ones have different lengths), close to the flat limit and for small enough perturbations.*

Proof. It follows directly from lemma 5.3 and properties L3 and L4 in section 2.3. □

Although the hypotheses on the polynomial R stated in corollary 5.3 are generic in the space of polynomials, they fail in some degenerate cases, where the corollary cannot be applied directly. Thus, further computations are necessary to obtain some information from the first non-vanishing Taylor coefficients of the functions $d, \tilde{d}, \hat{d} : \mathcal{P} \to \mathbb{R}$.

Some degenerate cases are listed below:

- If $R(s_1, s_2) = s_1$, then $\tilde{d}(\beta) = \beta_1^2(6\beta_1 - 8)\beta_2^3 + \mathrm{O}(\beta_2^4)$ and the xz-specular spatial bi-asymptotic orbits are transverse.
- If $R(s_1, s_2) = s_2^j$, $j \geqslant 1$, then $\hat{d}(\beta) = -2j^2(j+1)(1-\beta_1)\beta_1^{-2j}\lambda_1^{-1}\beta_2^{2j+3} + \mathrm{O}(\beta_2^{2j+4})$ and the y-axial spatial bi-asymptotic orbits are transverse.
- If $R(s_1, s_2) = s_1 s_2^j$ for some integer $j \geqslant 1$, then $d(\beta) = -\beta_1^{-j}\beta_2^{j+1} + \mathrm{O}(\beta_2^{j+2})$ and the xz-specular and y-axial spatial bi-asymptotic orbits have different lengths.

The computations in these degenerate cases are similar to those performed in appendix A.

To end this section, it is natural to ask whether some spatial version of theorem 4.3 holds; that is, whether under any non-quadratic analytic symmetric perturbation of the ellipsoid the symmetric spatial bi-asymptotic orbits become transverse close to the flat limit. We have tried to prove it for perturbations preserving the horizontal section of the ellipsoid, but the computations were too cumbersome.

5.9. A quartic perturbation

Here we focus our attention on the symmetric quartic perturbation of the ellipsoid (5.1) preserving its horizontal section $\mathcal{Q} \cap \{z = 0\}$ and its vertical one $\mathcal{Q} \cap \{y = 0\}$; that is,

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \varepsilon \frac{y^2}{b^2} \cdot \frac{z^2}{c^2} \right\}. \tag{5.9}$$

In spite of its simplicity, it turns out to be quite interesting.

The Melnikov potential (5.7) for this quartic perturbation is

$$L(r) = a \sum_{k \in \mathbb{Z}} \ell(\Lambda^k r) \qquad \ell(r) = \frac{\tau_y^2(r)\tau_z^2(r)}{\tau(\Lambda^{-1}r)\tau^2(r)\tau(\Lambda r)}. \tag{5.10}$$

The function $t = (t_1, t_2) \mapsto L(\mathrm{e}^{t_1}, \mathrm{e}^{t_2})$ is neither *hyperelliptic* nor *Abelian*, since it has only three (instead of four) complex periods independent over the reals:

$$\omega_1 := (\ln \lambda_1, \ln \lambda_2) \qquad \omega_2 := (\pi \mathrm{i}, 0) \qquad \omega_3 := (0, \pi \mathrm{i}).$$

This makes an important difference with the planar case, in which explicit computations can be easily performed using the theory of elliptic functions. In the spatial case, we are forced to consider some cases close to simpler limit cases in order to obtain analytic results. We have also carried out a numerical study.

Our results can be summarized as follows. Let $H = H(\beta)$ be the number of primary homoclinic orbits under the quartic perturbation (5.9) with parameter $\beta \in \mathcal{P}$. Then $H(\beta) = 16$ close to the flat, circular and oblate limits, but $H(\beta)$ undergoes infinitely many bifurcations when β approaches the prolate limit, oscillating between 16 and 32. Let us explain this. Suppose that we go from the initial parameter $\beta^{\mathrm{c}} = (0, 1)$ to some final parameter β^{p} on the hypotenuse of the right-angled triangle $\partial\mathcal{P}$ along a straight line. (The parameter β^{c} corresponds to a circle, whereas β^{p} corresponds to a prolate ellipsoid.) If β is close enough to β^{c}, $H(\beta) = 16$. After a certain bifurcation value β_1^+ of the parameter β is attained, $H(\beta) = 32$, and when a second bifurcation value β_1^- is crossed, $H(\beta) = 16$ again. These bifurcations occur infinitely many times when β approaches β^{p}; that is, there exist infinitely many bifurcation values $\beta_n^\pm$, $n \geqslant 1$ of the parameter β. The sequences $(\beta_n^\pm)_{n \geqslant 1}$ tend in a geometric way to $\beta^{\mathrm{p}} = (\beta_1^{\mathrm{p}}, \beta_2^{\mathrm{p}})$. To be more precise, if

$$\lambda^{\mathrm{p}} = \frac{1 + e^{\mathrm{p}}}{1 - e^{\mathrm{p}}} \qquad e^{\mathrm{p}} = \sqrt{1 - \beta_1^{\mathrm{p}}} = \sqrt{1 - \beta_2^{\mathrm{p}}} \tag{5.11}$$

is the (double) characteristic multiplier of the prolate ellipsoid associated with β^{p}, then

$$\beta^{\mathrm{p}} - \beta_n^\pm \sim \lambda^{\mathrm{p}}(\beta^{\mathrm{p}} - \beta_{n+1}^\pm) \qquad (n \to +\infty). \tag{5.12}$$

Moreover, if $\mathcal{P} = \mathcal{P}_- \cup \mathcal{P}_0 \cup \mathcal{P}_+$ is the decomposition of the parameter space given by

$$\mathcal{P}_\pm := \left\{ \beta \in \mathcal{P} : \pm \tilde{d}(\beta)\, \hat{d}(\beta) > 0 \right\} \qquad \mathcal{P}_0 := \left\{ \beta \in \mathcal{P} : \tilde{d}(\beta)\, \hat{d}(\beta) = 0 \right\} \tag{5.13}$$

then $\mathcal{P}_- = \{\beta \in \mathcal{P} : H(\beta) = 16\}$ and $\mathcal{P}_+ = \{\beta \in \mathcal{P} : H(\beta) = 32\}$, whereas $\mathcal{P}_0$ is formed by the infinitely many bifurcation curves which separate them; that is, $\mathcal{P}_0$ contains all the bifurcation values $\beta_n^\pm$, $n \geqslant 1$.

In figure 2 we have marked the points of $\mathcal{P}_-$ in white and those of $\mathcal{P}_+$ in black. These sets are formed by infinitely many strips connecting $\beta = (0, 0)$ and $\beta = (1, 1)$; that is, their extrema correspond to segments and spheres. We have represented only the first three black strips, although we have computed the first eight ones, using multiple-precision arithmetic to overcome some numerical difficulties. Nevertheless, the other black strips are so thin that they cannot been seen at the scale of that picture!

Using that $\mathcal{P}_- \cup \mathcal{P}_+ = \{\beta \in \mathcal{P} : \tilde{d}(\beta)\,\hat{d}(\beta) \neq 0\}$, we deduce that the perturbed symmetric bi-asymptotic orbits are transverse when β is inside a white or black strip.

Finally, inside each black strip we find a new bifurcation curve from $\beta = (0, 0)$ to $\beta = (1, 1)$ defined by $\{d(\beta) = 0\}$. The xz-specular and y-axial perturbed bi-asymptotic orbits have different lengths when β does not belong to some of these new curves.

We end this summary by stressing the main analogies and differences with respect to the planar case. First, the functions $d, \tilde{d}, \hat{d} : \mathcal{P} \to \mathbb{R}$ can be extended analytically to the flat limit; continuously to the spherical, oblate, circular and segment limits; but they cannot be extended to the prolate limit. Second, these functions are exponentially small close to the spherical limit. This makes it harder to perform the numerical computations for parameters $\beta \approx (1, 1)$; multiple-precision arithmetic is essential here. And last, close to the prolate limit, these functions undergo infinitely many changes of sign in the parameter space, which is the crucial difference with respect to the planar case. Let us recall that any quartic symmetric perturbation of a non-circular ellipse always has eight primary heteroclinic orbits: $H(\beta) \equiv 8$, see theorem 4.4. In the spatial case, this is false: the quantity $H(\beta)$ is no longer constant for quartic perturbations.

The rest of the section contains the proofs of the analytical results and the description of the numerical experiments. During the exposition it will become clear which results are analytical and which are numerical.

5.9.1. Analysis close to the flat limit. The quartic perturbation (5.9) fits into the frame of the previous subsection, since it preserves the horizontal section. Following the terminology of (5.8), it corresponds to the polynomial $R(s_1, s_2) = s_1$. In particular, the following relations hold:

$$d(\beta) = -\beta_2 + \mathrm{O}(\beta_2^2)$$

$$\tilde{d}(\beta) = \beta_1^2(6\beta_1 - 8)\beta_2^3 + \mathrm{O}(\beta_2^4)$$

$$\hat{d}(\beta) = 4\beta_1^{-1}\lambda_1^{-1}\beta_2^4 + \mathrm{O}(\beta_2^5).$$

The first and last ones are direct consequences of lemma 5.3, whereas the second one follows from the comments on degenerate cases after corollary 5.3.

Thus, all the symmetric bi-asymptotic orbits become transverse and the xz-specular and y-axial ones have different lengths, close to the flat limit and under the quartic perturbation (5.9), provided that ε is small enough. Finally, we cannot claim that close to the flat limit there are no more primary bi-asymptotic orbits, at least using only the above arguments. Nevertheless, we have checked numerically that this is the case.

5.9.2. Analysis close to the circular limit. The results obtained close to the flat limit also hold close to the circular limit: $\beta \to \beta^{\mathrm{c}} := (1, 0)$. The proof does not require any new ideas. Once the limits

$$\lim_{\beta\to\beta^{\mathrm{c}}} d(\beta)/\beta_2 = -1 \qquad \lim_{\beta\to\beta^{\mathrm{c}}} \tilde{d}(\beta)/\beta_2^3 = -2 \qquad \lim_{\beta\to\beta^{\mathrm{c}}} \hat{d}(\beta)/\beta_2^4 = 4$$

are established, all the results follow directly. The computations are standard.

