Jean Souchay
Rudolf Dvorak
Editors

LECTURE NOTES IN PHYSICS 790

Dynamics of Small Solar System Bodies and Exoplanets

Springer

Lecture Notes in Physics

The Lecture Notes in Physics

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching – quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic
- to serve as an accessible introduction to the field to postgraduate students and nonspecialist researchers from related areas
- to be a source of advanced teaching material for specialized seminars, courses and schools

Both monographs and multi-author volumes will be considered for publication. Edited volumes should, however, consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the managing editor at Springer:

Christian Caron
Springer Heidelberg
Physics Editorial Department I
Tiergartenstrasse 17
69121 Heidelberg / Germany
christian.caron@springer.com

J. Souchay
R. Dvorak (Eds.)

Dynamics of Small Solar System Bodies and Exoplanets

Springer

Editors
Jean Souchay
Observatoire de Paris
Dépt. Astronomie
Fondamental
61 avenue de l' Observatoire
75014 Paris
France
Jean.Souchay@obspm.fr

Rudolf Dvorak
Universität Wien
Astronomisches Institut
Sternwarte
Türkenschanzstr. 17
1180 Wien
Austria
dvorak@astro.univie.ac.at

Souchay J., Dvorak R. (Eds.): *Dynamics of Small Solar System Bodies and Exoplanets*, Lect. Notes Phys. 790 (Springer, Berlin Heidelberg 2010), DOI 10.1007/978-3-642-04458-8

Lecture Notes in Physics ISSN 0075-8450 e-ISSN 1616-6361
ISBN 978-3-642-04457-1 e-ISBN 978-3-642-04458-8
DOI 10.1007/978-3-642-04458-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009939566

Cover design: Integra Software Services Pvt. Ltd., Pondicherry

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book on recent investigations of the dynamics of celestial bodies in the solar and extra-Solar System is based on the elaborated lecture notes of a thematic school on the topic, held as a result of cooperation between the SYRTE Department of Paris Observatory and the section of astronomy of the Vienna University. Each chapter corresponds to a lecture of several hours given by its author(s). The book therefore represents a necessary and very precious document for teachers, students, and researchers in the field.

The first two chapters by A. Lemaître and H. Skokos deal with standard topics of celestial mechanics: the first one explains the basic principles of resonances in mechanics and their studies in the case of the Solar System. The differences between the various cases of resonance (mean motion, secular, etc.) are emphasized together with resonant effects on celestial bodies moving around the Sun. The second one deals with approximative methods of describing chaos. These methods, some of them being classical, as the Lyapounov exponents, other ones being developed in the very recent past, are explained in full detail. The second one explains the basic principles of resonances in mechanics and their studies in the case of the Solar System. The differences between the various cases of resonance (mean motion, secular, etc.) are emphasized together with resonant effects on celestial bodies moving around the Sun.

The following three chapters by A. Cellino, by P. Robutel and J. Souchay, and by M. Birlan deal with the recent improvements in the knowledge of the celestial mechanics of the Solar System and of the extra-Solar System. The discovery and the determination of various asteroid families in the two last decades from both their dynamical features and their physical characteristics constitute a tremendous step in the understanding of the constitution and the evolution of the asteroid belt as well as the Trojans. We explain how numerical integrations at a very large time scale can associate various bodies from a single parent one.

The astrometric space mission Gaia, to be launched in early 2012, will constitute a revolution in the precision of position and velocity determinations of celestial objects among which are the asteroids and the comets. In an extensive chapter Hestroffer, A. dell'Oro, A. Cellino, and P. Tanga present our current understanding relating to astro-photometric measurements and the dynamical properties of these bodies, as well as the dramatic improvements expected from the Gaia mission.

Comets are still a subject of deep investigation concerning their origins and the characteristics of the Oort's cloud, from which they are assumed to originate. Their dynamical evolution inside the Solar System strongly depends on their possible interactions with the large planets, in particular with Jupiter. A complete review on these objects is given by H. Rickmann. This is followed by a chapter by M. Fouchard explaining in full detail the way by which a perturbation from galactic tides and passing stars can trigger a mechanism leading to deviation of comets toward the inner Solar System.

A large part of the studies in celestial mechanics and dynamical astronomy is based on numerical integration. In an extensive chapter S. Eggl and R. Dvorak present various numerical methods used for solving the gravitational *N*-body problem and discuss their main properties.

Finally, the always-increasing number of recorded stellar systems with their escort of exoplanets leads to the fundamental questions of their dynamical stability as well as the existence of the zones in which conditions for life are gathered. E. Lohinger presents abundant examples of such systems and shows how their dynamical stability can be addressed.

We are sure that the present book will be very useful for any graduate student or specialist aiming at an up-to-date review of the most exciting topics in the fields of celestial mechanics and planetology of solar and extra-Solar Systems.

Both editors thank very strongly the Springer Editorial Board as well as the authors for their acceptance of the work and their nice contributions.

Paris, France — J. Souchay
Vienna, Austria — R. Dvorak
July 2009

Contents

Resonances: Models and Captures

A. Lemaître

Abstract The resonances in the Solar System are present everywhere and can be represented by simple models. This chapter presents a review of the main cases: mean motion, secondary, secular, spin orbit and gravitational resonances are introduced and modelled up by pendulum like or more sophisticated models. Dissipative mechanisms introducing slow variations of the parameters can produce capture into, jumps over or escapes from resonances. Hamiltonian dynamics and adiabatic invariant are combined to reproduce and understand these behviours.

1 Introduction

This chapter presents the basic models of resonance, playing a role (as first approximations) in the main situations of planetary systems: mean motion resonances, secular resonances, secondary resonances, spin–orbit resonances, and gravitational resonances. It shows how to reduce a complex problem to its most important resonant contribution and how to calculate captures into resonances or escapes from these resonances.

Of course most of these models are far too simple to describe the complex reality of a resonant N-body problem; however, they can give a qualitative idea about the dominant dynamics; the superposition of the various levels of resonance creates chaotic zones to be estimated and located.

This chapter is written in Hamiltonian formalism and intends to give the main tools to manipulate and model up a resonance in any context; it is not a review of the present state of the art of the resonances and their present knowledge. The references chosen in this chapter correspond to this peculiar and specific approach.

A. Lemaître (✉)
Unité de Systèmes Dynamiques, Département de Mathématique, FUNDP, Rempart de la Vierge 8, B5000 Namur, Belgium, anne.lemaitre@fundp.ac.be

Lemaître, A.: *Resonances: Models and Captures*. Lect. Notes Phys. **790**, 1–62 (2010)
DOI 10.1007/978-3-642-04458-8_1

2 The Hamiltonian Theory

First of all, let us remind the fundamental characteristics of the Hamiltonian formalism. A one degree of freedom Hamiltonian system is defined by a function $\mathcal{H}$ (called the *Hamiltonian*), function of q, p, and t, where q designates the variable, p the momentum, and t the time,

$$\mathcal{H} = \mathcal{H}(q, p, t)$$

and an associated set of two differential equations:

$$\dot{q} = \frac{\partial \mathcal{H}}{\partial p}(q, p, t),$$
$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial q}(q, p, t).$$

The Hamiltonian $\mathcal{H}$ is called *autonomous* if it does not depend explicitly on the time: in this case, it is a first integral or a constant of the motion:

$$\frac{d\mathcal{H}}{dt} = \frac{\partial \mathcal{H}}{\partial q}\,\dot{q} + \frac{\partial \mathcal{H}}{\partial p}\,\dot{p} = 0.$$

We introduce a new set of variables (Q, P) depending on q, p, and t, defined by

$$Q = Q(q, p, t),$$
$$P = P(q, p, t).$$

We consider that the inverse of this time-dependent transformation $\sqcup$ is also defined symbolically by

$$q = q(Q, P, t),$$
$$p = p(Q, P, t).$$

This transformation is *canonical* if for any Hamiltonian $\mathcal{H}(q, p, t)$ there exists a function $\mathcal{K}(Q, P, t)$ so that the differential equations system associated to $\mathcal{H}$ is transformed into a new system with respect to $\mathcal{K}$ which is also Hamiltonian, i.e., which can be written as

$$\dot{Q} = \frac{\partial \mathcal{K}}{\partial P}(Q, P, t) \qquad \dot{P} = -\frac{\partial \mathcal{K}}{\partial Q}(Q, P, t).$$

The Hamiltonian of the problem expressed in Q and P is given by

$$\mathcal{K}(Q, P, t) = \mu\, \mathcal{H}(q(Q, P, t), p(Q, P, t), t) + \mathcal{R}(Q, P, t),$$

where $\mu(t)$ is called the *multiplier* and $\mathcal{R}$ the *remaining function*; they depend on the transformation $\mathcal{T}$ and not on the initial Hamiltonian $\mathcal{H}$ (see [11] for more details). If the transformation $\mathcal{T}$ is independent of t, $\mathcal{R} = 0$.

The role of $\mu(t)$ is not fundamental; a simple scaling can easily eliminate this parameter:

$$Q' = \alpha\ Q \qquad \text{and} \qquad P' = \beta\ P \qquad \text{with} \qquad \frac{1}{\mu} = \alpha\ \beta.$$

The canonical transformations of parameter $\mu = 1$ are also called the *symplectic transformations*.

This new canonical set (Q, P) is not always given by an explicit relationship with q, p, and t; it can be introduced in a more implicit way, through a generic function $\mathcal{S} = \mathcal{S}(q, P, t)$ (function of the *old* variable q and of the *new* momentum P) which defines the canonical transformation by the partial differential equations:

$$p = \frac{\partial \mathcal{S}}{\partial q} \quad \text{and} \quad Q = \frac{\partial \mathcal{S}}{\partial P}.$$

For example, for the identical transformation, this generic function is simply $\mathcal{S}(q, P) = q\,P$.

We can generalize this one degree of freedom approach to n degrees of freedom; the *phase space* is then of dimension $2n$, n dimensions for the variables q_i or Q_i, and n dimensions for the momenta p_i or P_i.

3 The Action-Angle Variables

Let us first consider a one degree of freedom autonomous integrable Hamiltonian: $H(q, p) = h$. Even in simple models, the frequency associated to the variable q is not constant, it is dependent on the momentum p and on q itself: $\dot{q} = \frac{\partial H}{\partial p}(q, p)$.

Among all the possible canonical transformations, we are interested in the so-called *action-angle* ones, resulting in a Hamiltonian function depending only on the new momentum (and not on the new angle Ψ):

$$(q, p) \Rightarrow (\Psi, J) \quad \text{so that} \quad H(q, p) = \mathcal{K}(-, J) = K(J). \tag{1}$$

We introduce a generic function $\mathcal{S}(q, J)$ so that $p = \frac{\partial \mathcal{S}}{\partial q}$ and $\Psi = \frac{\partial \mathcal{S}}{\partial J}$, determined by the *Hamilton–Jacobi* equation:

$$H\left(q, \frac{\partial \mathcal{S}}{\partial q}(q, J)\right) = K(J).$$

If we impose that Ψ is an angular variable, increasing by 2π along a complete circuit on a periodic orbit, we can identify J (with a correcting factor 2π) with the *area* enclosed by the trajectory, and the (constant) frequency is now ω (see again [11]):

$$J = \frac{1}{2\pi} \oint p \, dq \quad \text{and} \quad \omega = \dot{\Psi} = \frac{\partial K}{\partial J}.$$

4 The Restricted Three-Body Problem

Let us remind first the physical context in which the different types of resonances will be encountered. For this first part, the bodies are considered as point masses and their motions are described by pure gravitational interactions.

The simplest model to encounter a resonance is the *restricted three-body problem*, starting with a classical two-body configuration, where the small mass is then perturbed by a larger external body on a simplified orbit.

4.1 Two-Body Hamiltonian Formulation

Let us start with the two-body problem, where the central mass is denoted by $\mathcal{M}$; we follow the motion of a test mass m. Its orbit is an ellipse, and the focus of this ellipse is the barycenter of both masses. With respect to an inertial frame, we introduce the classical and less classical elliptic elements (Fig. 1):

- a the semi-major axis
- e the eccentricity
- i the inclination
- ω the argument of the pericenter
- Ω the longitude of the ascending node
- M the mean anomaly (υ is the true anomaly)

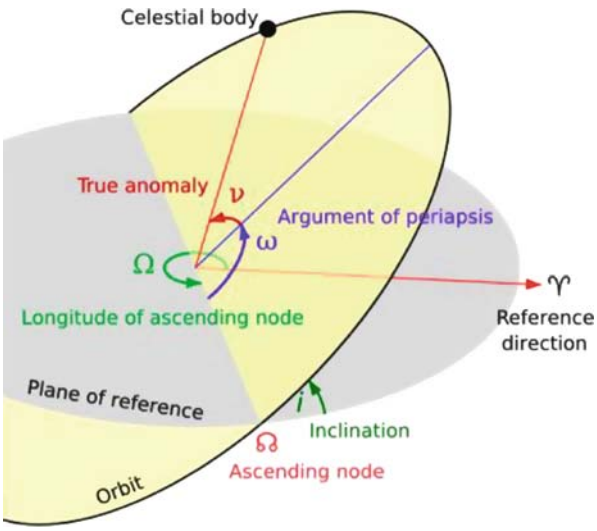

Fig. 1 Definition of the elliptic elements: a, e, i, ω, Ω, and υ

This set of elements is not canonical; to write them in a Hamiltonian formalism, we keep three angles as variables and we calculate the associated momenta (by a transformation of Matthieu) to get Delaunay elements:

Variables	Associated momenta	Dynamical equations	
M	$L = \sqrt{\mu a}$	$\dot{M} = \dfrac{\partial \mathcal{H}}{\partial L}$	$\dot{L} = -\dfrac{\partial \mathcal{H}}{\partial M}$
ω	$G = L\sqrt{1-e^2}$	$\dot{\omega} = \dfrac{\partial \mathcal{H}}{\partial G}$	$\dot{G} = -\dfrac{\partial \mathcal{H}}{\partial \omega}$
Ω	$H = G\cos i$	$\dot{\Omega} = \dfrac{\partial \mathcal{H}}{\partial H}$	$\dot{H} = -\dfrac{\partial \mathcal{H}}{\partial \Omega}$

where

$$\mu = \frac{\mathcal{G}\mathcal{M}^3}{(\mathcal{M}+m)^2} \simeq \mathcal{G}\mathcal{M} \quad \text{if} \quad \mathcal{M} >> m, \quad \text{with} \quad \mathcal{G} \text{ the gravitational constant.}$$

The Hamiltonian $\mathcal{H}$ coincides with the energy of the two-body problem called $\mathcal{H}_{2B}$:

$$\mathcal{H} = \mathcal{H}_{2B} = -\frac{\mu}{2a} = -\frac{\mu^2}{2L^2},$$

and we conclude obviously that L, G, H (and consequently a, e, and i), ω, and Ω are constants of motion; $M = nt + M_0$ where M_0 is the sixth initial condition and

$$n = \frac{\partial \mathcal{H}_{2B}}{\partial L} = \frac{\mu^2}{L^3} \quad \text{is the mean motion.}$$

This set is degenerate for $e = 0$ (no definition of ω) and for $i = 0$ (no definition of Ω). This is the reason for which we prefer another set of variables—momenta, called the *modified Delaunay elements* and defined as

Variables	Associated momenta	Dynamical equations	
$\lambda = M + \omega + \Omega$	L	$\dot{\lambda} = \dfrac{\partial \mathcal{H}}{\partial L}$	$\dot{L} = -\dfrac{\partial \mathcal{H}}{\partial \lambda}$
$p = -\omega - \Omega$	$P = L - G$	$\dot{p} = \dfrac{\partial \mathcal{H}}{\partial P}$	$\dot{P} = -\dfrac{\partial \mathcal{H}}{\partial p}$
$q = -\Omega$	$Q = G - H$	$\dot{q} = \dfrac{\partial \mathcal{H}}{\partial Q}$	$\dot{Q} = -\dfrac{\partial \mathcal{H}}{\partial q}$

The choice here is to keep the new momenta P and Q positive for the ellipses ($e < 1$), which induces the changes of signs in the angles p and q. If we choose to keep the initial signs of the angles, we have to pay attention to the negative signs of the momenta in the canonical transformation to cartesian coordinates.

The momentum P is proportional to the square of the eccentricity and Q to the square of (the sine of) the inclination.

4.2 The Third Body Perturbation

The potential generated by a third body (of mass m' and of position $\mathbf{s}$ in the chosen reference frame) introduces a perturbation on the motion of the small body (of mass m and position $\mathbf{r}$) which can be expressed by

$$V = -\mathcal{G}m' \left(\frac{1}{\Delta} - \frac{\mathbf{s} \cdot \mathbf{r}}{s^3} \right)$$

where $\Delta = \|\mathbf{r} - \mathbf{s}\|$.

We introduce the elliptic elements for both masses: a, e, i, ω, Ω, M, and λ have already been introduced for m, completed by υ the true anomaly and $\theta = \upsilon + \omega + \Omega$ the true longitude, and similar quantities but primed for the mass m': a' the semi-major axis, e' the eccentricity, i' the inclination, ω' the argument of pericenter, Ω' the longitude of the ascending node, M' the mean anomaly, $\lambda' = M' + \omega' + \Omega'$ the mean longitude, υ' the true anomaly, and $\theta' = \upsilon' + \omega' + \Omega'$ the true longitude. Of course the corresponding modified Delaunay angles are also defined for m', called p' and q', and are associated to the momenta P' and Q', linked to $L' = \sqrt{\mu' a'}$. n' denotes the mean motion of the third body, $n' = \frac{\mu'^2}{L'^3}$, with $\mu' = \frac{\mathcal{G}\mathcal{M}^3}{(\mathcal{M}+m')^2}$.

As we are interested in the motion of the small mass m, we consider in the *restricted problem* that the mass m' is not affected by m; consequently, all the primed variables are *known functions of time*, solutions of a two-body problem in the simplest cases, or of a full body planetary problem (not including the test mass m) in the most complete analyses.

We introduce the Legendre polynomials (here in the case of an outer perturber, with $s > r$):

$$V = -\frac{\mathcal{G}m'}{s} \sum_{l=2}^{\infty} \left(\frac{r}{s} \right)^l P_l(\cos \Psi) \quad \text{with} \quad \mathbf{r} \bullet \mathbf{s} = rs \cos \Psi,$$

where the symbol $\bullet$ designates the scalar product. Using the series expansions in e and i and Fourier developments (see classical references, like [37]), we can write the potential in the following form:

$$V = -\mathcal{G}m' \sum_{(\ell)=(\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6)} \mathcal{S}_{(\ell)}(a, a', e, e', i, i') \cos(\ell_1\lambda' + \ell_2\lambda + \ell_3\omega' + \ell_4\omega + \ell_5\Omega' + \ell_6\Omega), \tag{2}$$

which we express in Delaunay-modified canonical variables and momenta:

$$V = V(\lambda, p, q, L, P, Q, \lambda', p', q', L', P', Q')$$

to add to the two-body Hamiltonian $\mathcal{H}_{2B}$:

$$\mathcal{H} = -\frac{\mu^2}{2L^2} + V(\lambda, p, q, L, P, Q, \underbrace{\lambda', p', q', L', P', Q'}_{\text{known functions of } t})$$
$$= -\frac{\mu^2}{2L^2} + V(\lambda, p, q, L, P, Q, t). \tag{3}$$

A similar expression can be deduced for the case $s < r$ (see [37]).

In most problems of the Solar System, $\mathcal{M} >> m' >> m$ as in the problem *Sun-Jupiter-Asteroid* or *Earth-Moon-artificial satellite* or *Saturn-natural satellite-particle*, for which we assume that $\mu = \mathcal{G}(M + m) \simeq \mathcal{G}(M + m') \simeq \mathcal{G}M$.

4.3 The Angles and Their Frequencies

The next step is to classify the frequencies of the different angles with respect to each other. In our Solar System, we have clear separations of the frequencies, which means of the associated periods. Again, we are going to take this scale for this chapter, but it is obvious that other orders of magnitude could also be considered and treated by the same tools.

In this context we can conclude, from the Hamiltonian differential equations, that λ' and λ have larger frequencies (and then shorter periods) than p, p', q, and q'. We shall, therefore, refer to λ and λ' as short periodic angles and to p, p', q, and q' as long periodic or *secular* angles.

As an example, for an asteroid in the main belt, the period of λ is about a few years, while the periods of p and q are of the order of 10^4 or 10^5 years.

5 The Mean Motion Resonances

5.1 Simplifications

First of all, we are looking for resonances between the short periodic angles, which means λ and λ'; their frequencies are given in first approximations by the mean motions n and n'. We introduce a series of hypotheses which lead to a simplified model, describing this type of resonance.

As first simplification, we consider that both orbits are coplanar (we choose $i = i' = 0$) which means that $Q = Q' = 0$ and that q and q' do not appear anymore (by D'Alembert characteristic) which leads to a reduced potential V:

$$V = -\frac{\mu m'}{\mathcal{M}} \sum_{k,i_1,i_2,j_1,j_2} P^k_{i_1,i_2,j_1,j_2}(a, a')\, e^{2i_1+|j_1|} e'^{2i_2+|j_2|} \cos\left[(k + j_1)\lambda - (k - j_2)\lambda' + j_1 p + j_2 p'\right]. \tag{4}$$

The coefficients $P^k_{i_1,i_2,j_1,j_2}$ are functions of a and a' combinations of Laplace coefficients.

This expression is written with explicit reference to the powers of the two eccentricities, e and e', because it checks the D'Alembert characteristic: it means that the powers of the eccentricities are always greater than the corresponding multiples of the pericenter and that these two integers have the same parity:

$$2i_s + |j_s| \geq j_s, \quad 2i_s + |j_s| \text{ has the same parity as } j_s \quad \text{for } s = 1, 2.$$

Let us remind that, to refer to the Hamiltonian variables and momenta, $e = e(L, P)$ and $a = a(L)$.

In the general case, $a' = a'(t)$, $e' = e'(t)$, and $p' = p'(t)$, but in this present simplified context, we shall consider that the third body evolves on a fixed (planar) Keplerian orbit, $a' = a'_0$, $e' = e'_0$ and $p' = p'_0$, characterized by a constant mean motion n'.

We obtain a two degree of freedom Hamiltonian, time dependent through the (known) motion of m' (through the mean longitude $\lambda' = n't + \lambda'_0$), with the disparition of two variables, λ and p, and two momenta, L and P.

Finally, for the most simplified case, we consider that the perturbing body m' lies on a *circular* orbit, which means $e'_0 = 0$ and the disappearing of p'_0, to get

$$\mathcal{H}(\lambda, L, p, P, t) = -\frac{\mu^2}{2L^2} - \frac{\mu m'}{\mathcal{M}} \sum_{k, i_1, j_1} P^k_{i_1, j_1}(a, a')\, e^{2i_1 + |j_1|} \cos[(k + j_1)\lambda - k\,\lambda' + j_1\, p], \tag{5}$$

in which e' has disappeared, as well as the longitude of the pericenter p'; the only angle still defined for the perturbing mass m' is λ', the mean longitude.

5.2 The Resonance

We can now introduce the *mean motion resonance*. The two resonant frequencies are here the mean motions of the masses m and m'. The motion is resonant if the ratio of the two frequencies is very close to the ratio of two small integers, i.e.,

$$\frac{n}{n'} = \frac{j + i}{j} \quad \text{with } (j + i) \text{ and } j \text{ incommensurable small integers.}$$

It means that the mass m performs $j + i$ revolutions, while the mass m' performs j revolutions. If $a < a'$ the resonance is *internal or inner*, the orbit of m is inside the orbit of m' and $j > 0$; if $a > a'$ the resonance is *external or outer*, the orbit of m is outside that of m' and $j < 0$ (Fig. 2).

This also means that the ratio of the two semi-major axes is blocked to a specific value given by

$$\frac{n}{n'} = \frac{\mu^{\frac{1}{2}}}{a^{\frac{3}{2}}} \frac{a'^{\frac{3}{2}}}{\mu^{\frac{1}{2}}} = \left(\frac{a'}{a}\right)^{\frac{3}{2}} = \frac{j + i}{j} \quad \Rightarrow \quad a_{res} = \left(\frac{j}{j + i}\right)^{\frac{2}{3}} a'. \tag{6}$$

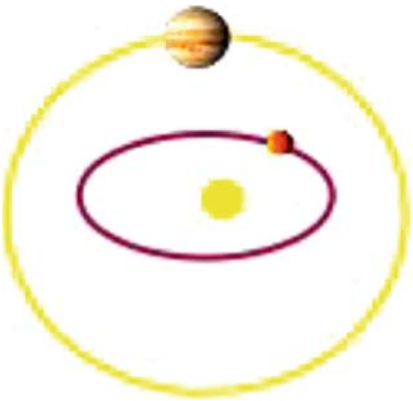

Fig. 2 Schematic view of a mean motion resonance: the inner planet performs $(j + i)$ revolutions when the outer (perturbing) one performs j revolutions

The integer $i = (j + i) - j$ (difference between the numerator and denominator of the quotient) is called the *order of the resonance*.

The main resonances are given by the following ratios of the mean motions (as for the Galilean satellites represented in Fig. 3)

n/n'	2/1	3/2	3/1	5/2	7/3	1/2	2/3	1/3
j	1	2	1	2	3	−2	−3	−3
i = order	1	1	2	3	4	1	1	2 .

The coefficients $P^k_{i_1,j_1}(a, a')$ are functions of the ratio $\frac{a}{a'}$ or $\frac{a'}{a}$ (corrected by a power of $a' = a'_0$) following the type of the resonance, inner or outer. This develop-

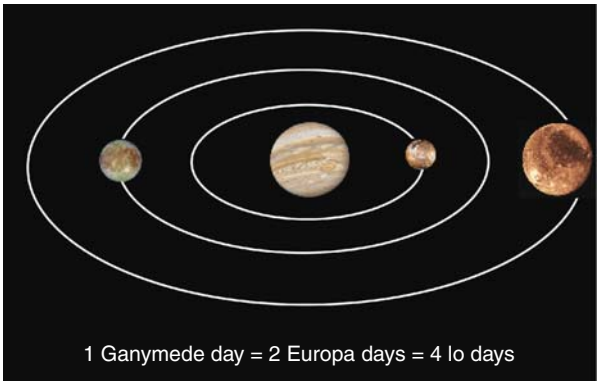

Fig. 3 Three Galilean satellites (Io, Europa, and Ganymede) blocked in mean motion resonances: the periods of revolution are in the ratios 1, 2, and 4

ment comes from an expansion, which means that this ratio has to be smaller than 1. The case of the resonance 1/1, like the Trojans asteroids, is analyzed through a different approach which will not be presented in this chapter (see Chap. 4).

5.3 The Time Dependence

The two degree of freedom Hamiltonian (5) is time dependent through the variable $\lambda' = n't + \lambda'_0$; it means that any canonical transform performed on this Hamiltonian will depend on time and will introduce corrective terms in the Hamiltonian function. However, because this dependence in time is purely linear, the usual way of tackling the problem is to introduce a third variable λ' combined to an artificial corresponding momentum Λ' so to get an autonomous three degree of freedom Hamiltonian $\mathcal{H}' = \mathcal{H} + n'\Lambda'$:

$$\mathcal{H}'(\lambda, L, p, P, \lambda', \Lambda') = -\frac{\mu^2}{2L^2} + n'\Lambda' - \frac{\mu m'}{\mathcal{M}} \sum_{k, i_1, j_1} P^k_{i_1, j_1}(L)\, e(L, P)^{2i_1 + |j_1|} \cos\left[(k + j_1)\lambda - k\,\lambda' + j_1\, p\right].$$

The third degree of freedom is associated with the differential equations:

$$\dot{\lambda}' = \frac{\partial \mathcal{H}'}{\partial \Lambda'} = n' \quad \text{(already known)} \quad \text{and} \quad \dot{\Lambda}' = -\frac{\partial \mathcal{H}'}{\partial \lambda'} \quad \text{(never used)}.$$

5.4 The Resonant Angle

In a specific region, where the semi-major axis $a \simeq (\frac{j}{j+i})^{\frac{1}{3}} a'$, there is a resonant combination of the two angles λ and λ' which has a smaller frequency (close to 0) than all the other linear combinations of these two angles, which should induce a long periodic motion.

The idea of a resonant model is then to isolate this specific frequency, to follow its long-term dynamics and to forget about all the other small short periodic variations. The technique consists in isolating this combination in a specific canonical variable, in averaging over all the other angles except the selected one and in reducing the problem to a one degree of freedom averaged resonant problem.

Let us define the resonant angle σ to be introduced in the canonical transformation:

$$\sigma = \frac{j+i}{i}\lambda' - \frac{j}{i}\lambda + p.$$

σ has a frequency close to 0, because of the resonance between λ and λ', and because $p = -\omega - \Omega$ is a slow angle.

We introduce a canonical transformation:

$$(\lambda, L, p, P, \lambda', \Lambda') \Rightarrow (\lambda, N, \sigma, S, \lambda', \Gamma'),$$

with the three new momenta

$$N = L + \frac{j}{i}P, \quad S = P, \quad \text{and} \quad \Gamma' = \Lambda' - \frac{j+i}{i}P$$

easily deduced from the (sufficient) condition of canonicity, i.e., the conservation of a differential form:

$$N\, d\lambda + S\, d\sigma + \Gamma'\, d\lambda' = L\, d\lambda + P\, dp + \Lambda'\, d\lambda'.$$

The next step is to introduce the new variables and momenta in the (autonomous three degree of freedom) Hamiltonian $\mathcal{H}'$:

$$\mathcal{H}'(\lambda, N, \sigma, S, \lambda', \Gamma') = -\frac{\mu^2}{2L^2} + n'\Gamma' - n'\frac{j}{i}S - \mu\,\frac{m'}{\mathcal{M}} \sum_{k, i_1, j_1} P^k_{i_1, j_1}\, e^{2i_1+|j_1|} \cos\phi_{k, j_1},$$

where $L = N - \frac{j}{i}P$, $e = e(N, S)$ and

$$\begin{aligned}
\phi_{k, j_1} &= (k + j_1)\lambda - k\,\lambda' + j_1\, p \\
&= (k + j_1)\lambda - k\,\lambda' + j_1 \left(\sigma - \frac{j+i}{i}\lambda + \frac{j}{i}\lambda'\right) \\
&= j_1\,\sigma + \left(k + j_1 - j_1\frac{j+i}{i}\right)\lambda + \left(j_1\frac{j}{i} - k\right)\lambda' \\
&= j_1\,\sigma + \left(k - j_1\frac{j}{i}\right)\lambda + \left(j_1\frac{j}{i} - k\right)\lambda'.
\end{aligned}$$

If $k = k^\star = j_1\frac{j}{i}$ in the summation (if this value is an integer), all the short periodic terms are eliminated. For this particular value of $k = k^\star$ (and after elimination of all the other angular combinations by averaging), we end up with the Hamiltonian:

$$\mathcal{H}^\star(\lambda, N, \sigma, S, \lambda', \Gamma') = -\frac{\mu^2}{2L^2} + n'\Gamma' - n'\frac{j}{i}S - \mu\,\frac{m'}{\mathcal{M}} \sum_{i_1, j_1} P^{k^\star}_{i_1, j_1}\, e^{2i_1+|j_1|} \cos(j_1\,\sigma). \tag{7}$$

The variables and momenta present in this Hamiltonian $\mathcal{H}^\star$ are now *averaged* quantities; for the sake of simplicity, we designate them by the same letters as the corresponding non-averaged ones, but formally we should designate them by $\bar\lambda$, $\bar N$, $\bar\sigma$, $\bar S$, $\bar\lambda'$, and $\bar\Gamma'$.

The artificially introduced third degree of freedom (connected to λ') does not play any role anymore; consequently, the term $n'\Gamma'$ can be dropped. Let us also

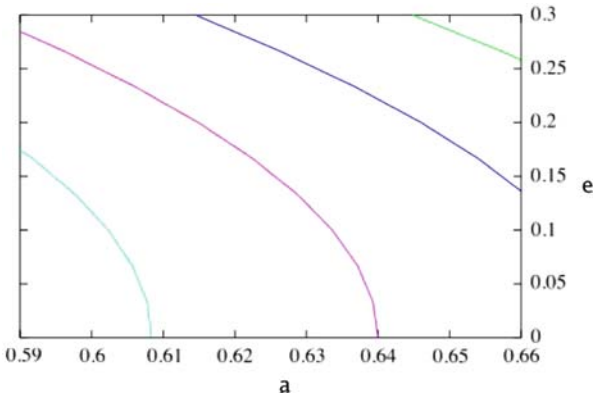

Fig. 4 The curves $N =$ constant in the plane (a, e) for the resonance 2:1

note that the variable λ is not present anymore in the Hamiltonian. Then its conjugated momentum N is a constant (Fig. 4). Finally, the degree of freedom variable—momentum (σ, S) describes the whole (averaged) dynamics.

As $N = L + \frac{j}{i} P$, each constant value of N corresponds to a set of coupled values of a and e; however, very often, it is associated to a specific value of a, the value $a^\star$ corresponding to the circular orbit on the plane identified by $N = N_0$: $N_0 = N(a, e) = N(a^\star, 0)$. With this convention, we can designate by N_{res} the value of N defined by $N_{res} = N(a_{res}, 0)$.

5.5 Position of the Mean Motion Resonances

The important mean motion resonances between Jupiter and an asteroid are known since 1866, on the diagrams of Kirkwood, with a few hundreds of asteroids; they are obvious in Fig. 5.

5.6 The Models of Mean Motion Resonances

If we analyze the value of $k^\star$, we see that the order of the resonance plays a role in its calculation. Indeed, if $i = 1$, $k^\star = \pm j\ j_1$ is always an integer; then, the first resonant term appears for $j_1 = \pm 1$ and the first power of e (obtained for $i_1 = 0$) is 1.

On the opposite, if $i = 2$, $k^\star = \pm \frac{j}{2}\ j_1$; let us now remind that $j + i$ and j should be incommensurable integers, it means that here j is odd. So $k^\star$ is half an integer for $j_1 = \pm 1$, which is impossible for a summation index. In consequence, the first possible value of j_1 is ± 2 and the first corresponding value of the eccentricity power is 2.

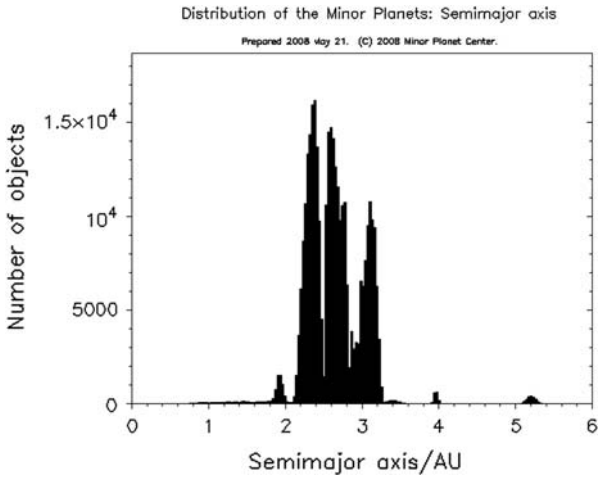

Fig. 5 Histogram of the minor planets of the main belt (May 2008) produced by the Minor Planet Center

If we generalize to any order of resonance, if the order of the resonance is i, the first value of j_1 acceptable is $\pm i$ and the first power of the eccentricity is i. The first resonant term is always proportional to $e^i \cos i\sigma$.