5.9.3. Analysis close to the oblate limit. In the oblate limit we shall establish a stronger result: the symmetric spatial bi-asymptotic orbits are transverse and *there are no more primary ones*. The key is to realize that the Melnikov potential can be continuously extended up to values of the parameters corresponding to oblate ellipsoids and, in addition, it becomes a function of separate variables for these degenerate parameters. This simplifies the analysis.

The ultimate reason for the separation of variables is the parabolic character of the two-periodic orbits associated with the diameters of oblate ellipsoids; that is, these orbits have some characteristic multiplier equal to one. Concretely, if the parameter belongs to the cathetus

$$\mathcal{P}^{\mathrm{o}} = \left\{\beta^{\mathrm{o}} = (\beta^{\mathrm{o}}_1, \beta^{\mathrm{o}}_2) : \beta^{\mathrm{o}}_1 = 1, 0 < \beta^{\mathrm{o}}_2 < 1\right\}$$

of the right-angled triangle $\partial\mathcal{P}$, its associated characteristic multipliers are

$$\lambda^{\mathrm{o}}_1 = 1 \qquad \lambda^{\mathrm{o}}_2 = \frac{1+e^{\mathrm{o}}_2}{1-e^{\mathrm{o}}_2} \qquad e^{\mathrm{o}}_2 = \sqrt{1-\beta^{\mathrm{o}}_2}. \tag{5.14}$$

The tau-polynomials $\tau, \tau_{\mathrm{x}}, \tau_{\mathrm{y}}, \tau_{\mathrm{z}}$ defined in lemma 5.1 have simpler expressions in the oblate limit. Concretely,

$$\tau^{\mathrm{o}}(r) = (1+r_1^2)(1+r_2^2) \qquad \tau^{\mathrm{o}}_{\mathrm{x}} = (1-r_1^2)(1-r_2^2)$$
$$\tau^{\mathrm{o}}_{\mathrm{y}}(r) = 2r_1(1-r_2^2) \qquad \tau^{\mathrm{o}}_{\mathrm{z}}(r) = 2r_2(1+r_1^2).$$

Then the Melnikov potential (5.10) becomes a function of separate variables, namely

$$L^{\mathrm{o}} : (0,+\infty)^2 \rightarrow (0,+\infty) \qquad L^{\mathrm{o}}(r) := \lim_{\beta\rightarrow\beta^{\mathrm{o}}} L(r) = aL^{\mathrm{o}}_1(r_1)L^{\mathrm{o}}_2(r_2) \tag{5.15}$$

where $L^{\mathrm{o}}_1 : (0,+\infty) \rightarrow (0,+\infty)$ is the rational function

$$L^{\mathrm{o}}_1(r_1) = \left(\frac{2r_1}{1+r_1^2}\right)^2$$

and $L^{\mathrm{o}}_2 : (0,+\infty) \rightarrow (0,+\infty)$ is given by the series

$$L^{\mathrm{o}}_2(r_2) = \sum_{k\in\mathbb{Z}} \ell^{\mathrm{o}}_2(\lambda^{\mathrm{o}k}_2 r_2) \qquad \ell^{\mathrm{o}}_2(r_2) = \frac{4(1-r_2^2)^2 r_2^2}{\left(1+r_2^2/\lambda^{\mathrm{o}}_2\right)(1+r_2^2)^2\left(1+\lambda^{\mathrm{o}}_2 r_2^2\right)}. \tag{5.16}$$

These expressions make clear that the functions $d, \tilde{d}, \hat{d} : \mathcal{P} \rightarrow \mathbb{R}$ can be continuously extended to the oblate limit, giving rise to the extensions $d^{\mathrm{o}}, \tilde{d}^{\mathrm{o}}, \hat{d}^{\mathrm{o}} : \mathcal{P}^{\mathrm{o}} \rightarrow \mathbb{R}$ defined by

$$d^{\mathrm{o}}(\beta^{\mathrm{o}}) = a^{-1}\left(L^{\mathrm{o}}(\tilde{r}^{\mathrm{o}}) - L^{\mathrm{o}}(\hat{r}^{\mathrm{o}})\right) = L^{\mathrm{o}}_1(1)\left(L^{\mathrm{o}}_2(\tilde{r}^{\mathrm{o}}_2) - L^{\mathrm{o}}_2(\hat{r}^{\mathrm{o}}_2)\right)$$
$$\tilde{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = a^{-2}\det[\mathrm{Hess}\, L^{\mathrm{o}}(\tilde{r}^{\mathrm{o}})] = L^{\mathrm{o}}_1(1)\frac{\mathrm{d}^2L^{\mathrm{o}}_1}{\mathrm{d}r_1^2}(1)L^{\mathrm{o}}_2(\tilde{r}^{\mathrm{o}}_2)\frac{\mathrm{d}^2L^{\mathrm{o}}_2}{\mathrm{d}r_2^2}(\tilde{r}^{\mathrm{o}}_2)$$
$$\hat{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = a^{-2}\det[\mathrm{Hess}\, L^{\mathrm{o}}(\hat{r}^{\mathrm{o}})] = L^{\mathrm{o}}_1(1)\frac{\mathrm{d}^2L^{\mathrm{o}}_1}{\mathrm{d}r_1^2}(1)L^{\mathrm{o}}_2(\hat{r}^{\mathrm{o}}_2)\frac{\mathrm{d}^2L^{\mathrm{o}}_2}{\mathrm{d}r_2^2}(\hat{r}^{\mathrm{o}}_2)$$

where

$$\tilde{r}^{\mathrm{o}} = (\tilde{r}^{\mathrm{o}}_1, \tilde{r}^{\mathrm{o}}_2) := \lim_{\beta\rightarrow\beta^{\mathrm{o}}} \tilde{r} = (1,1) \qquad \hat{r}^{\mathrm{o}} = (\hat{r}^{\mathrm{o}}_1, \hat{r}^{\mathrm{o}}_2) := \lim_{\beta\rightarrow\beta^{\mathrm{o}}} \hat{r} = \left(1, \sqrt{\lambda^{\mathrm{o}}_2}\right)$$

are the critical points in the oblate limit.

The study of the series (5.16) is the main difficulty in computing these extensions. This series can be expressed in terms of Jacobian elliptic functions, as we did in lemma 4.4. To such an end, we now recall some classical notation, which is borrowed from [WW27].

Given two quantities $k, k' \in (0, 1)$ such that $k^2 + k'^2 = 1$, k is called the *modulus* and k' is known as the *complementary modulus*. Then $K = \int_0^{\pi/2}(1 - k^2 \sin u)^{-1/2}\,\mathrm{d}u$ is the *complete elliptic integral of the first kind*, whereas $K' = \int_0^{\pi/2}(1 - k'^2 \sin u)^{-1/2}\,\mathrm{d}u$, so that K' is the same function of k' as K is of k. Finally, $q = \mathrm{e}^{-\pi K'/K}$ is the *nome*. If any of the numbers k, k', K, K' or q is given, all the rest are determined. For instance, in [WW27, p 479] we find the relations

$$\sqrt{2kK/\pi} = \sum_{n\in\mathbb{Z}} q^{(n+1/2)^2} \qquad \sqrt{2K/\pi} = \sum_{n\in\mathbb{Z}} q^{n^2} \qquad \sqrt{2k'K/\pi} = \sum_{n\in\mathbb{Z}} (-q)^{n^2} \tag{5.17}$$

which are useful to compute the modulus k, the complementary modulus k', and the complete elliptic integral of the first kind K, when the nome q is given. From now on, we shall consider that the quantities q, k, k' and K are determined by the parameter $\beta^{\mathrm{o}} \in \mathcal{P}^{\mathrm{o}}$ via the formulae (5.14), the identification

$$q = \mathrm{e}^{-\pi^2/\ln\lambda_2^{\mathrm{o}}}$$

and relations (5.17). Under these notation and assumptions, it turns out that

$$L_2^{\mathrm{o}}(r_2) = \text{constant} - \frac{4\lambda_2^{\mathrm{o}}}{(\lambda_2^{\mathrm{o}} - 1)^2}\left(\frac{2K}{\ln\lambda_2^{\mathrm{o}}}\right)^2 \mathrm{dn}^2\left(\frac{2K\log r_2}{\ln\lambda_2^{\mathrm{o}}}, k\right) \tag{5.18}$$

where $\mathrm{dn}(u) = \mathrm{dn}(u, k)$ is one of the 12 Jacobian elliptic functions (see appendix B). The exact value of the unknown additive constant is immaterial for our present purposes, although it could be expressed explicitly in terms of complete elliptic integrals of the first and second kinds, if necessary.

Now, we are ready to compute the extensions $d^{\mathrm{o}}, \tilde{d}^{\mathrm{o}}, \hat{d}^{\mathrm{o}} : \mathcal{P}^{\mathrm{o}} \to \mathbb{R}$ explicitly. The result is contained in the following lemma (compare with lemma 4.4 for the planar case.)

Lemma 5.4. *The critical points of the function (5.15) are just the points in the set* $\{1\} \times (\lambda^{\mathrm{o}})^{\mathbb{Z}/2}$, *all of them being non-degenerate. The extensions* $d^{\mathrm{o}}, \tilde{d}^{\mathrm{o}}, \hat{d}^{\mathrm{o}} : \mathcal{P}^{\mathrm{o}} \to \mathbb{R}$ *associated with the quartic perturbation (5.9) are*

$$d^{\mathrm{o}}(\beta^{\mathrm{o}}) = \frac{-4\pi^2\lambda_2^{\mathrm{o}}}{(\lambda_2^{\mathrm{o}} - 1)^2 \ln^2\lambda_2^{\mathrm{o}}}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4$$

$$\tilde{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = \frac{-16\pi^4\lambda_2^{\mathrm{o}} L_2^{\mathrm{o}}(\tilde{r}_2^{\mathrm{o}})}{(\lambda_2^{\mathrm{o}} - 1)^2 \ln^4\lambda_2^{\mathrm{o}}}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4 \left(\sum_{n\in\mathbb{Z}} q^{n^2}\right)^4$$

$$\hat{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = \frac{16\pi^4 L_2^{\mathrm{o}}(\hat{r}_2^{\mathrm{o}})}{(\lambda_2^{\mathrm{o}} - 1)^2 \ln^4\lambda_2^{\mathrm{o}}}\left(\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2}\right)^4 \left(\sum_{n\in\mathbb{Z}} (-q)^{n^2}\right)^4$$

where $q = \mathrm{e}^{-\pi^2/\ln\lambda_2^{\mathrm{o}}}$. *In particular, the extensions* $d^{\mathrm{o}}, \tilde{d}^{\mathrm{o}}, \hat{d}^{\mathrm{o}} : \mathcal{P}^{\mathrm{o}} \to \mathbb{R}$ *never vanish.*

Proof. The rational function $L_1^{\mathrm{o}}(r_1)$ has one (non-degenerate) critical point: $r_1 = 1$. From the properties of the Jacobian elliptic function $\mathrm{dn}(u, k)$, we deduce that the only critical points of the modular function $L_2^{\mathrm{o}}(r_2)$ are the points in the set $(\lambda_2^{\mathrm{o}})^{\mathbb{Z}/2}$, all of them being non-degenerate. This proves the first part of the lemma.