For small values of the eccentricity, this means that the amplitude of the resonant terms decreases rapidly, with the order of the resonance. In order words, only low-order resonances really play a significant role in the dynamics.

A last point before describing the fundamental reference models is to develop the first part of the Hamiltonian with the same level of approximation as the perturbation term; let us analyze the L momentum in the new set of momenta.

$$L = N - \frac{j}{i} S \quad \text{where } N \text{ is a constant and } S = P \text{ is proportional to } e^2.$$

Having used series expansions of e in the perturbation, it seems logical to perform and truncate the term of the two-body problem in the same way:

$$-\frac{\mu^2}{2L^2} - n'\frac{j}{i} S = -\frac{\mu^2}{2}\left(N - \frac{j}{i} S\right)^{-2} + n'\frac{j+i}{i} S \tag{8}$$

$$= -\frac{\mu^2}{2N^2}\left(1 - \frac{j}{i}\frac{S}{N}\right)^{-2} + n'\frac{j+i}{i} S \tag{9}$$

$$= -\frac{\mu^2}{2N^2}\left(1 + 2\,\frac{j}{i}\,\frac{S}{N} + 3\left(\frac{j}{i}\right)^2\left(\frac{S}{N}\right)^2 + \cdots\right) + n'\frac{j+i}{i} S$$

$$= C(N) + \alpha(N)\, S + \beta(N)\, S^2 + \cdots$$

where $C(N)$ is a constant term which can be dropped, $\alpha(N) = -\frac{\mu^2}{N^3}\ \frac{j}{i} + n' \frac{j+i}{i}$ and $\beta(N) = -\frac{3\mu^2}{2N^4}\ (\frac{j}{i})^2$; $\alpha(N)$ measures the distance between the mean motion and the exact resonance; there is a value of the first integral N which exactly coincides with the resonance:

$$\alpha(N) = 0 \Leftrightarrow \frac{\mu^2}{N^3} = n' \frac{j+i}{j} \Leftrightarrow N^3 = \frac{\mu^2}{n'}\ \frac{j}{i+j} = N_{res}^3.$$

The value of N obtained by putting $\alpha = 0$ is N_{res}, corresponding to the value of the exact resonance in semi-major axis given by (6).

In summary, with the help of (7) and (10), for a first-order resonance, we obtain the *second fundamental model of resonance* [9] or *Andoyer model*:

$$\begin{aligned} \mathcal{H}_1(N, \sigma, S) &= \alpha(N)\ S + \beta(N)\ S^2 - \mu\ \frac{m'}{\mathcal{M}}\ P_{0,1}^{k\star}\ e \cos\sigma \\ &= \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N)\ \sqrt{2S}\ \cos\sigma \end{aligned} \tag{10}$$

where

$$\epsilon(N) = -\mu\ \frac{m'}{\mathcal{M}}\ P_{0,1}^{k\star}(N)\ \frac{1}{\sqrt{N}} \quad \text{and} \quad e^2 \simeq \frac{2S}{N}.$$

For a second-order resonance, we obtain

$$\begin{aligned} \mathcal{H}_2(N, \sigma, S) &= \alpha(N)\ S + \beta(N)\ S^2 - \mu\ \frac{m'}{\mathcal{M}}\ P_{0,2}^{k\star}\ e^2 \cos 2\sigma \\ &= \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N)\ 2S\ \cos 2\sigma, \end{aligned}$$

where

$$\epsilon(N) = -\mu\ \frac{m'}{\mathcal{M}}\ P_{0,2}^{k\star}(N)\ \frac{1}{N}.$$

For a third-order resonance:

$$\begin{aligned} \mathcal{H}_3(N, \sigma, S) &= \alpha(N)\ S + \beta(N)\ S^2 - \mu\ \frac{m'}{\mathcal{M}}\ P_{0,3}^{k\star}\ e^3 \cos 3\sigma \\ &= \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N)\ (\sqrt{2S})^3\ \cos 3\sigma. \end{aligned}$$

All these models can still be simplified if the variation of S is supposed to be very small around a specific value S_0; in that case all these models reduce to a simple (translated) pendulum called the *First fundamental model of resonance*:

$$\mathcal{H}_0(N, \Psi, S) = \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N, S_0) \cos\Psi, \tag{11}$$

where $\Psi = i\,\sigma$, i being the order of the resonance.

On the opposite, some of these models suffer from the sharp truncation in S (which means in eccentricity) to represent correctly the physical or topological situations. This is why we keep a term further in the expansion of the perturbation:

$$\mathcal{H}_1^c(N, \sigma, S) = \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N)\ \cos\sigma + \eta(N)\ \cos 2\sigma. \qquad (12)$$

The angle σ in all these models becomes a *slow* angle in comparison with the fast angles λ and λ' (see [1, 18]).

6 The Secondary Resonances

The secondary resonances appear in a primary resonance problem, described by an angle σ, when a second degree of freedom, characterized by an angle ν or one of its multiples, enters in resonance with σ. We present here the secondary resonances inside a mean motion resonance; it is clear that the situation described here can be adapted to any other resonance case.

Let us consider, in our mean motion resonance hypothesis, that the third body evolves on a non-circular Keplerian coplanar orbit, characterized by non-zero values of e' and p':

$$V = -\frac{\mu m'}{\mathcal{M}} \sum_{k,i_1,i_2,j_1,j_2} P^k_{i_1,i_2,j_1,j_2}(a, a')\ e^{2i_1+|j_1|} e'^{\,2i_2+|j_2|}$$
$$\cos\ [(k + j_1)\lambda - (k - j_2)\lambda' + j_1 p + j_2 p'].$$

In that case, we introduce a second resonant angle ν taking into account p':

$$\nu = -\frac{j+i}{i}\lambda' + \frac{j}{i}\lambda - p'.$$

We introduce a canonical transformation:

$$(\lambda, L, p, P, \lambda', \Lambda') \Rightarrow (\nu, \mathcal{N}, \sigma, S, \lambda', \Gamma''),$$

with the three new momenta:

$$\mathcal{N} = \frac{i}{j}L + P, \quad S = P, \quad \text{and} \quad \Gamma'' = \Lambda' + \frac{j+i}{i}P.$$

The linear transformation is easily checked by the conservation of the differential form:

$$d\lambda\ L + dp\ P + d\lambda'\ \Lambda' = d\nu\ \mathcal{N} + d\sigma\ S + d\lambda'\ \Gamma''.$$

The momentum $\mathcal{N}$ is related to N by a simple factor: $\mathcal{N} = \frac{i}{j} N$. A function of N is then a function of $\mathcal{N}$.

The argument of the cosine can be written as

$$\begin{aligned}
&(k + j_1)\lambda - (k - j_2)\lambda' + j_1 p + j_2 p' \\
&= (k + j_1)\lambda - (k - j_2)\lambda' + j_1(\sigma - \frac{j+i}{i}\lambda + \frac{j}{i}\lambda') + j_2(-\nu - \frac{j+i}{i}\lambda + \frac{j}{i}\lambda') \\
&= j_1\sigma - j_2\nu + (k - j_1\frac{j}{i} - j_2\frac{j+i}{i})\lambda + (-k + j_2 + j_1\frac{j}{i} + j_2\frac{j}{i})\lambda'.
\end{aligned}$$

To eliminate the short periodic terms in λ and λ', we choose the value $k^\star$ of k:

$$k^\star = j_1\frac{j}{i} - j_2\frac{j+i}{i}.$$

In that case, for example, an averaged first-order resonant model, with a third body on a elliptic orbit, is characterized by the Hamiltonian:

$$\mathcal{H}(\sigma, S, \nu, \mathcal{N}) = \alpha(\mathcal{N}) + \beta(\mathcal{N})S^2 + \sqrt{2S}\sum_{i_2, j_2} \epsilon_{i_2, j_2}(\mathcal{N})\, e'^{2i_2+|j_2|}\, \cos(\sigma - j_2\nu).$$

The angles σ and ν are both long periodic; a secondary resonance can occur where the angle σ (the resonant angle of the primary resonance) enters into resonance with a multiple of ν: $\dot{\sigma} = j_2\dot{\nu}$, i.e.,

$$\begin{aligned}
\frac{j+i}{i}\dot{\lambda} - \frac{j}{i}\dot{\lambda}' + \dot{p} &= j_2\left(-\frac{j+i}{i}\dot{\lambda} + \frac{j}{i}\dot{\lambda}' - \dot{p}\right) \\
\dot{p} + j_2\dot{p}' &= (1 + j_2)\left[\frac{j}{i}\dot{\lambda}' - \frac{j+i}{i}\dot{\lambda}\right] \\
\frac{1}{1+j_2}[\dot{p} + j_2\dot{p}'] &= \frac{j}{i}\dot{\lambda}' - \frac{j+i}{i}\dot{\lambda}.
\end{aligned}$$

It means that the primary resonance is characterized by a frequency as small as some combination of the pericenter frequencies. We consider that the pericenters are not in resonance.

The secondary resonances can be represented locally by a pendulum-like model, the angle of which is $\Psi = \sigma - j_2\nu$. We can indeed perform a canonical transformation:

$$(\sigma, S, \nu, \mathcal{N}) \Rightarrow (\Psi, S, \nu, \mathcal{N}),$$

and by the sufficient condition of canonicity $d\sigma\, S + d\nu\, \mathcal{N} = d\Psi\, S + d\nu\, \mathcal{N}'$, we can deduce that $\mathcal{N}' = \mathcal{N} + j_2 S$.

After an averaging over ν, considered now as a fast variable in comparison with Ψ, the dynamics is given by $[\Psi, S)$ on a plane $\mathcal{N}' = \mathcal{N}'_0$. Indeed, $\mathcal{N}$ is constant because its associated variable ν is not present anymore in the Hamiltonian after this last averaging process.

Again, we should have replaced all the variables and momenta by their mean values, and describe the dynamics in $(\bar{\Psi}, \bar{S})$ on planes $\bar{\mathcal{N}}$ constant; however, we use the same notations as before, even if they designate other quantities.

The position of the secondary resonances inside the mean motion resonances has been analyzed by several authors, in particular [10, 29, 30].

7 The Secular Resonances

The secular resonances concern the slow angles, like the arguments of the pericenters or the longitudes of the nodes (Fig. 6). We first make the hypothesis that there is no efficient mean motion resonance in the neighborhood of the small mass m.

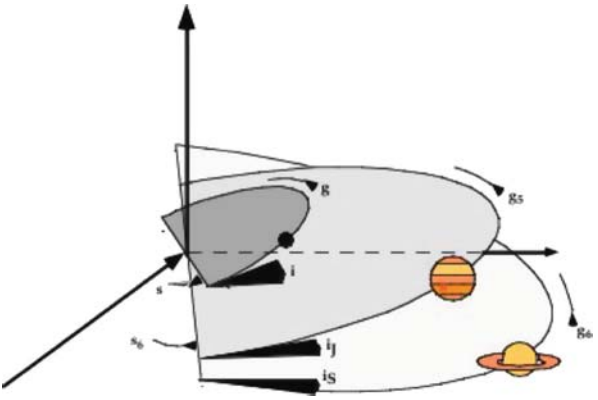

Fig. 6 Schematic graphic of a secular resonance with Jupiter or Saturn

7.1 The Keplerian Case: Kozai Resonances

We can rewrite the three-dimensional non-resonant Hamiltonian (3) in the Keplerian case, where L', P', Q', p', and q' are constants. Let us assume that the perturbing body is on a circular and planar orbit.

The only remaining function of time is λ':

$$\mathcal{H} = -\frac{\mu^2}{2L^2} + V(\lambda, p, q, L, P, Q, \lambda').$$

We average over the two short periodic angles λ and λ'. The resulting Hamiltonian is a two degree of freedom function given by $\bar{V}(p, P, q, Q)$. Developing this function in Fourier's series, we can write [20]

$$\bar{V}(p, P, q, Q) = \sum_i F_{2i}(P, Q) \cos 2i(p - q),$$

where $q - p = -\Omega + \omega + \Omega = \omega$ the argument of the pericenter.

We define a new set of canonical variables: $\Psi = p - q$ (conjugated to P) and q (conjugated to a new momentum M) so that

$$Pdp + Qdq = Pd\Psi + Mdq \quad \Leftrightarrow \quad Pdp + Qdq = Pdp - Pdq + Mdq,$$

which gives $M = P + Q = L - G + G - H = L - H$.

In this set of variables, the Hamiltonian reduces to

$$\bar{V}(p, P, q, Q) = \bar{V}(\Psi, P, q, M) = \sum F_{2i}(P, M) \cos 2i\Psi.$$

Let us note that $P = L - G \simeq L\frac{e^2}{2}$.

The dynamics is given by the differential equations:

$$\dot{\Psi} = \frac{\partial \bar{V}}{\partial P}, \qquad \dot{P} = -\frac{\partial \bar{V}}{\partial \Psi}, \quad \text{and} \quad M = M_0.$$

This is a one degree of freedom Hamiltonian system, in the phase space (Ψ, P), parametrized by the values of M.

M is a function of e and i given by $M = L(1 - \sqrt{1 - e^2} \cos i)$. It can be associated with a maximum value of the inclination, i_{max}, corresponding to $e = 0$:

$$M = M_0 = L(1 - \sqrt{1 - e^2} \cos i) = L(1 - \cos i_{max}).$$

Indeed, on the same plane $M = M_0$, any positive value of e will give a smaller value of $\sqrt{1 - e^2}$ to be compensated by a larger value of $\cos i$, which means a smaller value of i. So i_{max} is obtained for $e = 0$. The case $i = 0$ gives the maximal value of e called e_{max} on the plane $M = M_0$:

$$M = M_0 = L(1 - \sqrt{1 - e^2} \cos i) = L\left(1 - \sqrt{1 - e_{max}^2}\right).$$

The numerical integration of the differential equations shows that for values of i_{max} quite small, the phase space (see Figs. 7a and b) has a simple target-like look, with a stable equilibrium at $e = 0$ and circulation of the argument of pericenter; for higher values of i_{max} (see Fig. 7c), the circular orbit is unstable, and two stable equilibria appear for $\omega = \frac{\pi}{2}$ and $\omega = \frac{3\pi}{2}$. A separatrix (called *Kozai separatrix*) separates the circulation zone of the pericenter from the two librating (North and South) regions.

We can also say that an exact Kozai resonance is characterized by $\dot{\omega} = 0 = \dot{q} - \dot{p}$ or $\dot{p} = \dot{q}$.

In the two libration regions, the pericenter is blocked in a *Kozai resonance*. Let us remind that this behavior only concerns orbits with high values of i_{max} and e_{max}.

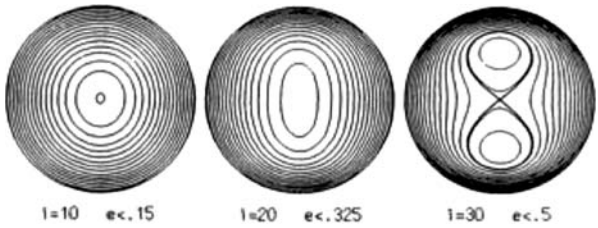

Fig. 7 Phase spaces in Cartesian coordinates related to the eccentricity and longitude of the pericenter for three different values of M_0 corresponding to $i_{max} = 10^\circ$ and $e_{max} = 0.15$ for the *left figure*, $i_{max} = 20^\circ$ and $e_{max} = 0.325$ for the *central one*, and $i_{max} = 30^\circ$ and $e_{max} = 0.5$ for the *right one*. The *border circle* corresponds to $i = 0^\circ$ and the *center* to $e = 0$ in the three cases (taken from [35]

7.2 The Non-Keplerian Case

Let us assume now that $p' = \omega' - \Omega'$ and $q' = -\Omega'$ are time dependent and given by linear functions of the time:

$$p' = g't + p'_0,$$
$$q' = s't + q'_0.$$

To avoid to work with a time-dependent Hamiltonian, we introduce p' and q' as two supplementary degrees of freedom to which we associate artificial momenta called P' and Q'. After averaging the potential (2) over the two fast angles λ and λ', the autonomous Hamiltonian is given by

$$\mathcal{K}(p, P, q, Q, p', P', q', Q') = \bar{V}(p, q, P, Q, p', q') + g'P' + s'Q',$$

with

$$\bar{V} = -\mathcal{G}m' \sum_{(\ell)=(\ell_3,\ell_4,\ell_5,\ell_6)} \mathcal{S}_{(\ell)}(a, a', e, e', i, i') \cos(\ell_3 p' + \ell_4 p + \ell_5 q' + \ell_6 q).$$

A linear secular resonance is present when the frequency associated to the angle $\Psi_{(l)} = \ell_3 p' + \ell_4 p + \ell_5 q' + \ell_6 q$ becomes very close to zero. It means that the motion of the pericenter or of the node (or a linear combination of the two) of the massless body follows the linear motion of the pericenter or the node of one of the perturbers (or a linear combination of the two) and remains blocked in this configuration for long periods of time. For an asteroid, we can identify secular resonances like $g = g_5$, $g = g_6$, $h = s_6$, $g + s = g_5 + s_6$, $g + s = g_6 + s_6$, $g - s = g_5 - s_6$, $g - s = g_6 - s_6$, and $2g = g_5 + g_6$ for the most important ones, where g_6 and s_6 are g' and s' when the perturber is Saturn, and g_5 is g' when it is Jupiter (Fig. 8).

g and s represent the frequencies of p and q, which are not constant: they depend on the values of the eccentricity and of the inclination. Strictly speaking, g and s

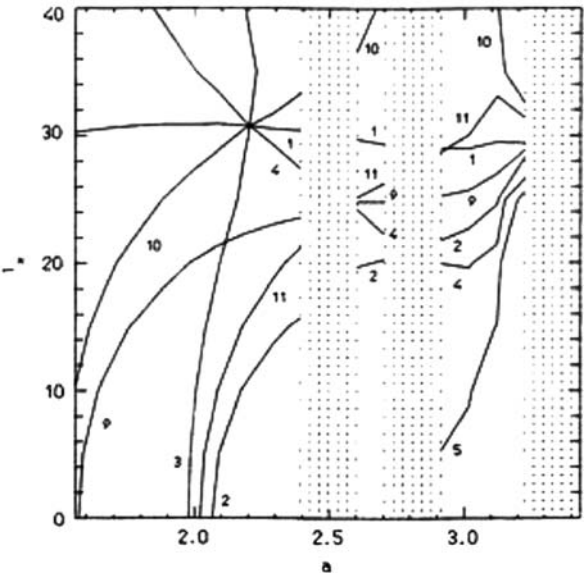

Fig. 8 Position of the main secular resonances with Jupiter or Saturn in the plane (a, i) (taken from [35]) for $e = 0.1$. The numbers correspond to the following secular resonances: $1 \equiv g = g_5$, $2 \equiv g = g_6$, $3 \equiv h = s_6$, $4 \equiv g + s = g_5 + s_6$, $5 \equiv g + s = g_6 + s_6$, $10 \equiv g - s = g_5 - s_6$, $11 \equiv g - s = g_6 - s_6$, and $9 \equiv 2g = g_5 + g_6$

are the frequencies of the angles, after local transformation of the variables into action-angle variables (in which the angle is always a linear function of time). g and s can be considered as *mean* values of the frequencies of p and q; they are also called the *proper frequencies* of p and q.

A secular resonance means that a critical combination of two (or four) slow angles becomes a very slow angle, ten or hundred times slower than the two secular initial ones.

Once the resonant combination is isolated (for a specific set of values for ℓ_i, $i = 3, \ldots, 6$, denoted by ℓ_i^*), after having defined a resonant angle σ by

$$\sigma = \ell_3^* \omega' + \ell_4^* \omega + \ell_5^* \Omega' + \ell_6^* \Omega,$$

we can average over all the other combinations of the secular angles and obtain again a pendulum-like Hamiltonian as first approximation or one of the classical *models of resonance* following the terms kept in the expansion.

For example, Morbidelli and Henrard [35], for the resonance $g = g_6$, obtained a phase space (limited by a circle) very similar (topologically) to the second fundamental model of resonance. The resonant angle σ represents, after passage to angle variables and averaging, the difference $\varpi - \varpi_6$ (Fig. 9).

The calculation of the positions of the secular resonances inside the main mean motion resonances (in the inner and outer minor planet belt) has contributed to

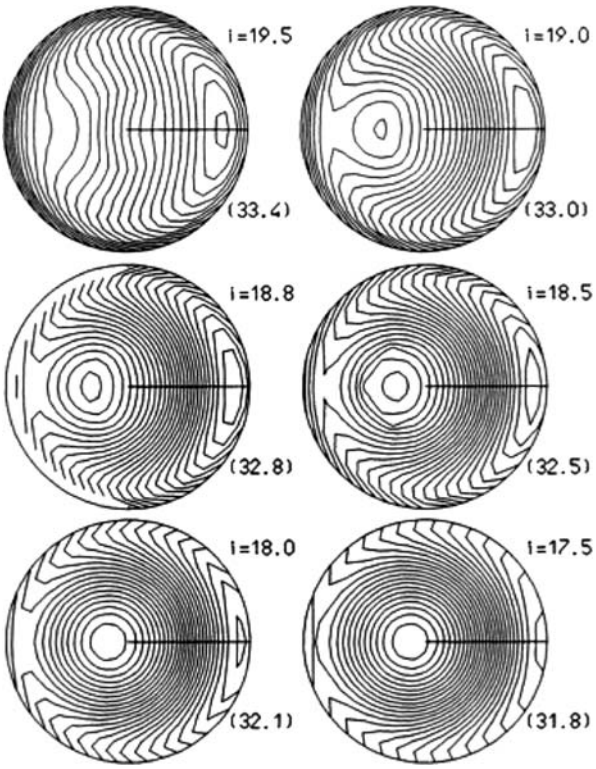

Fig. 9 Phase space for the secular resonance g_6 for a specific value of the semi-major axis $a = 2.6$ AU and for different values of i_{max} [36]

understand much better the dynamics of the small bodies of the Solar System; it has been determined by the tools described in this chapter but with many more variables and degrees of freedom, by semi-numerical techniques to avoid series expansions in eccentricities and inclinations (see [32–34] or [31]).

8 The Pendulum

As we have seen, many problems of resonances (mean motion, secondary, or secular) have a pendulum-like dynamics as first approximation. Let us formulate the pendulum differential equations and comment on the associated motions.

8.1 Formulation and Scaling

The first model of resonance is the pendulum, as already mentioned in (11); in celestial mechanics cases (development of the third body perturbation, spin–orbit

resonance, geostationary resonances), it is generally given first in the following Hamiltonian form:

$$\mathcal{H}(\sigma, S) = \alpha S^2 + \beta S + \epsilon \cos(\sigma - \sigma_0).$$

We introduce a change of phase to make σ_0 disappear ($r = \sigma - \sigma_0$) and a translation on S to get rid of the linear term: $R = S - S_0$. $\mathcal{H}$ becomes

$$\begin{aligned}\mathcal{H}(r, R) &= \alpha\ (R + S_0)^2 + \beta\ (R + S_0) + \epsilon\ \cos r \\ &= \alpha R^2 + 2\alpha R\ S_0 + \alpha S_0^2 + \beta R + \beta S_0 + \epsilon \cos r \\ &= \alpha R^2 + \epsilon \cos r + C_0,\end{aligned}$$

if we choose $S_0 = -\frac{\beta}{2\alpha}$. C_0 is a constant term depending on S_0. After addition of a constant and a scaling, the Hamiltonian $\mathcal{H}$ is replaced by K given by

$$K(r, R) = \frac{1}{2\alpha}(\mathcal{H}(r, R) - C_0) = \frac{R^2}{2} - b \cos r,$$

with $b = -\frac{\epsilon}{2\alpha}$ the unique parameter of the model.

8.2 Equilibria and Phase Space

The dynamics of the system is given by the differential equations:

$$\begin{aligned}\dot{r} &= \frac{\partial K}{\partial R} = R \\ \dot{R} &= -\frac{\partial K}{\partial r} = -b\ \sin r.\end{aligned}$$

The equilibria are characterized by $R = 0$ and $\sin r = 0$, which means $r = 2k\pi$ or $r = \pi + 2k\pi$, k is an integer. The stable equilibria (corresponding to a minimum of K) are $r = 2k\pi$ and $R = 0$, the unstable ones $R = 0$ and $r = (2k + 1)\pi$. We choose the interval $[-\pi, \pi[$ to represent the periodic motion.

Starting from $-\pi$ and arriving to π ($R > 0$) or starting from π and going to $-\pi$ ($R > 0$) in an infinite time, the two separatrices (called C_1 and C_2) divide the phase space into three distinct regions called, in reference with the classical pendulum in mechanics, *positive circulation*, *negative circulation*, and *resonance* (Fig. 10).

The equation of the separatrix (corresponding to the level curves $K = b$) is given by

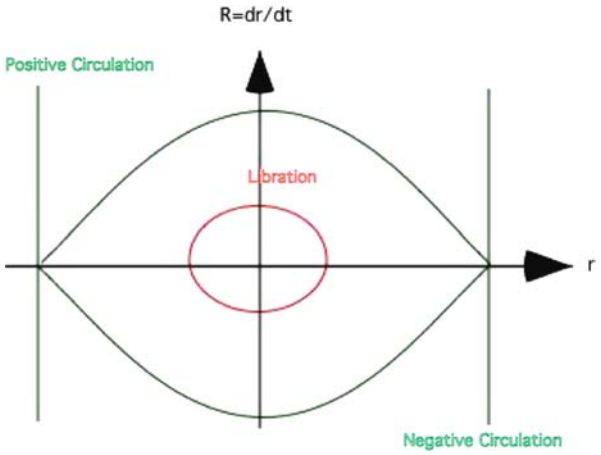

Fig. 10 The three regions of the pendulum

$$b = \frac{R^2}{2} - b\cos r$$
$$\frac{R^2}{2} = b\,(1 + \cos r)$$
$$R^2 = 4b\ \cos^2\frac{r}{2}.$$

The corresponding action (the area enclosed by the two separatrices, C_1 and C_2, divided by 2π) can be calculated as follows:

$$J_{\text{Separatrix}} = J_S = \frac{1}{2\pi}\int_{C_1\cup C_2} R\,dr = \frac{4}{2\pi}\int_0^{\pi} R\,dr = \frac{4\sqrt{b}}{\pi}\int_0^{\pi}\cos\frac{r}{2}\,dr = \frac{8\sqrt{b}}{\pi}.$$

8.3 Action-Angle Variables

Action-angles variables can be introduced in the two types of dynamics: circulation or resonance. Both cases are characterized by values of $K = h$, giving the curve implicit equation

$$h = \frac{R^2}{2} - b\,\cos r \quad \text{or} \quad R^2 = 2(h + b\cos r).$$

The idea, as already described in (1), is to introduce for each level curve, a canonical transformation from (r, R) to (Ψ, J) so that the Hamiltonian only depends on J, which means that Ψ is automatically a linear function of time.

8.3.1 Circulation case : $h > b$

The action J is the area (divided by 2π) enclosed between the level curve h and the r-axis and vertically between the axes $r = -\pi$ and $r = \pi$ (Fig. 11).

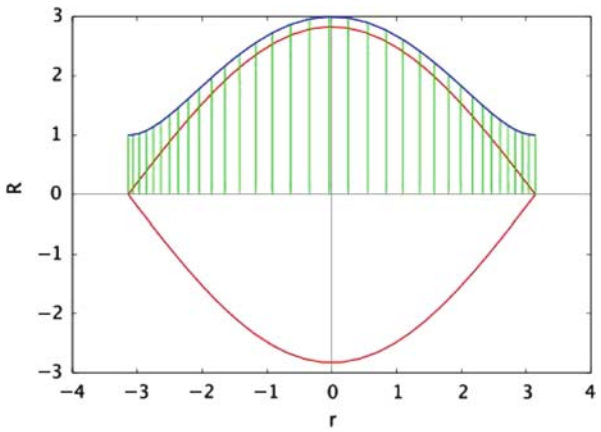

Fig. 11 The positive circulation case

By obvious symmetry, it can be reduced to twice the calculation on the interval $[0, \pi[$,

$$
\begin{aligned}
2\pi\, J &= 2\int_0^{\pi} \sqrt{2(h + b\,\cos r)}\, dr \\
&= 2\int_0^{\pi} \sqrt{2(h + b - 2b\,\sin^2\frac{r}{2})}\quad dr \\
&= 4\sqrt{2}\int_0^{\frac{\pi}{2}} \sqrt{h + b - 2b\sin^2 u}\quad du \qquad \text{where } r = 2u \\
&= 4\sqrt{2}\int_0^{\frac{\pi}{2}} \sqrt{h + b - 2b\sin^2 u}\quad du \\
&= 4\sqrt{2(h+b)}\int_0^{\frac{\pi}{2}} \sqrt{1 - k^2\sin^2 u}\, du \qquad \text{where } k^2 = \frac{2b}{h+b} \\
&= 4\sqrt{2(h+b)}\quad \mathbb{E}(k),
\end{aligned}
$$

where $\mathbb{E}(k) = \int_0^{\frac{\pi}{2}} \sqrt{1 - k^2\sin^2 u)}\, du$ is the classical complete elliptic integral.

We can easily deduce $\frac{\partial J}{\partial h}$:

$$
\begin{aligned}
2\pi \frac{\partial J}{\partial h} &= \frac{4}{\sqrt{2(h+b)}}\, \mathbb{E}(k) + 4\sqrt{2(h+b)}\, \frac{d\mathbb{E}(k)}{dk}\, \frac{dk}{dh} \\
&= \frac{4}{\sqrt{2(h+b)}}\, \mathbb{E}(k) + 4\sqrt{2(h+b)}\, [\frac{\mathbb{E}(k) - \mathbb{K}(k)}{k}]\, \frac{-b}{k(h+b)^2} \\
&= \frac{4}{\sqrt{2(h+b)}}\, \mathbb{K}(k),
\end{aligned}
$$

where $\mathbb{K}(k) = \int_0^{\frac{\pi}{2}} \frac{1}{\sqrt{1-k^2 \sin^2 u)}}\, du$ is the other complete elliptic integral. The following relation is easy to check by derivation with respect to k:

$$
\frac{d\mathbb{E}(k)}{dk} = \frac{d}{dk} \int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 u}\, du = \frac{\mathbb{E}(k) - \mathbb{K}(k)}{k}.
$$

It means that each trajectory is characterized by a different frequency $\omega = \frac{\partial K}{\partial J} = \frac{1}{\frac{\partial J}{\partial h}}$.

The angle Ψ is defined as

$$
\begin{aligned}
\Psi - \Psi_0 &= \frac{\partial K}{\partial J}\, t \\
&= \frac{1}{\frac{\partial J}{\partial h}} \int_0^t dt = \frac{1}{\frac{\partial J}{\partial h}} \int_0^r \frac{dt}{dr}\, dr == \frac{1}{\frac{\partial J}{\partial h}} \int_0^r \frac{1}{\dot{R}}\, dr, \\
&= \frac{\pi}{\mathbb{K}(k)}\, \mathbb{F}(\frac{r}{2}, k)
\end{aligned}
$$

where $\mathbb{F}(x, k) = \int_0^x \frac{1}{\sqrt{1-k^2 \sin^2 u}}\, du$ is the incomplete elliptic integral ($\mathbb{F}(\frac{\pi}{2}, k) = \mathbb{K}(k)$).

8.3.2 Resonance case: $-b < h < b$

The action J is here related to the area (divided by 2π) enclosed by a complete closed curve C and by symmetry, four times the same area between $r = 0$ and $r = r_0$. r_0 is the value of r when $R = 0$, which means $\cos r_0 = -\frac{h}{b}$ (Fig. 12),

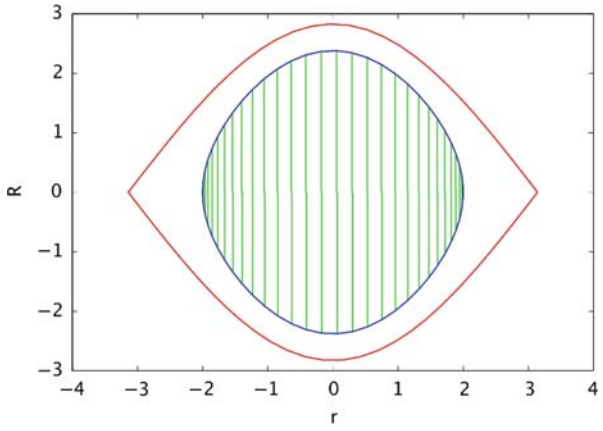

Fig. 12 The resonance case

$$
\begin{aligned}
2\pi\, J &= \oint_C \sqrt{2(h + b\,\cos r)}\quad dr \\
&= 4\sqrt{2}\int_0^{r_0} \sqrt{h + b - 2b\,\sin^2\frac{r}{2}}\quad dr \\
&= 8\sqrt{2}\int_0^{\frac{r_0}{2}} \sqrt{h + b - 2b\,\sin^2 u}\quad du \\
&= 8\sqrt{2}\int_0^{\frac{\pi}{2}} \sqrt{h + b - 2b\,\sin^2\frac{r_0}{2}\sin^2 v}\quad \frac{du}{dv}\,dv
\end{aligned}
\tag{13}
$$

where

$$
\sin u = \sin\frac{r_0}{2}\sin v. \tag{14}
$$

Let us calculate $\frac{du}{dv}$,

$$
\begin{aligned}
\cos u\; du &= \sin\frac{r_0}{2}\;\cos v\; dv, \\
\sqrt{1 - \sin^2\frac{r_0}{2}\;\sin^2 v}\; du &= \sin\frac{r_0}{2}\;\cos v\; dv, \\
\frac{du}{dv} &= \sin\frac{r_0}{2}\;\frac{\cos v}{\sqrt{1 - \sin^2\frac{r_0}{2}\;\sin^2 v}}\; dv,
\end{aligned}
\tag{15}
$$

and the square root in the integral,

$$
h + b - 2b\,\sin^2\frac{r_0}{2}\sin^2 v = h + b - 2b\,(1 - \frac{h^2}{b^2})\;\sin^2 v. \tag{16}
$$

We introduce (15) and (16) in (13) to obtain

$$2\pi\, J = \frac{8}{\pi}\,\sqrt{b}\,[\mathbb{K}(k)(k^2-1)+\mathbb{E}(k)].$$

This gives J as a function of k, and then as a function of h, from which we can extract the inverse of the frequency:

$$\frac{\partial J}{\partial h} = \frac{2\mathbb{K}(k)}{\pi\sqrt{b}}.$$

The angle Ψ is obtained by the same way as in the circulation case:

$$\begin{aligned}\Psi - \Psi_0 &= \frac{\partial K}{\partial J}\, t \\ &= \frac{\pi}{2\mathbb{K}(k)}\mathbb{F}(s,k), \quad \text{where } \sin s = k \sin\frac{r}{2}.\end{aligned}$$

8.4 The Harmonic Oscillator

When h is close to $-b$, which means that the curve is close to the stable equilibrium, the pendulum can be approximated by a simple harmonic oscillator; it also corresponds to the approximation $\cos r = 1 - \frac{r^2}{2}$. The Hamiltonian writes

$$K = \frac{R^2}{2} - b(1-\frac{r^2}{2}) \quad \text{or} \quad K + b = \frac{R^2}{2} + b\,\frac{r^2}{2}. \tag{17}$$

In that case, the action-angle canonical coordinates Ψ and J are introduced in an easy way, without any elliptic integral:

$$\begin{aligned}R &= \nu\,\sqrt{2J}\cos\Psi, \\ r &= \frac{1}{\nu}\,\sqrt{2J}\sin\Psi,\end{aligned}$$

where ν is a constant to be determined by the ellipsoid equation (17). Indeed, let us replace these new variables in the Hamiltonian:

$$\begin{aligned}\mathcal{K} = K + b &= \frac{R^2}{2} + b\,\frac{r^2}{2} \\ &= \nu^2 J\cos^2\Psi + b\,\frac{1}{\nu^2}\,J\sin^2\Psi \\ &= \nu^2\,J \quad \text{if} \quad \nu^2 = b\,\frac{1}{\nu^2} \quad \text{or} \quad \nu^4 = b \\ &= \sqrt{b}\,J.\end{aligned}$$

This gives automatically the *fundamental frequency of the resonance*:

$$\dot{\Psi} = \frac{\partial \mathcal{K}}{\partial J} = \omega = \sqrt{b}.$$

8.5 Generalization

This procedure can be generalized and applied about any stable equilibrium (center); let us start with a Hamiltonian $\mathcal{H} = \mathcal{H}(q, p)$ and consider a stable equilibrium (q_0, p_0) solution of $\frac{\partial \mathcal{H}}{\partial q} = 0 = \frac{\partial \mathcal{H}}{\partial p}$.

In the neighborhood of (q_0, p_0), we develop the Hamiltonian:

$$\begin{aligned}\mathcal{H} &= \mathcal{H}(q, p)\\ &= \mathcal{H}(q_0, p_0) + \frac{\partial \mathcal{H}}{\partial q}_{q_0, p_0}(q - q_0) + \frac{\partial \mathcal{H}}{\partial p}_{q_0, p_0}(p - p_0)\\ &\quad + \frac{1}{2}\left(\frac{\partial^2 \mathcal{H}}{\partial q^2}_{q_0, p_0}(q - q_0)^2 + 2\frac{\partial^2 \mathcal{H}}{\partial q \partial p}_{q_0, p_0}(q - q_0)(p - p_0)\right.\\ &\quad \left. + \frac{\partial^2 \mathcal{H}}{\partial p^2}_{q_0, p_0}(p - p_0)^2\right) + \cdots,\end{aligned}$$

which gives

$$\mathcal{H} - \mathcal{H}(q_0, p_0) = \frac{1}{2}\,(a\,(\Delta q)^2 + 2b\;\Delta q\,\Delta p + c\,(\Delta p)^2 + \cdots),$$

with a, b, and c representing the second partial derivatives: $\Delta q = q - q_0$ and $\Delta p = p - p_0$.

The problem is now a simple reduction of conic; we introduce first a rotation (of angle θ) to get rid of the $\Delta q\,\Delta p$ term, followed by a scaling transformation similar as that of the harmonic oscillator, to obtain

$$\begin{aligned}p' &= \Delta q\;\sin\theta + \Delta p\;\cos\theta = \nu\,\sqrt{2J}\;\cos\Psi,\\ q' &= \Delta q\;\cos\theta - \Delta p\;\sin\theta = \frac{1}{\nu}\,\sqrt{2J}\;\sin\Psi.\end{aligned}$$

The angle θ is defined by the equation

$$(a - c)\,\sin 2\theta + 2b\,\cos 2\theta = 0.$$

The Hamiltonian is then

$$\mathcal{H} - \mathcal{H}(q_0, p_0) = \frac{1}{2}\,(A\,q'^2 + C\;p'^2) = A\;\frac{1}{\nu^2}\;J\;\sin^2\Psi + C\;\nu^2\;J\;\cos^2\Psi = C\;\nu^2\;J,$$

where $\frac{A}{\nu^2} = C\ \nu^2$ or $\frac{A}{C} = \nu^4$. A and C are related to a, b, c, and θ by the relations:

$$A = a\ \cos^2\theta - 2b\ \sin\theta\ \cos\theta + c\ \sin^2\theta,$$
$$C = a\ \sin^2\theta + 2b\ \sin\theta\ \cos\theta + c\ \cos^2\theta.$$

The fundamental frequency associated with this stable equilibrium is then

$$C\ \nu^2 = \sqrt{AC} = \dot{\Psi} \quad \text{and} \quad \mathcal{H} - \mathcal{H}(q_0, p_0) = \sqrt{AC}\ J.$$

The passage to action-angle variables is the only correct way of calculating such a frequency, in the case of libration; it replaces a librating angle by a circulating angle Ψ, for which it is meaningful to calculate a frequency (or a period) around the center.

9 The Second Fundamental Model of Resonance

The following toy model of resonance (called the second fundamental model of resonance or also Andoyer's model) allows us to introduce a non-constant amplitude in front of the cosine term (10); this amplitude is here dependent on the momentum S. The dynamics of such a model is less symmetric than the pendulum one. Let us write the simplest formulation of this Hamiltonian:

$$\mathcal{H} = \alpha S + \beta S^2 + \epsilon\sqrt{2S}\ \cos\sigma.$$

As already mentioned, the parameters α, β, and ϵ are functions of N, constant in this context. To perform the analysis of the model, it seems adequate to reduce those three parameters to a unique one. The procedure consists in changing the time and the momentum scales and modifying the sign and the phase of the angle σ.

9.1 Reduction to One Parameter

Let us describe this procedure for this model; we introduce a *new time* τ, a *new momentum* R, and a *new angle* r related to our initial set through the relations

$$\tau = a\ t,$$
$$R = b\ S,$$
$$r = c\sigma + d.$$

The constants $a \geq 0$, $b \geq 0$, $c = \pm 1$, and d have to be calculated to keep a Hamiltonian formulation in the new set of variable—momentum (r, R) with reference to a new time τ (see [9]).

It also induces a rescaling of the Hamiltonian; in other words, the related transformation is not *completely canonical*, but only *canonical of parameter* μ). The new Hamiltonian is called K and is linked to the initial one by the relation:

$$K(r, R) = \mu\ \mathcal{H}(\sigma(r), S(R)) = \mu\ \mathcal{H}\left(\frac{r-d}{c}, \frac{R}{b}\right),$$

where $\mu = \mu(a, b, c, d)$ and the differential equations associated are as follows:

$$\begin{aligned}\frac{dr}{d\tau} &= \frac{\partial K}{\partial R},\\ \frac{dR}{d\tau} &= -\frac{\partial K}{\partial r}.\end{aligned}$$

If we connect this system to the initial one, we obtain, through a succession of partial derivatives

$$\begin{aligned}\frac{dr}{d\sigma}\ \frac{d\sigma}{dt}\frac{dt}{d\tau} &= \frac{dK}{d\mathcal{H}}\ \frac{\partial\mathcal{H}}{\partial S}\frac{dS}{dR},\\ \text{which means}\quad c\ \frac{d\sigma}{dt}\ \frac{1}{a} &= \mu\ \frac{\partial\mathcal{H}}{\partial S}\frac{1}{b},\\ \frac{dR}{dS}\ \frac{dS}{dt}\frac{dt}{d\tau} &= -\frac{dK}{d\mathcal{H}}\ \frac{\partial\mathcal{H}}{\partial\sigma}\frac{d\sigma}{dr},\\ \text{which means}\quad b\ \frac{dS}{dt}\frac{1}{a} &= -\mu\ \frac{\partial\mathcal{H}}{\partial\sigma}\frac{1}{c}.\end{aligned}$$

We obtain $\mu = \frac{bc}{a}$.

We choose a very simple form for K, with a unique constant Δ:

$$K(r, R) = \Delta R + R^2 - 2\sqrt{2R}\,\cos r. \tag{18}$$

In this formulation we keep the unique parameter in the linear term in R; this is purely arbitrary and other transformations would have introduced a unique parameter in front of the quadratic term or in front of the trigonometric contribution. We are now going to prove that there exists a canonical transformation arriving to such an Hamiltonian.

We can write

$$\begin{aligned}K(r, R) &= \mu\ \mathcal{H}(\sigma(r), S(R))\\ &= \mu\ \mathcal{H}\left(\frac{r-d}{c}, \frac{R}{b}\right)\\ &= \mu\ \alpha\frac{R}{b} + \mu\ \beta\left(\frac{R}{b}\right)^2 + \mu\ \epsilon\sqrt{\frac{2R}{b}}\cos\left(\frac{r-d}{c}\right)\end{aligned}$$

$$= \frac{\mu\,\alpha}{b} R + \frac{\mu\,\beta}{b^2} R^2 + \frac{\mu\,\epsilon}{\sqrt{b}} \sqrt{2R} \cos(r-d) \text{ because } c = \pm 1$$
$$= \Delta R + R^2 - 2\sqrt{2R}\, \cos r$$

and consequently,

$$\Delta = \frac{\mu\,\alpha}{b} = \frac{bc}{a}\frac{\alpha}{b} = c\,\frac{\alpha}{a},$$
$$1 = \frac{\mu\,\beta}{b^2} = \frac{bc}{a\,b^2}\frac{\beta}{} = c\,\frac{\beta}{a\,b},$$
$$-2\cos r = \frac{\mu\,\epsilon}{\sqrt{b}} \cos(r-d) = \frac{bc\,\epsilon}{a\,\sqrt{b}} \cos(r-d),$$
$$= c\,\frac{\sqrt{b}\,\epsilon}{a} \cos(r-d).$$

As $a \geq 0$ and $b \geq 0$, the second equation allows to choose $c = \pm 1$; if $\beta > 0$, $c = +1$, and if $\beta < 0$, $c = -1$, i.e.,

$$c = \text{ sign } \beta.$$

The third equation determines d according to the sign of $(\beta\epsilon)$:

$$-2\cos r = \text{ sign } \beta\ \frac{\sqrt{b}\,\epsilon}{a} \cos(r-d) = \text{ sign } (\beta\epsilon)\ \frac{\sqrt{b}\,|\epsilon|}{a} \cos(r-d).$$

If $\beta\epsilon < 0$, $d = 0$ and if $\beta\epsilon > 0$, $d = \pi$.
To find $a \geq 0$ and $b \geq 0$, we use $ab = |\beta|$ and $\frac{\sqrt{b}\,|\epsilon|}{a} = 2$ to obtain

$$b = \left(2\,|\frac{\beta}{\epsilon}|\right)^{\frac{2}{3}} \quad \text{and} \quad a = \left(\frac{|\beta|\epsilon^2}{4}\right)^{\frac{1}{3}},$$

which gives

$$\mu = \frac{bc}{a} = \text{ sign } \beta\ \left(\frac{16\beta^2}{\epsilon^4}\right)^{\frac{1}{3}}$$

and

$$\Delta = \text{ sign } \beta\ \alpha\ \left(\frac{4}{|\beta|\epsilon^2}\right)^{\frac{1}{3}} = \alpha\ \left(\frac{4}{\beta\epsilon^2}\right)^{\frac{1}{3}}.$$

9.2 Equilibria

We calculate the equilibrium of the Hamiltonian $K(r, R)$ by equating the partial derivatives of K to zero. To avoid the well-known singularities in polar coordinates, we introduce a canonical transformation to introduce Cartesian coordinates:

$$x = \sqrt{2R}\,\cos r \quad \text{the momentum,}$$
$$y = \sqrt{2R}\,\sin r \quad \text{the variable,}$$

and the Hamiltonian becomes

$$K(x, y) = \left(\frac{x^2 + y^2}{2}\right)^2 + \Delta\,\left(\frac{x^2 + y^2}{2}\right) - 2x. \tag{19}$$

The associated dynamical system is equating to zero to find the equilibria:

$$\frac{dy}{d\tau} = \frac{\partial K}{\partial x} = \left(x^2 + y^2\right)\,x + \Delta\,x - 2 = 0,$$
$$\frac{dx}{d\tau} = -\frac{\partial K}{\partial y} = -\left(x^2 + y^2\right)\,y - \Delta\,y = 0.$$

From the second equation, we deduce

$$y = 0 \quad \text{or} \quad x^2 + y^2 = -\Delta.$$

Replacing the second solution in the first equation leaves to $-2 = 0$; so all the equilibria are characterized by the condition $y = 0$ and x satisfies a cubic equation:

$$x^3 + \Delta x - 2 = 0,$$

which gives 1 or 3 real roots, following the values of Δ.

To distinguish clearly the cases with 1 or 3 roots, we introduce a new parameter δ to replace Δ:

$$\Delta = -3(\delta + 1).$$

The case $\delta < 0$ will correspond to the phase spaces with only one equilibrium, the case $\delta > 0$ to the phase spaces with three equilibria.

Let us remind here the expression of the first root of a cubic equation given in the following general formulation: $a_3x^3 + a_2x^2 + a_1x + a_0 = 0$,

$$x = \left(R + \sqrt{r^2 + Q^3}\right)^{\frac{1}{3}} + \left(R - \sqrt{r^2 + Q^3}\right)^{\frac{1}{3}} \quad \text{if } Q^3 + R^2 \geq 0,$$
$$x = 2\sqrt{-Q}\cos\left(\tfrac{1}{3}\arccos\frac{R}{\sqrt{-Q^3}}\right) - \frac{a_2}{3a_3} \qquad \text{if } Q^3 + R^2 < 0,$$

with

$$Q = -\frac{a_2^2}{9a_3^2} + \frac{a_1}{3a_3} \quad \text{and} \quad R = -\frac{a_0}{2a_3} + \frac{a_1a_2}{6a_3^2} - \frac{a_2^3}{27a_3^2}.$$

Applying these expressions to our case, $a_3 = 1$, $a_2 = 0$, $a_1 = \Delta = -3(\delta + 1)$, and $a_0 = -2$, we obtain the analytical formulation of the equilibria.

When $\delta < 0$ the unique real equilibrium x_1 is given by

$$x_1 = (1+\gamma)^{\frac{1}{3}} + (1-\gamma)^{\frac{1}{3}} \tag{20}$$

with $\gamma^2 = 1 - (\delta+1)^2$, and when $\delta > 0$ the three real equilibria x_1, x_2, and x_3 are explicitly calculated as follows:

$$\begin{aligned} x_1 &= 2\,s\ \cos\Delta, \\ x_2 &= -s\cos\Delta - \sqrt{3}\,s\ \sin\Delta, \\ x_3 &= -s\cos\Delta + \sqrt{3}\,s\ \sin\Delta, \end{aligned}$$

with $s = \sqrt{\delta+1}$ and $\cos 3\Delta = \frac{1}{s^3}$.

The stability of these equilibria is easily obtained by the calculation of the second partial derivatives of K for each of them: x_1 is stable for any δ, x_2 is stable, and x_3 unstable for $\delta > 0$.

9.3 The Phase Space

The Hamiltonian (19) is a function of x and y and of the parameter δ; for each value of δ, we can draw the curves K = constant, in the Cartesian phase space (x, y). Different cases are represented in Fig. 13: $\delta = -3$, $\delta = -1$, $\delta = 2$, and $\delta = 5$. The first two cases correspond to negative values of δ, with only one stable real equilibrium. The level curves are almost ellipses for $\delta = -3$, far from the resonance, giving a target-like global picture; for $\delta = -1$ this is not the case anymore and the curves are different, their behavior already showing the proximity of the resonance. For the last two cases, corresponding to positive values of δ, we see clearly the three equilibria and the two separatrices (C_1 and C_2) dividing the phase space in three regions: an internal, a resonant, and an external region. For large values of δ, the separatrices are far away from the origin ($x = 0$, $y = 0$) and the phase space, near the origin, looks again like a target.

The case $\delta = 0$ is characterized by the apparition of two separatrices C_1 and C_2 emerging from the unstable equilibrium x_3; for any positive value of δ, these separatrices are present. Their intersection points with the axis $y = 0$, x_4 and x_5 can be calculated ([2]) as functions of δ:

$$\begin{aligned} x_4 &= \pm\sqrt{2R_{min}} \quad \text{and} \quad R_{min} = \frac{s^2}{2}(6-t^2) - 2\sqrt{st}, \\ x_5 &= \sqrt{2R_{max}} \quad \text{and} \quad R_{max} = \frac{s^2}{2}(6-t^2) + 2\sqrt{st}, \end{aligned}$$

with $t = \cos\Delta + \sqrt{3}\ \sin\Delta$.

We can also explicitly calculate the area enclosed by these two curves C_1 and C_2, which are functions of the parameter δ: the function $A_1(\delta)$ corresponds to the area

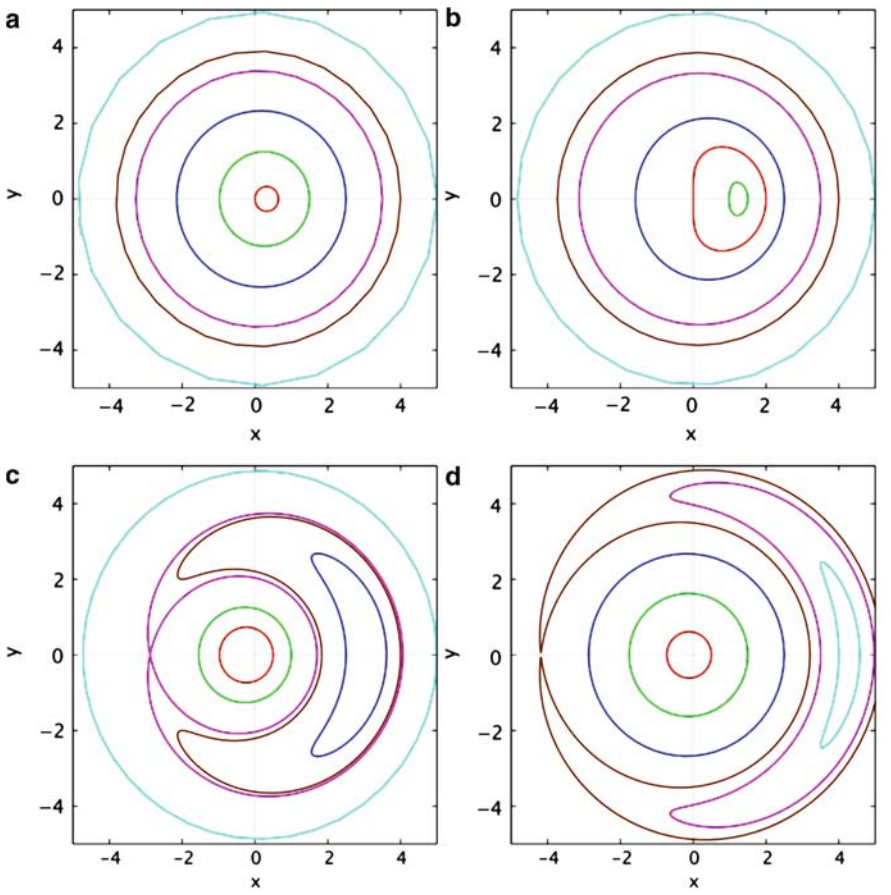

Fig. 13 The phase spaces for $\delta = -3$, $\delta = -1$, $\delta = 2$, and $\delta = 5$

(divided by 2π as in the action-angle variables) of the curve $\mathcal{C}_1$ starting from x_3, crossing the axis $y = 0$ at x_4 and encircling the stable equilibrium x_2; the function $A_2(\delta)$ corresponds to the area (divided by 2π) of the curve $\mathcal{C}_2$ starting from x_3, crossing the axis $y = 0$ at x_5 and encircling the two stable equilibria x_1 and x_2 (Figs. 14 and 15).

Their analytical expressions are quite simple ([2, 21]):

$$A_1(\delta) = 6\,s^2\,\left(\frac{\pi}{2} - (\arcsin st)^{-\frac{3}{2}}\right) - \frac{6}{st}\,\sqrt{(st)^3 - 1},$$
$$A_2(\delta) = 6\,s^2\,\left(\frac{\pi}{2} - (\arcsin st)^{-\frac{3}{2}}\right) + \frac{6}{st}\,\sqrt{(st)^3 - 1}.$$

We can note that if $\delta = 0$, $s = 1$, then $A_1(0) = 0$ and $A_2(0) = 6\pi$.

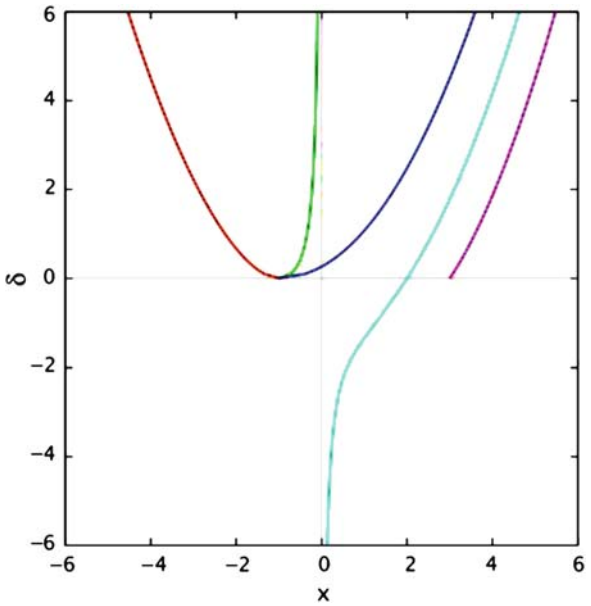

Fig. 14 The three equilibria (x_1, x_2, and x_3) and the two limits of the separatrix on the x-axis (x_4 and x_5) on the *horizontal axis*, for each value of δ on the *vertical axis*

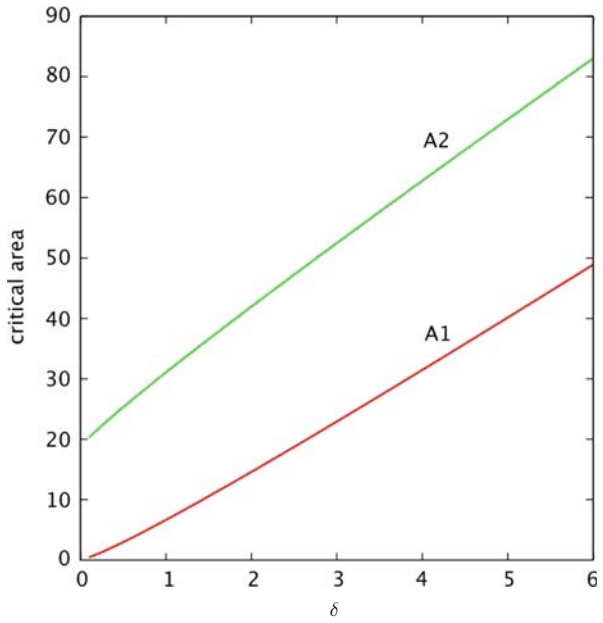

Fig. 15 The two critical areas included in both parts of the separatrix, $A_1(\delta)$ and $A_2(\delta)$

The derivatives of the two functions A_1 and A_2 with respect to δ which we shall need later, $\frac{dA_1}{d\delta}$ and $\frac{dA_2}{d\delta}$, can be calculated and are also functions of δ. We can also notice that in the most interesting region for resonant motions (between, for example, $\delta = 2$ and $\delta = 4$, i.e., for positive but not too large values of δ) these derivatives are quasi-constants but not equal:

$$\frac{dA_1}{d\delta} \simeq 8.4 \quad \text{and} \quad \frac{dA_2}{d\delta} \simeq 10.4.$$

9.4 The Three Zones

These two separatrices C_1 and C_2 divide the phase space into three zones or regions: the *internal zone* inside the curve C_1, the *external zone* outside the curve C_2, and the *resonant zone* between the two curves. Many authors refer to this last zone as the *libration zone* which emphasizes the fact that the angle σ is (in most cases) not circulating, but oscillating between two extrema. However, let us remark that some orbits in the internal zone could also librate, and that this property of libration can be easily destroyed by a simple translation of the origin. On the opposite, the resonance zone definition is a topological characteristic, invariant by change of reference frames.

9.5 The Resonant Frequency

For the stable equilibrium x_1 in the resonant region (for $\delta \geq 0$), we can calculate the fundamental resonant frequency by the passage to local action-angle variables, as described in (1), using the Cartesian canonical variables x and y, around $x = x_1$ and $y = 0$ (Fig. 16).

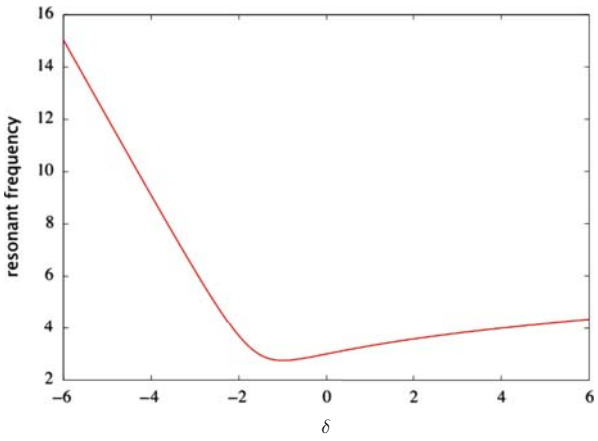

Fig. 16 The fundamental frequency associated to the resonant equilibrium x_1

It leads to a frequency ω_1 given by

$$\frac{d\Psi}{d\tau} = \omega = \sqrt{ac},$$

where

$$a = -3(\delta + 1) + x_1^2,$$
$$c = -3(\delta + 1) + 3\,x_1^2.$$

Let us give an example on the plane $\delta = 2$, we select a curve very close to the equilibrium x_1; we calculate its period (here, 1.75 units of the time called τ); if we check the fundamental frequency for $\delta = 2$, we obtain $\omega_1 = 3.58$ (per unit of time). It is easy to check (Fig. 17) that

$$\text{Period} = \frac{2\pi}{\omega_1} = \frac{2\pi}{3.58} = 1.75 \text{ units of time.}$$

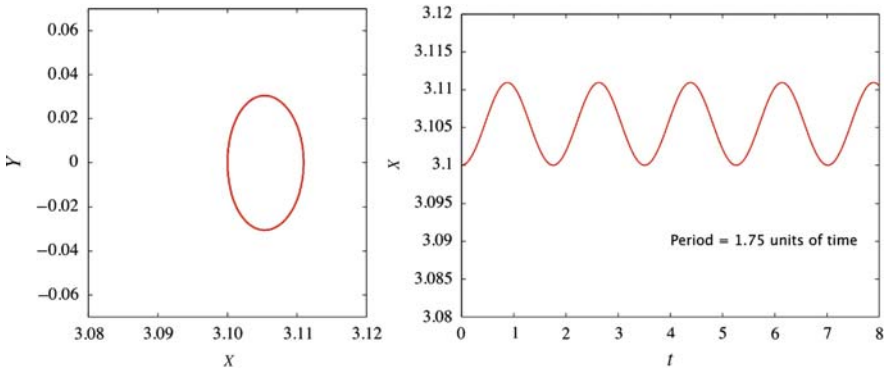

Fig. 17 A curve close to the exact resonance on the plane $\delta = 2$: (x, y) on the *left*, (t, x) on the *right*

10 The Probability of Capture

Let us start with a one degree of freedom Hamiltonian, expressed in coordinates σ (the variable) and S (the momentum) and depending on a parameter δ, slowly varying with time:

$$H(\sigma, S, \delta) \quad \text{with} \quad \dot{\sigma} = \frac{\partial H}{\partial S} \quad \text{and} \quad \dot{S} = -\frac{\partial H}{\partial \sigma}.$$

10.1 Conservative Dynamics

We first consider that, for δ constant, the problem is integrable ($H = h$) and that suitable action-angle variables (Ψ, J) can be introduced through a generating function $\mathcal{F}(\sigma, J, \delta)$, such that the new Hamiltonian K is only dependent on the action J and not on the angle Ψ:

$$h = H(\sigma, S, \delta) = K(-, J, \delta) \quad \text{with} \quad S = \frac{\partial \mathcal{F}}{\partial \sigma} \quad \text{and} \quad \Psi = \frac{\partial \mathcal{F}}{\partial J}.$$

The Hamilton–Jacobi equation becomes

$$h = H(\sigma, \frac{\partial \mathcal{F}}{\partial \sigma}, \delta) = K(-, J, \delta), \tag{21}$$

with its implicit solution

$$S = \frac{\partial \mathcal{F}}{\partial \sigma} = S(\sigma, J, \delta),$$

or starting with an initial value σ_0:

$$\mathcal{F}(\sigma, J, \delta) = \int_{\sigma_0}^{\sigma} S(\sigma', J, \delta)\, d\sigma'.$$

Ψ is chosen such that it makes a revolution of 2π along any closed trajectory characterized by $H = h$. We introduce the generating function $\mathcal{G}$ corresponding to $\mathcal{F}$ on a closed trajectory:

$$\mathcal{G}(J, \delta) = \oint S(\sigma', J, \delta)\, d\sigma'.$$

Because $\Psi = \frac{\partial \mathcal{F}}{\partial J}$, on a closed trajectory it becomes

$$2\pi = \frac{\partial \mathcal{G}}{\partial J} \quad \text{or} \quad J = \frac{\mathcal{G}}{2\pi} = \frac{1}{2\pi} \oint S(\sigma', J, \delta)\, d\sigma',$$

which corresponds to the *oriented area* of the closed trajectory divided by 2π.

We call *oriented area* the area of the trajectory when it is followed clockwise and minus the area of the trajectory when it is followed counterclockwise.

Let us remark that $\dfrac{\partial \mathcal{G}}{\partial J} = \oint \dfrac{\partial S}{\partial J}(\sigma', J, \delta)\, d\sigma'$ and that S depends on J through the Hamilton–Jacobi equation (21), which means through $K = K(J)$:

$$\frac{\partial S}{\partial J} = \frac{\partial S}{\partial K}\,\frac{\partial K}{\partial J} = \frac{1}{\frac{\partial H}{\partial S}}\,\frac{\partial K}{\partial J} = \frac{1}{\dot{\sigma}}\,\frac{\partial K}{\partial J},$$

$\dot{\sigma} = 0$ at the unstable equilibrium, which means that this calculation as no meaning for closed orbits near the separatrices.

Another useful expression for J and consequently for the oriented area is given by

$$J = \frac{1}{2\pi} \oint S d\sigma = -\frac{1}{2\pi} \oint \sigma \, dS = \frac{1}{4\pi} \oint (S d\sigma - \sigma dS).$$

10.2 Dissipative Dynamics

Let us now remind that δ is a parameter *slowly varying with time* which means, in mathematical context:

$$|\dot{\delta}| \leq \zeta \quad \text{and} \quad |\ddot{\delta}| \leq \zeta^2,$$

where ζ is a slow parameter with respect to the characteristic periods of the closed trajectories (at least a factor 10 slower).

We perform the same canonical transformation to action-angle variables as in the conservative case; however, the Hamiltonian H and the generating function $\mathcal{F}$ are time dependent through δ; this means that the new Hamiltonian K is obtained by a corrected formula with respect to H:

$$H(\sigma, S, \delta) - \frac{\partial \mathcal{F}}{\partial t} = K(-, J, \delta).$$

The Hamiltonian writes

$$\begin{aligned} \mathcal{K}(\Psi, J, \delta) &= H(\sigma(\Psi, J, \delta), S(\Psi, J, \delta), \delta) \\ &= K(-, J, \delta) + \frac{\partial \mathcal{F}}{\partial t} \\ &= K(-, J, \delta) + \dot{\delta} \, \frac{\partial \mathcal{F}}{\partial \delta}(\Psi, J, \delta) \end{aligned}$$

and is associated to the dynamics:

$$\dot{\Psi} = \frac{\partial \mathcal{K}}{\partial J} = \frac{\partial K}{\partial J} + O(\eta) \quad \text{and} \quad \dot{J} = -\frac{\partial \mathcal{K}}{\partial \Psi} = O(\eta).$$

We can conclude that for small η the area J is quasi-constant, as long as we avoid the separatrices regions, and for times smaller than $\frac{1}{\eta}$. We follow the behavior of the dynamical system by the help of a *guiding trajectory*, the area of which is quasi-constant; δ is slowly changing with time, but at each time we fix its value and calculate the enclosed area, we get a quasi-constant quantity (with variations smaller than η for times smaller than $\frac{1}{\eta}$). We give in Fig. 18 an example of this behavior, for a pendulum of parameter δ, evolving from $\delta = \delta_0$ to $\delta = \delta_1$: the *eye*

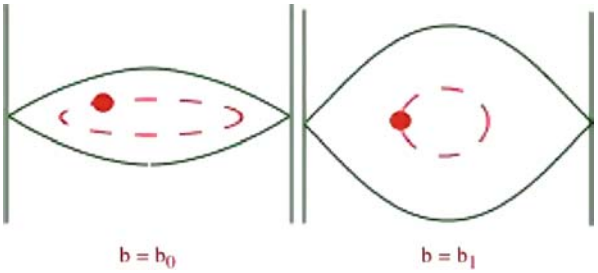

Fig. 18 The area enclosed by the guiding trajectory is kept constant on different phase spaces

of the cat is growing, the guiding trajectory evolves, but keeps a constant enclosed area, which allows to identify it from phase space to phase space.

10.3 Crossing of Separatrices

The interesting situation for capture (or escape) in resonance is obvious when the guiding trajectory approaches a separatrix.

Let us take any of the models that we have presented before, with an unstable equilibrium, from which two critical curves C_1 and C_2 start, dividing the space phase in three distinct regions, called 1, 2, or 3; they correspond to the positive circulation, the libration, and the negative circulation regions for the pendulum, and to the internal, the resonant and the external zones for the second fundamental model of resonance (Fig. 19).

Here we assume that the area enclosed by the guiding trajectory coincides, for a specific value of δ, with the area enclosed by one of the separatrices.