The computation of the first extension is immediate:

$$d^{\mathrm{o}}(\beta^{\mathrm{o}}) = L_1^{\mathrm{o}}(1)\left(L_2^{\mathrm{o}}(\tilde{r}_2^{\mathrm{o}}) - L_2^{\mathrm{o}}(\hat{r}_2^{\mathrm{o}})\right) = \frac{-4\lambda_2^{\mathrm{o}}}{(\lambda_2^{\mathrm{o}} - 1)^2}\left(\frac{2kK}{\ln\lambda_2^{\mathrm{o}}}\right)^2$$

where we have used that $\mathrm{dn}(0,k) = 1$ and $\mathrm{dn}(K,k) = k' = \sqrt{1-k^2}$. (Note that the unknown additive constant disappears.)

The second extension is obtained as follows:

$$\tilde{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = L_1^{\mathrm{o}}(1)\frac{\mathrm{d}^2 L_1^{\mathrm{o}}}{\mathrm{d}r_1^2}(1)L_2^{\mathrm{o}}(\tilde{r}_2^{\mathrm{o}})\frac{\mathrm{d}^2 L_2^{\mathrm{o}}}{\mathrm{d}r_2^2}(\tilde{r}_2^{\mathrm{o}}) = \frac{-16\lambda_2^{\mathrm{o}} L_2^{\mathrm{o}}(\tilde{r}_2^{\mathrm{o}})}{(\lambda_2^{\mathrm{o}}-1)^2}\left(\frac{2kK}{\ln\lambda_2^{\mathrm{o}}}\right)^2\left(\frac{2K}{\ln\lambda_2^{\mathrm{o}}}\right)^2$$

using the well known derivative rules for the Jacobian elliptic functions

$$\mathrm{sn}' = \mathrm{cn}\cdot\mathrm{dn} \qquad \mathrm{cn}' = -\,\mathrm{sn}\cdot\mathrm{dn} \qquad \mathrm{dn}' = -k^2\cdot\mathrm{sn}\cdot\mathrm{cn}$$

together with the properties

$$\mathrm{sn}(0,k) = 0 \qquad \mathrm{cn}(0,k) = 1 \qquad \mathrm{dn}(0,k) = 1$$

which can be found in [WW27, pp 492–3].

The computation of the last extension is very similar:

$$\hat{d}^{\mathrm{o}}(\beta^{\mathrm{o}}) = L_1^{\mathrm{o}}(1)\frac{\mathrm{d}^2 L_1^{\mathrm{o}}}{\mathrm{d}r_1^2}(1)L_2^{\mathrm{o}}(\hat{r}_2^{\mathrm{o}})\frac{\mathrm{d}^2 L_2^{\mathrm{o}}}{\mathrm{d}r_2^2}(\hat{r}_2^{\mathrm{o}}) = \frac{16 L_2^{\mathrm{o}}(\hat{r}_2^{\mathrm{o}})}{(\lambda_2^{\mathrm{o}}-1)^2}\left(\frac{2kK}{\ln\lambda_2^{\mathrm{o}}}\right)^2\left(\frac{2k'K}{\ln\lambda_2^{\mathrm{o}}}\right)^2.$$

Finally, the lemma follows from relations (5.17). □

By an argument of continuity, the properties *in* the oblate limit are still valid *close to* the oblate limit. In particular, the Melnikov potential (5.10) has only a couple of critical points close to the oblate limit and $d(\beta)$, $\tilde{d}(\beta)$, $\hat{d}(\beta)$ are non-zero if β is close enough to $\mathcal{P}^{\mathrm{o}}$. Consequently, we obtain a complete description of the situation close to the oblate limit, which is summarized in the following theorem.

Theorem 5.3. *The billiard inside the quartic perturbation (5.9) has exactly 16 primary bi-asymptotic orbits (the xz-specular and y-axial ones) close to the oblate limit, for small enough perturbations. Moreover, these orbits are transverse and the xz-specular and y-axial spatial bi-asymptotic orbits have different lengths.*

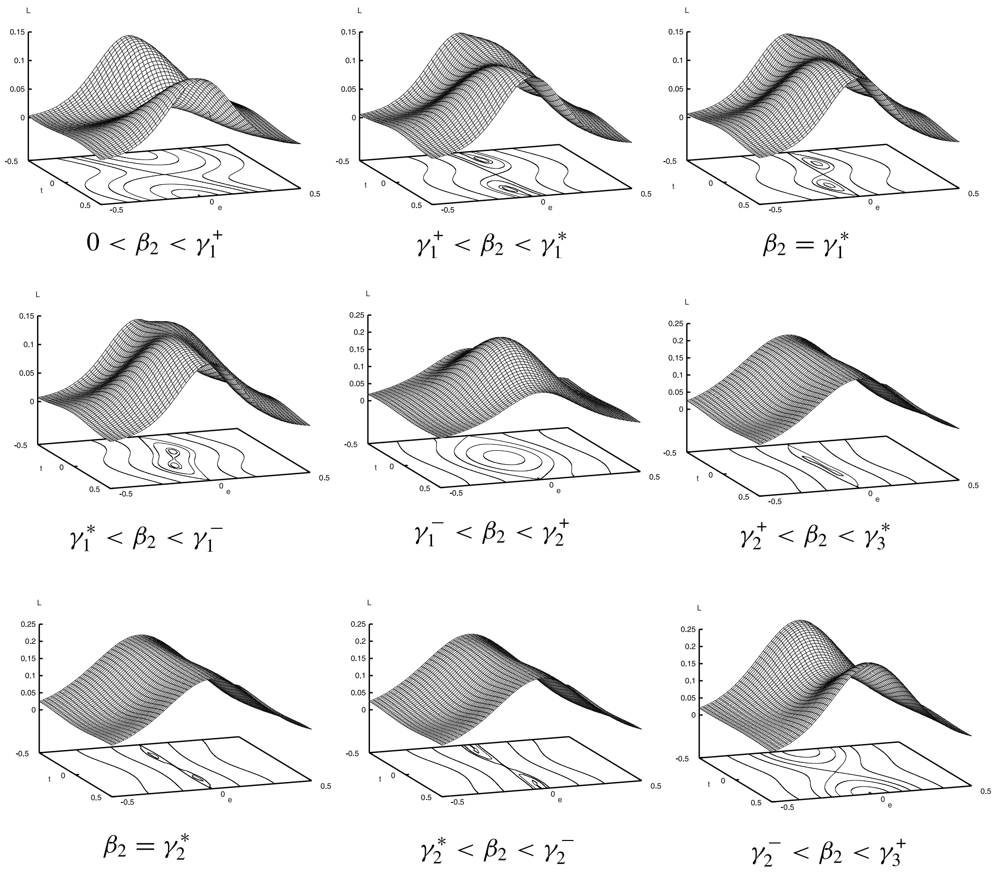
5.9.4. Experiments close to the prolate limit. Once we know what happens close to the flat, circular and oblate limits, we focus on the rest of the parameter space, emphasizing what happens close to the prolate limit.

The first experiment is to draw the graph and the level curves of the Melnikov potential (5.10) for several values of the parameter β to detect all of its critical points and to study the possible bifurcations. Hence, it is useful to introduce some variables best suited for the pictures. The time–energy variables $(t,e) \in \mathbb{R}^2$ defined by

$$r_1 = \lambda_1^t \lambda_2^e \qquad r_2 = \lambda_2^t/\lambda_1^e \tag{5.19}$$

are a good choice for several reasons. On the one hand, in these variables the Melnikov potential $L(t,e)$ is one-periodic in t, because the linear map $r = (r_1, r_2) \mapsto \Lambda r = (\lambda_1 r_1, \lambda_2 r_2)$ reads as $(t,e) \mapsto (t+1, e)$ in the time–energy variables. (This motivates the terminology. The action of the map increases the *time* t by one unit, but does not change the *energy* e.) On the other hand, $L(t,e)$ tends to zero exponentially fast when $|e| \to \infty$. This has to do with the fact that the quartic perturbation (5.9) preserves the horizontal section $\mathcal{Q}_{\mathrm{xy}} = \mathcal{Q} \cap \{z = 0\}$ and the vertical one $\mathcal{Q}_{\mathrm{xz}} = \mathcal{Q} \cap \{y = 0\}$ of the ellipsoid.

In figure 10 we have displayed the graphs and some level curves of the Melnikov potential $L(t,e)$ for $a = 1$, $\beta_1 = 0.5$, and several increasing values of β_2. In all the pictures, the range



$0 < \beta_2 < \gamma_1^+$ $\gamma_1^+ < \beta_2 < \gamma_1^*$ $\beta_2 = \gamma_1^*$

$\gamma_1^* < \beta_2 < \gamma_1^-$ $\gamma_1^- < \beta_2 < \gamma_2^+$ $\gamma_2^+ < \beta_2 < \gamma_3^*$

$\beta_2 = \gamma_2^*$ $\gamma_2^* < \beta_2 < \gamma_2^-$ $\gamma_2^- < \beta_2 < \gamma_3^+$

Figure 10. Graphs and level curves of the Melnikov potential for $a = 1$, $\beta_1 = 0.5$ and $\beta_2 = 0.2, 0.3, \gamma_1^*, 0.325, 0.41, 0.456, \gamma_2^*, 0.4575, 0.48$. The variables $(t, e) \in \mathbb{R}^2$ are given by $r_1 = \lambda_1^t \lambda_2^e$ and $r_2 = \lambda_2^t / \lambda_1^e$. The Melnikov potential is one-periodic in t and tends to zero when $|e| \to \infty$.

in the time–energy variables (t, e) is the square $[-0.5, 0.5]^2$. Nothing interesting falls out of this range, since L is one-periodic in t and tends to zero as $|e| \to \infty$.