The adiabatic invariant approach fails and has to be replaced by a calculation of jumps from a region i to a region j, associated with a probability. Let us denote the unstable equilibrium by $\sigma^\star$ and $S^\star$ and let us consider a new Hamiltonian B relative to this equilibrium:

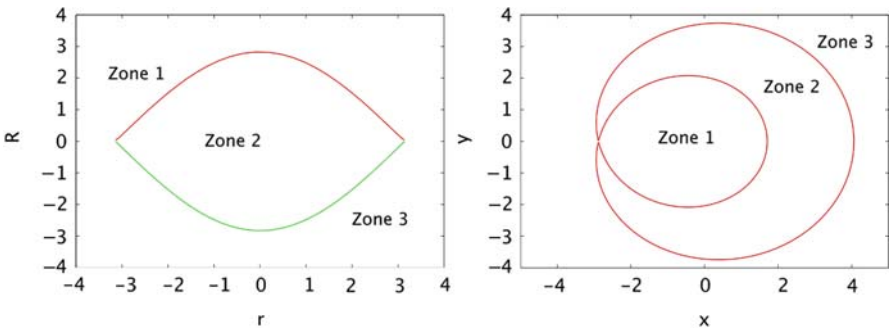

Fig. 19 The zones 1, 2, and 3 for the pendulum and the second fundamental model of resonance

$$B(\sigma, S, \delta) = H(\sigma, S, \delta) - H(\sigma^\star, S^\star, \delta).$$

It means that the two separatrices (joining at the unstable point) are characterized by a level $B = 0$:

$$\frac{dB}{dt} = \dot{\delta}\,\frac{dB}{d\delta}.$$

We can associate a sign, denoted by s_i, to B for each of the regions i, with the following characteristic: $s_1\ s_3 > 0$ and $s_1\ s_2 < 0$.

We also introduce a curve γ_i in each region i, which corresponds to a closed curve following the separatrix in this region (Fig. 20).

The energy (lost or gained) along a revolution on the curve γ_i can be approximated by the quantity B_i given by

$$\begin{aligned}
B_i &= \oint_{\gamma_i} dB = \int_{-\infty}^{\infty} \dot{\delta}\,\frac{dB}{d\delta}\,dt \\
&= \int_{-\infty}^{\infty} \dot{\delta}\,\left(\frac{\partial B}{\partial \sigma}\,\frac{\partial \sigma}{\partial \delta} + \frac{\partial B}{\partial S}\,\frac{\partial S}{\partial \delta} + \frac{\partial B}{\partial \delta}\right) dt \\
&\simeq \dot{\delta}\,\int_{-\infty}^{\infty} \left(\frac{\partial B}{\partial \sigma}\,\frac{\partial \sigma}{\partial \delta} + \frac{\partial B}{\partial S}\,\frac{\partial S}{\partial \delta} + \frac{\partial B}{\partial \delta}\right) dt \\
&\simeq \dot{\delta}\,\frac{\partial}{\partial \delta}\Big[\int_{-\infty}^{\infty} \left(\frac{\partial B}{\partial \sigma}\,\sigma + \frac{\partial B}{\partial S}\,S + B\right) dt\Big] + O(\dot{\delta}^2) \\
&\simeq \dot{\delta}\,\frac{\partial}{\partial \delta}\Big[\int_{-\infty}^{\infty} (-\dot{S}\,\sigma + \dot{\sigma}\,S)\,dt\Big] + O(\dot{\delta}^2) \qquad (B = 0 \ \text{ on } \gamma_i) \\
&= \dot{\delta}\,\frac{\partial}{\partial \delta}\Big[\oint_{\gamma_i} (-dS\,\sigma + d\sigma\,S)\,dt\Big] + O(\dot{\delta}^2) \\
&= 2\pi\dot{\delta}\,\frac{\partial \mathcal{A}_i}{\partial \delta},
\end{aligned}$$

where $\mathcal{A}_i$ is the oriented area of γ_i.

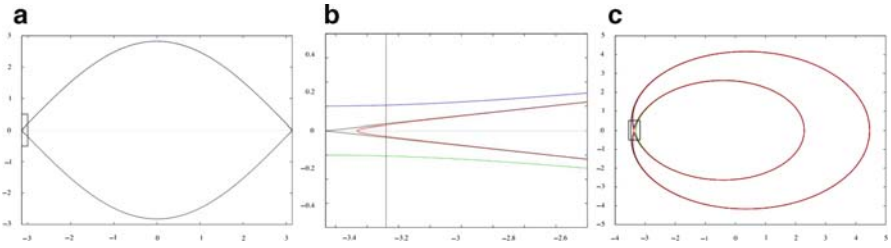

Fig. 20 The three curves $\gamma_i (i = 1, 2, 3)$ following the separatrices (C_1 and/or C_2) in the regions 1, 2, or 3 for the pendulum model and for the SFMR model

For the pendulum (A is the area of the libration region):

$$\mathcal{A}_1 = \frac{A}{2}, \quad \mathcal{A}_2 = A, \quad \mathcal{A}_1 = -\frac{A}{2},$$

and for the second fundamental model of resonance:

$$\mathcal{A}_1 = A_1, \quad \mathcal{A}_2 = -(A_2 - A_1), \quad \mathcal{A}_1 = -A_2.$$

The balance of energy gives

$$s_1 B_1 + s_2 B_2 + s_3 B_3 = 0.$$

10.4 Probability of Capture

We start with a trajectory in region i; we calculate the local situation of the energy: $s_i B_i$.

- $s_i B_i > 0$: it means that either we gain energy ($B_i > 0$) in a region where the separatrix has the minimum energy level ($s_i > 0$) or we loose energy ($B_i < 0$) in a region where the separatrix has the maximum energy level ($s_i > 0$); in both cases, we leave the separatrix and enter deeper in region i; no crossing of separatrix occurs and the orbit does not leave the region i.
- $s_i B_i < 0$: a gain or a loss of energy corresponds to getting closer to the separatrix; the crossing of the separatrix is then obvious, but to which other region, j or k ? (i, j, k being different and $\in \{1, 2, 3\}$). From the balance of energy, we know that $s_j B_j + s_k B_k > 0$ and then two cases are possible:
  - $s_j B_j > 0$ and $s_k B_k < 0$: it means that if we enter the region k, by our previous discussion, we are going out quite immediately. So the capture in the region j is the only possibility.
  - $s_j B_j > 0$ and $s_k B_k > 0$: the capture in regions j and k is possible, both behaviors are associated with *probabilities of capture* given by the following expressions:

$$Pr_{i\to j} = \frac{s_j B_j}{s_j B_j + s_k B_k} \tag{22}$$

$$= -\frac{s_j B_j}{s_i B_i}$$

$$Pr_{i\to k} = \frac{s_k B_k}{s_j B_j + s_k B_k} \tag{23}$$

$$= -\frac{s_k B_k}{s_i B_i}. \tag{24}$$

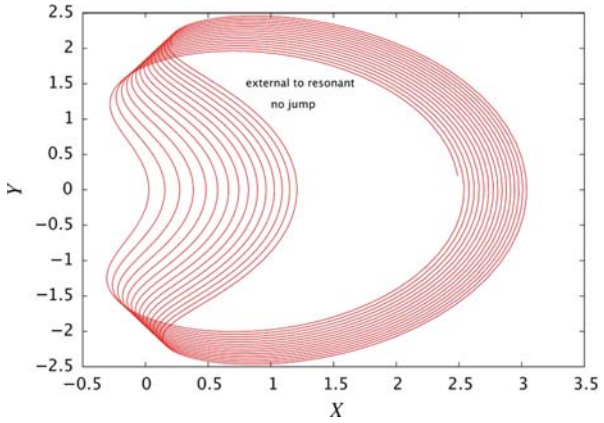

Fig. 21 A soft capture into resonance, without jump, observed in the second fundamental model of resonance from an initial external orbit: the trajectory for negative values of δ is a (deformed) ellipse, and when δ increases and reaches positive values, the trajectory is continuously moving to the resonant region, keeping its enclosed area constant

> Another way of describing these probabilities is to say that the phase of the system at the capture is unknown, and that the transition to region j or k depends on the phase of the system at that moment. Using a *probability argument* means that we assume that all phases are equiprobable.

We see in (24) that the probability of capture is not dependent on $\dot{\delta}$ (at first order) and is directly linked to the increasing or decreasing of the critical areas as functions of δ.

When the capture does not occur, the same formulae could explain a *jump* from an external to an internal orbit or from a positive to a negative circulation (Figs. 21

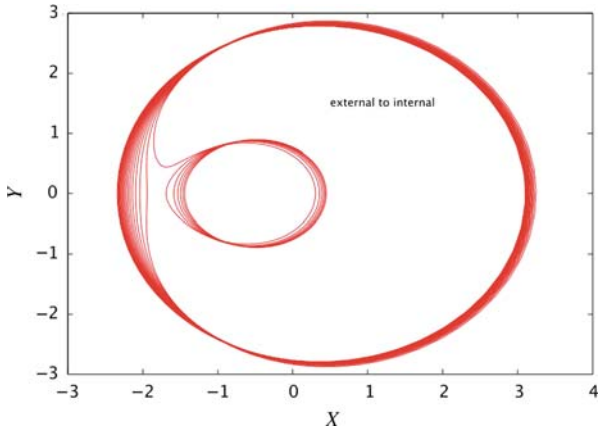

Fig. 22 A jump observed in the second fundamental model of resonance from an initial external orbit to an internal one

and 22). These parts of the theory have been used, in particular, to explain the depletion of the Kirkwood gaps in the asteroid main belt [22] and the differences between several resonances [23].

This is the simplest formula of capture into resonance that we can give; it can be applied and completed in several more specific contexts. For example, Malhotra calculated probabilities of capture in a secondary resonance and applied it to the case of Miranda and Umbriel in the Uranian system [26, 27]. Let us also mention the sweeping of secular resonances analyzed by the same model [24].

11 More Complicated Models: SFMRAS

Similar studies can be performed in more complicated models in which the number of equilibria, their stability, and the number of topological zones can be more important (Fig. 23).

For the resonances of order 2 and 3, the models can be described by a unique parameter (the δ parameter) and simple expressions for the probabilities of capture can be deduced [21].

For the order 4 a second parameter (called b) is already introduced, which complicates the topology of the phase space; however, this parameter is, for many applications, much more stable than δ and can be assigned to a specific constant value in local approaches.

For the second fundamental model of resonance with asymmetric equilibria, there are also two parameters, δ and b: indeed, if we start with the expression given by (12)

$$\mathcal{H}_1^c(N, \sigma, S) = \alpha(N)\ S + \beta(N)\ S^2 + \epsilon(N)\ \cos(\sigma) + \eta(N)\ \cos 2\sigma,$$

we can introduce the same scalings (of time and momentum) and change of phase as for the classical case with symmetric equilibria (see (18)):

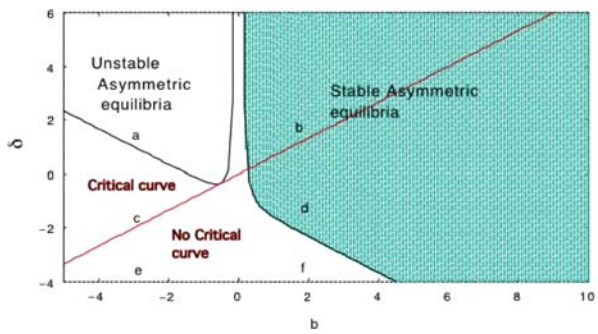

Fig. 23 Stability of symmetric and asymmetric equilibria in the second fundamental model of resonance with asymmetric equilibria (taken from [19])

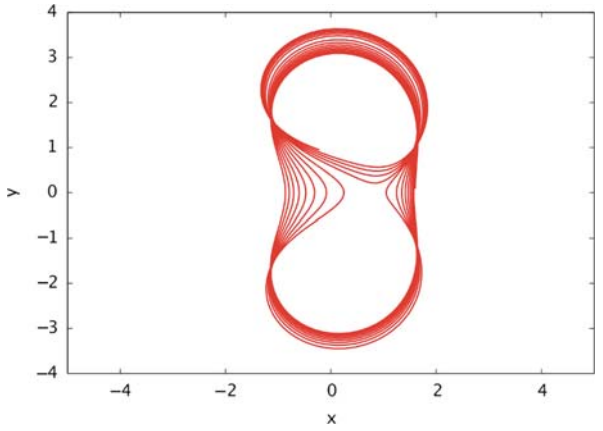

Fig. 24 Capture of a trajectory in an asymmetric equilibrium; δ varies from -1 to -0.75

$$K(r, R) = -3(\delta + 1)R + R^2 - 2\sqrt{2R}\cos r + 2\,b\,R\cos 2r.$$

The stability of the symmetric and asymmetric equilibria can be very different following their sign and values as functions of δ and b [19]. If a slow dissipation is introduced in the model, the parameters δ and b slowly change with respect to the time, giving adiabatic behavior of the trajectories. The areas are conserved as far as no critical curve is encountered. For the crossings of the separatrices, appropriate formulas of probability of capture are calculated (see Jancart 2004). As an illustration, we give a case of capture in asymmetric equilibrium; the initial value of δ is -1, b is kept constant ($b = 2$), and the dissipation is introduced by the coefficient $\dot{\delta} = 0.05$. The parameter δ evolves from -1 to -0.75. Figure 24 clearly shows the capture in the upper asymmetric equilibrium.

12 The Spin–Orbit Resonance

A very interesting class of resonances concerns the synchronous rotations, like the Moon, the Jovian, or the Saturnian satellites, and also the unique case of spin–orbit resonance 3:2, Mercury (Fig. 25).

12.1 The Rotation Variables

Let us assume that the body is not a point mass anymore. It is here considered as a rigid body, of mass M with three momenta of inertia A, B, and C, chosen as $A \leq B \leq C$.

We will use Andoyer's variables [5] to describe the rotation of the rigid body around its center of mass. They are based on two linked sets of Euler's angles

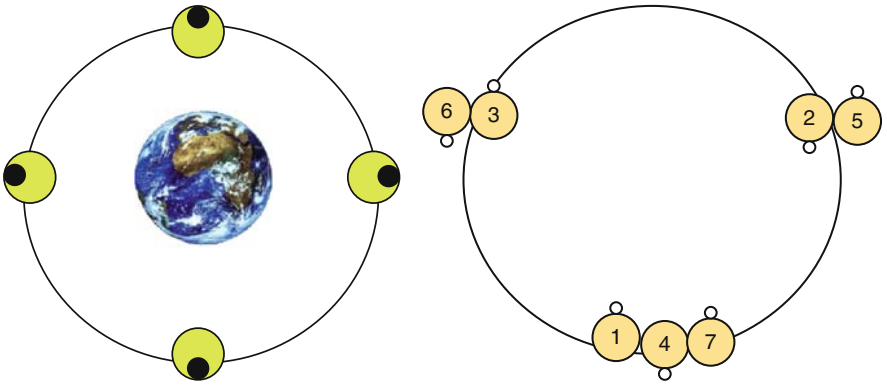

Fig. 25 A schematic view of a synchronous rotation, like the Moon or the Galilean satellites where the period of rotation equals the period of revolution around the primary, and the case of Mercury, where the period of rotation is 2/3 of the period of revolution

(Fig. 26). The first set (h, K, g) locates the position of the angular momentum vector **G** in an inertial frame of reference (the ecliptic plane at some epoch, for example); the second Euler's set (g, J, l) locates the body frame (the axis of inertia) in the

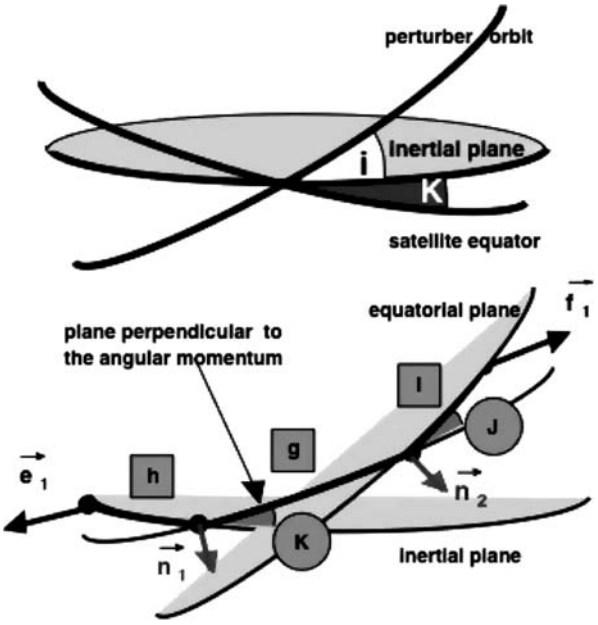

Fig. 26 The linked sets of Euler angles (l, K, g) and (g, J, h) from which the Andoyer's angular variables are defined. They locate the body frame $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$ with respect to the inertial frame $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$

previous frame tied to the angular momentum. The origin of both frames is the center of mass as origin, and the axes are the principal axes of inertia of the body.

The canonical set of Andoyer's variables consists of the three angular variables l, g, h and of their conjugated momenta defined by the norm G of the angular momentum and two of its projections: L is its projection onto the axis of figure and H onto the inertial axis.

Therefore, we define the following set of Andoyer's variables:

Variables	Momenta
l	$L = G\cos J$
g	G
h	$H = G\cos K$

With these variables the vectors $\boldsymbol{\omega}$ (the instantaneous rotation vector) and $\mathbf{G}$ (the angular momentum vector referred to the center of mass) can be computed. Their components in the frame of the principal axis of the body are

$$\begin{aligned}\boldsymbol{\omega} &= (A^{-1}G\sin J\sin l,\, B^{-1}G\sin J\cos l,\, C^{-1}G\cos J),\\ \mathbf{G} &= (G\sin J\sin l,\, G\sin J\cos l,\, G\cos J).\end{aligned}$$

The kinetic energy of the rotation is thus

$$\begin{aligned}T &= \frac{1}{2}(\,\boldsymbol{\omega}\mid\mathbf{G}\,)\\ &= \frac{1}{2}\,G^2\sin^2 J\left[\frac{\sin^2 l}{A}+\frac{\cos^2 l}{B}\right]+\frac{G^2\cos^2 J}{2C}\\ &= \frac{1}{2}(G^2-L^2)\left[\frac{\sin^2 l}{A}+\frac{\cos^2 l}{B}\right]+\frac{L^2}{2C}.\end{aligned}\tag{25}$$

Notice that the only angular variable appearing in it is l. Hence the dynamics of the free motion is reduced to a one degree of freedom (l, L) problem, the phase space of which is described in [5].

Andoyer's variables present so-called *virtual singularities*; when $J = 0$ the angular variables l and g are undefined but their sum is well defined; when $K = 0$, the angles g and h are not defined, although their sum is well defined. In order to avoid these singularities, we shall thus use the following *modified Andoyer's variables*:

$$\begin{aligned}\lambda_1 &= l+g+h, & \Lambda_1 &= G,\\ \lambda_2 &= -l, & \Lambda_2 &= G-L = G(1-\cos J) = 2G\sin^2\frac{J}{2},\\ \lambda_3 &= -h, & \Lambda_3 &= G-H = G(1-\cos K) = 2G\sin^2\frac{K}{2}.\end{aligned}$$

This set of variables—momenta (λ_i, Λ_i) is canonical and called the set of *modified Andoyer's elements*, which partially solves the problem of virtual singularities.

The fast spin motion is given by the variable g in Andoyer's variables, which means by the first variable λ_1 in the modified set of variables. The spin velocity is given by the main contribution:

$$\dot{\lambda}_1 = \frac{\partial T}{\partial \Lambda_1} =\simeq \frac{\Lambda_1}{C}. \tag{26}$$

We can associate canonical Cartesian coordinates to (λ_2, Λ_2) by the usual transformation: $(\xi = \sqrt{2\Lambda_2}\sin\lambda_2, \eta = \sqrt{2\Lambda_2}\cos\lambda_2)$. Then the Hamiltonian takes the following form:

$$\mathcal{H} = \frac{\Lambda_1^2}{2C} + \frac{4\Lambda_1 - \xi^2 - \eta^2}{8C}\left[\frac{\gamma_1+\gamma_2}{1-\gamma_1-\gamma_2}\xi^2 + \frac{\gamma_1-\gamma_2}{1-\gamma_1+\gamma_2}\eta^2\right], \tag{27}$$

where

$$\gamma_1 = (2C - A - B)/2C \quad \text{and} \quad \gamma_2 = (B - A)/2C. \tag{28}$$

12.2 Perturbation

To introduce a spin–orbit resonance, we need to mix the rotation dynamics (mainly the angle λ_1) with the orbital motion (mainly the mean longitude).

Let us consider that the orbital dynamics of the rigid body of mass M is perfectly known; in the simplest cases, it is given by a Keplerian orbit, in a two-body configuration with a point mass m (the Earth for the lunar motion, the Sun for Mercury's orbit). In the reference frame linked to M, the orbit of m is described by elliptic elements $(a, e, i, \omega, \Omega, \ell)$ defined as always: a the semi-major axis, e the eccentricity, i the inclination, ℓ the mean anomaly, ω the argument of the pericenter, and Ω the longitude of the node, defined with respect to the selected inertial frame.

We also introduce the mean motion of m denoted by $n = \sqrt{\frac{\mathcal{G}(M+m)}{a^3}}$.

The gravitational potential due to the presence of m can be expressed by

$$V = -\mathcal{G}\, m \iiint_W \frac{\rho\, dW}{r'},$$

where ρ is the density inside the volume W of the body and r' is the distance between m and any volume element dW inside the body.

Using the usual expansion of the potential in spherical harmonics, we find

$$V = -\frac{\mathcal{G}m}{r}\left\{1 + \sum_{n\geq 1}\frac{1}{r^n}\sum_{m=0}^{n} P_n^m(\sin\phi)\left[C_n^m\cos m\Psi + S_n^m\sin m\Psi\right]\right\},$$

where Ψ and ϕ are the longitude and latitude of m in the body frame, and r is the distance between m and the center of mass of the body.

We limit the expansion to the second order terms; the first term $\mathcal{G}\frac{m}{r}$ will be taken into account later on, it has no direct effect on the rotation. Then we limit the potential to the following formula:

$$V = \frac{3\mathcal{G}m}{2r^3}\left[C_2^0(x^2+y^2) - 2C_2^2(x^2-y^2)\right]. \tag{29}$$

where (x, y, z) are the components, in the body frame, of the unit vector pointing to m ($x^2+y^2+z^2=1$). The unscaled coefficients $C_2^0 = \frac{A+B-2C}{2}$ and $C_2^2 = \frac{B-A}{4}$ in the potential are related to their usual scaled coefficients J_2 and C_2 by $C_2^0 = -MR^2J_2$ and $C_2^2 = MR^2C_{22}$.

The potential now reads:

$$V = n^2C\left(\frac{a}{r}\right)^3\left[\delta_1(x^2+y^2) + \delta_2(x^2-y^2)\right], \tag{30}$$

with

$$\delta_1 = -\frac{3}{2}\,\frac{m\,\gamma_1}{m+M} = -\frac{3}{2}\,\frac{m}{m+M}\,\frac{MR^2}{C}\,J_2,$$

$$\delta_2 = -\frac{3}{2}\,\frac{m\,\gamma_2}{m+M} = 3\frac{m}{m+M}\,\frac{MR^2}{C}\,C_{22},$$

The Hamiltonian is time dependent through the orbital motion of m; we introduce a new angular variable, the mean longitude of m, $\lambda = \ell + \omega + \Omega = n\ t + \lambda_0$ to which we associate a momentum Λ.

The complete Hamiltonian, obtained by the summation of the kinetic energy (25), the orbital motion (classical two-body potential), and the perturbing potential (30), becomes

$$\mathcal{H} = n\Lambda + \frac{\Lambda_1^2}{2C} + \frac{4\Lambda_1 - \xi^2 - \eta^2}{8C}\left[\frac{\gamma_1+\gamma_2}{1-\gamma_1-\gamma_2}\,\xi^2 + \frac{\gamma_1-\gamma_2}{1-\gamma_1+\gamma_2}\,\eta^2\right]$$
$$+ n^2C\left(\frac{a}{r}\right)^3\left[\delta_1(x^2+y^2) + \delta_2(x^2-y^2)\right]. \tag{31}$$

The first term is the two-body energy, in which the term $\mathcal{G}\dfrac{m}{r}$ of the gravitational potential is inserted.

12.3 The 1:1 Resonance

We then introduce the set of resonant canonical variables, with the apparition of the difference between the two quasi-synchronous variables following [12] but also [28]:

$$\begin{array}{ll} \sigma = \lambda_1 - \lambda & S = \Lambda_1 \\ \lambda & \Gamma = \Lambda + \Lambda_1 \end{array}$$

The spin velocity of the body, given by its first approximation (26), i.e. by S/C, is assumed to be almost equal to its orbital velocity given by n. λ is then a *fast variable* while σ in a 1 : 1 spin-orbit case, becomes much slower.

The next step in the theory of the rotation is to perform an *averaging* canonical transformation in order to eliminate the fast variable (which is hidden in x and y) from the expression of the Hamiltonian and to follow the dynamics of the resonant angle σ. As it is well known and already applied in the previous sections, the effect of a first-order averaging transformation is simply to remove all the terms which contain this variable. We assume that this step has been performed and we finally obtain the *averaged* Hamiltonian (33).

In other words, the original Hamiltonian contains periodic terms with linear combinations of the angles σ, λ, and $\lambda_3 + \Omega$; we average over the short periods, which means over λ. The remaining Hamiltonian (for a circular orbit and neglecting the terms of fourth order in ξ and η) becomes

$$\begin{aligned} \mathcal{H} = {} & n\Gamma - nS + \frac{S^2}{2C} + \frac{S}{2C}\left[\frac{\gamma_1+\gamma_2}{1-\gamma_1-\gamma_2}\,\xi^2 + \frac{\gamma_1-\gamma_2}{1-\gamma_1+\gamma_2}\,\eta^2\right] \\ & + n^2 C\left[\delta_1(x^2+y^2) + \delta_2(x^2-y^2)\right], \end{aligned} \tag{32}$$

with

$$\begin{aligned} x^2+y^2 = {} & F_0 + F_1\cos\nu + F_2\cos 2\nu \\ & - (\xi^2+\eta^2)\,[G_0 + G_1\cos\nu + G_2\cos 2\nu] \\ & - (\xi^2-\eta^2)\sum_{i=0}^{5} B_i\,\cos(2\sigma + i\nu) \\ & + 2\,\xi\,\eta\sum_{i=0}^{5} A_i\,\sin(2\sigma + i\nu), \\ x^2-y^2 = {} & (2-\xi^2-\eta^2)\sum_{i=0}^{5} C_i\,\cos(2\sigma + i\nu) \\ & - (\xi^2-\eta^2)\,[H_0 + H_1\cos\nu + H_2\cos 2\nu] \\ & + 2\,\xi\,\eta\sum_{i=0}^{5} D_i\,\sin(2\sigma + i\nu), \end{aligned} \tag{33}$$

where $\nu = \lambda_3 + \Omega$. The functions F_i, G_i, H_i, A_i, B_i, c_i, and D_i are polynomials in $\cos K$, $\sin K$, $\cos i$, and $\sin i$ (see [12] for explicit formulations), which also means functions of S and Λ_3. Γ is now a constant because its conjugate variable λ (the fast variable) is not present anymore in the averaged Hamiltonian $\mathcal{H}$; then it can be forgotten.

12.4 Precessing Motion

To give an immediate and interesting generalization, we can assume that m is on a slowly precessing circular orbit, with a precession frequency $\dot{\Omega}$ measured on the same inertial frame, centered on M.

In that case, we also introduce a fictitious momentum P associated to Ω and a term $\dot{\Omega}\ P$ in the Hamiltonian. The variables and momenta are now:

$$(\sigma, \xi, \lambda_3, \Omega, S, \eta, \Lambda_3, P),$$

and if we use $\nu = \lambda_3 + \Omega$ instead of λ_3 and if we introduce the momentum $P' = P - \Lambda_3$, we obtain the following new set of variables:

$$(\sigma, \xi, \nu, \Omega, S, \eta, \Lambda_3, P').$$

The corresponding Hamiltonian becomes

$$\mathcal{H} = \dot{\Omega}\,(P' + \Lambda_3) - nS + \frac{S^2}{2C} + \frac{S}{2C}\left[\frac{\gamma_1+\gamma_2}{1-\gamma_1-\gamma_2}\,\xi^2 + \frac{\gamma_1-\gamma_2}{1-\gamma_1+\gamma_2}\,\eta^2\right] + n^2C\left[\delta_1(x^2+y^2) + \delta_2(x^2-y^2)\right].$$

12.5 The Equilibrium

Writing up the differential equations generated by this Hamiltonian, we calculate the equilibria by putting them to zero.

The interesting stable equilibrium (the exact spin–orbit resonance) is characterized by

- $\sigma = 0$: the axis of smallest moment of inertia points toward the perturber.
- $\xi = 0 = \eta$: the axis of largest moment of inertia is aligned with the angular momentum.
- $\nu = 0$: the lines of node of the orbit and of the equator are aligned.
- The equation $\frac{\partial \mathcal{H}}{\partial \Lambda_3} = 0$ fixes the value of the obliquity $K^\star$ of the equilibrium by the equation:

$$4\,\dot{\Omega}\sin K^\star - \frac{n^2\,C}{S}\Big((2\,\delta_1 + \delta_2)\ \sin(2K^\star - 2\,i) + 2\delta_2 \sin(K^\star - i)\Big) = 0. \quad (34)$$

- The equation $\frac{\partial \mathcal{H}}{\partial S} = 0$ gives the value of S at the equilibrium:

$$\frac{S^\star}{C} = n - \frac{n^2\, C\,(1-\cos K^\star)}{4S}\,\Big((2\,\delta_1+\delta_2)\,\sin(2K^\star - 2\,i) + 2\delta_2 \sin(K^\star - i)\Big). \tag{35}$$

Let us notice that if we neglect the precession rate ($\dot{\Omega} = 0$) in (34), the obliquity at the equilibrium $K^\star$ coincides with the inclination i, and the two frequencies are exactly equal: $S^\star = nC$.

On the opposite, if $\dot{\Omega} \neq 0$, there is a difference between $K^\star$ and i and (35) shows the correction to add on the exact commensurability.

If the inclination i is very small ($\sin i \simeq i$ and $\cos i \simeq 1$), the equilibrium equation gives an analytical solution for $K^\star$:

$$K^\star = \frac{\delta_1+\delta_2}{\delta_1+\delta_2 - \frac{S^\star\,\dot{\Omega}}{n^2\,C}}\; i.$$

The sign of this quantity is the sign of its denominator: let us first remark that at first order, we can write

$$\frac{S^\star\,\dot{\Omega}}{n^2\,C} = \frac{S^\star}{nC}\,\frac{\dot{\Omega}}{n} \simeq \frac{\dot{\Omega}}{n}.$$

Consequently,

- if $\delta_1+\delta_2 < \frac{\dot{\Omega}}{n}$ the value of $K^\star < 0$, as for the Moon, for example, where $\dot{\Omega}$ is large;
- if $\delta_1+\delta_2 > \frac{\dot{\Omega}}{n}$ the value of $K^\star > 0$, as in the case of Europa, where the precession rate $\dot{\Omega}$ is smaller.

12.6 The Models

If we want a very simple one degree of freedom model of resonance for the spin–orbit motion, it is quite easy by eliminating the precession ($\dot{\Omega} = 0$) and keeping two degrees of freedom to their values at the equilibrium: we simply assume that $\xi = 0 = \eta$, $\nu = 0$ and $K^\star = i$.

We can show that $x^2 + y^2$ is then reduced to a constant term (that we drop) and $x^2 - y^2$ is proportional to $\cos 2\sigma$. The only degree of freedom is the couple (σ, S).

The Hamiltonian (after elimination of the constants) is reduced to

$$\begin{aligned}\mathcal{H} &= -nS + \frac{S^2}{2C} + n^2 C\ \delta_2\ \epsilon\ \cos 2\sigma \\ &= -nS + \frac{S^2}{2C} + n^2 C \frac{3m}{m+M} \frac{MR^2}{C}\ C_{22}\ \epsilon\ \cos 2\sigma \\ &= -nS + \frac{S^2}{2C} + n^2\ \frac{3m}{m+M} MR^2 C_{22}\ \epsilon\ \cos 2\sigma \\ &= -nS + \frac{S^2}{2C} + n^2\ \frac{3m}{m+M} \frac{(B-A)}{4}\ \epsilon\ \cos 2\sigma,\end{aligned}$$

where ϵ is a numerical factor depending on the functions defined in (33).

We obtain a classical pendulum model, which describes the *averaged* motion in the case of the spin–orbit resonance. The linear term is easily eliminated by a translation around $S^\star = nC$ and the final momentum is $\Delta S = S - S^\star$:

$$\begin{aligned}-nS + \frac{S^2}{2C} &= -n\ (\Delta S + S^\star) + \frac{(\Delta S + S^\star)^2}{2C} \\ &= -n\ \Delta S - n\ S^\star + \frac{\Delta S^2}{2C} + \frac{2\ \Delta S\ S^\star}{2C} + \frac{(S^\star)^2}{2C} \\ &= \frac{\Delta S^2}{2C} + \Delta S\ (-n + \frac{S^\star}{C}) + \text{constant terms} \\ &\equiv \frac{\Delta S^2}{2C}.\end{aligned}$$

If we reintroduce the short periodic terms, we get time-dependent contributions which affect the dynamics of the pendulum, especially in the region of the separatrices. These models have been developed and analyzed by several authors, especially [4, 3], based on the differential equation (perturbed pendulum):

$$\ddot{\sigma} + \epsilon \sum_j A_j(e)\ \sin(2\sigma - jnt) = 0.$$

12.7 The Fundamental Frequencies

Let us calculate the three fundamental proper frequencies (often called *free frequencies*) associated to the three-dimensional equilibrium. We assume to be very close to the equilibrium, and we expand the Hamiltonian in powers of six small quantities; they represent the distances from each variable or momentum to its value at the equilibrium:

$$\begin{aligned}\Delta_\sigma &= \sigma \qquad \Delta_S = S - S^\star \\ \Delta_\nu &= \nu \qquad \Delta_\Lambda = \Lambda_3 - \Lambda_3^\star \\ \Delta_\xi &= \xi \qquad \Delta_\eta = \eta\end{aligned}$$

and the expansion gives

$$\begin{aligned} 2\mathcal{H} \simeq = & \, a_1 \, \Delta_\sigma^2 + 2\, a_2 \, \Delta_\sigma \, \Delta_\nu + a_3 \, \Delta_\nu^2 \\ & + b_1 \, \Delta_S^2 + 2\, b_2 \, \Delta_S \, \Delta_\Lambda + b_3 \, \Delta_\Lambda^2 \\ & + c_1 \, \Delta_\xi^2 + c_2 \, \Delta_\eta^2, \end{aligned}$$

where a_i, b_i, and c_i are the second partial derivatives of the Hamiltonian evaluated at the equilibrium coordinates (see [12] for explicit expressions).

An untangling transformation [16] is necessary to dissociate the contributions of the first and second degrees of freedom; a final scaling allows to write the Hamiltonian (written in action-angle variables) as a summation of three momenta multiplied by a frequency:

$$\mathcal{H} = \omega_1 \, J_1 + \omega_2 \, J_2 + \omega_3 \, J_3.$$

The three frequencies ω_i (and the corresponding periods T_i) are the fundamental frequencies of the rotation: the free motion is a quasi-periodic function of these three motions, perturbed by forced terms coming from external bodies or interactions. Of course, in non-averaged models, the orbital period is also present as the fourth period.

This formalism has been developed not only for the Galilean satellites, Europa and Io [13–15, 17], but also for Titan [38].

12.8 The Case of Mercury

The case of Mercury is slightly different from these mentioned above: it is blocked in a 3:2 spin-orbit resonance, which means that the basic (kernel) model depends on the eccentricity, which is not the case for the 1:1 commensurability. The influence of the precession of the orbit ($\dot{\Omega}$) is much less important, the equilibrium obliquity $K^\star$ is moved by a quantity of the order of $2'$ from the inclination of Mercury, which is about 7° with respect to the ecliptic. One of the first analyses of this 3 : 2 spin-orbit was done by [41] with very recent contributions [40, 39].

A complete Hamiltonian analysis of this spin-orbit resonance can be found in [6] for the first two frequencies, [7] for the third frequency, and [42] for the coupling (and untangling) of the first two degrees of freedom (σ, S) and (ν, Λ_3). The three fundamental periods for Mercury are (depending on the values for C_{22}), and calculated in the case of a rigid body, of the order of

- For σ, the *longitude of the libration in longitude*: between 10 and 15 years
- For ν, the *nodes commensurability*: around 1060 years
- For ξ, the *wobble*: around 585 years

and the fourth period is, of course, 88 days which corresponds to the orbital period. For a more complete analysis, let us refer to [25]. The frequencies change drastically if Mercury is assumed to have a liquid core [8].

13 The Gravitational Resonances

Another interesting family of resonances concerns the commensurabilities between the orbital period of a first object (called here *the satellite*) orbiting a second body (called here *the planet*) with the rotation period of this planet; typically the geostationary situation of an artificial satellite in gravitational resonance 1:1.

13.1 The Potential of the Earth

The apparition of such a resonance comes from the fact that the planets are neither spherical nor homogeneous. The potential function induced by such a body on an external body is written as

$$U(\mathbf{r}) = \mu \int_V \frac{\rho(\mathbf{r}_p)}{\|\mathbf{r} - \mathbf{r}_p\|}\, dV\,,$$

where $\rho(\mathbf{r}_p)$ stands for the density at some position $\mathbf{r}_p$ inside the planet, $\|\mathbf{r} - \mathbf{r}_p\|$ is the distance between the body and any particular volume element located at $\mathbf{r}_p$, and $\mu = \mathcal{G}\, M$, with $\mathcal{G}$ is the gravitational constant and M the mass of the planet.

This potential is developed in several steps. First, we introduce the Legendre polynomials:

$$\frac{1}{\|\mathbf{r} - \mathbf{r}_p\|} = \frac{1}{r} \sum_{n=0}^{\infty} \left(\frac{r_p}{r}\right)^n \mathcal{P}_n(\cos\,\Psi) \qquad \text{where} \quad \frac{\|\mathbf{r}_p\|}{\|\mathbf{r}\|} = \frac{r_p}{r},$$

where Ψ is the geocentric angle between $\mathbf{r}$ and $\mathbf{r}_p$ and $\mathcal{P}_n$ are the *Legendre polynomials* of degree n.

Second, by introducing the spherical coordinates in the planet-fixed reference frame, i.e., the longitude λ and the latitude ϕ of the body of position $\mathbf{r}$ and of coordinates x, y, and z:

$$\begin{aligned} x &= r\,\cos\phi\,\cos\lambda \\ y &= r\,\cos\phi\,\sin\lambda \\ z &= r\,\sin\phi\,, \end{aligned}$$

as well as the corresponding quantities λ_p and ϕ_p for the volume element at $\mathbf{r}_p$, and by using the decomposition formula, the Legendre polynomials can be expanded into spherical harmonics

$$\mathcal{P}_n(\cos\,\Psi) = \sum_{m=0}^{n}(2-\delta_{0m})\,\frac{(n-m)!}{(n+m)!}\,\mathcal{P}_n^m(\sin\phi_p)\,\cos(m(\lambda-\lambda_p))\,,$$

where $\delta_{ij} = 1$ for $i = j$ and zero otherwise. $\mathcal{P}_n^m$ are the so-called *associated Legendre functions*. We write the gravity potential in the form

$$U(r,\lambda,\phi) = -\frac{\mu}{r}\sum_{n=0}^{\infty}\sum_{m=0}^{n}\left(\frac{R_e}{r}\right)^n\,\mathcal{P}_n^m(\sin\phi)(C_{nm}\cos\,m\lambda + S_{nm}\sin\,m\lambda)\,, \quad (36)$$

where R_e is the equatorial radius of the planet and where the quantities C_{nm} and S_{nm} are the spherical harmonics coefficients which are given by

$$C_{nm} = \frac{2-\delta_{0m}}{M_\oplus}\,\frac{(n-m)!}{(n+m)!}\int_V\left(\frac{r_p}{R_e}\right)^n\mathcal{P}_n^m(\sin\phi_p)\cos(m\lambda_p)\,\rho(r_p)\,dV,$$

$$S_{nm} = \frac{2-\delta_{0m}}{M_\oplus}\,\frac{(n-m)!}{(n+m)!}\int_V\left(\frac{r_p}{R_e}\right)^n\mathcal{P}_n^m(\sin\phi_p)\sin(m\lambda_p)\,\rho(r_p)\,dV\,.$$

The coefficient C_{00} is equal to 1; all terms S_{n0} are obviously zero, the coefficients C_{10}, C_{11}, and S_{11} correspond to the center of mass coordinates divided by the equatorial radius. Therefore, these coefficients are zero if the coordinate system refers to the planet center of mass. Similarly, the coefficients C_{21} and S_{21} are zero if the z-axis is aligned with the planet main axis of inertia. Finally, it can be shown that

$$J_2 = -C_{20} = \frac{2C-B-A}{2\,M\,R_e^2} \qquad \text{and} \quad C_{22} = \frac{B-A}{4\,M\,R_e^2}\,,$$

where A, B, and C (with $A < B < C$) are the principal moments of inertia of the planet.

With these choices, the potential is expressed in the following form, with a single cosine term, a phase difference λ_{nm} as well as a new J_{nm} coefficient:

$$U(r,\lambda,\phi) = -\frac{\mu}{r} + \frac{\mu}{r}\sum_{n=2}^{\infty}\sum_{m=0}^{n}\left(\frac{R_e}{r}\right)^n\,\mathcal{P}_n^m(\sin\phi)\,J_{nm}\,\cos m(\lambda-\lambda_{nm}),$$

using the definitions for $n \geq m \geq 0$

$$C_{nm} = -J_{nm}\,\cos(m\lambda_{nm}),\;\; S_{nm} = -J_{nm}\,\sin(m\lambda_{nm}),$$
$$J_{nm} = \sqrt{C_{nm}^2 + S_{nm}^2},\;\; m\,\lambda_{nm} = \arctan\left(\frac{-S_{nm}}{-C_{nm}}\right).$$

The next step is to develop the gravity field in terms of the satellite orbital elements $(a, e, i, \Omega, \omega, M)$

$$U(r,\lambda,\phi) = -\frac{\mu}{r} - \sum_{n=2}^{\infty}\sum_{m=0}^{n}\sum_{p=0}^{n}\sum_{q=-\infty}^{+\infty}\frac{\mu}{a}\left(\frac{R_e}{a}\right)^n F_{nmp}(i)\,G_{npq}(e)\,S_{nmpq}(\Omega,\omega,M,\theta)\,,$$

where the functions S_{nmpq} depend on the geopotential coefficients C_{nm} and S_{nm}

$$S_{nmpq}(\Omega,\omega,M,\theta) = \begin{bmatrix} +C_{nm} \\ -S_{nm} \end{bmatrix}_{n-m\,\text{odd}}^{n-m\,\text{even}} \cos\,\Theta_{nmpq}(\Omega,\omega,M,\theta) + \begin{bmatrix} +S_{nm} \\ +C_{nm} \end{bmatrix}_{n-m\,\text{odd}}^{n-m\,\text{even}} \sin\,\Theta_{nmpq}(\Omega,\omega,M,\theta)\,,$$

and the angle is defined by

$$\Theta_{nmpq}(\Omega,\omega,M,\theta) = (n-2p)\,\omega + (n-2p+q)\,M + m(\Omega-\theta)\,.$$

where θ is the sidereal time. The subscript indexes represented by n, m, p, q are integers that identify the terms in the so-called *inclination functions* $F_{nmp}(i)$ and *eccentricity functions* $G_{npq}(e)$ for a particular harmonic (n, m).

13.2 Resonance with the Rotation of the Planet

The orbital period of an object in orbit is said to be *in resonance* with the rotation of the planet if a small integer number q_1 of sidereal days of the planet is equal to a small integer number q_2 of revolution periods of the object, that is,

$$\frac{P_R}{P_{obj}} = \frac{q_1}{q_2}\,,$$

where P_R is the rotational period of the planet, that is, $2\pi/n_R = 1$ planetary day ($n_R = \dot{\theta}$) and P_{obj} is the orbital period of the satellite orbiting the planet.

These resonances occur when the rate of the Kaula gravitational argument is close to zero, that is,

$$\dot{\Theta}_{nmpq}(\dot{\Omega},\dot{\omega},\dot{M},\dot{\theta}) = (n-2p)\,\dot{\omega} + (n-2p+q)\,\dot{M} + m(\dot{\Omega}-\dot{\theta}) \simeq 0\,.$$

Typically, when the condition $q = 0$ is satisfied (when we consider a zero-order expansion with respect to the eccentricity), we have

$$(n-2p)\,(\dot{\omega}+\dot{M}) \simeq m\,(\dot{\theta}-\dot{\Omega})\,,$$

or similarly

$$\frac{\dot{\omega} + \dot{M}}{\dot{\theta} - \dot{\Omega}} \simeq \frac{q_1}{q_2}. \tag{37}$$

Such resonances are also said to be *Repeat Ground-Track Resonances*. The rates of both ω and Ω are small and the simplified resonance condition reads

$$\frac{\dot{M}}{\dot{\theta}} \simeq \frac{\dot{\lambda}}{\dot{\theta}} \simeq \frac{q_1}{q_2}.$$

When the ratio q_1/q_2 is close to 1, the resonance is clearly associated with the geostationary orbit whereas it is close to 2 for the GPS satellites.

13.3 Resonant Hamiltonian Formalism—the Resonance Angle

Let us write the potential truncated at the second order and degree harmonic, denoted by $\mathcal{U}_{J_{22}}$:

$$\mathcal{U}_{J_{22}} = 3\,\frac{\mu^4\,R_e^2}{L^6}\left[C_{22}\,(\bar{x}^2 - \bar{y}^2) + S_{22}\,(2\bar{x}\,\bar{y})\right],$$

where $\bar{x} = x/r$ and $\bar{y} = y/r$ and let us confine ourselves to the circular orbits in the equatorial plane ($i = 0$ and $e = 0$). Within these assumptions and in order to outline the main features of the 1:1 resonance, we consider the following "minimum" resonant Hamiltonian $\mathcal{H}$ including the two-body problem, the (simplified) potential $\mathcal{U}_{J_{22}}$ and a contribution coming from the *external* angle θ, the sidereal time, which introduces the rotation of the planet in the dynamics and is associated to a momentum Λ:

$$\mathcal{H}(\lambda, L, \theta, \Lambda) = -\frac{\mu^2}{2L^2} + \dot{\theta}\,\Lambda + \mathcal{U}_{J_{22}}(\lambda, L, \theta, \Lambda),$$

where λ is the mean longitude and L is the Delaunay-associated momentum, $L = \sqrt{\mathcal{G}\,Ma}$.

In the case of a $1:1$ gravitational resonance, we define the resonant angle σ by $\sigma = \lambda - \theta$.

In order to keep a canonical set of variables with L associated to σ, we use the following symplectic transformation (see [43]) :

$$d\sigma\,L' + d\theta'\,\Lambda' = d\lambda\,L + d\theta\,\Lambda,$$

leading to the new set of canonical variables

$$\sigma = \lambda - \theta, \qquad L' = L, \qquad \theta' = \theta, \qquad \Lambda' = \Lambda + L,$$

and the new Hamiltonian formulation including the resonant angle

$$\mathcal{H}(\sigma, L, \theta, \Lambda') = -\frac{\mu^2}{2L^2} + \dot{\theta}(\Lambda' - L) + \mathcal{U}_{J_{22}}(\sigma, L, \theta, \Lambda').$$

13.4 Simplified Analytical Averaged Model

To get the final model of resonance, we average the Hamiltonian function over the fast angular variable θ, and we obtain the following result:

$$\bar{\mathcal{H}}(\bar{L}, \bar{\sigma}, \bar{\Lambda}) = -\frac{\mu^2}{2\bar{L}^2} - \dot{\theta}\bar{L} + \frac{1}{\bar{L}^6}\left[\alpha_1 \cos 2\bar{\sigma} + \alpha_2 \sin 2\bar{\sigma}\right],$$

in which the quantities are all averaged; for simplicity we shall use again the same letters (without bars) in the calculations of the equilibria.

The numerical values of α_1 and α_2 come from the coefficients C_{22} and S_{22}. For the Earth, their values are

$$\alpha_1 \simeq 0.1077 \times 10^{-6}, \qquad \alpha_2 \simeq -0.6204 \times 10^{-7}.$$

Two stable equilibria $(\sigma_{11}^*, L_{11}^*)$, $(\sigma_{12}^*, L_{12}^*)$ as well as two unstable equilibria $(\sigma_{21}^*, L_{21}^*)$, $(\sigma_{22}^*, L_{22}^*)$ are found to be solutions of

$$\frac{\partial \mathcal{H}}{\partial L} = \frac{\partial \mathcal{H}}{\partial \sigma} = 0,$$

where

$$\begin{aligned} \sigma_{11}^* &= \lambda^* & \sigma_{12}^* &= \lambda^* + \pi \\ \sigma_{21}^* &= \lambda^* + \frac{\pi}{2} & \sigma_{22}^* &= \lambda^* + \frac{3\pi}{2}, \end{aligned}$$

as well as

$$L_{11}^* = L_{12}^* = 0.99999971, \qquad L_{21}^* = L_{22}^* = 1.00000029,$$

where the distance unit has been set to the exact resonant position, namely 42, 164 km for the Earth. Again for the Earth, the angular value λ^* is the first quadrant solution of

$$\tan 2\lambda^* = \frac{S_{22}}{C_{22}} = \frac{\alpha_2}{\alpha_1},$$

that is, $\lambda^* \simeq 75.07^\circ$. (Fig. 27)

13.5 The Resonant Frequency

The Hamiltonian is reduced to a quadratic form in a neighborhood of the stable equilibrium point.

Let us introduce the resonant Cartesian coordinates ($x = \sqrt{2L}\cos\sigma$, $y = \sqrt{2L}\sin\sigma$) and at any equilibrium ($x^* = \sqrt{2L^*}\cos\sigma^*$, $y^* = \sqrt{2L^*}\sin\sigma^*$). Developing the Hamiltonian function in Taylor series around one of the stable equilibria (x^*, y^*), up to the second order, we find, after having dropped the constant additive terms and setting $X = (x - x^*)$ and $Y = (y - y^*)$:

$$\mathcal{H}^*(X, Y) = \frac{1}{2}(aX^2 + 2bXY + cY^2) + \cdots .$$

The values a, b, and c stand for the second-order derivatives

$$a = \left.\frac{\partial^2\mathcal{H}}{\partial x^2}\right|_{(L^*,\sigma^*)}, \qquad b = \left.\frac{\partial^2\mathcal{H}}{\partial x\partial y}\right|_{(L^*,\sigma^*)}, \qquad c = \left.\frac{\partial^2\mathcal{H}}{\partial y^2}\right|_{(L^*,\sigma^*)},$$

where (L^*, σ^*) are the values of (L, σ) evaluated at the first stable equilibrium. We use the *reducing transformation* from (X, Y) to (q, p) by means of the rotation angle Ψ:

$$X = p\cos\Psi + q\sin\Psi \quad \text{and} \quad Y = -p\sin\Psi + q\cos\Psi ,$$

where Ψ is solution of $(a - c)\sin 2\Psi + 2b\cos 2\Psi = 0$.

As a consequence, we find the new Hamiltonian formulation

$$\mathcal{H}^*(p, q) = \frac{1}{2}\left[A\,p^2 + C\,q^2\right] ,$$

with $A = a\cos 2\Psi - 2b\sin\Psi\cos\Psi + c\sin 2\Psi$ and $C = a\sin 2\Psi + 2b\sin\Psi\cos\Psi + c\cos 2\Psi$.

A last scaling canonical transformation of the form $p = \alpha\,p'$ and $q = \frac{1}{\alpha}q'$ obtained by solving the following equation $A\,\alpha^2 = \dfrac{C}{\alpha^2}$ allows us to write the new Hamiltonian as

$$\mathcal{H}(J, \phi, \Lambda) = \sqrt{AC}\,J, \quad \text{where} \quad p' = \sqrt{2J}\cos\phi \quad \text{and} \quad q' = \sqrt{2J}\cos\phi$$

J and ϕ are the corresponding *action-angle* variables.

Subsequently, we find the resonant fundamental frequency ν_f at equilibrium.

$$\nu_f = \frac{\partial\mathcal{H}}{\partial J} = \sqrt{AC} .$$

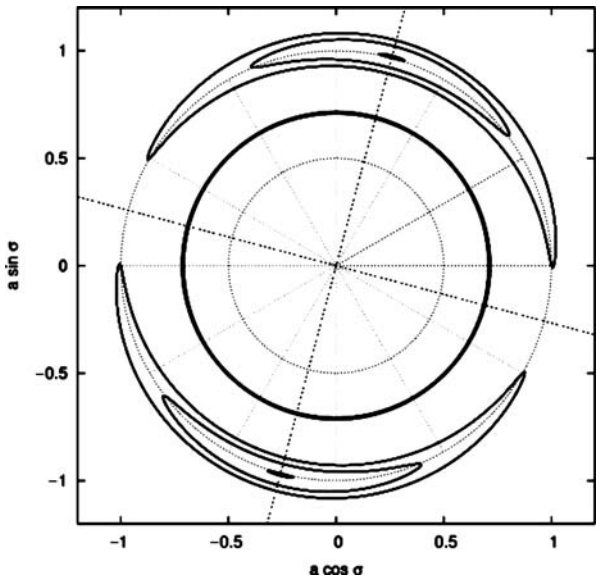

Fig. 27 The resonant phase space in the case of the Earth

Numerical computation for the Earth leads to the following value $\nu_f = 7.674 \times 10^{-3}$/days, that is a period of 818.7 days.

13.6 Width of the Resonance

By a similar approach, we can easily estimate the width of the resonant zone; we take the Hamiltonian level curve corresponding to one of the unstable equilibria L_u and σ_u

$$\mathcal{H}(L_u, \sigma_u) = H_u = -\frac{\mu^2}{2L^2} - \dot{\theta} L + \frac{1}{L^6} [\alpha_1 \cos 2\sigma + \alpha_2 \sin 2\sigma] ,$$

and we find the maxima and minima of this "banana curve," corresponding to the values of σ at the stable equilibria; by a quadratic approximation about L_u, we obtain the width of the banana at the stable points, i.e., the width Δ of the resonant zone. It can be approached by

$$\Delta = \sqrt{\frac{\gamma^2 + 8\delta\beta}{\beta^2}} , \quad \delta = \frac{\alpha_1}{L_u^6 \cos 2\sigma_u} , \quad \beta = -\frac{3}{2}\frac{\mu^2}{L_u^4} , \quad \gamma = \frac{\mu^2}{L_u^3} - \dot{\theta} . \qquad (38)$$

The numerical value of the width of the geostationary resonant zone is of the order of 69 km.

References

1. Beaugé, C.: Celest. Mech. Dynam. Astron. **60**, 225 (1994) 15
2. Borderies, N., Goldreich, P.: Celest. Mech. Dynam. Astron. **32**, 127 (1984) 33, 34
3. Celletti, A., Falcolini, C.: Celest. Mech. Dynam. Astron. **78**, 227 (2000) 53
4. Celletti, A., Chierchia, L.: Celest. Mech. Dynam. Astron. **76**, 229 (2000) 53
5. Deprit, A., Am. J. Phys. **35**, 424 (1967) 45, 47
6. D'Hoedt, S., Lemaitre, A.: Celest. Mech. Dynam. Astron. **89**, 267 (2004) 54
7. D'Hoedt, S., Lemaitre, A.: In: Kurtz, D.W. (ed.) Transits of Venus: New Views of the Solar System and Galaxy: IAU Colloquium 196, pp. 263–270. Cambridge University Press, Cambridge (2006) 54
8. Dufey, J., Lemaitre, A. Rambaux, N.: Celest. Mech. Dynam. Astron. **101**, 141 (2008) 55
9. Henrard, J., Lemaitre, A.: Celest. Mech. Dynam. Astron. **30**, 197 (1983) 14, 29
10. Henrard, J., Lemaitre, A.: Celest. Mech. Dynam. Astron. **39**, 213 (1986) 17
11. Henrard, J.: Dynam. Report. **2**, 117 (1993) 3
12. Henrard, J., Schwanen, G.: Celest. Mech. Dynam. Astron. **89**, 181 (2004) 50, 51, 54
13. Henrard, J.: Celest. Mech. Dynam. Astron. **91**, 131 (2005) 54
14. Henrard, J.: Celest. Mech. Dynam. Astron. **93**, 101 (2005) 54
15. Henrard, J.: Icarus **178**, 144 (2005) 54
16. Henrard, J., Lemaitre, A.: Astro. J. **130**, 2415 (2005) 54
17. Henrard, J., In: Souchay, J. (ed.): *Dynamics of Extended Celestial Bodies and Rings*, Lect. Notes Phys. **682**, 159. Springer, Berlin Germany (2006) 54
18. Jancart, S., Lemaitre, A., Istace, A.: Celest. Mech. Dynam. Astron. **84**, 197 (2002) 15
19. Jancart, S., Lemaitre, A.: Celest. Mech. Dynam. Astron. **81**, 75 (2001) 44, 45
20. Kozai, Y.: Astron. J. **67**, 591 (1963) 17
21. Lemaitre, A.: Celest. Mech. Dynam. Astron. **32**, 109 (1984) 34, 44
22. Lemaitre, A.: Celest. Mech. Dynam. Astron. **34**, 329 (1984) 44
23. Lemaitre, A., Henrard, J.: Celest. Mech. Dynam. Astron. **43**, 91 (1988) 44
24. Lemaitre, A., Dubru, P.: Celest. Mech. Dynam. Astron. **52**, 57 (1991) 44
25. Lemaitre, A., D'Hoedt, S., Rambaux, N.: Celest. Mech. Dynam. Astron. **95**, 213 (2006) 55
26. Malhotra, R., Dermott, S.F.: Icarus **85**, 444 (1990) 44
27. Malhotra, R.: Icarus **87**, 249 (1990) 44
28. Moons, M.: Moon Planets **27**, 257 (1982) 50
29. Moons, M., Morbidelli, A.: Celest. Mech. Dynam. Astron. **56**, 273 (1993) 17
30. Moons, M., Morbidelli, A.: Celest. Mech. Dynam. Astron. **57**, 99 (1993) 17
31. Moons, M., Morbidelli, A.: Icarus **114**, 33 (1995) 21
32. Morbidelli, A., Moons, M.: Icarus, **102**, 316 (1993) 21
33. Morbidelli, A., Moons, M.: Icarus, **103**, 99 (1993) 21
34. Morbidelli, A., Moons, M.: Icarus **115**, 60 (1995) 21
35. Morbidelli, A., Henrard, J.: Celest. Mech. Dynam. Astron. **51**, 131 (1991) 19, 20
36. Morbidelli, A., Henrard, J.: Celest. Mech. Dynam. Astron. **51**, 169 (1991) 21
37. Murray, C.D., Dermott, S.F.: Solar System Dynamics. Cambridge University Press, Cambridge, UK (2000) 6, 7
38. Noyelles, B., Lemaitre, A., Vienne, A.: Astron. Astrophys. **478**, 959 (2008) 54
39. Peale, S.J.: Icarus **178**, 4 (2005) 54
40. Peale, S.J.: Icarus **181**, 338 (2006) 54
41. Peale, S.J.: Astron. J. **79**, 722 (1974) 54
42. Rambaux, N., Lemaitre, A., D'Hoedt, S.: Astron. Astrophys. **470**, 741 (2007) 54
43. Valk, S., Lemaitre, A., Anselmo, L.: Adv. Space Res. **41**, 1077 (2007) 58

The Lyapunov Characteristic Exponents and Their Computation

Ch. Skokos

For want of a nail the shoe was lost.
For want of a shoe the horse was lost.
For want of a horse the rider was lost.
For want of a rider the battle was lost.
For want of a battle the kingdom was lost.
And all for the want of a horseshoe nail.
For Want of a Nail (proverbial rhyme)

Abstract We present a survey of the theory of the Lyapunov Characteristic Exponents (LCEs) for dynamical systems, as well as of the numerical techniques developed for the computation of the maximal, of few and of all of them. After some historical notes on the first attempts for the numerical evaluation of LCEs, we discuss in detail the multiplicative ergodic theorem of Oseledec [102], which provides the theoretical basis for the computation of the LCEs. Then, we analyze the algorithm for the computation of the maximal LCE, whose value has been extensively used as an indicator of chaos, and the algorithm of the so-called standard method, developed by Benettin et al. [14], for the computation of many LCEs. We also consider different discrete and continuous methods for computing the LCEs based on the QR or the singular value decomposition techniques. Although we are mainly interested in finite-dimensional conservative systems, i.e., autonomous Hamiltonian systems and symplectic maps, we also briefly refer to the evaluation of LCEs of dissipative systems and time series. The relation of two chaos detection techniques, namely the fast Lyapunov indicator (FLI) and the generalized alignment index (GALI), to the computation of the LCEs is also discussed.

Ch. Skokos (✉)
Astronomie et Systèmes Dynamiques, IMCCE, Observatoire de Paris, 77 avenue Denfert–Rochereau, F-75014 Paris, France; Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Strasse 38, D-01187 Dresden, Germany, hskokos@pks.mpg.de

Skokos, Ch.: *The Lyapunov Characteristic Exponents and Their Computation*. Lect. Notes Phys. **790**, 63–135 (2010)
DOI 10.1007/978-3-642-04458-8_2

1 Introduction

One of the basic information in understanding the behavior of a dynamical system is the knowledge of the spectrum of its *Lyapunov Characteristic Exponents (LCEs)*. The LCEs are asymptotic measures characterizing the average rate of growth (or shrinking) of small perturbations to the solutions of a dynamical system. Their concept was introduced by Lyapunov when studying the stability of nonstationary solutions of ordinary differential equations [96] and has been widely employed in studying dynamical systems since then. The value of the maximal LCE (mLCE) is an indicator of the chaotic or regular nature of orbits, while the whole spectrum of LCEs is related to entropy (Kolmogorov-Sinai entropy) and dimension-like (Lyapunov dimension) quantities that characterize the underlying dynamics.

By *dynamical system* we refer to a physical and/or mathematical system consisting of (a) a set of l real state variables $x_1, x_2 \ldots, x_l$, whose current values define the state of the system, and (b) a well-defined rule from which the evolution of the state with respect to an independent real variable (which is usually referred as the time t) can be derived. We refer to the number l of state variables as the *dimension* of the system and denote a state using the vector $\mathbf{x} = (x_1, x_2 \ldots, x_l)$, or the matrix $\mathbf{x} = [\, x_1 \; x_2 \ldots x_l \,]^{\mathrm{T}}$ notation, where $(^{\mathrm{T}})$ denotes the transpose matrix. A particular state $\mathbf{x}$ corresponds to a point in an l-dimensional space $\mathcal{S}$, the so-called *phase space* of the system, while a set of states $\mathbf{x}(t)$ parameterized by t is referred as an *orbit* of the dynamical system.

Dynamical systems come in essentially two types:

1. *Continuous dynamical systems* described by differential equations of the form

$$\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t),$$

   with dot denoting derivative with respect to a continuous time t and $\mathbf{f}$ being a set of l functions $f_1, f_2 \ldots, f_l$ known as the *vector field*.
2. *Discrete dynamical systems* or *maps* described by difference equations of the form

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n),$$

   with $\mathbf{f}$ being a set of l functions $f_1, f_2 \ldots, f_l$ and $\mathbf{x}_n$ denoting the vector $\mathbf{x}$ at a discrete time $t = n$ (integer).

Let us now define the term *chaos*. In the literature there are many definitions. A brief and concise presentation of them can be found, for example, in [90]. We adopt here one of the most famous definitions of chaos due to Devaney [35, p. 50], which is based on the topological approach of the problem.

Definition 1. Let V be a set and $\mathbf{f} : V \to V$ a map on this set. We say that $\mathbf{f}$ is `chaotic` on V if

1. **f** has sensitive dependence on initial conditions.
2. **f** is topologically transitive.
3. periodic points are dense in V.

Let us explain in more detail the hypothesis of this definition.

Definition 2. $\mathbf{f} : V \to V$ has *sensitive dependence on initial conditions* if there exists $\delta > 0$ such that, for any $\mathbf{x} \in V$ and any neighborhood Δ of $\mathbf{x}$, there exist $\mathbf{y} \in \Delta$ and $n \geq 0$, such that $|\mathbf{f}^n(\mathbf{x}) - \mathbf{f}^n(\mathbf{y})| > \delta$, where $\mathbf{f}^n$ denotes n successive applications of **f**.

Practically this definition implies that there exist points arbitrarily close to **x** which eventually separate from **x** by at least δ under iterations of **f**. We point out that not all points near **x** need eventually move away from **x** under iteration, but there must be at least one such point in every neighborhood of **x**.

Definition 3. $\mathbf{f} : V \to V$ is said to be *topologically transitive* if for any pair of open sets $U,\ W \subset V$ there exists $n > 0$ such that $\mathbf{f}^n(U) \cap W \neq \emptyset$.

This definition implies the existence of points which eventually move under iteration from one arbitrarily small neighborhood to any other. Consequently, the dynamical system cannot be decomposed into two disjoint invariant open sets.

From Definition 1 we see that a chaotic system possesses three ingredients: (a) unpredictability because of the sensitive dependence on initial conditions, (b) indecomposability because it cannot be decomposed into noninteracting subsystems due to topological transitivity, and (c) an element of regularity because it has periodic points which are dense.

Usually, in physics and applied sciences, people focus on the first hypothesis of Definition 1 and use the notion of chaos in relation to the sensitive dependence on initial conditions. The most commonly employed method for distinguishing between regular and chaotic motion, which quantifies the sensitive dependence on initial conditions, is the evaluation of the mLCE χ_1. If $\chi_1 > 0$ the orbit is chaotic. This method was initially developed at the late 1970s based on theoretical results obtained at the end of the 1960s.

The concept of the LCEs has been widely presented in the literature from a practical point of view, i.e., the description of particular numerical algorithms for their computation [54, 44, 62, 92, 36]. Of course, there also exist theoretical studies on the LCEs, which are mainly focused on the problem of their existence, starting with the pioneer work of Oseledec [102]. In that paper the Multiplicative Ergodic Theorem (MET), which provided the theoretical basis for the numerical computation of the LCEs, was stated and proved. The MET was the subject of several theoretical studies afterward [108, 114, 76, 141]. A combination of important theoretical and numerical results on LCEs can be found in the seminal papers of Benettin et al. [13, 14], written almost 30 years ago, where an explicit method for the computation of all LCEs was developed.

In the present report we focus our attention both on the theoretical framework of the LCEs and on the numerical techniques developed for their computation. Our

goal is to provide a survey of the basic results on these issues obtained over the last 40 years, after the work of Oseledec [102]. To this end, we present in detail the mathematical theory of the LCEs and discuss its significance without going through tedious mathematical proofs. In our approach, we prefer to present the definitions of various quantities and to state the basic theorems that guarantee the existence of the LCE, citing at the same time the papers where all the related mathematical proofs can be found. We also describe in detail the various numerical techniques developed for the evaluation of the maximal, of few or even of all LCEs, and explain their practical implementation. We do not restrict our presentation to the so-called *standard method* developed by Benettin et al. [14], as it is usually done in the literature (see e.g., [54, 44, 92]), but we include in our study modern techniques for the computation of the LCEs like the discrete and continuous methods based on the singular value decomposition (SVD) and the QR decomposition procedures.

In our analysis we deal with finite-dimensional dynamical systems and in particular with autonomous Hamiltonian systems and symplectic maps defined on a compact manifold, meaning that we exclude cases with escapes in which the motion can go to infinity. We do not consider the rather exceptional cases of completely chaotic systems and of integrable ones, i.e., systems that can be solved explicitly to give their variables as single-valued functions of time, but we consider the most general case of "systems with divided phase space" [30, p. 19] for which *regular*[1] (quasiperiodic) and *chaotic orbits* co-exist. In such systems one sees both regular and chaotic domains. But the regular domains contain a dense set of unstable periodic orbits, which are followed by small chaotic regions. On the other hand, the chaotic domains contain stable periodic orbits that are followed by small islands of stability. Thus, the regular and chaotic domains are intricately mixed. However, there are regions where order is predominant, and other regions where chaos is predominant.

Although in our report the theory of LCEs and the numerical techniques for their evaluation are presented mainly for *conservative systems*, i.e., system that preserve the phase space volume, these techniques are not valid only for such models. For completeness sake, we also briefly discuss at the end of the report the computation of LCEs for *dissipative systems*, for which the phase space volume decreases on average, and for time series.

We tried to make the paper self-consistent by including definitions of the used terminology and brief overviews of all the necessary mathematical notions. In addition, whenever it was considered necessary, some illustrative examples have been added to the text in order to clarify the practical implementation of the presented material. Our aim has been to make this review of use for both the novice and the more experienced practitioner interested in LCEs. To this end, the reader who is interested in reading up on detailed technicalities is provided with numerous signposts to the relevant literature.

---

[1] Regular orbits are often called *ordered orbits* (see, e.g., [30, p. 18]).

Throughout the text bold lowercase letters denote vectors, while matrices are represented, in general, by capital bold letters. We also note that the most frequently used abbreviations in the text are LCE(s), Lyapunov characteristic exponent(s); p-LCE, Lyapunov characteristic exponent of order p; mLCE, maximal Lyapunov characteristic exponent; p-mLCE, maximal Lyapunov characteristic exponent of order p; MET, multiplicative ergodic theorem; SVD, singular value decomposition; PSS, Poincaré surface of section; FLI, fast Lyapunov indicator; GALI, generalized alignment index.