In the time–energy variables, the Melnikov potential has the symmetries

$$L(-t, -e) = L(t, e) \qquad L\left(\tfrac{1}{2} - t, -e\right) = L\left(\tfrac{1}{2} + t, e\right)$$

which follow from the symmetry $L = L \circ \mathrm{I}$ (see lemma 5.2) and the periodicity in t. In particular, the points $(\tilde{t}, \tilde{e}) = (0, 0)$ and $(\hat{t}, \hat{e}) = \left(\frac{1}{2}, 0\right)$ are critical points, obtained from the already known critical points $\tilde{r} = (1, 1)$ and $\hat{r} = (\lambda_1^{1/2}, \lambda_2^{1/2})$ by the change of variables (5.19).

The first and last pictures in figure 10, which correspond, respectively, to $\beta_2 = 0.2$ and $\beta_2 = 0.48$, look the same from a qualitative point of view: $\tilde{r}$ is a saddle point, $\hat{r}$ is a global maximum and there are no more critical points. Thus, $H(0.5, 0.2) = H(0.5, 0.48) = 8 \times 2 = 16$. Nevertheless, when the parameter grows from $\beta_2 = 0.2$ up to $\beta_2 = 0.48$, the following bifurcations take place:

(a) at $\gamma_1^+ \approx 0.287\,199\,647\,426$, $\hat{r}$ becomes a saddle point and a couple (the first one) of global maxima are created from it;

(b) at $\gamma_1^* \approx 0.312\,364\,388\,028$, the saddle points $\tilde{r}$ and $\hat{r}$ lie on the same level curve of L: $L(\tilde{r}) = L(\hat{r})$, and the couple of maxima are halfway between them;
(c) at $\gamma_1^- \approx 0.335\,698\,528\,316$, the global maxima meet $\tilde{r}$, and they disappear, whereas $\tilde{r}$ becomes the only global maximum;
(d) at $\gamma_2^+ \approx 0.455\,587\,000\,258$, $\tilde{r}$ becomes again a saddle point and a couple (the second one) of global maxima are created from it;
(e) at $\gamma_2^* \approx 0.456\,737\,539\,206$, the saddle points $\tilde{r}$ and $\hat{r}$ are in the same level curve of L: $L(\tilde{r}) = L(\hat{r})$, and the second couple of maxima are halfway between them;
(f) at $\gamma_2^- \approx 0.457\,977\,789\,177$, the global maxima meet $\hat{r}$, and they disappear, whereas $\hat{r}$ becomes the only global maximum.

These bifurcations form a *cycle*, in the sense that after the last one we are just in the same situation as before the first one. (The three first bifurcations do not form a cycle, because the critical points $\tilde{r}$ and $\hat{r}$ have changed their types after them.) To gain more insight into this cycle, we have drawn in figure 11 the graphs of the functions d, $\tilde{d}$, and $\hat{d}$ for $\beta_1 = 0.5$ and $\beta_2 \in (0, \beta_1)$. Clearly, their changes of sign are associated with the above-described cycle bifurcations. On the one hand, a change in $\tilde{d}$ (respectively, $\hat{d}$), means that the type of $\tilde{r}$ (respectively, $\hat{r}$) changes between the global maximum and saddle point. On the other hand, $\tilde{r}$ and $\hat{r}$ lie on the same level curve of L if and only if d vanishes. Therefore, the bifurcation values γ_1^+, γ_1^* and γ_1^- are the first zeros of $\hat{d}$, d and $\tilde{d}$, whereas γ_2^+, γ_2^* and γ_2^- are the second ones of $\tilde{d}$, d and $\hat{d}$.

To gain more insight into the prolate limit, we introduce the variable $\eta > 0$, defined as

$$\beta_2 = \beta_1\big(1 - \lambda_1^{-2\eta}\big) \tag{5.20}$$

which tends to infinity when $\beta_2 \to \beta_1$. This logarithmic variable η is particularly well suited to elucidate the situation close to the prolate limit. For instance, the numerical computations strongly suggest that the functions d, $\tilde{d}$, and $\hat{d}$ tend to be one-periodic in η, whereas the distance between their zeros tend to $\frac{1}{2}$ as $\eta \to +\infty$ (see figure 11). This behaviour is general: it is observed for any fixed $\beta_1 \in (0, 1)$, and not only for $\beta_1 = 0.5$.

We now summarize the conclusions that can be obtained from this numerical study.

Numerical result 5.4. *Let $\mathcal{P} = \mathcal{P}_- \cup \mathcal{P}_0 \cup \mathcal{P}_+$ be the decomposition of the parameter space given in (5.13). Then the quartic perturbation (5.9) has exactly 16 (respectively, 32) primary bi-asymptotic orbits for $\beta \in \mathcal{P}_-$ (respectively, $\beta \in \mathcal{P}_+$). Moreover, the open sets $\mathcal{P}_\pm$ (respectively, the bifurcation set $\mathcal{P}_0$) are formed by infinitely many strips (respectively, curves) connecting the points $\beta = (0, 0)$ and $\beta = (1, 1)$, as displayed in figure 2. Finally, the bifurcation curves tend to the prolate limit at the geometric rate stated in formulae (5.11) and (5.12).*

To end this section, we present an amazing relation between the oblate and prolate limits. We hope to explain it in a further study. Let

$$\mathcal{P}^{\mathrm{p}} = \big\{\beta^{\mathrm{p}} = (\beta_1^{\mathrm{p}}, \beta_2^{\mathrm{p}}) : 0 < \beta_1^{\mathrm{p}} = \beta_2^{\mathrm{p}} < 1\big\}$$

be the hypotenuse of $\partial\mathcal{P}$. The function $d : \mathcal{P} \to \mathbb{R}$ has an oscillatory behaviour close to the prolate limit; that is, for $\beta \to \beta^{\mathrm{p}} \in \mathcal{P}^{\mathrm{p}}$. We have computed the limit amplitude of these oscillations as a function of the parameter β^{p}. This limit amplitude is related to the behaviour close to the oblate limit presented in lemma 5.4 as follows.

Numerical result 5.5. *Given a couple of degenerate parameters $\beta^{\mathrm{o}} = (1, \beta_2^{\mathrm{o}}) \in \mathcal{P}^{\mathrm{o}}$ and $\beta^{\mathrm{p}} = (\beta_1^{\mathrm{p}}, \beta_2^{\mathrm{p}}) \in \mathcal{P}^{\mathrm{p}}$ such that $\beta_2^{\mathrm{o}} = \beta_j^{\mathrm{p}}$, then*

$$\limsup_{\beta\to\beta^{\mathrm{p}}} d(\beta) = -d^{\mathrm{o}}(\beta^{\mathrm{o}})/2 \qquad \liminf_{\beta\to\beta^{\mathrm{p}}} d(\beta) = d^{\mathrm{o}}(\beta^{\mathrm{o}})/2.$$

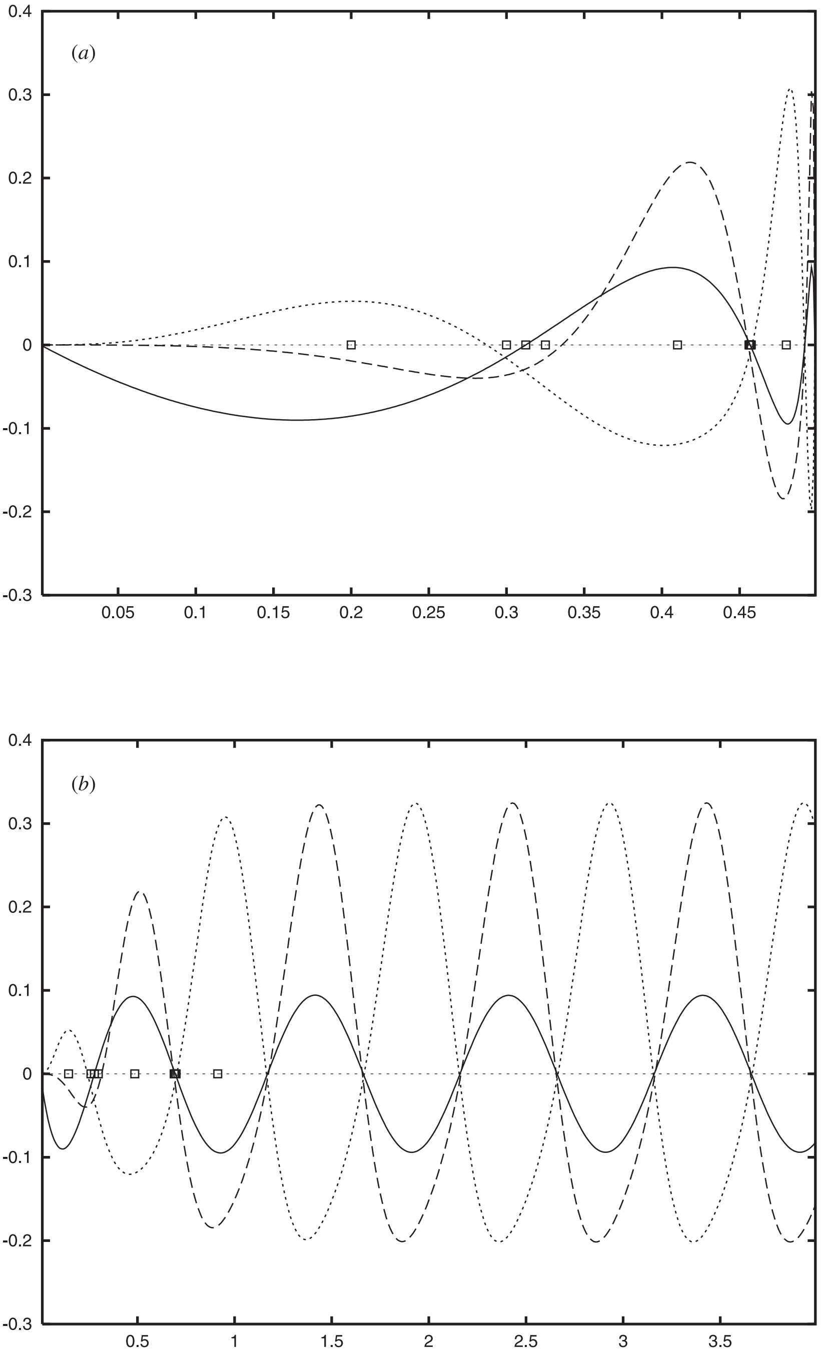


Figure 11. Graphs of d (full curve), $\tilde{d}$ (broken curve) and $\hat{d}$ (dotted curve) for $\beta_1 = 0.5$. (*a*) The horizontal variable is $\beta_2 \in (0, \beta_1)$. (*b*) The logarithmic variable defined in (5.20). The points marked with squares correspond to the values of β_2 for which the Melnikov potential have been shown in figure 10.