This chapter is organized as follows.

In Sect. 2 we present the basic concepts of Hamiltonian systems and symplectic maps, emphasizing on the evolution of orbits, as well as of deviation vectors about them. In particular, we define the so-called variational equations for Hamiltonian systems and the tangent map for symplectic maps, which govern the time evolution of deviation vectors. We also provide some simple examples of dynamical systems and derive the corresponding set of variational equations and the corresponding tangent map.

Section 3 contains some historical notes on the first attempts for the application of the theoretical results of Oseledec [102] for the actual computation of the LCEs. We recall how the notion of exponential divergence of nearby orbits was eventually quantified by the computation of the mLCE, and we refer to the papers where the mLCE or the spectrum of LCEs were computed for the first time.

The basic theoretical results on the LCEs are presented in Sect. 4 following mainly the milestone papers of Oseledec [102] and Benettin et al. [13, 14]. In Sect. 4.1 the basic definitions and theoretical results of LCEs of various orders are presented. The practical consequences of these results on the computation of the LCEs of order 1 and of order $p > 1$ are discussed in Sects. 4.2 and 4.3, respectively. Then, in Sect. 4.4 the MET of Oseledec [102] is stated in its various forms, while its consequences on the spectrum of LCEs for conservative dynamical systems are discussed in Sect. 4.5.

Section 5 is devoted to the computation of the mLCE χ_1, which is the oldest chaos indicator used in the literature. In Sect. 5.1 the method for the computation of the mLCE is discussed in great detail and the theoretical basis of its evaluation is explained. The corresponding algorithm is presented in Sect. 5.2, while the behavior of χ_1 for regular and chaotic orbits is analyzed in Sect. 5.3.

In Sect. 6 the various methods for the computation of part or of the whole spectrum of LCEs are presented. In particular, in Sect. 6.1 the standard method developed in [119, 14] is presented in great detail, while the corresponding algorithm is given in Sect. 6.2. In Sect. 6.3 the connection of the standard method with the discrete QR decomposition technique is discussed and the corresponding QR algorithm is given, while Sect. 6.4 is devoted to the presentation of other techniques for computing few or all LCEs, which are based on the SVD and QR decomposition algorithms.

In Sect. 7 we briefly refer to various chaos detection techniques based on the analysis of deviation vectors, as well as to a second category of chaos indicators based on the analysis of the time series constructed by the coordinates of the orbit

under consideration. The relation of two chaos indicators, namely the fast Lyapunov indicator (FLI) and the generalized alignment index (GALI), to the computation of the LCEs is also discussed.

Although the main topic of our presentation is the theory and the computation of the LCEs for conservative dynamical systems, in the last section of our report some complementary issues related to other types of dynamical systems are concisely presented. In particular, Sect. 8.1 is devoted to the computation of the LCEs for dissipative systems, while in Sect. 8.2 some basic features on the numerical computation of the LCEs from a time series are presented.

Finally, in Appendix we present some basic elements of the exterior algebra theory in connection to the evaluation of wedge products, which are needed for the computation of the volume elements appearing in the definitions of the various LCEs.

2 Autonomous Hamiltonian Systems and Symplectic Maps

In our study we consider two main types of conservative dynamical systems:

1. Continuous systems corresponding to an *autonomous Hamiltonian system* of N degrees (ND) of freedom having a Hamiltonian function

$$H(q_1, q_2, \ldots, q_N, p_1, p_2, \ldots, p_N) = h = \text{constant}, \tag{1}$$

where q_i and p_i, $i = 1, 2, \ldots, N$ are the generalized coordinates and conjugate momenta, respectively. An orbit in the $l = 2N$-dimensional phase space $\mathcal{S}$ of this system is defined by a vector:

$$\mathbf{x}(t) = (q_1(t), q_2(t), \ldots, q_N(t), p_1(t), p_2(t), \ldots, p_N(t)),$$

with $x_i = q_i$, $x_{i+N} = p_i$, $i = 1, 2, \ldots, N$. The time evolution of this orbit is governed by the Hamilton equations of motion, which in matrix form are given by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = \left[\frac{\partial H}{\partial \mathbf{p}} \; -\frac{\partial H}{\partial \mathbf{q}} \right]^{\mathrm{T}} = \mathbf{J}_{2N} \cdot \mathbf{DH}, \tag{2}$$

with $\mathbf{q} = (q_1(t), q_2(t), \ldots, q_N(t))$, $\mathbf{p} = (p_1(t), p_2(t), \ldots, p_N(t))$, and

$$\mathbf{DH} = \left[\frac{\partial H}{\partial q_1} \; \frac{\partial H}{\partial q_2} \; \cdots \; \frac{\partial H}{\partial q_N} \; \frac{\partial H}{\partial p_1} \; \frac{\partial H}{\partial p_2} \; \cdots \; \frac{\partial H}{\partial p_N} \right]^{\mathrm{T}}.$$

Matrix $\mathbf{J}_{2N}$ has the following block form:

$$\mathbf{J}_{2N} = \begin{bmatrix} \mathbf{0}_N & \mathbf{I}_N \\ -\mathbf{I}_N & \mathbf{0}_N \end{bmatrix},$$

with $\mathbf{I}_N$ being the $N \times N$ identity matrix and $\mathbf{0}_N$ being the $N \times N$ matrix with all its elements equal to zero. The solution of (2) is formally written with respect to the induced flow $\Phi^t : \mathcal{S} \to \mathcal{S}$ as

$$\mathbf{x}(t) = \Phi^t\left(\mathbf{x}(0)\right). \tag{3}$$

2. Symplectic maps of $l = 2N$ dimensions having the form

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n). \tag{4}$$

A *symplectic map* is an area-preserving map whose *Jacobian matrix*

$$\mathbf{M} = \mathbf{Df}(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_{2N}} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_{2N}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_{2N}}{\partial x_1} & \frac{\partial f_{2N}}{\partial x_2} & \cdots & \frac{\partial f_{2N}}{\partial x_{2N}} \end{bmatrix},$$

satisfies

$$\mathbf{M}^{\mathrm{T}} \cdot \mathbf{J}_{2N} \cdot \mathbf{M} = \mathbf{J}_{2N}. \tag{5}$$

The state of the system at the discrete time $t = n$ is given by

$$\mathbf{x}_n = \Phi^n\left(\mathbf{x}_0\right) = \left(\mathbf{f}\right)^n\left(\mathbf{x}_0\right), \tag{6}$$

where $\left(\mathbf{f}\right)^n\left(\mathbf{x}_0\right) = \mathbf{f}(\mathbf{f}(\cdots \mathbf{f}(\mathbf{x}_0)\cdots))$, n times.

2.1 Variational Equations and Tangent Map

Let us now turn our attention to the (continuous or discrete) time evolution of deviation vectors $\mathbf{w}$ from a given reference orbit of a dynamical system. These vectors evolve on the *tangent space* $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ of $\mathcal{S}$. We denote by $d_{\mathbf{x}}\Phi^t$ the linear mapping which maps the tangent space of $\mathcal{S}$ at point $\mathbf{x}$ onto the tangent space at point $\Phi^t(\mathbf{x})$, and so we have $d_{\mathbf{x}}\Phi^t : \mathcal{T}_{\mathbf{x}}\mathcal{S} \to \mathcal{T}_{\Phi^t(\mathbf{x})}\mathcal{S}$ with

$$\mathbf{w}(t) = d_{\mathbf{x}}\Phi^t\,\mathbf{w}(0), \tag{7}$$

where $\mathbf{w}(0)$, $\mathbf{w}(t)$ are deviation vectors with respect to the given orbit at times $t = 0$ and $t > 0$, respectively.

In the case of the Hamiltonian system (1) an initial deviation vector $\mathbf{w}(0) = (\delta x_1(0), \delta x_2(0), \ldots, \delta x_{2N}(0))$ from the solution $\mathbf{x}(t)$ (3) evolves on the tangent space $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ according to the so-called *variational equations*:

$$\dot{\mathbf{w}} = \mathbf{Df}(\mathbf{x}(t)) \cdot \mathbf{w} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}(t)) \cdot \mathbf{w} = \left[\mathbf{J}_{2N} \cdot \mathbf{D}^2\mathbf{H}(\mathbf{x}(t))\right] \cdot \mathbf{w} =: \mathbf{A}(t) \cdot \mathbf{w}\,, \quad (8)$$

with $\mathbf{D}^2\mathbf{H}(\mathbf{x}(t))$ being the Hessian matrix of Hamiltonian (1) calculated on the reference orbit $\mathbf{x}(t)$ (3), i.e.,

$$\mathbf{D}^2\mathbf{H}(\mathbf{x}(t))_{i,j} = \left.\frac{\partial^2 H}{\partial x_i \partial x_j}\right|_{\Phi^t(\mathbf{x}(0))}\,, \quad i, j = 1, 2, \ldots, 2N.$$

We underline that (8) represents a set of *linear differential equations* with respect to $\mathbf{w}$, having time-dependent coefficients, since matrix $\mathbf{A}(t)$ depends on the particular reference orbit, which is a function of time t. The solution of (8) can be written as

$$\mathbf{w}(t) = \mathbf{Y}(t) \cdot \mathbf{w}(0), \quad (9)$$

where $\mathbf{Y}(t)$ is the so-called *fundamental matrix* of solutions of (8), satisfying the following equation.

$$\dot{\mathbf{Y}}(t) = \mathbf{Df}(\mathbf{x}(t)) \cdot \mathbf{Y}(t) = \mathbf{A}(t) \cdot \mathbf{Y}(t)\,, \quad \text{with } \mathbf{Y}(0) = \mathbf{I}_{2N}. \quad (10)$$

In the case of the symplectic map (4) the evolution of a deviation vector $\mathbf{w}_n$, with respect to a reference orbit $\mathbf{x}_n$, is given by the corresponding *tangent map*:

$$\mathbf{w}_{n+1} = \mathbf{Df}(\mathbf{x}_n) \cdot \mathbf{w}_n = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_n) \cdot \mathbf{w}_n =: \mathbf{M}_n \cdot \mathbf{w}_n. \quad (11)$$

Thus, the evolution of the initial deviation vector $\mathbf{w}_0$ is given by

$$\mathbf{w}_n = \mathbf{M}_{n-1} \cdot \mathbf{M}_{n-2} \cdot \ldots \cdot \mathbf{M}_0 \cdot \mathbf{w}_0 =: \mathbf{Y}_n \cdot \mathbf{w}_0, \quad (12)$$

with $\mathbf{Y}_n$ satisfying the relation

$$\mathbf{Y}_{n+1} = \mathbf{M}_n \cdot \mathbf{Y}_n = \mathbf{Df}(\mathbf{x}_n) \cdot \mathbf{Y}_n, \quad \text{with } \mathbf{Y}_0 = \mathbf{I}_{2N}. \quad (13)$$

2.2 Simple Examples of Dynamical Systems

As representative examples of dynamical systems we consider (a) the well-known 2D Hénon–Heiles system [72], having the Hamiltonian function

$$H_2 = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(x^2 + y^2) + x^2 y - \frac{1}{3}y^3, \tag{14}$$

with equations of motion

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{p}_x \\ \dot{p}_y \end{bmatrix} = \mathbf{J}_4 \cdot \mathbf{DH}_2 = \mathbf{J}_4 \cdot \begin{bmatrix} x + 2xy \\ y + x^2 - y^2 \\ p_x \\ p_y \end{bmatrix} \Rightarrow \begin{cases} \dot{x} &= p_x \\ \dot{y} &= p_y \\ \dot{p}_x &= -x - 2xy \\ \dot{p}_y &= -y - x^2 + y^2 \end{cases}, \tag{15}$$

and (b) the 4-dimensional (4d) symplectic map

$$\begin{aligned} x_{1,n+1} &= x_{1,n} + x_{3,n} \\ x_{2,n+1} &= x_{2,n} + x_{4,n} \\ x_{3,n+1} &= x_{3,n} - \nu \sin(x_{1,n+1}) - \mu[1 - \cos(x_{1,n+1} + x_{2,n+1})] \\ x_{4,n+1} &= x_{4,n} - \kappa \sin(x_{2,n+1}) - \mu[1 - \cos(x_{1,n+1} + x_{2,n+1})] \end{aligned} \quad (\text{mod } 2\pi), \tag{16}$$

with parameters ν, κ, and μ. All variables are given (mod 2π), so $x_{i,n} \in [\pi, \pi)$, for $i = 1, 2, 3, 4$. This map is a variant of Froeschlé's 4d symplectic map [52] and its behavior has been studied in [31, 123]. It is easily seen that its Jacobian matrix satisfies Eq. (5).

2.3 Numerical Integration of Variational Equations

When dealing with Hamiltonian systems the variational equations (8) have to be integrated simultaneously with the Hamilton equations of motion (2). Let us clarify the issue by looking to a specific example. The variational equations of the 2D Hamiltonian (14) are the following:

$$\dot{\mathbf{w}} = \begin{bmatrix} \dot{\delta x} \\ \dot{\delta y} \\ \dot{\delta p}_x \\ \dot{\delta p}_y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 - 2y & -2x & 0 & 0 \\ -2x & -1 + 2y & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \delta x \\ \delta y \\ \delta p_x \\ \delta p_y \end{bmatrix} \Rightarrow$$
$$\begin{cases} \dot{\delta x} &= \delta p_x \\ \dot{\delta y} &= \delta p_y \\ \dot{\delta p}_x &= (-1 - 2y)\delta x + (-2x)\delta y \\ \dot{\delta p}_y &= (-2x)\delta x + (-1 + 2y)\delta y \end{cases}. \tag{17}$$

This system of differential equations is linear with respect to δx, δy, δp_x, δp_y, but it cannot be integrated independently of system (15) since the x and y variables appear explicitly in it. Thus, if we want to follow the time evolution of an initial deviation vector $\mathbf{w}(0)$ with respect to a reference orbit with initial condition $\mathbf{x}(0)$, we are obliged to integrate simultaneously the whole set of differential equations (15) and (17).

A numerical scheme for integrating the variational equations (8), which exploits their linearity and is particularly useful when we need to evolve more than one deviation vectors is the following. Solving the Hamilton equations of motion (2) by any numerical integration scheme we obtain the time evolution of the reference orbit (3). In practice this means that we know the values $\mathbf{x}(t_i)$ for $t_i = i\,\Delta t$, $i = 0, 1, 2, \ldots$, where Δt is the integration time step. Inserting this numerically known solution to the variational equations (8) we end up with a linear system of differential equations with constant coefficients for every time interval $[t_i, t_i + \Delta t)$, which can be solved explicitly.

For example, in the particular case of Hamiltonian (14), the system of variational equations (17) becomes

$$\begin{array}{l} \dot{\delta x} = \delta p_x \\ \dot{\delta y} = \delta p_y \\ \dot{\delta p}_x = [-1 - 2y(t_i)]\,\delta x + [-2x(t_i)]\,\delta y \\ \dot{\delta p}_y = [-2x(t_i)]\,\delta x + [-1 + 2y(t_i)]\,\delta y \end{array} \quad , \quad \text{for } t \in [t_i, t_i + \Delta t), \tag{18}$$

which is a linear system of differential equations with constant coefficients and thus, easily solved. In particular, (18) can by considered as the Hamilton equations of motion corresponding to the Hamiltonian function:

$$H_V(\delta x, \delta y, \delta p_x, \delta p_y) = \\ \frac{1}{2}\left(\delta p_x^2 + \delta p_y^2\right) + \frac{1}{2}\left\{[1 + 2y(t_i)]\,\delta x^2 + [1 - 2y(t_i)]\,\delta y^2 + 2\,[2x(t_i)]\,\delta x \delta y\right\}. \tag{19}$$

The Hamiltonian formalism (19) of the variational equations (18) is a specific example of a more general result. In the case of the usual Hamiltonian function

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\sum_{i=1}^{N} p_i^2 + V(\mathbf{q}), \tag{20}$$

with $V(\mathbf{q})$ being the potential function, the variational equations (8) for the time interval $[t_i, t_i + \Delta t)$ take the form (see, e.g., [12])

$$\dot{\mathbf{w}} = \begin{bmatrix} \dot{\boldsymbol{\delta}}\mathbf{q} \\ \dot{\boldsymbol{\delta}}\mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_N & \mathbf{I}_N \\ -\mathbf{D}^2\mathbf{V}(\mathbf{q}(t_i)) & \mathbf{0}_N \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\delta}\mathbf{q} \\ \boldsymbol{\delta}\mathbf{p} \end{bmatrix}$$

with $\boldsymbol{\delta}\mathbf{q} = (\delta q_1(t), \delta q_2(t), \ldots, \delta q_N(t))$, $\boldsymbol{\delta}\mathbf{p} = (\delta p_1(t), \delta p_2(t) \ldots, \delta p_N(t))$, and

$$\mathbf{D}^2\mathbf{V}(\mathbf{q}(t_i))_{jk} = \left.\frac{\partial^2 V(\mathbf{q})}{\partial q_j \partial q_k}\right|_{\mathbf{q}(t_i)} \quad , \quad j, k = 1, 2, \ldots, N.$$

Thus, the tangent dynamics of (20) is represented by the Hamiltonian function (see, e.g., [105])

$$H_V(\boldsymbol{\delta q}, \boldsymbol{\delta p}) = \frac{1}{2}\sum_{j=1}^{N} \delta p_i^2 + \frac{1}{2}\sum_{j,k}^{N} \mathbf{D}^2\mathbf{V}(\mathbf{q}(t_i))_{jk}\delta q_j \delta q_k .$$

2.4 Tangent Dynamics of Symplectic Maps

In the case of symplectic maps, the dynamics on the tangent space, which is described by the tangent map (11), cannot be considered separately from the phase space dynamics determined by the map (4) itself. This is because the tangent map depends explicitly on the reference orbit $\mathbf{x}_n$.

For example, the tangent map of the 4d map (16) is

$$\begin{aligned}
\delta x_{1,n+1} &= \delta x_{1,n} + \delta x_{3,n} \\
\delta x_{2,n+1} &= \delta x_{2,n} + \delta x_{4,n} \\
\delta x_{3,n+1} &= a_n \delta x_{1,n} + b_n \delta x_{2,n} + (1 + a_n)\delta x_{3,n} + b_n \delta x_{4,n} \\
\delta x_{4,n+1} &= b_n \delta x_{1,n} + c_n \delta x_{2,n} + b_n \delta x_{3,n} + (1 + c_n)\delta x_{4,n}
\end{aligned} , \tag{21}$$

with

$$\begin{aligned}
a_n &= -\nu \cos(x_{1,n+1}) - \mu \sin(x_{1,n+1} + x_{2,n+1}) \\
b_n &= -\mu \sin(x_{1,n+1} + x_{2,n+1}) \\
c_n &= -\kappa \cos(x_{2,n+1}) - \mu \sin(x_{1,n+1} + x_{2,n+1})
\end{aligned} ,$$

which explicitly depend on $x_{1,n}$, $x_{2,n}$, $x_{3,n}$, $x_{4,n}$. Thus, the evolution of a deviation vector requires the simultaneous iteration of both the map (16) and the tangent map (21).

3 Historical Introduction: The Early Days of LCEs

Prior to the discussion of the theory of the LCEs and the presentation of the various algorithms for their computation, it would be interesting to go back in time and see how the notion of LCEs, as well as the nowadays taken-for-granted techniques for evaluating them, were formed.

The LCEs are asymptotic measures characterizing the average rate of growth (or shrinking) of small perturbations to the orbits of a dynamical system, and their concept was introduced by Lyapunov [96]. Since then they have been extensively used for studying dynamical systems. As it has already been mentioned, one of the basic features of chaos is the sensitive dependence on initial conditions and the LCEs provide quantitative measures of response sensitivity of a dynamical system to small changes in initial conditions. For a chaotic orbit at least one LCE is positive, implying exponential divergence of nearby orbits, while in the case of regular orbits all LCEs are zero. Therefore, the presence of positive LCEs is a signature of chaotic

behavior. Usually the computation of only the mLCE χ_1 is sufficient for determining the nature of an orbit, because $\chi_1 > 0$ guarantees that the orbit is chaotic.

Characterization of the chaoticity of an orbit in terms of the divergence of nearby orbits was introduced by Hénon and Heiles [72] and further used by several authors (e.g., [48, 51, 52, 131, 22, 21]). In these studies two initial points were chosen very close to each other, having phase space distance of about $10^{-7} - 10^{-6}$, and were evolved in time. If the two initial points were located in a region of regular motion their distance increased approximately linearly with time, while if they were belonging to a chaotic region the distance exhibited an exponential increase in time (Fig. 1).

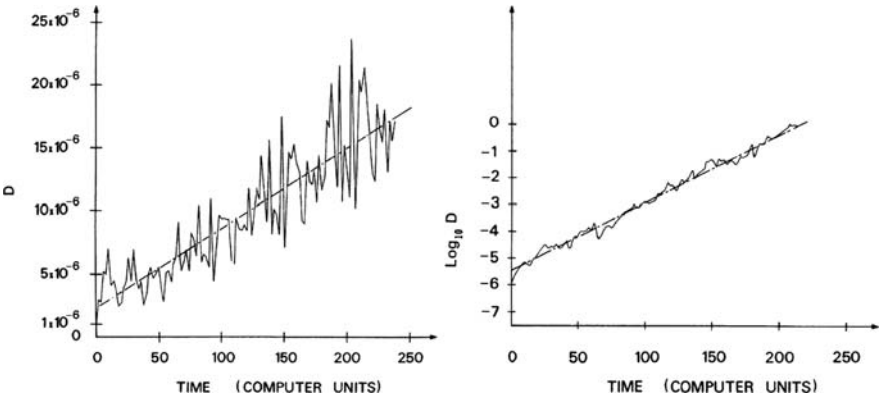

Fig. 1 Typical behavior of the time evolution of the distance D between two initially close orbits in the case of regular and chaotic orbits. The particular results are obtained for a 2D Hamiltonian system describing a Toda lattice of two particles with unequal masses (see [22] for more details). The initial Euclidian distance of the two orbits in the 4-dimensional phase space is $D_0 = 10^{-6}$. D exhibits a linear (on the average) growth when the two orbits are initially located in a region of regular motion (left panel), while it grows exponentially in the case of chaotic orbits (right panel). The big difference in the values of D between the two cases is evident since the two panels have the same horizontal (time) axis but different vertical ones. In particular, the vertical axis is linear in the left panel and logarithmic in the right panel (after [22])

Although the theory of LCEs was applied to characterize chaotic motion by Oseledec [102], quite some time passed until the connection between LCEs and exponential divergence was made clear [10, 106]. It is worth mentioning that Casartelli et al. [21] defined a quantity, which they called "stochastic parameter," in order to quantify the exponential divergence of nearby orbits, which was realized afterward in [10] to be an estimator of the mLCE for $t \to \infty$.

So, the mLCE χ_1 was estimated for the first time in [10], as the limit for $t \to \infty$ of an appropriate quantity $X_1(t)$, which was obtained from the evolution of the phase space distance of two initially close orbits. In this paper some nowadays well-established properties of $X_1(t)$ were discussed, like for example, the fact that $X_1(t)$ tends to zero in the case of regular orbits following a power law $\propto t^{-1}$, while it tends to nonzero values in the case of chaotic orbits (Fig. 2). The same algorithm

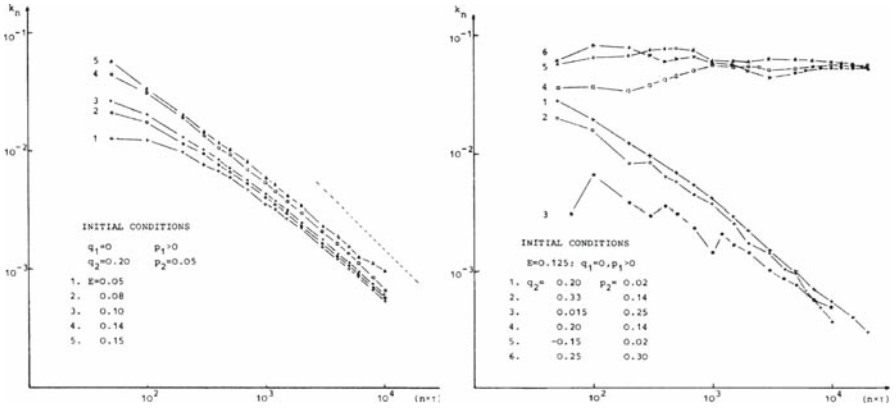

Fig. 2 Evolution of $X_1(t)$ (denoted as k_n) with respect to time t (denoted by $n \times \tau$) in log–log scale for several orbits of the Hénon–Heiles system (14). In the left panel $X_1(t)$ is computed for five different regular orbits at different energies H_2 (denoted as E) and it tends to zero following a power law $\propto t^{-1}$. A dashed straight line corresponding to a function proportional to t^{-1} is also plotted. In the right panel the evolution of $X_1(t)$ is plotted for three regular orbits (curves 1–3) and three chaotic ones (curves 4–6) for $H_2 = 0.125$. Note that the values of the initial conditions given in the two panels correspond to $q_1 = x$, $q_2 = y$, $p_1 = p_x$, $p_2 = p_y$ in (14) (after [10])

was immediately applied for the computation of the mLCE of a dissipative system, namely the Lorenz system [99].

The next improvement of the computational algorithm for the evaluation of the mLCE was introduced in [34], where the variational equations were used for the time evolution of deviation vectors instead of the previous approach of the simultaneous integration of two initially close orbits. This more direct approach constituted a significant improvement for the computation of the mLCE since it allowed the use of larger integration steps, diminishing the real computational time and also eliminated the problem of choosing a suitable initial distance between the nearby orbits.

In [11] a theorem was formulated, which led directly to the development of a numerical technique for the computation of some or even of all LCEs, based on the time evolution of more than one deviation vectors, which are kept linearly independent through a Gram-Schmidt orthonormalization procedure (see also [9]). This method was explained in more detail in [119], where it was applied to the study of the Lorenz system, and was also presented in [12], where it was applied to the study of an ND Hamiltonian system with N varying from 2 to 10.

The theoretical framework, as well as the numerical method for the computation of the maximal, some or even all LCEs were given in the seminal papers of Benettin et al. [13, 14]. In [14] the complete set of LCEs was calculated for several different Hamiltonian systems, including 4- and 6-dimensional maps. In Fig. 3 we show the results of [14] concerning the 3D Hamiltonian system of [34]. The importance of the papers of Benettin et al. [13, 14] is reflected by the fact that almost all methods for the computation of the LCEs are more or less based on them. Immediately the ideas

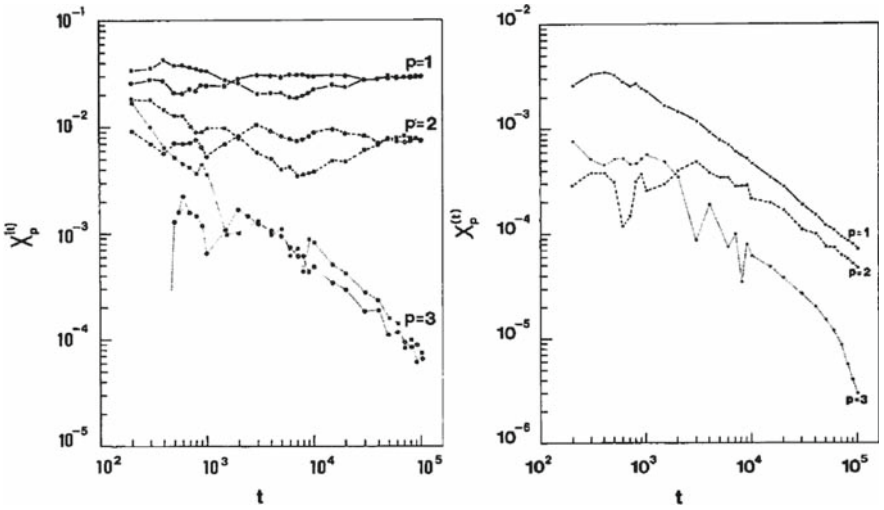

Fig. 3 Time evolution of appropriate quantities denoted by $X_p^{(t)}$, $p = 1, 2, 3$, having, respectively, as limits for $t \to \infty$ the first three LCEs χ_1, χ_2, χ_3, for two chaotic orbits (left panel) and one regular orbit (right panel) of the 3D Hamiltonian system initially studied in [34] (see [14] for more details). In both panels $X_3^{(t)}$ tends to zero implying that $\chi_3 = 0$. This is due to the fact that Hamiltonian systems have at least one vanishing LCE, namely the one corresponding to the direction along the flow (this property is explained in Sect. 4.5). On the other hand, χ_1 and χ_2 seem to get nonzero values (with $\chi_1 > \chi_2$) for chaotic orbits, while they appear to vanish for regular orbits (after [14])

presented in [13, 14] were used for the computation of the LCEs for a variety of dynamical systems like infinite-dimensional systems described by delay differential equations [46], dissipative systems [44], conservative systems related to Celestial Mechanics problems [53, 55], as well as for the determination of the LCEs from a time series [144, 118].

4 Lyapunov Characteristic Exponents: Theoretical Treatment

In this section we define the LCEs of various orders presenting also the basic theorems which guarantee their existence and provide the theoretical background for their numerical evaluation. In our presentation we basically follow the fundamental papers of Oseledec [102] and of Benettin et al. [13] where all the theoretical results of the current section are explicitly proved.

We consider a continuous or discrete dynamical system defined on a differentiable manifold $\mathcal{S}$. Let $\Phi^t(\mathbf{x})$ denote the state at time t of the system which at time $t = 0$ was at $\mathbf{x}$ (see (3) and (6) for the continuous and discrete case respectively). For the action of Φ^t over two successive time intervals t and s we have the following composition law:

$$\Phi^{t+s} = \Phi^t \circ \Phi^s.$$

The tangent space at $\mathbf{x}$ is mapped onto the tangent space at $\Phi^t(\mathbf{x})$ by the differential $d_{\mathbf{x}}\Phi^t$ according to (7). The action of $\Phi^t(\mathbf{x})$ is given by (9) for continuous systems and by (12) for discrete ones. Thus, the action of $d_{\mathbf{x}}\Phi^t$ on a particular initial deviation vector $\mathbf{w}$ of the tangent space is given by the multiplication of matrix $\mathbf{Y}(t)$ for continuous systems or $\mathbf{Y}_n$ for discrete systems with vector $\mathbf{w}$. From (9) and (12) we see that the action of $d_{\mathbf{x}}\Phi^t$ over two successive time intervals t and s satisfies the composition law:

$$d_{\mathbf{x}}\Phi^{t+s} = d_{\Phi^s(\mathbf{x})}\Phi^t \circ d_{\mathbf{x}}\Phi^s. \tag{22}$$

This equation can be written in the form

$$\mathbf{R}(t+s, \mathbf{x}) = \mathbf{R}(t, \Phi^s(\mathbf{x})) \cdot \mathbf{R}(s, \mathbf{x}), \tag{23}$$

where $\mathbf{R}(t, \mathbf{x})$ is the matrix corresponding to $d_{\mathbf{x}}\Phi^t$. We note that since $\mathbf{Y}(0) = \mathbf{Y}_0 = \mathbf{I}_{2N}$ we get $d_{\mathbf{x}}\Phi^0\mathbf{w} = \mathbf{w}$ and $\mathbf{R}(0, \mathbf{x}) = \mathbf{I}_{2N}$. A function $\mathbf{R}(t, \mathbf{x})$ satisfying relation (23) is called a *multiplicative cocycle* with respect to the dynamical system Φ^t.

Let $\mathcal{S}$ be a measure space with a normalized measure μ such that

$$\mu(\mathcal{S}) = 1 \ , \ \ \mu\left(\Phi^t\mathcal{A}\right) = \mu(\mathcal{A}) \tag{24}$$

for $\mathcal{A} \subset \mathcal{S}$. Suppose also that a smooth Riemannian metric $\| \ \|$ is defined on $\mathcal{S}$. We consider the multiplicative cocycle $\mathbf{R}(t, \mathbf{x})$ corresponding to $d_{\mathbf{x}}\Phi^t$ and we are interested in its asymptotic behavior for $t \to \pm\infty$. Since, as mentioned by Oseledec [102], the case $t \to +\infty$ is analogous to the case $t \to -\infty$, we restrict our treatment to the case $t \to +\infty$, where time is increasing. In order to clarify what we are practically interested in let us consider a nonzero vector $\mathbf{w}$ of the tangent space $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ at $\mathbf{x}$. Then the quantity

$$\lambda_t(\mathbf{x}) = \frac{\|d_{\mathbf{x}}\Phi^t\mathbf{w}\|}{\|\mathbf{w}\|}$$

is called the *coefficient of expansion in the direction of* $\mathbf{w}$. If

$$\limsup_{t\to\infty} \frac{1}{t} \ln \lambda_t(\mathbf{x}) > 0$$

we say that exponential diverge occurs in the direction of $\mathbf{w}$. Of course the basic question we have to answer is whether the *characteristic exponent* (also called *characteristic exponent of order* 1)

$$\lim_{t\to\infty} \frac{1}{t} \ln \lambda_t(\mathbf{x})$$

exists.

We will answer this question in a more general framework without restricting ourselves to multiplicative cocycles. So, the results presented in the following Sect. 4.1 are valid for a general class of matrix functions, a subclass of which contains the multiplicative cocycles which are of more practical interest to us, since they describe the time evolution of deviation vectors for the dynamical systems we study.

4.1 Definitions and Basic Theorems

Let $\mathbf{A}_t$ be an $n \times n$ matrix function defined either on the whole real axis or on the set of integers, such that $\mathbf{A}_0 = \mathbf{I}_n$, for each time t the value of function $\mathbf{A}_t$ is a nonsingular matrix and $\|\mathbf{A}_t\|$ the usual 2-norm of $\mathbf{A}_t$.[2] In particular, we consider only matrices $\mathbf{A}_t$ satisfying

$$\max\left\{\|\mathbf{A}_t\|, \|\mathbf{A}_t^{-1}\|\right\} \leq e^{ct} \tag{25}$$

with $c > 0$ a suitable constant.

Definition 4. Considering a matrix function $\boldsymbol{A}_t$ as above and a nonzero vector $\mathbf{w}$ of the Euclidian space $\mathbb{R}^n$ the quantity

$$\chi(\mathbf{A}_t, \mathbf{w}) = \limsup_{t\to\infty} \frac{1}{t} \ln \|\mathbf{A}_t \mathbf{w}\| \tag{26}$$

is called the *1-dimensional Lyapunov Characteristic Exponent* or the *Lyapunov Characteristic Exponent of order 1 (1-LCE)* of $\boldsymbol{A}_t$ with respect to vector $\mathbf{w}$.

For simplicity we will usually refer to 1-LCEs as LCEs.