6. The high-dimensional case

In this section we describe the extension to the high-dimensional case of some of the results already presented for the spatial case. For the sake of brevity, and to avoid unnecessary repetition, we have adopted a compact style. In particular, all the proofs have been omitted.

To begin with, let us consider the generic ellipsoid

$$\mathcal{Q} = \left\{ q \in \mathbb{R}^{n+1} : \langle q, D^{-2} q \rangle = 1 \right\} \qquad D = \operatorname{diag}(d_0, \ldots, d_n) \qquad d_0 > \cdots > d_n > 0. \tag{6.1}$$

The chord joining the vertices $(-d_0, 0, \ldots, 0)$ and $(d_0, 0, \ldots, 0)$ is the diameter of the ellipsoid $\mathcal{Q}$. The set formed by the two-periodic points associated with the diameter

$$\mathcal{M}^{\mathrm{h}} = \left\{ m_+^{\mathrm{h}}, m_-^{\mathrm{h}} \right\} \qquad m_\pm^{\mathrm{h}} = (q_\pm^{\mathrm{h}}, v_\pm^{\mathrm{h}}) \qquad q_\pm^{\mathrm{h}} = (\pm d_0, 0, \ldots, 0) \qquad v_\pm^{\mathrm{h}} = (1, 0, \ldots, 0)$$

is a hyperbolic periodic set of the elliptic billiard map f whose unstable and stable invariant manifolds are doubled.

The sets $\mathcal{W}, \mathcal{W}^{\mathrm{u}}, \mathcal{W}^{\mathrm{s}}, \mathcal{W}_\pm^{\mathrm{u}}, \mathcal{W}_\pm^{\mathrm{s}}$ are defined as in the previous sections. Let

$$\mathcal{Q}(\kappa) = \left\{ q \in \mathbb{R}^{n+1} : \langle q, D(\kappa)^{-2} q \rangle = 1 \right\} \qquad D(\kappa)^2 = D^2 - \kappa^2 \operatorname{Id} \qquad \kappa \neq d_0, \ldots, d_n$$

be the family of non-degenerate quadrics confocal to the ellipsoid, and

$$\mathcal{Q}_j = \left\{ q = (q_0, \ldots, q_n) \in \mathbb{R}^{n+1} : q_j = 0, \qquad \sum_{i \neq j} \frac{{q_i}^2}{{d_i}^2 - {d_j}^2} = 1 \right\} \qquad j = 1, \ldots, n$$

be the family of *degenerate focal quadrics* of the ellipsoid. When $\kappa \to d_j$, the quadric $\mathcal{Q}(\kappa)$ flattens into a region of the hyperplane $\mathcal{H}_j := \{q \in \mathbb{R}^{n+1} : q_j = 0\}$ enclosed by (or outside) the degenerate focal quadric $\mathcal{Q}_j$; that is,

$$\mathcal{Q}_j = \mathcal{Q}^+(d_j) \cap \mathcal{Q}^-(d_j) \subset \mathcal{H}_j \qquad \mathcal{Q}^\pm(d_j) = \lim_{\kappa \to d_j^\pm} \mathcal{Q}(\kappa).$$

As already mentioned in the introduction, *any segment (or its prolongation) of a billiard trajectory inside the ellipsoid $\mathcal{Q} = \mathcal{Q}(0)$ is tangent to n fixed confocal quadrics $\mathcal{Q}(\kappa_1), \ldots, \mathcal{Q}(\kappa_n)$*. The first integrals of the family $\kappa_1, \ldots, \kappa_n$ can be computed by means of their relation with the family of involutive first integrals

$$I_j(m) = \frac{\prod_{i=1}^n (\kappa_i^2(m) - d_j^2)}{\prod_{i \neq j} (d_i^2 - d_j^2)} = p_j^2 + \sum_{i \neq j} \frac{(q_j p_i - q_i p_j)^2}{d_j^2 - d_i^2} \qquad j = 0, \ldots, n \tag{6.2}$$

where $m = (q, p)$, $q = (q_0, \ldots, q_n)$ and $p = (p_0, \ldots, p_n)$.

In a similar way to the spatial case, we shall say that a line is tangent to the degenerate focal quadric $\mathcal{Q}_j$ when it is contained in the hyperplane $\mathcal{H}_j$ or it intersects $\mathcal{Q}_j$. In particular, a line is tangent to all the degenerate focal quadrics if and only if it intersects all of them. All of these intersections and tangencies are understood in a projective sense.

Now, we can give the geometric characterization of the bi-asymptotic set, which is the tool to see that the invariant manifolds are doubled.

Proposition 6.1. $\mathcal{W} = \mathcal{W}^{\mathrm{u}} = \mathcal{W}^{\mathrm{s}} = \mathcal{M}_{(d_1, \ldots, d_n)}$, *where*

$$\begin{aligned} \mathcal{M}_{(d_1, \ldots, d_n)} &= \{ m \in \mathcal{M} : \kappa_j(m) = d_j \textit{ for all } 1 \leqslant j \leqslant n \} \\ &= \{ m \in \mathcal{M} : I_j(m) = 0 \textit{ for all } 1 \leqslant j \leqslant n \} \\ &= \{ m = (q, p) \in \mathcal{M} : q + \langle p \rangle \textit{ intersects } \mathcal{Q}_j \textit{ for all } 1 \leqslant j \leqslant n \}. \end{aligned}$$

Once we know the geometry of the bi-asymptotic set, we focus on its dynamics. In order to describe that dynamics, we must introduce the following notation.

Let $\hat{\mathbb{R}}$ be the extended real line, and i be the imaginary unit. Let $\mathrm{I} : \hat{\mathbb{R}}^n \to \hat{\mathbb{R}}^n$ be the involution $\mathrm{I}(r_1, \ldots, r_n) = (1/r_1, \ldots, 1/r_n)$, where $0^{-1} = \infty$ and $\infty^{-1} = 0$. If $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $s \in \mathbb{R}$, and m is a map defined on $\hat{\mathbb{R}}^n$ or $\mathbb{R}^n$, we denote by $m \circ \Lambda^s$ the map $r \mapsto m(\Lambda^s r) = m(\lambda_1^s r_1, \ldots, \lambda_n^s r_n)$.

We shall also adopt the standard multinomial notation $r^\epsilon = \prod_i {r_i}^{\epsilon_i}$ and $|\epsilon| = \sum_i \epsilon_i$, for multi-indices $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{N}^n$ and vectors $r = (r_1, \ldots, r_n) \in \hat{\mathbb{R}}^n$. In addition, if $r = (r_1, \ldots, r_n)$ and $1 \leqslant j \leqslant n$, we set $r_{\neq j} = (r_1, \ldots, r_{j-1}, r_{j+1}, \ldots, r_n)$. If A is an $n \times n$ matrix, $A_{\neq j}$ denotes the $(n-1) \times (n-1)$ matrix obtained by deleting the jth row and the jth column of A. Finally, given a multi-index $\epsilon \in \mathbb{N}^n$ and an $n \times n$ matrix $A = (\alpha_{ij})$, let $\Pi(A, \epsilon) := \prod_{\epsilon_i \neq \epsilon_j} \alpha_{ij}$, if $\{(i, j) : \epsilon_i \neq \epsilon_j\} \neq \emptyset$, and $\Pi(A, \epsilon) := 1$, otherwise.

Lemma 6.1. *Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ be the diagonal matrix whose entries are the characteristic multipliers*

$$\lambda_j = \frac{1 + e_j}{1 - e_j} \qquad e_j = \sqrt{1 - \beta_j} \qquad \beta_j = d_j^2/d_0^2.$$

Let $A = (\alpha_{ij})$ be the $n \times n$ symmetric matrix of positive elements defined by

$$\alpha_{ij}^2 = \begin{cases} (e_i + e_j)/(e_i - e_j) & \text{if} \quad i > j \\ \alpha_{ji}^2 & \text{if} \quad i < j \\ 1 & \text{if} \quad i = j. \end{cases}$$

Let $\tau, \tau_0, \tau_1, \ldots, \tau_n \in \mathbb{R}[r] = \mathbb{R}[r_1, \ldots, r_n]$ be the tau-polynomials

$$\tau(r) = \sum_{\epsilon \in \{0,2\}^n} \Pi(A, \epsilon) r^\epsilon$$

$$\tau_0(r) = \tau(\mathrm{i}r) = \sum_{\epsilon \in \{0,2\}^n} (-1)^{|\epsilon|/2} \Pi(A, \epsilon) r^\epsilon$$

$$\tau_j(r) = \Pi_{i=1}^n \alpha_{ij} \cdot r_j \cdot \sum_{\epsilon \in \{0,2\}^n} (-1)^{|\epsilon_{>j}|/2} \Pi(A_{\neq j}, \epsilon_{\neq j}) {r_{\neq j}}^{\epsilon_{\neq j}} \qquad 1 \leqslant j \leqslant n.$$

Let $\chi = (\tau_0/\tau, \ldots, \tau_n/\tau) : \hat{\mathbb{R}}^n \to \hat{\mathbb{R}}^{n+1}$. Let $q = D\chi : \mathbb{R}^n \to \mathcal{Q}$, $p = \chi \circ \Lambda^{-1/2} : \mathbb{R}^n \to \mathbb{S}^n$ and $m = (q, p) : \mathbb{R}^n \to \mathcal{M}$.

Then the maps $m_\pm^{\mathrm{u,s}} : \mathbb{R}^n \to \mathcal{M}$ defined by $m_\pm^{\mathrm{u}} = \pm m$ and $m_\pm^{\mathrm{s}} = \pm(-1)^n m \circ \mathrm{I}$ are natural parametrizations of the invariant manifolds $\mathcal{W}_\pm^{\mathrm{u,s}}$; that is, $m_\pm^{\mathrm{u,s}} : \mathbb{R}^n \to \mathcal{W}_\pm^{\mathrm{u,s}}$ are analytic diffeomorphisms such that

$$m_\pm^{\mathrm{u,s}}(0, \ldots, 0) = m_\pm^{\mathrm{h}} \qquad f \circ m_\pm^{\mathrm{u}} = m_\mp^{\mathrm{u}} \circ \Lambda \qquad f \circ m_\pm^{\mathrm{s}} = m_\mp^{\mathrm{s}} \circ \Lambda^{-1}.$$

Let us define the *dimension* of a billiard orbit inside an ellipsoid of $\mathbb{R}^{n+1}$ as the dimension of the vectorial subspace of $\mathbb{R}^{n+1}$ generated by its velocities. Billiard orbits have generically dimension $n + 1$, but two-periodic orbits have dimension one, and there exists a complete hierarchy of orbits between them.