We note that the value of the norm $\|\mathbf{w}\|$ does not influence the value of $\chi(\mathbf{A}_t, \mathbf{w})$. For example, considering a vector $\beta\mathbf{w}$, with $\beta \in \mathbb{R}$ a nonzero constant, instead of $\mathbf{w}$ in Definition 4, we get the extra term $\ln|\beta|/t$ (with $|\ |$ denoting the absolute value) in (26) whose limiting value for $t \to \infty$ is zero and thus does not change the value of $\chi(\mathbf{A}_t, \mathbf{w})$. More importantly, the value of the LCE is independent of the norm appearing in (26). This can be easily seen as follows: Let us consider a second norm $\|\ \|'$ satisfying the inequality

$$\beta_1\|\mathbf{w}\| \leq \|\mathbf{w}\|' \leq \beta_2\|\mathbf{w}\|$$

---

[2] The 2-norm $\|\mathbf{A}\|$ of an $n\times n$ matrix $\mathbf{A}$ is induced by the 2-norm of vectors, i.e., the usual Euclidean norm $\|\mathbf{x}\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$, by

$$\|\mathbf{A}\| = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|},$$

and is equal to the largest eigenvalue of matrix $\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}$.

for some positive real numbers β_1, β_2, and for all vectors $\mathbf{w}$. Such norms are called *equivalent* (see, e.g., [73, Sect. 5.4.7]). Then, by the above-mentioned argument it is easily seen that the use of norm $\|\ \|'$ in (26) leaves unchanged the value of $\chi(\mathbf{A}_t, \mathbf{w})$. Since all norms of finite-dimensional vector spaces are equivalent, we conclude that the LCEs do not depend on the chosen norm.

Let $\mathbf{w}_i$, $i = 1, 2, \ldots, p$ be a set of linearly independent vectors in $\mathbb{R}^n$, E^p be the subspace generated by all $\mathbf{w}_i$ and $\mathrm{vol}_p(\mathbf{A}_t, E^p)$ be the volume of the p-parallelogram having as edges the p vectors $\mathbf{A}_t\mathbf{w}_i$. This volume is computed as the norm of the wedge product of these vectors (see Appendix for the definition of the wedge product and the actual evaluation of the volume)

$$\mathrm{vol}_p(\mathbf{A}_t, E^p) = \|\mathbf{A}_t\mathbf{w}_1 \wedge \mathbf{A}_t\mathbf{w}_2 \wedge \cdots \wedge \mathbf{A}_t\mathbf{w}_p\|.$$

Let also $\mathrm{vol}_p(\mathbf{A}_0, E^p)$ be the volume of the initial p-parallelogram defined by all $\mathbf{w}_i$, since $\mathbf{A}_0$ is the identity matrix. Then the quantity

$$\lambda_t(E^p) = \frac{\mathrm{vol}_p(\mathbf{A}_t, E^p)}{\mathrm{vol}_p(\mathbf{A}_0, E^p)}$$

is called the *coefficient of expansion in the direction of* E^p and it depends only on E^p and not on the choice of the linearly independent set of vectors. Obviously for an 1-dimensional subspace E^1 the coefficient of expansion is $\|\mathbf{A}_t\mathbf{w}_1\|/\|\mathbf{w}_1\|$. If the limit

$$\lim_{t\to\infty} \frac{1}{t} \ln \lambda_t(E^p)$$

exits it is called the *characteristic exponent of order p in the direction of* E^p.

Definition 5. Considering the linearly independent set $\mathbf{w}_i$, $i = 1, 2, \ldots, p$ and the corresponding subspace E^p of $\mathbb{R}^n$ as above, the *p-dimensional Lyapunov Characteristic Exponent* or the *Lyapunov Characteristic Exponent of order p (p-LCE)* of $\boldsymbol{A}_t$ with respect to subspace E^p is defined as

$$\chi(\mathbf{A}_t, E^p) = \limsup_{t\to\infty} \frac{1}{t} \ln \mathrm{vol}_p(\mathbf{A}_t, E^p). \tag{27}$$

Similarly to the case of the 1-LCE, the value of the initial volume $\mathrm{vol}_p(\mathbf{A}_0, E^p)$, as well as the used norm, do not influence the value of $\chi(\mathbf{A}_t, E^p)$.

From (25) and the Hadamard inequality (see, e.g., [102]), according to which the Euclidean volume of a p-parallelogram does not exceed the product of the lengths of its sides, we conclude that the LCEs of (26) and (27) are finite.

From the definition of the LCE it follows that

$$\chi(\mathbf{A}_t, c_1\mathbf{w}_1 + c_2\mathbf{w}_2) \le \max\{\chi(\mathbf{A}_t, \mathbf{w}_1), \chi(\mathbf{A}_t, \mathbf{w}_2)\}$$

for any two vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$ and $c_1, c_2 \in \mathbb{R}$ with $c_1, c_2 \neq 0$, while the Hadamard inequality implies that if $\mathbf{w}_i$, $i = 1, 2, \ldots, n$ is a basis of $\mathbb{R}^n$ then

$$\sum_{i=1}^{n} \chi(\mathbf{A}_t, \mathbf{w}_i) \geq \limsup_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t|, \tag{28}$$

where $\det \mathbf{A}_t$ is the determinant of matrix $\mathbf{A}_t$.

It can be shown that for any $r \in \mathbb{R}$ the set of vectors $\{\mathbf{w} \in \mathbb{R}^n : \chi(\mathbf{A}_t, \mathbf{w}) \leq r\}$ is a vector subspace of $\mathbb{R}^n$ and that the function $\chi(\mathbf{A}_t, \mathbf{w})$ with $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{w} \neq 0$ takes at most n different values, say

$$\nu_1 > \nu_2 > \cdots > \nu_s \text{ with } 1 \leq s \leq n. \tag{29}$$

For the subspaces

$$L_i = \left\{\mathbf{w} \in \mathbb{R}^n : \chi(\mathbf{A}_t, \mathbf{w}) \leq \nu_i\right\}, \tag{30}$$

we have

$$\mathbb{R}^n = L_1 \supset L_2 \supset \cdots \supset L_s \supset L_{s+1} \stackrel{\text{def}}{=} \{0\}, \tag{31}$$

with $L_{i+1} \neq L_i$ and $\chi(\mathbf{A}_t, \mathbf{w}) = \nu_i$ if and only if $\mathbf{w} \in L_i \setminus L_{i+1}$ for $i = 1, 2, \ldots, s$. So in descending order each LCE "lives" in a space of dimensionality less than that of the preceding exponent. Such a structure of linear spaces with decreasing dimension, each containing the following one, is called a *filtration*.

Definition 6. A basis $\mathbf{w}_i$, $i = 1, 2, \ldots, n$ of $\mathbb{R}^n$ is called normal if $\sum_{i=1}^{n} \chi(\boldsymbol{A}_t, \mathbf{w}_i)$ attains a minimum at this basis. In other words, the basis $\mathbf{w}_i$, is a `normal basis` if

$$\sum_{i=1}^{n} \chi(\mathbf{A}_t, \mathbf{w}_i) \leq \sum_{i=1}^{n} \chi(\mathbf{A}_t, \mathbf{g}_i),$$

where $\mathbf{g}_i$, $i = 1, 2, \ldots, n$ is any other basis of $\mathbb{R}^n$.

A normal basis $\mathbf{w}_i$, $i = 1, 2, \ldots, n$ is not unique but the numbers $\chi(\boldsymbol{A}_t, \mathbf{w}_i)$ depend only on $\boldsymbol{A}_t$ and not on the particular normal basis and are called the LCEs of function $\boldsymbol{A}_t$. By a possible permutation of the vectors of a given normal basis we can always assume that $\chi(\boldsymbol{A}_t, \mathbf{w}_1) \geq \chi(\boldsymbol{A}_t, \mathbf{w}_2) \geq \cdots \geq \chi(\boldsymbol{A}_t, \mathbf{w}_n)$.

Definition 7. Let $\mathbf{w}_i$, $i = 1, 2, \ldots, n$ be a normal basis of $\mathbb{R}^n$ and $\chi_1 \geq \chi_2 \geq \cdots \geq \chi_n$, with $\chi_i \equiv \chi(\boldsymbol{A}_t, \mathbf{w}_i)$, $i = 1, 2, \ldots, n$, the LCEs of these vectors. Assume that value ν_i, $i = 1, 2, \ldots, s$ appears exactly $k_i = k_i(\nu_i) > 0$ times among these numbers. Then k_i is called the `multiplicity` of value ν_i and the collection (ν_i, k_i) $i = 1, 2, \ldots, s$ is called the `spectrum of LCEs`.

In order to clarify the used notation we stress that χ_i, $i = 1, 2, \ldots, n$ are the n (possibly nondistinct) LCEs, satisfying $\chi_1 \geq \chi_2 \geq \cdots \geq \chi_n$, while ν_i, $i = 1, 2, \ldots, s$ represent the s ($1 \leq s \leq n$), different values the LCEs have, with $\nu_1 > \nu_2 > \cdots > \nu_s$.

Definition 8. The matrix function $\boldsymbol{A}_t$ is called `regular` as $t \to \infty$ if for each normal basis $\mathbf{w}_i$, $i = 1, 2, \ldots, n$ it holds that

$$\sum_{i=1}^{n} \chi(\mathbf{A}_t, \mathbf{w}_i) = \liminf_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t|,$$

which, due to (28) leads to

$$\liminf_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t| = \limsup_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t|$$

guaranteeing that the limit

$$\lim_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t|$$

exists, is finite, and is equal to

$$\lim_{t\to\infty} \frac{1}{t} \ln |\det \mathbf{A}_t| = \sum_{i=1}^{n} \chi(\mathbf{A}_t, \mathbf{w}_i) = \sum_{i=1}^{s} k_i \nu_i.$$

We can now state a very important theorem for the LCEs:

Theorem 1. *If the matrix function $\boldsymbol{A}_t$ is regular then the LCEs of all orders are given by (26) and (27) where the* $\limsup_{t\to\infty}$ *is substituted by* $\lim_{t\to\infty}$

$$\chi(\boldsymbol{A}_t, \mathbf{w}) = \lim_{t\to\infty} \frac{1}{t} \ln \|\boldsymbol{A}_t \mathbf{w}\| \tag{32}$$

$$\chi(\boldsymbol{A}_t, E^p) = \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{vol}_p(\boldsymbol{A}_t, E^p). \tag{33}$$

In particular, for any p-dimensional subspace $E^p \subseteq \mathbb{R}^n$ we have

$$\chi(\boldsymbol{A}_t, E^p) = \sum_{j=1}^{p} \chi_{i_j}, \tag{34}$$

with a suitable sequence $1 \leq i_1 \leq i_2 \leq \cdots \leq i_p \leq n$.

The part of the theorem concerning equations (32) and (33) was proved by Oseledec in [102], while (34), although was not explicitly proved in [102], can be considered as a rather easily proven byproduct of the results presented there. Actually, the validity of (34) was shown in [13].

4.2 Computing LCEs of Order 1

Let us now discuss how we can use Theorem 1 for the numerical computation of LCEs, starting with the computation of LCEs of order 1.

As we have already mentioned in (29), the LCE takes at most n different values ν_i , $i = 1, 2, \ldots, s$, $1 \leq s \leq n$. If we could know a priori the sequence (31) of subspaces L_i $i = 1, 2, \ldots, s$ of $\mathbb{R}^n$ we would, in principle, be able to compute the values ν_i of all LCEs. This could be done by taking an initial vector $\mathbf{w}_i \in L_i \setminus L_{i+1}$ and compute

$$\nu_i = \lim_{t \to \infty} \frac{1}{t} \ln \|\mathbf{A}_t \mathbf{w}_i\| \quad , \quad i = 1, 2, \ldots, s. \tag{35}$$

Now apart from $L_1 = \mathbb{R}^n$ all the remaining subspaces L_i, $i = 2, 3, \ldots, s$ have positive codimension $\text{codim}(L_i)$ $(= \dim \mathbb{R}^n - \dim L_i > 0)$ and thus, vanishing Lebesgue measure. Then a random choice of $\mathbf{w} \in \mathbb{R}^n$ would lead to the computation of χ_1 from (35), because, in principle $\mathbf{w}$ will belong to L_1 and not to the subspaces L_i $i = 2, \ldots, s$. Let us consider a simple example in order to clarify this statement.

Suppose that L_1 is the usual 3-dimensional space $\mathbb{R}^3$, $L_2 \subset L_1$ is a particular 2-dimensional plane of $\mathbb{R}^3$, e.g., the plane $z = 0$, $L_3 \subset L_2$ is a particular 1-dimensional line, e.g., the x axis (Fig. 4a) and the corresponding LCEs are $\chi_1 > \chi_2 > \chi_3$ with multiplicities $k_1 = k_2 = k_3 = 1$. For this case we have $\dim L_1 = 3$, $\dim L_2 = 2$, $\dim L_3 = 1$ and $\text{codim}(L_1) = 0$, $\text{codim}(L_2) = 1$, $\text{codim}(L_3) = 2$. Concerning the measures μ of these subspaces of $\mathbb{R}^3$, it is obvious that $\mu(L_2) = \mu(L_3) = 0$, since the measure of a surface or of a line in the 3-dimensional space $\mathbb{R}^3$ is zero.

If we randomly choose a vector $\mathbf{w} \in \mathbb{R}^3$ it will belong to L_1 and not to L_2, i.e., having its z coordinate different from zero and thus, (35) would lead to the computation of the mLCE χ_1. Vector $\mathbf{w}_1$ in Fig. 4(b) represents such a random choice. In order to compute χ_2 from (35) we should choose vector $\mathbf{w}$ not randomly but in a specific way. In particular, it should belong to L_2 but not to L_3, so its z coordinate should be equal to zero. Thus this vector should have the form $\mathbf{w} = (w_1, w_2, 0)$ with $w_1, w_2 \in \mathbb{R}$, $w_2 \neq 0$, like vector $\mathbf{w}_2$ in Fig. 4b. Our choice will become even more specific if we would like to compute χ_3 because in this case $\mathbf{w}$ should be of the form $\mathbf{w} = (w_1, 0, 0) \neq \mathbf{0}$ with $w_1 \in \mathbb{R}$. Vector $\mathbf{w}_3$ of Fig. 4b is a choice of this kind.

From this example it becomes evident that a random choice of vector $\mathbf{w}$ in (35) will lead to the computation of the largest LCE χ_1 with probability one. One more comment concerning the numerical implementation of (35) should be added here.

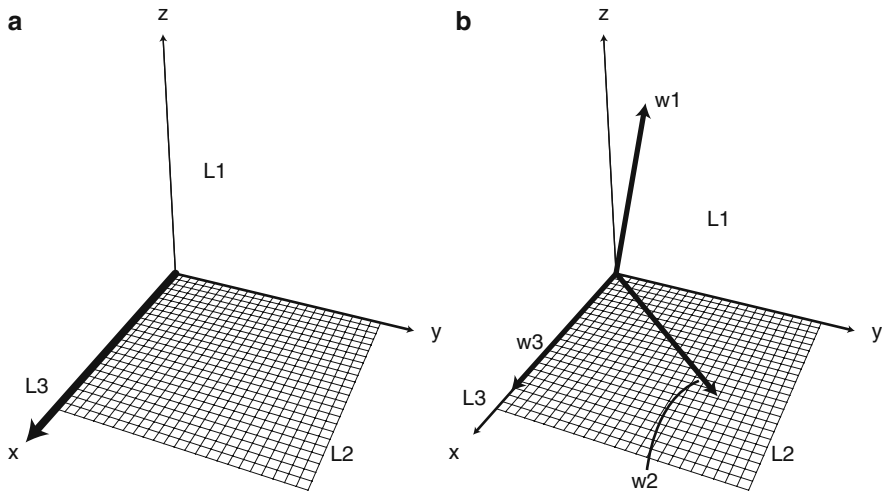

Fig. 4 (**a**) A schematic representation of the sequence of subspaces (31) where L_1 identifies with $\mathbb{R}^3$, $L_2 \subset L_1$ is represented by the xy plane and the x axis is considered as the final subspace $L_3 \subset L_2$. (**b**) A random choice of a vector in $L_1 \equiv \mathbb{R}^3$ will result with probability one to a vector belonging to L_1 and not to L_2, like vector $\mathbf{w}_1$. Vectors $\mathbf{w}_2$, $\mathbf{w}_3$ belonging, respectively, to $L_2 \setminus L_3$ and to L_3 are not random since their coordinates should satisfy certain conditions. In particular, the z coordinate of $\mathbf{w}_2$ should be zero, while both the z and y coordinate of $\mathbf{w}_3$ should vanish. The use of $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ in (35) leads to the computation of χ_1, χ_2, and χ_3, respectively

Even if in some special examples one could happen to know a priori the subspaces L_i $i = 1, 2, \ldots, s$, so that one could choose $\mathbf{w} \in L_i \setminus L_{i+1}$ with $i \neq 1$ then the computational errors would eventually lead to the numerical computation of χ_1. Such an example was presented in [14].

4.3 Computing LCEs of Order $p > 1$

Let us now turn our attention to the computation of p-LCEs with $p > 1$. Equation (34) of Theorem 1 actually tells us that the p-LCE $\chi(\mathbf{A}_t, E^p)$ can take at most $\binom{n}{p}$ distinct values, i.e., as many as all the possible sums of p 1-LCEs out of n are. Now, as the choice of a random vector $\mathbf{w} \in \mathbb{R}^n$, or in other words, of a random 1-dimensional subspace of $\mathbb{R}^n$ produced by $\mathbf{w}$, leads to the computation of the maximal 1-LCE, the random choice of a p-dimensional subspace E^p of $\mathbb{R}^n$, or equivalently the random choice of p linearly independent vectors $\mathbf{w}_i$ $i = 1, 2, \ldots, p$, leads to the computation of the *maximal p-LCE (p-mLCE)* which is equal to the sum of the p largest 1-LCEs

$$\chi(\mathbf{A}_t, E^p) = \sum_{i=1}^{p} \chi_i. \tag{36}$$

This relation was formulated explicitly in [11, 9] and proved in [13] but was implicitly contained in [102]. The practical importance of (36) was also clearly explained in [119]. Benettin et al. [13] gave a more rigorous form to the notion of the random choice of E^p, which is essential for the derivation of (36), by introducing a condition that subspace E^p should satisfy. They named this condition *Condition R* (at random). According to Condition R a p-dimensional space $E^p \subset \mathbb{R}^n$ is chosen at random if for all $j = 2, 3, \ldots, s$ we have

$$\dim(E^p \cap L_j) = \max\left\{0,\, p - \sum_{i=1}^{j-1} k_i\right\}, \tag{37}$$

where L_j belongs to the sequence of subspaces (31) and k_i is the multiplicity of the LCE ν_i (Definition 7).

In order to clarify these issues let us consider again the example presented in Fig. 4, where we have three distinct values for the 1-LCEs $\chi_1 > \chi_2 > \chi_3$ with multiplicities $k_1 = k_2 = k_3 = 1$. In this case the 2-LCE can take one of the three possible values $\chi_1 + \chi_2$, $\chi_2 + \chi_3$, $\chi_1 + \chi_3$, while the 3-LCE takes only one possible value, namely $\chi_1 + \chi_2 + \chi_3$.

The computation of the 2-LCE requires the choice of two linearly independent vectors $\mathbf{w}_1$, $\mathbf{w}_2$ and the application of (33). The two vectors $\mathbf{w}_1$, $\mathbf{w}_2$ define a 2-dimensional plane E^2 in $\mathbb{R}^3$ and $\chi(\mathbf{A}_t, E^2)$ practically measures the time rate of the coefficient of expansion of the surface of the parallelogram having as edges the vectors $\mathbf{A}_t\mathbf{w}_1$, $\mathbf{A}_t\mathbf{w}_2$.

By choosing the two vectors $\mathbf{w}_1$, $\mathbf{w}_2$ randomly we define a random plane E^2 in $\mathbb{R}^3$ which intersects the subspace L_2 (plane xy) along a line, i.e., $\dim(E^2 \cap L_2) = 1$ and the subspace L_3 (x axis) at a point, i.e., $\dim(E^2 \cap L_3) = 0$ (Fig. 5a). This random choice of plane E^2 satisfies Condition R (37) and thus, (33) leads to the computation of the 2-mLCE, namely $\chi_1 + \chi_2$. This result can be also understood in the following way. Plane E^2 in Fig. 5a can be considered to be spanned by two vectors $\mathbf{w}_1$, $\mathbf{w}_2$ such that $\mathbf{w}_1 \in L_1$ but not in its subspace L_2 and $\mathbf{w}_2 \in L_2$ but not in its subspace L_3. Then the expansion of $\mathbf{w}_1 \in L_1 \setminus L_2$ is determined by the LCE χ_1 and the expansion of $\mathbf{w}_2 \in L_2 \setminus L_3$ by the LCE χ_2. These 1-dimensional expansion rates result to an expansion rate equal to $\chi_1 + \chi_2$ for the surface defined by the two vectors.

Other more carefully designed choices of the E^2 subspace lead to the computation of the other possible values of the 2-LCE. If for example $\mathbf{w}_1 \in L_2 \setminus L_3$ and $\mathbf{w}_2 \in L_3$ (Fig. 5b) we have $E^2 = L_2$ with $\dim(E^2 \cap L_2) = 2$ and $\dim(E^2 \cap L_3) = 1$. In this case the expansion of $\mathbf{w}_1$ is determined by the LCE χ_2 and of $\mathbf{w}_2$ by χ_3, and so the computed 2-LCE is $\chi_2 + \chi_3$. Finally, a choice of E^2 of the form presented in Fig. 5c leads to the computation of $\chi_1 + \chi_3$. In this case the plane E^2 is defined by $\mathbf{w}_1 \in L_1 \setminus L_2$ and $\mathbf{w}_2 \in L_3$ and intersects subspaces L_2 and L_3 along the line corresponding to L_3, i.e., $\dim(E^2 \cap L_2) = 1$ and $\dim(E^2 \cap L_3) = 1$. It can be easily checked that for the last two choices of E^2 (Fig. 5b, c) for which the computed 2-LCE does not take its maximal possible value, Condition R (37) is not satisfied,

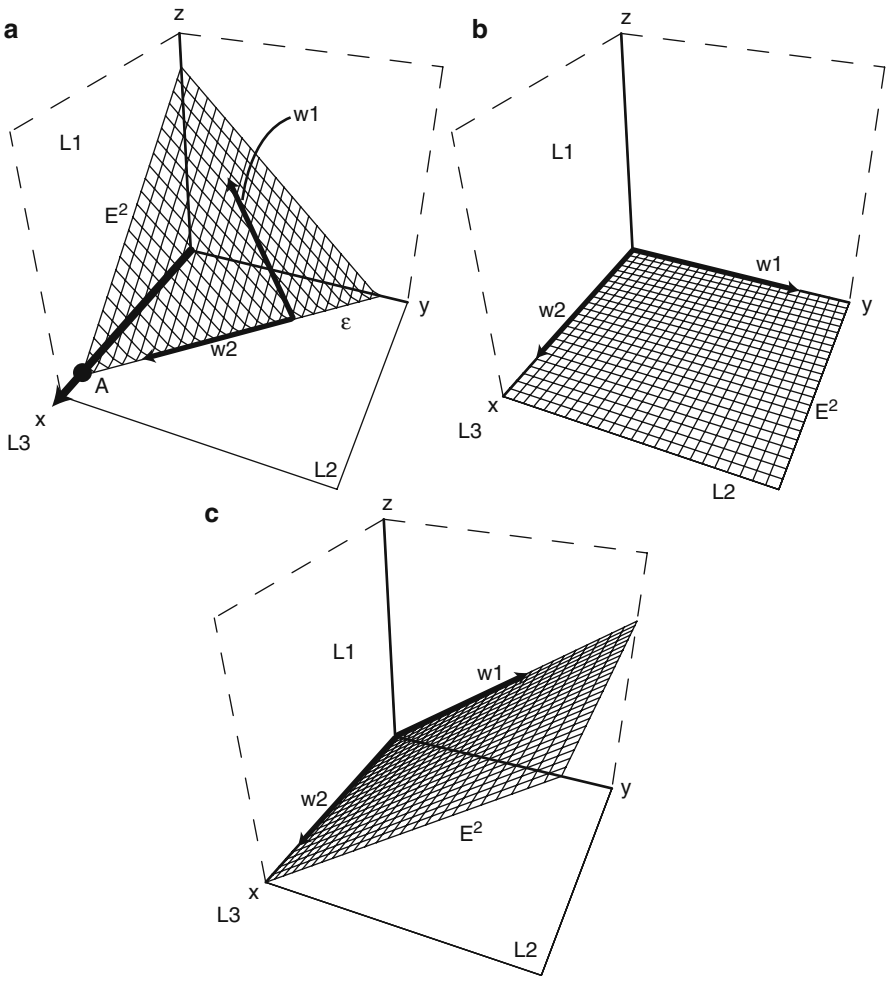

Fig. 5 Possible choices of the 2-dimensional space E^2 for the computation of the 2-LCE in the example of Fig. 4, where $\mathbb{R}^3$ is considered as the tangent space of a hypothetical dynamical system. In each panel the chosen "plane" E^2 is drawn, as well as one of its possible basis constituted of vectors $\mathbf{w}_1$, $\mathbf{w}_2$. (**a**) A random choice of E^2 leads to a plane intersecting L_2 along line ϵ ($\dim(E^2 \cap L_2) = 1$) and L_1 at point A ($\dim(E^2 \cap L_3) = 0$). In this case (33) gives $\chi(\mathbf{A}_t, E^2) = \chi_1 + \chi_2$. More carefully made choices of E^2 (which are obviously not made at random) results to configurations leading to the computation of $\chi_2 + \chi_3$ (**b**) and $\chi_1 + \chi_3$ (**c**) from (33). In these cases E^2 does not satisfy Condition R (37) since $\dim(E^2 \cap L_2) = 2$, $\dim(E^2 \cap L_3) = 1$ in (**b**) and $\dim(E^2 \cap L_2) = 1$, $\dim(E^2 \cap L_3) = 1$ in (**c**)

as one should have expected from the fact that these choices correspond to carefully designed configurations and not to a random process.

Similarly to the case of the computation of the 1-LCEs we note that, even if in some exceptional case one could know a priori the subspaces L_i $i = 1, 2, \ldots, s$, so that one could choose $\mathbf{w}_i$ $i = 1, 2, \ldots, p$ to span a particular subspace E^p in order to

compute a specific value of the p-LCE, smaller than $\sum_{i=1}^{p} \chi_i$ (like in Fig. 5b c), the inevitable computational errors would eventually lead to the numerical computation of the maximal possible value of the p-LCE.

Summarizing we point out that the practical implementation of Theorem 1 guarantees that a random choice of p initial vectors $\mathbf{w}_i$ $i = 1, 2, \ldots, p$ with $1 \leq p \leq n$ generates a space E^p which satisfies Condition R (37) and leads to the actual computation of the corresponding p-mLCE, namely $\chi_1 + \chi_2 + \ldots + \chi_p$. This statement, which was originally presented in [11, 9], led to the standard algorithm for the computation of all LCEs presented in [14]. This algorithm is analyzed in Sect. 6.1.

4.4 The Multiplicative Ergodic Theorem

After presenting results concerning the existence and the computation of the LCEs of all orders for a general matrix function $\mathbf{A}_t$, let us restrict our study to the case of multiplicative cocycles $\mathbf{R}(t, \mathbf{x})$, which are matrix functions satisfying (23). The multiplicative cocycles arise naturally in discrete and continuous dynamical systems as was explained in the beginning of Sect. 4.

In particular, we consider the multiplicative cocycle $d_{\mathbf{x}}\Phi^t$ which maps the tangent space at $\mathbf{x} \in \mathcal{S}$ to the tangent space at $\Phi^t(\mathbf{x}) \in \mathcal{S}$ for a dynamical system defined on the differentiable manifold $\mathcal{S}$. We recall that $\mathcal{S}$ is a measure space with a normalized measure μ and that Φ^t is a diffeomorphism on $\mathcal{S}$, i.e., Φ^t is a measurable bijection of $\mathcal{S}$ which preserves the measure μ (24) and whose inverse is also measurable. We remark that in measure theory we disregard sets of measure 0. In this sense Φ^t is called measurable if it becomes measurable upon disregarding from $\mathcal{S}$ a set of measure 0. Quite often we will use the expression "*for almost all* $\mathbf{x}$ *with respect to measure* μ" for the validity of a statement, implying that the statement is true for all points $\mathbf{x}$ with the possible exception of a set of points with measure 0.

A basic property of the multiplicative cocycles is their regularity, since Theorem 1 guarantees the existence of characteristic exponents and the finiteness of the LCEs of all orders for regular multiplicative cocycles. Thus, it is important to determine specific conditions that multiplicative cocycles should fulfill in order to be regular. Such conditions were first provided by Oseledec [102] who also formulated and proved the so-called *Multiplicative Ergodic Theorem* (MET), which is often referred as *Oseledec's theorem*.

The MET gives information about the dynamical structure of a multiplicative cocycle $\mathbf{R}(t, \mathbf{x})$ and its asymptotic behavior for $t \to \infty$. The application of the MET for the particular multiplicative cocycle $d_{\mathbf{x}}\Phi^t$ provides the theoretical framework for the computation of the LCEs for dynamical systems. The MET is one of the milestones in the study of ergodic properties of dynamical systems and it can be considered as a sort of a spectral theorem for random matrix products [113]. As a testimony to the importance of this theorem one can find several alternative proofs for it in the literature. The original proof of Oseledec [102] applies to both continuous and discrete systems. In view to the application to algebraic groups, Raghunathan [108] devised a simple proof of the MET, which nevertheless could

not guarantee the finiteness of all LCEs. Although Raghunathan's results apply only to maps, an extension to flows, following the ideas of Oseledec, was given by Ruelle [114]. Benettin et al. [13] proved a somewhat different version of the theorem being mainly interested in its application to Hamiltonian flows and symplectic maps. Alternative proofs can also be found in [76, 141].

In [102] Oseledec proved that a multiplicative cocycle $\mathbf{R}(t, \mathbf{x})$ is regular and thus, the MET is applicable to it, if it satisfies the condition

$$\sup_{|t|\leq 1} \ln^+ \|\mathbf{R}^{\pm}(t, \mathbf{x})\| \in L^1(\mathcal{S}, \mu)^3, \tag{38}$$

where $\ln^+ a = \max\{0, \ln a\}$. From (38) we obtain the estimate

$$\|\mathbf{R}(t, \mathbf{x})\| \leq e^{J(\mathbf{x})|t|}, \tag{39}$$

for $t \to \pm\infty$ for almost all $\mathbf{x}$ with respect to μ, where $J(\mathbf{x})$ is a measurable function. From (39) it follows that $\mathbf{R}(t, \mathbf{x})$, considered as a function of t for fixed $\mathbf{x}$, satisfies (25). Benettin et al. [13] considered a slightly different version of the MET with respect to the one presented in [102]. Their version was adapted to the framework of a continuous or discrete dynamical system with Φ^t being a diffeomorphism of class C^1, i.e., both Φ^t and its inverse are continuously differentiable. They formulated the MET for the particular multiplicative cocycle $d_{\mathbf{x}}\Phi^t$, which they proved to be regular. Since our presentation is mainly focused on autonomous Hamiltonian systems and symplectic maps we will also state the MET for the specific cocycle $d_{\mathbf{x}}\Phi^t$. The version of the MET we present is mainly based on [102, 114, 13] and combines different formulations of the theorem given by various authors over the years.

Theorem 2 (Multiplicative Ergodic Theorem—MET). *Consider a dynamical system as follows: Let its phase space $\mathcal{S}$ be an n-dimensional compact manifold with a normalized measure μ, $\mu(\mathcal{S}) = 1$, and a smooth Riemannian metric $\| \|$. Consider also a measure–preserving diffeomorphism Φ^t of class C^1 satisfying*

$$\Phi^{t+s} = \Phi^t \circ \Phi^s \ ,$$

with t denoting time and having real (continuous system) or integer (discrete system) values. Then for almost all $\mathbf{x} \in \mathcal{S}$, with respect to measure μ we have:

1. *The family of multiplicative cocycles $d_{\mathbf{x}}\Phi^t : \mathcal{T}_{\mathbf{x}}\mathcal{S} \to \mathcal{T}_{\Phi^t(\mathbf{x})}\mathcal{S}$, where $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ denotes the tangent space of $\mathcal{S}$ at point $\mathbf{x}$, is regular.*
2. *The LCEs of all orders exist and are independent of the choice of the Riemannian metric of $\mathcal{S}$.*

---

[3] We recall that a measurable function $f : \mathcal{S} \to \mathbb{R}$ (or $\mathbb{C}$) of the measure space $(\mathcal{S}, \mu)$ belongs to the space $L^1(\mathcal{S}, \mu)$ if its absolute value has a finite Lebesgue integral, i.e.,

$$\int |f| d\mu < \infty.$$

In particular, for any $\mathbf{w} \in \mathcal{T}_{\mathbf{x}}\mathcal{S}$ *the finite limit*

$$\chi(\mathbf{x}, \mathbf{w}) = \lim_{t\to\infty} \frac{1}{t} \ln \|d_{\mathbf{x}}\Phi^t \mathbf{w}\| \tag{40}$$

exists and defines the LCE of order 1 (1-LCE). There exists at least one normal basis $\mathbf{v}_i$, $i = 1, 2, \ldots, n$ *of* $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ *for which the corresponding (possibly nondistinct) 1-LCEs* $\chi_i(\mathbf{x}) = \chi(\mathbf{x}, \mathbf{v}_i)$ *are ordered as*

$$\chi_1(\mathbf{x}) \geq \chi_2(\mathbf{x}) \geq \cdots \geq \chi_n(\mathbf{x}). \tag{41}$$

Assume that the value $\nu_i(\mathbf{x})$, $i = 1, 2, \ldots, s$ *with* $s = s(\mathbf{x})$, $1 \leq s \leq n$ *appears exactly* $k_i(\mathbf{x}) = k_i(\mathbf{x}, \nu_i) > 0$ *times among these numbers. Then the spectrum of LCEs* $(\nu_i(\mathbf{x}), k_i(\mathbf{x}))$, $i = 1, 2, \ldots, s$ *is a measurable function of* $\mathbf{x}$, *and as* $\mathbf{w} \neq 0$ *varies in* $\mathcal{T}_{\mathbf{x}}\mathcal{S}$, $\chi(\mathbf{x}, \mathbf{w})$ *takes one of these s different values*

$$\nu_1(\mathbf{x}) > \nu_2(\mathbf{x}) > \cdots > \nu_s(\mathbf{x}). \tag{42}$$

It also holds

$$\sum_{i=1}^{s} k_i(\mathbf{x})\nu_i(\mathbf{x}) = \lim_{t\to\infty} \frac{1}{t} \ln |\det d_{\mathbf{x}}\Phi^t|. \tag{43}$$

For any p-dimensional $(1 \leq p \leq n)$ *subspace* $E^p \subseteq \mathcal{T}_{\mathbf{x}}\mathcal{S}$, *generated by a linearly independent set* $\mathbf{w}_i$, $i = 1, 2, \ldots, p$ *the finite limit*

$$\chi(\mathbf{x}, E^p) = \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{vol}_p(d_{\mathbf{x}}\Phi^t, E^p), \tag{44}$$

where $\mathrm{vol}_p(d_{\mathbf{x}}\Phi^t, E^p)$ *is the volume of the p-parallelogram having as edges the vectors* $d_{\mathbf{x}}\Phi^t \mathbf{w}_i$, *exists, and defines the LCE of order p (p-LCE). The value of* $\chi(\mathbf{x}, E^p)$ *is equal to the sum of p 1-LCEs* $\chi_i(\mathbf{x})$, $i = 1, 2, \ldots, n$.