The interest of this concept is twofold. On the one hand, a billiard orbit inside a generic ellipsoid of $\mathbb{R}^{n+1}$ and bi-asymptotic to the diameter of the ellipsoid is on the separatrix $\mathcal{S}$ if and only if it has dimension $n + 1$. On the other hand, a billiard orbit is homoclinic (respectively, heteroclinic) for the square of the billiard map f^2 if and only if its dimension is odd (respectively, even). Both claims generalize results given in the spatial case. For instance, a bi-asymptotic billiard orbit inside a spatial generic ellipsoid is homoclinic (respectively, heteroclinic) for the map f^2 if and only if it is spatial (respectively, planar).

As a corollary of the characterization of the separatrix given above, we find that:

Proposition 6.2. *The separatrix of the elliptic billiard has* 2^{n+1} *connected components:*

$$\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_- \qquad \mathcal{S}_\pm := \bigcup_{\sigma\in\{+,-\}^n} \mathcal{S}^\sigma_\pm \qquad \mathcal{S}^\sigma_\varsigma := \{\varsigma m(r) : \sigma_j r_j > 0 \textit{ for all } 1 \leqslant j \leqslant n\}.$$

The length of the orbits in the separatrix is

Proposition 6.3. Length $\mathcal{O} = -2\sum_{j=1}^n \sqrt{d_0^2 - d_j^2}$, *for all* $\mathcal{O} \subset \mathcal{S}$.

Next, we tackle the persistence of the symmetric bi-asymptotic orbits under symmetric perturbations. A hypersurface of $\mathbb{R}^{n+1}$ is *symmetric* when it is symmetric with regard to all of the coordinate axes of the Euclidean space $\mathbb{R}^{n+1}$. A billiard orbit inside a symmetric hypersurface will be called *symmetric* when its billiard configuration is symmetric with regard to some coordinate subspace of $\mathbb{R}^{n+1}$. There are two kinds of symmetric bi-asymptotic orbits of dimension $n+1$ inside a generic ellipsoid of $\mathbb{R}^{n+1}$. They are the *even-symmetric* ones and the *odd-symmetric* ones, which are symmetric with regard to the subspaces

$$E = \left\{q = (q_0, \dots, q_n) \in \mathbb{R}^{n+1} : q_j = 0 \text{ for all even } j\right\}$$

$$O = \left\{q = (q_0, \dots, q_n) \in \mathbb{R}^{n+1} : q_j = 0 \text{ for all odd } j\right\}$$

respectively. The following results are the high-dimensional versions of theorem 5.1 and corollaries 5.1 and 5.2.

Theorem 6.1. *Inside a generic ellipsoid of* $\mathbb{R}^{n+1}$ *there are* 2^{n+1} *even-symmetric (and* 2^{n+1} *odd-symmetric) billiard orbits bi-asymptotic to the diameter of dimension* $n+1$. *They persist under symmetric perturbations.*

Corollary 6.1. *The even-symmetric (respectively, odd-symmetric) bi-asymptotic orbits persist under any small perturbation, preserving the symmetry with regard to the subspace* E *(respectively,* O*).*

Corollary 6.2. *Inside a generic ellipsoid of* $\mathbb{R}^{n+1}$ *there are* $4(3^n - 1)$ *symmetric billiard orbits bi-asymptotic with the diameter. They persist under symmetric perturbations.*

In order to explain the corollary, let us consider all the coordinate sections of the ellipsoid that contain its diameter; that is, the sections of the form

$$\mathcal{Q}_J = \left\{q = (q_0, \dots, q_n) \in \mathcal{Q} : q_j = 0 \text{ for all } j \notin J\right\} \qquad J \subset \{1, \dots, n\}.$$

If two consecutive impact points are on $\mathcal{Q}_J$, the same happens to all the impact points, so $\mathcal{Q}_J$ gives rise to an invariant sub-system of the billiard map with the same properties as a billiard inside a generic ellipsoid of $\mathbb{R}^{m+1}$, where $m = \#J$. Hence, there are 2^{m+2} persistent symmetric bi-asymptotic orbits inside $\mathcal{Q}_J$ of dimension $m+1$ (see theorem 6.1). Then the total number of persistent symmetric bi-asymptotic orbits is $\sum_{m=1}^n \binom{n}{m} 2^{m+2} = 4(3^n - 1)$, since there are $\binom{n}{m}$ subsets $J \subset \{1, \dots, n\}$ such that $\#J = m$.

Finally, we consider the symmetric perturbations of the ellipsoid (6.1) defined by means of an implicit equation such as

$$\mathcal{Q} = \left\{ q \in \mathbb{R}^{n+1} : \langle q, D^{-2} q \rangle = 1 + \varepsilon P(q_1^2/d_1^2, \ldots, q_n^2/d_n^2) \right\} \tag{6.3}$$

for some function $P : \mathbb{R}^n \to \mathbb{R}$ such that $P(0, \ldots, 0) = 0$. We shall call this perturbation *entire* (respectively, *quadratic*) if the function P is entire (respectively, linear).

Lemma 6.2. *The Melnikov potential associated with the billiard inside (6.3) consists in* 2^{n+1} *copies of the function* $L : (0, +\infty)^n \to \mathbb{R}$ *defined by*

$$L(r) = d_0 \sum_{k \in \mathbb{Z}} \ell(\Lambda^k r) \qquad \ell(r) = \frac{\tau^2(r)}{\tau(\Lambda^{-1/2} r) \cdot \tau(\Lambda^{1/2} r)} P\left(\frac{\tau_1^2(r)}{\tau^2(r)}, \ldots, \frac{\tau_n^2(r)}{\tau^2(r)} \right) \tag{6.4}$$

where the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ *and the tau-polynomials* $\tau, \tau_1, \ldots, \tau_n$ *are defined in lemma 6.1. Moreover,* $L \circ \Lambda = L = L \circ \mathrm{I}$. *In particular,* $\tilde{r} = (1, \ldots, 1)$ *and* $\hat{r} = (\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ *are critical points of L.*

Our confidence in the *extended Birkhoff conjecture* has been strengthened in view of the following theorem.

Theorem 6.2. *An entire symmetric perturbation of a generic ellipsoid of* $\mathbb{R}^{n+1}$ *gives rise to a uniformly integrable billiard if and only if it is quadratic.*

The proof of this theorem follows the same lines as in the planar or spatial case. The crux of the argument is that the Melnikov potential (6.4) is analytic at the point $r^* = (1, \ldots, 1, \mathrm{i}) \in \mathbb{C}^n$ if and only if the perturbation (6.3) is quadratic. This also implies that the separatrix splits under any non-quadratic entire symmetric perturbation.

Acknowledgments

This work has been partially supported by the INTAS grants 97-10771 and 00-221. Two of the authors (AD and RR-R) have also been partially supported by the Spanish grant DGICYT BFM2000-0805 and the Catalan grant CIRIT 2000SGR–00027. One of the authors (YuF) was also supported by a CICYT grant. The authors wish to express their appreciation to S Bolotin, H Lomelí and V Rothos for several private communications. We are also grateful to the referee who pointed out to us the work of Lerman and Umanskiĭ. It is also a pleasure to thank H Koch, R de la Llave and A Olvera for very stimulating discussions and fruitful remarks.

Appendix A. Computations close to the flat limit

By linearity, it suffices to consider the monomial perturbations $R(s) = R(s_1, s_2) = s_1^i s_2^j$, for non-negative integers i and j. Then the perturbation (5.8) takes the form

$$\mathcal{Q}_\varepsilon = \left\{ q = (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 + \varepsilon (y/b)^{2i} (z/c)^{2j+2} \right\} \tag{A.1}$$

and the formulae used to prove lemma 5.3 are equivalent to the following ones:

(i) $a^{-1} L(\tilde{r}) = \delta_{i0} \beta_1^j \beta_2 + \mathrm{O}(\beta_2^2)$,

(ii) $a^{-1} \partial_{11} L(\tilde{r}) = \mathrm{O}(\beta_2^2)$,

(iii) $a^{-1} \partial_{12} L(\tilde{r}) = a^{-1} \partial_{21} L(\tilde{r}) = 2j \delta_{i0} e_1 \beta_1^j \beta_2 + \mathrm{O}(\beta_2^2)$,

(iv) $a^{-1} \partial_{22} L(\tilde{r}) = \mathrm{O}(\beta_2)$,

(i′) $a^{-1}L(\hat{r}) = \delta_{j0}\beta_2 + \mathrm{O}(\beta_2^2)$,
(ii′) $a^{-1}\partial_{11}L(\hat{r}) = -2\mathrm{i}\delta_{j0}\lambda_1^{-1}\beta_2 + \mathrm{O}(\beta_2^2)$,
(iii′) $a^{-1}\partial_{12}L(\hat{r}) = a^{-1}\partial_{21}L(\hat{r}) = \mathrm{O}(\beta_2^{5/2})$, and
(iv′) $a^{-1}\partial_{22}L(\hat{r}) = 2[\delta_{j1} - \mathrm{i}\delta_{j0}]\beta_1^{-1}\beta_2^3 + \mathrm{O}(\beta_2^4)$,

where δ_{ij} denotes the Kronecker delta.

The rest of the appendix is devoted to sketching the proofs of the formulae (i)–(iv), the other proofs being very similar.

To begin with, let $\nu = (\nu_1, \nu_2) = (1/\lambda_1, 1/\lambda_2)$. Then

$$\alpha^2 = \frac{e_2 + e_1}{e_2 - e_1} = \frac{\lambda_1\lambda_2 - 1}{\lambda_2 - \lambda_1} = \frac{1 - \nu_1\nu_2}{\nu_1 - \nu_2}$$

and the Melnikov potential associated with the perturbation (A.1) can be written as

$$L(r) = L(r;\nu;a) = a\sum_{k\in\mathbb{Z}} \ell(\nu_1^k r_1, \nu_2^k r_2;\nu) \qquad \ell(r) = \ell(r;\nu) = \frac{h(\nu)f(r_1)g(r_2)}{q(r;\nu)}$$

where

$$h(\nu) = 2^m\nu_1\nu_2(\nu_1 - \nu_2)^{m/2}(1 - \nu_1\nu_2)^{m/2}$$
$$f(r_1) = \left(1 + r_1^2\right)^{2j+2} r_1^{2i}$$
$$g(r_2) = \left(1 - r_2^2\right)^{2i} r_2^{2j+2}$$
$$q(r;\nu) = q^-(r;\nu)\left(q^0(r;\nu)\right)^{m-2} q^+(r;\nu)$$
$$q^+(r;\nu) = (\nu_1 - \nu_2)\left(1 + \nu_1\nu_2 r_1^2 r_2^2\right) + (1 - \nu_1\nu_2)\left(\nu_1 r_1^2 + \nu_2 r_2^2\right)$$
$$q^0(r;\nu) = (\nu_1 - \nu_2)\left(1 + r_1^2 r_2^2\right) + (1 - \nu_1\nu_2)\left(r_1^2 + r_2^2\right)$$
$$q^-(r;\nu) = (\nu_1 - \nu_2)\left(\nu_1\nu_2 + r_1^2 r_2^2\right) + (1 - \nu_1\nu_2)\left(\nu_2 r_1^2 + \nu_1 r_2^2\right)$$

and $m = 2i + 2j + 2 \geqslant 2$ is the order of the perturbation.

For further reference, we recall the relations

$$\beta_n = \frac{4\nu_n}{(1+\nu_n)^2} = 4\nu_n + \mathrm{O}(\nu_n^2) \qquad e_n = \frac{1-\nu_n}{1+\nu_n} \qquad n = 1, 2 \tag{A.2}$$

between the parameters $\beta = (\beta_1, \beta_2)$, $e = (e_1, e_2)$ and $\nu = (\nu_1, \nu_2)$.

Besides, for briefness, we will use the notation $\nu^k = (\nu_1^k, \nu_2^k)$, $h = h(\nu)$, $f_k = f(\nu_1^k)$, $f_k' = f'(\nu_1^k)$, $f_k'' = f''(\nu_1^k)$, $g_k = g(\nu_2^k)$, $g_k' = g'(\nu_2^k)$, $g_k'' = g''(\nu_2^k)$, $q_k = q(\nu^k;\nu)$, $\partial_n q_k = \partial_n q(\nu^k;\nu)$, $\partial_{n_1n_2}q_k = \partial_{n_1n_2}q(\nu^k;\nu)$, $\ell_k = \ell(\nu^k;\nu)$, $\partial_n\ell_k = \partial_n\ell(\nu^k;\nu)$, and $\partial_{n_1n_2}\ell_k = \partial_{n_1n_2}\ell(\nu^k;\nu)$, for $k \in \mathbb{Z}$ and $n, n_1, n_2 \in \{1, 2\}$.

In analogy with the proof of lemma 4.3 in the planar case, there are two main points in the current computations for the spatial case: to estimate the leading terms $\ell_0 = \ell(\tilde{r})$ and $\partial_{n_1n_2}\ell_0 = \partial_{n_1n_2}\ell(\tilde{r})$, and to bound the series $a^{-1}L(\tilde{r}) - \ell_0 = \sum_{k\neq 0}\ell_k$ and $a^{-1}\partial_{n_1n_2}L(\tilde{r}) - \partial_{n_1n_2}\ell_0 = \sum_{k\neq 0}\partial_{n_1n_2}\ell_k$.

In order to obtain the principal term in the small parameter ν_2 of ℓ_0 and the four partial derivatives $\partial_{n_1n_2}\ell_0$, we study, separately, the factors involved in the computations.

The factor $h = h(\nu)$ can be easily analysed:

$$h = 2^m\nu_1^{1+m/2}\nu_2 + \mathrm{O}(\nu_2^2).$$

Evaluating the functions $f(r_1)$ and $g(r_2)$ (together with their first and second derivatives) at the points $r_1 = \nu_1^0 = 1 = \tilde{r}_1$ and $r_2 = \nu_2^0 = 1 = \tilde{r}_2$, respectively, we arrive at

$$f_0 = 2^{2j+2} \qquad f_0' = m2^{2j+2} \qquad f_0'' = (m^2 - 2\mathrm{i})2^{2j+2}$$

and

$$g_0 = \delta_{i0} \qquad g_0' = m\delta_{i0} \qquad g_0'' = (m^2 - m)\delta_{i0} + 8\delta_{i1}.$$

Finally, we study the factor $q(r) = q(r;\nu)$ and its first partial derivatives at the point $r = \nu^0 = (1,1) = \tilde{r}$. Since q is the denominator, its partial derivatives will appear in the numerator. Due to that, it is interesting to express each partial derivative as a multiple of the original function q, at least in a first approximation in the small parameter ν_2. This can be accomplished after some long computations, yielding the following results:

$$\begin{aligned}
q_0 &= 2^m \nu_1^2 (1+\nu_1)^{m-2} + \mathrm{O}(\nu_2) \\
\partial_1 q_0 &= m q_0 \\
\partial_2 q_0 &= m q_0 + \mathrm{O}(\nu_2) \\
\partial_{11} q_0 &= m^2 q_0 \\
\partial_{12} q_0 &= \left(m^2 - m + 2 + \frac{2(m-2)\nu_1}{1+\nu_1} \right) q_0 + \mathrm{O}(\nu_2) \\
\partial_{22} q_0 &= (m^2 - 2) q_0 + \mathrm{O}(\nu_2).
\end{aligned}$$

Using all of these formulae, jointly with relations (A.2), we obtain

$$\begin{aligned}
\ell_0 &= \frac{h f_0 g_0}{q_0} = 4\delta_{i0} \left(\frac{4\nu_1}{(1+\nu_1)^2} \right)^j \nu_2 + \mathrm{O}(\nu_2^2) = \delta_{i0}\beta_1^j \beta_2 + \mathrm{O}(\beta_2^2) \\
\partial_{11}\ell_0 &= \frac{h g_0}{q_0^3} \left(q_0^2 f_0'' - 2 q_0 \partial_1 q_0 f_0' + \left[2(\partial_1 q_0)^2 - q_0 \partial_{11} q_0 \right] f_0 \right) \\
&= \frac{h g_0}{q_0} \left(f_0'' - 2m f_0' + m^2 f_0 \right) = 0 \\
\partial_{12}\ell_0 &= \frac{h}{q_0^3} \left(q_0^2 f_0' g_0' - q_0 \partial_1 q_0 f_0 g_0' - q_0 \partial_2 q_0 f_0' g_0 + [2\partial_1 q_0 \partial_2 q_0 - q_0 \partial_{12} q_0] f_0 g_0 \right) \\
&= \frac{h}{q_0} \left(f_0' g_0' - m(f_0' g_0 + f_0 g_0') + \left[m^2 + m - 2 - \frac{2(m-2)\nu_1}{1+\nu_1} \right] f_0 g_0 \right) + \mathrm{O}(\nu_2^2) \\
&= 8j\delta_{i0} \frac{1-\nu_1}{1+\nu_1} \left(\frac{4\nu_1}{(1+\nu_1)^2} \right)^j \nu_2 + \mathrm{O}(\nu_2^2) = 2j\delta_{i0} e_1 \beta_1^j \beta_2 + \mathrm{O}(\beta_2^2) \\
\partial_{22}\ell_0 &= \frac{h f_0}{q_0^3} \left(q_0^2 g_0'' - 2 q_0 \partial_2 q_0 g_0' + \left[2(\partial_2 q_0)^2 - q_0 \partial_{22} q_0 \right] g_0 \right) \\
&= \frac{h f_0}{q_0} \left(g_0'' - 2m g_0' + (m^2 + 2) g_0 \right) + \mathrm{O}(\nu_2^2) = \mathrm{O}(\beta_2).
\end{aligned}$$

Consequently, formulae (i)–(iv) follow from the estimates

$$a^{-1} L(\tilde{r}) = \ell_0 + \mathrm{O}(\beta_2^2) \qquad a^{-1}\,\mathrm{Hess}\, L(\tilde{r}) = \begin{pmatrix} \partial_{11}\ell_0 & \partial_{12}\ell_0 \\ \partial_{21}\ell_0 & \partial_{22}\ell_0 \end{pmatrix} + \mathrm{O}(\beta_2^2). \tag{A.3}$$

To check that the first estimate holds, we note that $a^{-1}L(\tilde{r}) = \ell_0 + 2\sum_{k>0}\ell_k$, since $\ell(r) = \ell(r^{-1})$. Moreover, using the bounds

$$0 < h(\nu) \leqslant 2^m \nu_2 \qquad 0 < f_k \leqslant 2^{2j+2} \qquad 0 < g_k \leqslant \nu_2^{2k(j+1)} \qquad q_k \geqslant 2^{-m}\nu_1^{m+1}\nu_2$$

we find that $0 < \ell_k \leqslant 2^{2m+2j+2}\nu_1^{-(m+1)}\nu_2^{2k(j+1)}$, for all $k > 0$. Therefore,

$$\left|a^{-1}L(\tilde{r}) - \ell_0\right| = 2\sum_{k\geqslant 1}\ell_k \leqslant 2^{2m+2j+3}\nu_1^{-(m+1)}\sum_{k\geqslant 1}\nu_2^{2k(j+1)} = \mathrm{O}(\nu_2^{2j+2}) \leqslant \mathrm{O}(\beta_2^2).$$

This completes the proof of the first estimate in (A.3). The second estimate can be proved in an analogous way, although with more work. We omit the details.

Appendix B. Computations with Jacobian elliptic functions

Here, we address the explicit computation of the series (4.9) and (5.16) in terms of the square of the Jacobian elliptic function $\mathrm{dn}(u) = \mathrm{dn}(u, k)$. It is possible to present a unified treatment of these series, using that both can be written as

$$L : (0, +\infty) \to \mathbb{R} \qquad L(r) = \sum_{k\in\mathbb{Z}} \ell(\lambda^k r) \qquad \ell(r) = \frac{16\mu r^4 + 4\nu(1+r^2)^2 r^2}{(1+r^2/\lambda)(1+r^2)^2(1+\lambda r^2)}$$

for some $\mu, \nu \in \mathbb{R}$ and $\lambda > 1$. In (4.9), $\mu = 1$ and $\nu = 0$. In (5.16), $\mu = -1$ and $\nu = 1$.

To fit this computation within the framework of elliptic functions, it is convenient to make the change of variables $r = \mathrm{e}^t$, so that the series above is transformed into

$$\bar{L}(t) = L(\mathrm{e}^t) = \sum_{k\in\mathbb{Z}} \bar{\ell}(t + kh) \qquad \bar{\ell}(t) = \ell(\mathrm{e}^t) = \frac{\mu + \nu\cosh^2 t}{\cosh(t - h/2)\cosh^2 t\cosh(t + h/2)}$$

with $h = \ln\lambda$. The function $\bar{L}(t)$ is elliptic; that is, it is meromorphic in the whole complex plane and has two complex periods independent over the reals: h and $\pi\mathrm{i}$.

We now recall that *elliptic functions are characterized (modulo additive constants) by their periods, poles and principal parts*: the difference of two elliptic functions with the same periods, poles and principal parts is a bounded entire function, and hence a constant function by Liouville's theorem. Therefore, we are naturally led to the search for the poles (and their principal parts) of the series $\bar{L}(t) = \sum_{k\in\mathbb{Z}}\bar{\ell}(t + kh)$.

First, the poles of $\bar{\ell}(t)$ are the points in the sets $\pi\mathrm{i}/2 + \pi\mathrm{i}\mathbb{Z}$ and $\pi\mathrm{i}/2 \pm h/2 + \pi\mathrm{i}\mathbb{Z}$. The poles $t_0 \in \pi\mathrm{i}/2 + \pi\mathrm{i}\mathbb{Z}$ are double, whereas the poles $t_0^{\pm} \in \pi\mathrm{i}/2 \pm h/2 + \pi\mathrm{i}\mathbb{Z}$ are simple. Moreover, $a_{-1}(t_0) = 0$, $a_{-2}(t_0) = -4\mu\lambda/(\lambda - 1)^2$ and $a_{-1}(t_0^+) + a_{-1}(t_0^-) = 0$, where $a_s(\tau)$ denotes the coefficient of the term $(t - \tau)^s$ in the Laurent expansion of $\bar{\ell}(t)$ around $t = \tau$. Hence, the elliptic function $\bar{L}(t) = \sum_{k\in\mathbb{Z}}\bar{\ell}(t + kh)$ is characterized (modulo an additive constant) by the following properties:

(i) its periods are h and $\pi\mathrm{i}$;
(ii) its poles are the points in the set $\pi\mathrm{i}/2 + h\mathbb{Z} + \pi\mathrm{i}\mathbb{Z}$; and
(iii) its principal part around any pole t_0 is $-4\mu\lambda(\lambda - 1)^{-2}(t - t_0)^{-2}$.

On the other hand, the square of the Jacobian elliptic function $\mathrm{dn}(u) = \mathrm{dn}(u, k)$ is characterized (modulo an additive constant) by the properties:

(i′) its periods are $2K$ and $2K'\mathrm{i}$;
(ii′) its poles are the points in the set $K'\mathrm{i} + 2K\mathbb{Z} + 2K'\mathrm{i}\mathbb{Z}$,

(iii′) the principal part around any pole u_0 is $-(u-u_0)^{-2}$, see [WW27, section 22]. Here, k is the *modulus*, $K = \int_0^{\pi/2}(1-k^2\sin u)^{-1/2}\,\mathrm{d}u$ is the *complete elliptic integral of the first kind*, and $K' = \int_0^{\pi/2}(1-k'^2\sin u)^{-1/2}\,\mathrm{d}u$, where k' is the *complementary modulus*: $k^2+k'^2=1$. Finally, the quantity $q = \mathrm{e}^{-\pi K'/K}$ is called the *nome*.

Therefore, if we take $q = \mathrm{e}^{-\pi^2/\ln\lambda}$, then $K' = K\pi/h$ and

$$\bar{L}(t) = \text{constant } + 4\mu\lambda(\lambda-1)^{-2}\,(2K/h)^2\,\mathrm{dn}^2\,(2Kt/h,k).$$

Finally, the formulae (4.10) and (5.18) follow using that $\mu = 1$ for the quartic planar perturbation and $\mu = -1$ for the quartic spatial one.

References

[AM78] Abraham R and Marsden J E 1978 *Foundations of Mechanics* (Menlo Park, CA: Benjamin-Cummings)

[AvM89] Adler M and van Moerbeke P 1989 The complex geometry of the Kowalewski–Painlevé analysis *Invent. Math.* **97** 3–51

[Bir27] Birkhoff G D 1927 *Dynamical Systems (Am. Math. Soc. Coll. Pub. vol 9)* (Providence, RI: American Mathematical Society)

[BK96] Berglund N and Kunz H 1996 Integrability and ergodicity of classical billiards in a magnetic field *J. Stat. Phys.* **83** 81–126

[BLS98] Bobenko A I, Lorbeer B and Suris Yu B 1998 Integrable discretizations of the Euler top *J. Math. Phys.* **39** 6668–83

[BM00] Bolotin S V and MacKay R S 2000 Periodic and chaotic trajectories of the second species for the n-centre problem *Celestial Mech. Dynam. Astron.* **77** 49–75

[Cus78] Cushman R 1978 Examples of nonintegrable analytic Hamiltonian vector fields with no small divisors *Trans. Am. Math. Soc.* **238** 45–55

[DLR01] Delshams A, Lomelí H and Ramírez-Ros R 2001 Canonical Poincaré–Melnikov method for maps *Preprint* in progress

[DR96] Delshams A and Ramírez-Ros R 1996 Poincaré–Melnikov–Arnold method for analytic planar maps *Nonlinearity* **9** 1–26

[DR97] Delshams A and Ramírez-Ros R 1997 Melnikov potential for exact symplectic maps *Commun. Math. Phys.* **190** 213–45

[DR98] Delshams A and Ramírez-Ros R 1998 Homoclinic orbits of twist maps and billiards *Symmetry and Perturbation Theory* ed D Bambusi and G Gaeta (Florence: Quaderni GNFM) pp 46–86

[DR99] Delshams A and Ramírez-Ros R 1999 Singular separatrix splitting and the Melnikov method: an experimental study *Exp. Math.* **8** 29–48

[Fed99] Fedorov Yu 1999 Classical integrable systems and billiards related to generalized Jacobians *Acta Appl. Math.* **55** 251–301

[Fed00] Fedorov Yu 2000 Discrete versions of some algebraic integrable systems related to generalized Jacobians *SIDE III—Symmetries and Integrability of Difference Equations (Sabaudia, 1998)* (Providence, RI: American Mathematical Society) pp 147–60

[GPB89] Glasser M L, Papageorgiou V G and Bountis T C 1989 Melnikov's function for two-dimensional mappings *SIAM J. Appl. Math.* **49** 692–703

[KMOC95] Koiller J, Markarian R, Oliffson-Kamphorst S and Pinto de Carvalho S 1995 Time-dependent billiards *Nonlinearity* **8** 983–1003

[KMOC96] Koiller J, Markarian R, Oliffson-Kamphorst S and Pinto de Carvalho S 1996 Static and time-dependent perturbations of the classical elliptical billiard *J. Stat. Phys.* **83** 127–143

[Koz98] Kozlova T V 1998 Nonintegrability of a rotating elliptic billiard *J. Appl. Math. Mech.* **62** 81–5

[KT91] Kozlov V V and Treshchëv D V 1991 *Billiards: a Genetic Introduction to the Dynamics of Systems with Impacts (Transl. Math. Monographs vol 89)* (Providence, RI: American Mathematical Society)

[Lev97] Levallois P 1997 Calcul d'une fonction de Melnikov et de ses zéros pour une perturbation algébrique du billard elliptique *Ergod. Theor. Dynam. Syst.* **17** 435–44

[Lom96] Lomelí H E 1996 Perturbations of elliptic billiards *Physica* D **99** 59–80

[Lom97] Lomelí H E 1997 Applications of the Melnikov method to twist maps in higher dimensions using the variational approach *Ergod. Theor. Dynam. Syst.* **17** 445–62

[LT93] Levallois P and Tabanov M B 1993 Séparation des séparatrices du billard elliptique pour une perturbation algébrique et symétrique de l'ellipse *C.R. Acad. Sci., Paris I* **316** 589–92

[LU94] Lerman L M and Umanskiĭ Ya L 1994 Classification of four-dimensional integrable Hamiltonian systems and Poisson actions of $\mathbb{R}^2$ in extended neighborhoods of simple singular points. II *Russ. Acad. Sci. Sb. Math.* **78** 479–506

[Mei92] Meiss J D 1992 Symplectic maps, variational principles and transport *Rev. Mod. Phys.* **64** 795–848

[MMP84] MacKay R S, Meiss J D and Percival I C 1984 Transport in Hamiltonian systems *Physica* D **13** 55–81

[Mum84] Mumford D 1984 *Tata Lectures on Theta II Jacobian Theta Functions and Differential Equations* (Boston, MA: Birkhäuser)

[MV91] Moser J and Veselov A P 1991 Discrete versions of some classical integrable systems and factorization of matrix polynomials *Commun. Math. Phys.* **139** 217–43

[RB85] Robnik M and Berry M V 1985 Classical billiards in magnetic fields *J. Phys. A: Math. Gen.* **18** 1361–78

[Sma65] Smale S 1965 Diffeomorphisms with many periodic points *Differential and Combinatorial Topology (A Symposium in Honor of Marston Morse)* ed S S Cairns (Princeton, NJ: Princeton University Press) pp 63–80

[Tab94] Tabanov M B 1994 Separatrices splitting for Birkhoff's billiard in symmetric convex domain, closed to an ellipse *Chaos* **4** 595–606

[Tab95] Tabachnikov S 1995 Billiards *Panor. Synth.* **1** vi+142

[Ves88] Veselov A P 1988 Integrable systems with discrete time and difference operators *Funct. Anal. Appl.* **22** 83–93

[WT85] Wojciechowski S and Tsiganov A V 1985 Integrable one-particle potentials related to the Neumann system and the Jacobi problem of geodesic motion on an ellipsoid *Phys. Lett.* A **107** 106–11

[WW27] Whittaker E T and Watson G N 1927 *A Course of Modern Analysis* (Cambridge: Cambridge University Press)