3. *The set of vectors*

$$L_i(\mathbf{x}) = \{\mathbf{w} \in \mathcal{T}_{\mathbf{x}}\mathcal{S} : \chi(\mathbf{x}, \mathbf{w}) \leq \nu_i(\mathbf{x})\} \quad , \quad 1 \leq i \leq s$$

is a linear subspace of $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ *satisfying*

$$\mathcal{T}_{\mathbf{x}}\mathcal{S} = L_1(\mathbf{x}) \supset L_2(\mathbf{x}) \supset \cdots \supset L_s(\mathbf{x}) \supset L_{s+1}(\mathbf{x}) \stackrel{def}{=} \{0\}. \tag{45}$$

If $\mathbf{w} \in L_i(\mathbf{x}) \setminus L_{i+1}(\mathbf{x})$ *then* $\chi(\mathbf{x}, \mathbf{w}) = \nu_i(\mathbf{x})$ *for* $i = 1, 2, \ldots, s$. *The multiplicity* $k_i(\mathbf{x})$ *of values* $\nu_i(\mathbf{x})$ *is given by* $k_i(\mathbf{x}) = \dim L_i(\mathbf{x}) - \dim L_{i+1}(\mathbf{x})$.

4. *The symmetric positive-defined matrix*

$$\mathbf{\Lambda}_{\mathbf{x}} = \lim_{t\to\infty} \left(\boldsymbol{Y}^{\mathrm{T}}(t) \cdot \boldsymbol{Y}(t)\right)^{1/2t}$$

exists. $\mathbf{Y}(t)$ *is the matrix corresponding to* $d_{\mathbf{x}}\Phi^t$ *and is defined by* (10) *and* (13) *for continuous and discrete dynamical systems, respectively. The logarithms of the eigenvalues of* $\mathbf{\Lambda}_{\mathbf{x}}$ *are the* s *distinct 1-LCEs* (42) *of the dynamical system. The corresponding eigenvectors are orthogonal (since* $\mathbf{\Lambda}_{\mathbf{x}}$ *is symmetric), and for the corresponding eigenspaces* $V_1(\mathbf{x}), V_2(\mathbf{x}), \ldots, V_s(\mathbf{x})$ *we have*

$$k_i(\mathbf{x}) = \dim V_i(\mathbf{x}) \ , \quad L_i(\mathbf{x}) = \bigoplus_{r=i}^{s} V_r(\mathbf{x}) \ \textit{for} \ i = 1, 2, \ldots, s.$$

Thus, $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ *is decomposed as*

$$\mathcal{T}_{\mathbf{x}}\mathcal{S} = V_1(\mathbf{x}) \oplus V_2(\mathbf{x}) \oplus \cdots \oplus V_s(\mathbf{x}),$$

and for every nonzero vector $\mathbf{w} \in V_i(\mathbf{x})$*,* $i = 1, 2, \ldots, s$*, we get*

$$\chi(\mathbf{x}, \mathbf{w}) = \nu_i(\mathbf{x}).$$

A short remark is necessary here. The regularity of $d_{\mathbf{x}}\Phi^t$, which guarantees the validity of (40) and (44) and the finiteness of the LCEs of all orders, should not be confused with the regular nature of orbits of the dynamical system. Regular orbits have all their LCEs equal to zero (see also the discussion in Sect. 5.3).

Benettin et al. [11, 13] have formulated also the following theorem which provides the theoretical background for the numerical algorithm they presented in [14] for the computation of all LCEs.

Theorem 3. *Under the assumptions of the MET, the* p*-LCE of any* p*-dimensional subspace* $E^p \subseteq \mathcal{T}_{\mathbf{x}}\mathcal{S}$ *satisfying Condition R* (37) *is equal to the sum of the* p *largest 1-LCEs* (41)*:*

$$\chi(\mathbf{x}, E^p) = \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{vol}_p(d_{\mathbf{x}}\Phi^t, E^p) = \sum_{i=1}^{p} \chi_i(\mathbf{x}). \tag{46}$$

4.5 Properties of the Spectrum of LCEs

Let us now turn our attention to the structure of the spectrum of LCEs for ND autonomous Hamiltonian systems and for $2N$d symplectic maps, which are the main dynamical systems we are interested in. Such systems preserve the phase space volume, and thus, the r. h. s. of (43) vanishes. So for the sum of all the 1-LCEs we have

$$\sum_{i=1}^{2N} \chi_i(\mathbf{x}) = 0. \tag{47}$$

The symplectic nature of these systems gives indeed more. It has been proved in [13] that the spectrum of LCEs consists of pairs of values having opposite signs

$$\chi_i(\mathbf{x}) = -\chi_{2N-i+1}(\mathbf{x}) \ , \ \ i = 1, 2, \ldots, N. \tag{48}$$

Thus, the spectrum of LCEs becomes

$$\chi_1(\mathbf{x}) \geq \chi_2(\mathbf{x}) \geq \cdots \geq \chi_N(\mathbf{x}) \geq -\chi_N(\mathbf{x}) \geq \cdots \geq -\chi_2(\mathbf{x}) \geq -\chi_1(\mathbf{x}).$$

For autonomous Hamiltonian flows we can say something more. Let us first recall that for a general differentiable flow on a compact manifold without stationary points at least one LCE must vanish [13, 70]. This follows from the fact that, in the direction along the flow a deviation vector grows only linearly in time. So, in the case of a Hamiltonian flow, due to the symmetry of the spectrum of LCEs (48), at least two LCEs vanish, i.e.,

$$\chi_N(\mathbf{x}) = \chi_{N+1}(\mathbf{x}) = 0,$$

while the presence of any additional independent integral of motion leads to the vanishing of another pair of LCEs.

Let us now study the particular case of a periodic orbit of period T, such that $\Phi^T(\mathbf{x}) = \mathbf{x}$, following [9, 12]. In this case $d_{\mathbf{x}}\Phi^T$ is a linear operator on the tangent space $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ so that for any deviation vector $\mathbf{w}(0) \in \mathcal{T}_{\mathbf{x}}\mathcal{S}$ we have

$$\mathbf{w}(T) = \mathbf{Y} \cdot \mathbf{w}(0), \tag{49}$$

where $\mathbf{Y}$ is the constant matrix corresponding to $d_{\mathbf{x}}\Phi^T$. Suppose that $\mathbf{Y}$ has $2N$ (possibly complex) eigenvalues λ_i, $i = 1, 2, \ldots, 2N$, whose magnitudes can be ordered as

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{2N}|.$$

Let $\hat{\mathbf{w}}_i$, $i = 1, 2, \ldots, 2N$, denote the corresponding unitary eigenvectors. Then for $\mathbf{w}(0) = \hat{\mathbf{w}}_i$ (49) implies

$$\mathbf{w}(kT) = \lambda_i^k \hat{\mathbf{w}}_i \ , \ \ k = 1, 2, \ldots \tag{50}$$

and so we conclude from (40) that

$$\chi(\mathbf{x}, \hat{\mathbf{w}}_i) = \frac{1}{T} \ln |\lambda_i| = \chi_i(\mathbf{x}), \ \ i = 1, 2, \ldots, 2N.$$

Furthermore for a deviation vector

$$\mathbf{w}(0) = c_1\hat{\mathbf{w}}_1 + c_2\hat{\mathbf{w}}_2 + \ldots + c_{2N}\hat{\mathbf{w}}_{2N}$$

with $c_i \in \mathbb{R}$, $i = 1, 2, \ldots, 2N$, it follows from (50) that the first nonvanishing coefficient c_i eventually dominates the evolution of $\mathbf{w}(t)$ and we get $\chi(\mathbf{x}, \mathbf{w}) = \chi_i$. In this case we can define a filtration similar to the one presented in (45) by defining $L_1 = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_{2N}] = \mathcal{T}_{\mathbf{x}}\mathcal{S}$, $L_2 = [\hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_{2N}]$, $\ldots$, $L_{2N} = [\hat{\mathbf{w}}_{2N}]$, $L_{2N+1} = [\mathbf{0}]$, where $[\ ]$ denotes the linear space spanned by vectors $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_{2N}$ and so on. It becomes evident that a random choice of an initial deviation vector $\mathbf{w}(0) \in \mathcal{T}_{\mathbf{x}}\mathcal{S}$ will lead to the computation of the mLCE $\chi_1(\mathbf{x})$ since, in general, $\mathbf{w}(0) \in L_1 \setminus L_2$.

So, in the case of an unstable periodic orbit where $|\lambda_1| > 1$ we get $\chi_1(\mathbf{x}) > 0$, which implies that nearby orbits diverge exponentially from the periodic one. This orbit is not called chaotic, although its mLCE is larger than zero, but simply "unstable". In fact, unstable periodic orbits exist also in integrable systems. Since the measure of periodic orbits in a general dynamical system has zero measure, periodic orbits (stable and unstable) are rather exceptional.

In the general case of a nonperiodic orbit we are no more allowed to use concepts as eigenvectors and eigenvalues because the linear operator $d_{\mathbf{x}}\Phi^t$ maps $\mathcal{T}_{\mathbf{x}}\mathcal{S}$ into $\mathcal{T}_{\Phi^t(\mathbf{x})}\mathcal{S} \neq \mathcal{T}_{\mathbf{x}}\mathcal{S}$, while eigenvectors are intrinsically defined only for linear operators of a linear space into itself. Nevertheless, in the case of nonperiodic orbits the MET proves the existence of the LCEs and of filtration (45). In a way, the MET provides an extension of the linear stability analysis of periodic orbits to the case of nonperiodic ones, although one should always keep in mind that the LCEs are related to the real and positive eigenvalues of the symmetric, positive-defined matrix $\mathbf{Y}^{\mathrm{T}}(t) \cdot \mathbf{Y}(t)$ [63, 98]. On the other hand, linear stability analysis involves the computation of the eigenvalues of the nonsymmetric matrix $\mathbf{Y}(t)$, which solves the linearized equations of motion (10) for Hamiltonian flows or (13) for maps. These eigenvalues are real or come in pairs of complex conjugate pairs and, in general, they are not directly related to the LCEs which are real numbers.

An important property of the LCEs is that they are constant in a connected chaotic domain. This is due to the fact that every nonperiodic orbit in the same connected chaotic domain covers densely this domain, thus, two different orbits of the same domain are in a sense dynamically equivalent. The unstable periodic orbits in this chaotic domain have in general LCEs that are different from the constant LCEs of the nonperiodic orbits. This is due to the fact that the periodic orbits do not visit the whole domain, thus, they cannot characterize its dynamical behavior. In fact, different periodic orbits have different LCEs.

5 The Maximal LCE

From this point on, in order to simplify our notation, we will not explicitly write the dependence of the LCEs on the specific point $\mathbf{x} \in \mathcal{S}$. So, in practice, considering that we are referring to a specific point $\mathbf{x} \in \mathcal{S}$, we denote by χ_i the LCEs of order 1 and by $\chi_i^{(p)}$ the LCEs of order p.

For the practical determination of the chaotic nature of orbits a numerical computation of the mLCE χ_1 can be employed. If the studied orbit is regular $\chi_1 = 0$, while if it is chaotic $\chi_1 > 0$, implying exponential divergence of nearby orbits. The computation of the mLCE has been used extensively as a chaos indicator after the introduction of numerical algorithms for the determination of its value at late 1970s [10, 99, 8, 34, 14].

Apart from using the mLCE as a criterion for the chaoticity or the regularity of an orbit its value also attains a "physical" meaning and defines a specific timescale for the considered dynamical system. In particular, the inverse of the mLCE, which is called *Lyapunov time*,

$$t_L = \frac{1}{\chi_1}, \tag{51}$$

gives an estimate of the time needed for a dynamical system to become chaotic and in practice measures the time needed for nearby orbits of the system to diverge by e (see e.g, [30, p. 508]).

5.1 Computation of the mLCE

The mLCE can be computed by the numerical implementation of (40). In Sect. 4.2 we showed that a random choice of the initial deviation vector $\mathbf{w}(0) \in \mathcal{T}_{\mathbf{x}}\mathcal{S}$ leads to the numerical computation of the mLCE. We recall that the deviation vector $\mathbf{w}(t)$ at time $t > 0$ is determined by the action of the operator $d_{\mathbf{x}}\Phi^t$ on the initial deviation vector $\mathbf{w}(0)$ according to (7)

$$\mathbf{w}(t) = d_{\mathbf{x}}\Phi^t \, \mathbf{w}(0). \tag{52}$$

This equation represents the solution of the variational equations (8) or the evolution of a deviation vector under the action of the tangent map (11) and takes the form (9) and (12), respectively. We emphasize that, both the variational equations and the equations of the tangent map are linear with respect to the tangent vector $\mathbf{w}$, i.e.,

$$d_{\mathbf{x}}\Phi^t \, (a \, \mathbf{w}) = a \, d_{\mathbf{x}}\Phi^t \mathbf{w}, \quad \text{for any } \; a \in \mathbb{R}. \tag{53}$$

In order to evaluate the mLCE of an orbit with initial condition $\mathbf{x}(0)$, one has to follow simultaneously the time evolution of the orbit itself and of a deviation vector $\mathbf{w}$ from this orbit with initial condition $\mathbf{w}(0)$. In the case of a Hamiltonian flow (continuous time) we solve simultaneously the Hamilton equations of motion (2) for the time evolution of the orbit and the variational equations (8) for the time evolution of the deviation vector. In the case of a symplectic map (discrete time) we iterate the map (4) for the evolution of the orbit simultaneously with the tangent map (11), which determines the evolution of the tangent vector. The mLCE is then computed as the limit for $t \to \infty$ of the quantity

$$X_1(t) = \frac{1}{t} \ln \frac{\|d_{\mathbf{x}(0)} \Phi^t \, \mathbf{w}(0)\|}{\|\mathbf{w}(0)\|} = \frac{1}{t} \ln \frac{\|\mathbf{w}(t)\|}{\|\mathbf{w}(0)\|}, \tag{54}$$

often called *finite time mLCE*. So, we have

$$\chi_1 = \lim_{t \to \infty} X_1(t). \tag{55}$$

The direct numerical implementation of (54) and (55) for the evaluation of χ_1 meets a severe difficulty. If, for example, the orbit under study is chaotic, the norm $\|\mathbf{w}(t)\|$ increases exponentially with increasing time t, leading to numerical overflow, i.e., $\|\mathbf{w}(t)\|$ attains very fast extremely large values that cannot be represented in the computer. This difficulty can be overcome by a procedure which takes advantage of the linearity of $d_{\mathbf{x}} \Phi^t$ (53) and of the composition law (22). Fixing a small time interval τ we express time t with respect to τ as $t = k\tau$, $k = 1, 2, \ldots$. Then for the quantity $X_1(t)$ we have

$$\begin{aligned} X_1(k\tau) =& \frac{1}{k\tau} \ln \frac{\|\mathbf{w}(k\tau)\|}{\|\mathbf{w}(0)\|} \\ =& \frac{1}{k\tau} \ln \left(\frac{\|\mathbf{w}(k\tau)\|}{\|\mathbf{w}((k-1)\tau)\|} \frac{\|\mathbf{w}((k-1)\tau)\|}{\|\mathbf{w}((k-2)\tau)\|} \cdots \frac{\|\mathbf{w}(2\tau)\|}{\|\mathbf{w}(\tau)\|} \frac{\|\mathbf{w}(\tau)\|}{\|\mathbf{w}(0)\|} \right) \\ =& \frac{1}{k\tau} \sum_{i=1}^{k} \ln \frac{\|\mathbf{w}(i\tau)\|}{\|\mathbf{w}((i-1)\tau)\|} \Rightarrow \\ & X_1(k\tau) = \frac{1}{k\tau} \sum_{i=1}^{k} \ln \frac{\|d_{\mathbf{x}(0)} \Phi^{i\tau} \, \mathbf{w}(0)\|}{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|}. \end{aligned} \tag{56}$$

Denoting by D_0 the norm of the initial deviation vector $\mathbf{w}(0)$

$$D_0 = \|\mathbf{w}(0)\|,$$

we get for the evolved deviation vector at time $t = k\tau$

$$\begin{aligned} d_{\mathbf{x}(0)} \Phi^{i\tau} \, \mathbf{w}(0) \;=\; & d_{\mathbf{x}(0)} \Phi^{\tau + (i-1)\tau} \, \mathbf{w}(0) \overset{(22)}{=} d_{\Phi^{(i-1)\tau}(\mathbf{x}(0))} \Phi^{\tau} \left(d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \mathbf{w}(0) \right) \\ \overset{(53)}{=} & \frac{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|}{D_0} \, d_{\Phi^{(i-1)\tau}(\mathbf{x}(0))} \Phi^{\tau} \left(\frac{d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \mathbf{w}(0)}{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|} D_0 \right) \Rightarrow \\ & \frac{d_{\mathbf{x}(0)} \Phi^{i\tau} \, \mathbf{w}(0)}{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|} = \frac{d_{\Phi^{(i-1)\tau}(\mathbf{x}(0))} \Phi^{\tau} \left(\frac{d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \mathbf{w}(0)}{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|} D_0 \right)}{D_0}. \end{aligned} \tag{57}$$

Let us now denote by

$$\hat{\mathbf{w}}((i-1)\tau) = \frac{d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \mathbf{w}(0)}{\|d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \, \mathbf{w}(0)\|} D_0,$$

the deviation vector at point $\Phi^{(i-1)\tau}(\mathbf{x}(0))$ having the same direction with $\mathbf{w}((i-1)\tau)$ and norm D_0, and by D_i its norm after its evolution for τ time units

$$D_i = \| d_{\Phi^{(i-1)\tau}(\mathbf{x}(0))} \Phi^{\tau} \hat{\mathbf{w}}((i-1)\tau) \|.$$

Using this notation we derive from (57)

$$\ln \frac{\| d_{\mathbf{x}(0)} \Phi^{i\tau} \mathbf{w}(0) \|}{\| d_{\mathbf{x}(0)} \Phi^{(i-1)\tau} \mathbf{w}(0) \|} = \ln \frac{D_i}{D_0} = \ln \alpha_i, \tag{58}$$

with α_i being the local coefficient of expansion of the deviation vector for a time interval of length τ when the corresponding orbit evolves from position $\Phi^{(i-1)\tau}(\mathbf{x}(0))$ to position $\Phi^{i\tau}(\mathbf{x}(0))$ ($\ln \alpha_i / \tau$ is also called *stretching number* [135], [30, p. 257]).

From (55), (56), and (58) we conclude that the mLCE χ_1 can be computed as

$$\chi_1 = \lim_{k\to\infty} X_1(k\tau) = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \frac{D_i}{D_0} = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \alpha_i. \tag{59}$$

Since the initial norm D_0 can have any arbitrary value, one usually sets it to $D_0 = 1$. Equation (59) implies that practically χ_1 is the limit value, for $t \to \infty$, of the mean of the stretching numbers along the studied orbit [14, 57, 135].

5.2 The Numerical Algorithm

In practice, for the evaluation of the mLCE we follow the evolution of a unitary initial deviation vector $\hat{\mathbf{w}}(0) = \mathbf{w}(0)$, $\|\mathbf{w}(0)\| = D_0 = 1$ and every $t = \tau$ time units we replace the evolved vector $\mathbf{w}(k\tau)$, $k = 1, 2, \ldots$, by vector $\hat{\mathbf{w}}(k\tau)$ having the same direction but norm equal to 1 ($\|\hat{\mathbf{w}}(k\tau)\| = 1$). Before each new renormalization the corresponding α_k is computed and χ_1 is estimated from (59).

More precisely at $t = \tau$ we have $\alpha_1 = \|\mathbf{w}(\tau)\|$. Then we define a unitary vector $\hat{\mathbf{w}}(\tau)$ by renormalizing $\mathbf{w}(\tau)$ and using it as an initial deviation vector we evolve it along the orbit from $\mathbf{x}(\tau)$ to $\mathbf{x}(2\tau)$ according to (52), having $\mathbf{w}(2\tau) = d_{\mathbf{x}(\tau)} \Phi^{\tau}\, \hat{\mathbf{w}}(\tau)$. Then we define $\alpha_2 = \|\mathbf{w}(2\tau)\|$ and we estimate χ_1 (see Fig. 6). We iteratively apply the above-described procedure until a good approximation of χ_1 is achieved. The algorithm for the evaluation of the mLCE χ_1 is described in pseudo-code in Table 1.

Instead of utilizing the variational equations or the tangent map for the evolution of a deviation vector in the above-described algorithm, one could integrate (2) or iterate (4) for two orbits starting nearby and estimate $\mathbf{w}(t)$ by difference. Indeed, this approach, influenced by the rough idea of divergence of nearby orbits introduced

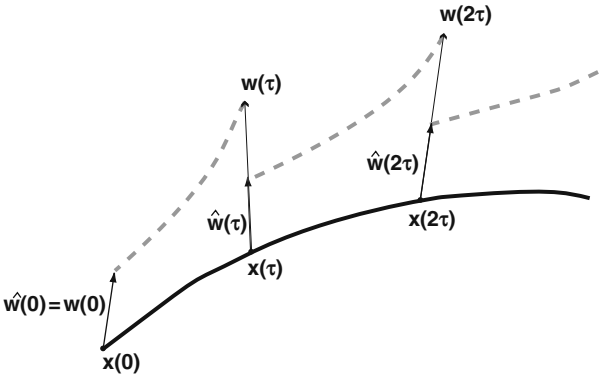

Fig. 6 Numerical scheme for the computation of the mLCE χ_1. The unitary deviation vector $\hat{\mathbf{w}}((i-1)\tau)$, $i = 1, 2, \ldots$, is evolved according to the variational equations (8) (continuous time) or the equations of the tangent map (11) (discrete time) for $t = \tau$ time units. The evolved vector $\mathbf{w}(i\tau)$ is replaced by a unitary vector $\hat{\mathbf{w}}(i\tau)$ having the same direction with $\mathbf{w}(i\tau)$. For each successive time interval $[(i-1)\tau, i\tau]$ the quantity $\alpha_i = \|\mathbf{w}(i\tau)\|$ is computed and χ_1 is estimated from (59)

Table 1 The algorithm for the computation of the mLCE χ_1 as the limit for $t \to \infty$ of $X_1(t)$ according to (59). The program computes the evolution of $X_1(t)$ as a function of time t up to a given upper value of time $t = T_M$ or until $X_1(t)$ attains a very small value, smaller than a low threshold value X_{1m}

Input:	1. Hamilton equations of motion (2) and variational equations (8), or equations of the map (4) and of the tangent map (11). 2. Initial condition for the orbit $\mathbf{x}(0)$. 3. Initial *unitary* deviation vector $\mathbf{w}(0)$. 4. Renormalization time τ. 5. Maximal time: T_M and minimum allowed value of $X_1(t)$: X_{1m}.
Step 1	**Set** the stopping flag, SF $\leftarrow 0$, and the counter, $k \leftarrow 1$.
Step 2	**While** (SF $= 0$) **Do** **Evolve** the orbit and the deviation vector from time $t = (k-1)\tau$ to $t = k\tau$, i. e. **Compute** $\mathbf{x}(k\tau)$ and $\mathbf{w}(k\tau)$.
Step 3	**Compute** current value of $\alpha_k = \|\mathbf{w}(k\tau)\|$. **Compute** and **Store** current value of $X_1(k\tau) = \frac{1}{k\tau}\sum_{i=1}^{k} \ln \alpha_i$.
Step 4	Renormalize deviation vector by **Setting** $\mathbf{w}(k\tau) \leftarrow \mathbf{w}(k\tau)/\alpha_k$.
Step 5	**Set** the counter $k \leftarrow k + 1$.

Step 6	**If** $[(k\tau > T_M)$ *or* $(X_1((k-1)\tau) < X_{1m})]$ **Then**
	Set SF $\leftarrow 1$.
	End If

	End While
Step 7	**Report** the time evolution of $X_1(t)$.

in [72], was initially adopted for the computation of the mLCE [10, 99, 8]. This technique was abandoned after a while as it was realized that the use of explicit equations for the evolution of deviation vectors was more reliable and efficient [34, 119, 14], although in some cases it is used also nowadays (see, e.g., [145]).

5.3 Behavior of $X_1(t)$ for Regular and Chaotic Orbits

Let us now discuss in more detail the behavior of the computational scheme for the evaluation of the mLCE for the cases of regular and chaotic orbits.

The LCE of regular orbits vanish [10, 23] due to the linear increase with time of the norm of deviation vectors. We illustrate this behavior in the case of an ND Hamiltonian system, but a similar analysis can be easily carried out for symplectic maps. In such systems regular orbits lie on N-dimensional tori. If such tori are found around a stable periodic orbit, they can be accurately described by N formal integrals of motion in involution, so that the system would appear locally integrable. This means that we could perform a local transformation to action-angle variables, considering as actions $J_1, J_2, \ldots, J_N$ the values of the N formal integrals, so that Hamilton's equations of motion, locally attain the form

$$\dot{J}_i = 0, \quad \dot{\theta}_i = \omega_i(J_1, J_2, \ldots, J_N), \quad i = 1, 2, \ldots, N. \tag{60}$$

These equations can be easily integrated to give

$$J_i(t) = J_{i0}, \quad \theta_i(t) = \theta_{i0} + \omega_i(J_{10}, J_{20}, \ldots, J_{N0})\, t, \quad i = 1, 2, \ldots, N,$$

where $J_{i0}, \theta_{i0}, i = 1, 2, \ldots, N$ are the initial conditions of the studied orbit.

By denoting as $\xi_i, \eta_i, i = 1, 2, \ldots, N$ small deviations of J_i and θ_i respectively, the variational equations (8) of system (60) describing the evolution of a deviation vector are as follows:

$$\dot{\xi}_i = 0, \quad \dot{\eta}_i = \sum_{j=1}^{N} \omega_{ij} \cdot \xi_j, \quad i = 1, 2, \ldots, N,$$

where

$$\omega_{ij} = \left. \frac{\partial \omega_i}{\partial J_j} \right|_{\mathbf{J}_0}, \quad i, j = 1, 2, \ldots, N,$$

and $\mathbf{J}_0 = (J_{10}, J_{20}, \ldots, J_{N0}) =$ constant represents the N–dimensional vector of the initial actions. The solution of these equations is

$$\begin{aligned} &\xi_i(t) = \xi_i(0) \\ &\eta_i(t) = \eta_i(0) + \left[\textstyle\sum_{j=1}^{N} \omega_{ij} \xi_j(0)\right] t, \end{aligned} \quad i = 1, 2, \ldots, N. \tag{61}$$

From (61) we see that an initial deviation vector $\mathbf{w}(0)$ with coordinates $\xi_i(0)$, $i = 1, 2, \ldots, N$ in the action variables and $\eta_i(0)$, $i = 1, 2, \ldots, N$ in the angles, i.e., $\mathbf{w}(0) = (\xi_1(0), \xi_2(0), \ldots, \xi_N(0), \eta_1(0), \eta_2(0), \ldots, \eta_N(0))$, evolves in time in such a way that its action coordinates remain constant, while its angle coordinates increase linearly in time. This behavior implies an almost linear increase of the norm

of the deviation vector. To see this, let us assume that vector $\mathbf{w}(0)$ has initially unit magnitude, i.e.,

$$\sum_{i=1}^{N} \xi_i^2(0) + \sum_{i=1}^{N} \eta_i^2(0) = 1$$

whence the time evolution of its norm is given by

$$\|\mathbf{w}(t)\| = \left\{ 1 + \left[\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \omega_{ij} \xi_j(0) \right)^2 \right] t^2 + \left[2 \sum_{i=1}^{N} \left(\eta_i(0) \sum_{j=1}^{N} \omega_{ij} \xi_j(0) \right) \right] t \right\}^{1/2}.$$

This implies that the norm for long times grows linearly with t:

$$\|\mathbf{w}(t)\| \propto t. \tag{62}$$

So, from (54) we see that for long times $X_1(t)$ is of the order $\mathcal{O}(\ln t/t)$, which means that $X_1(t)$ tends asymptotically to zero, as $t \to \infty$ like t^{-1}. This asymptotic behavior is evident in numerical computations of the mLCE of regular orbits, as we can see, for example, in the left panel of Fig. 2.

The asymptotic behavior of $X_1(t)$ for regular orbits, described above, represents a particular case of a more general estimation presented in [63]. In particular, Goldhirsch et al. [63] showed that, in general, after some initial transient time the value of the mLCE χ_1 is related to its finite time estimation by

$$X_1(t) = \chi_1 + \frac{b + z(t)}{t}, \tag{63}$$

where b is a constant and $z(t)$ is a "noise" term of zero mean. According to their analysis, this approximate formula is valid for both regular and chaotic orbits. It is easily seen that from (63) we retrieve again the asymptotic behavior $X_1(t) \propto t^{-1}$ for the case of regular orbits ($\chi_1 = 0$).

In the case of chaotic orbits the variation of $X_1(t)$ is usually irregular for relatively small t and only for large t the value of $X_1(t)$ stabilizes and tends to a constant positive value which is the mLCE χ_1. If, for example, the value of χ_1 is very small then initially, for small and intermediate values of t, the term proportional to t^{-1} dominates the r.h.s. of (63) and $X_1(t) \propto t^{-1}$. As t grows the significance of term $(b + z(t))/t$ diminishes and eventually the value of χ_1 becomes dominant and $X_1(t)$ stabilizes. It becomes evident that for smaller values of χ_1 the larger is the time required for $X_1(t)$ to reach its limiting value, and consequently $X_1(t)$ behaves as in the case of regular orbits, i.e., $X_1(t) \propto t^{-1}$ for larger time intervals. This behavior is clearly seen in Fig. 7 where the evolution of $X_1(t)$ of chaotic orbits with small mLCE is shown. In particular, the values of the mLCE are $\chi_1 \approx 8 \times 10^{-3}$ (left panel) and $\chi_1 \approx 1.6 \times 10^{-7}$ (right panel). In both panels the evolution of $X_1(t)$ of

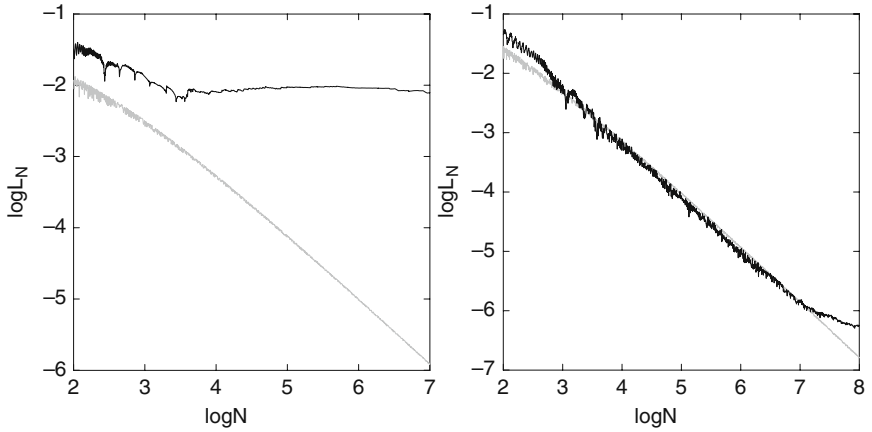

Fig. 7 Evolution of $X_1(t)$ (denoted as L_N) with respect to the discrete time t (denoted as N) in log–log scale for regular (grey curves) and chaotic (black curves) orbits of the 4d map (16) (left panel) and of a 4d map composed of two coupled 2d standard maps (right panel) (see [122] for more details). For regular orbits $X_1(t)$ tends to zero following a power law decay, $X_1(t) \propto t^{-1}$. For chaotic orbits $X_1(t)$ exhibits for some initial time interval the same power law decay before stabilizing to the positive value of the mLCE χ_1. The length of this time interval is larger for smaller values of χ_1. The chaotic orbits have $\chi_1 \approx 8 \times 10^{-3}$ (left panel) and $\chi_1 \approx 1.6 \times 10^{-7}$ (right panel) (after [122])

regular orbits (following the power law $\propto t^{-1}$) is also plotted in order to facilitate the comparison between the two cases.

6 Computation of the Spectrum of LCEs

While the knowledge of the mLCE χ_1 can be used for determining the regular ($\chi_1 = 0$) or chaotic ($\chi_1 > 0$) nature of orbits, the knowledge of part, or of the whole spectrum of LCEs, provides additional information on the underlying dynamics and on the statistical properties of the system and can be used for measuring the fractal dimension of strange attractors in dissipative systems.

In Sect. 4.5 it was stated that for Hamiltonian systems the existence of an integral of motion results to a pair of zero values in the spectrum of LCEs. As an example of such case we refer to the Hamiltonian system studied in [12]. This system has one more integral of motion apart from the Hamiltonian function and so four LCEs were always found to be equal to zero. Thus, the determination of the number of LCEs that vanish can be used as an indicator of the number of the independent integrals of motion that a dynamical system has.

It has been also stated in Sect. 4.5 that the spectrum of the LCEs of orbits in a connected chaotic region is independent of their initial conditions. So, we have a strong indication that two chaotic orbits belong to connected chaotic regions if they exhibit the same spectrum. As an example of this situation we refer to the case

studied in [3] of two chaotic orbits of a 16D Hamiltonian system having similar spectra of LCEs but very different initial conditions.

Vice versa, the existence of different LCEs spectra of chaotic orbits provides strong evidence that these orbits belong to different chaotic regions of the phase space that do not communicate. In [14] two chaotic orbits, previously studied in [34], were found to have significantly different spectra of LCEs and they were considered to belong to different chaotic regions which were called the "big" (corresponding to the largest χ_1) and the "small" chaotic sea. It is worth mentioning that the numerical results of [14] suggested the possible existence of an additional integral of motion for the "small" chaotic sea, since χ_2 seemed to vanish. This assumption was in accordance to the results of [34] where such an integral was formally constructed.

The spectrum of LCEs is also related to two important quantities namely, the metric entropy, also called *Kolmogorov–Sinai (KS) entropy* h, and the *information dimension* D_1, which are trying to quantify the statistical properties of dynamical systems. For the explicit definition of these quantities, as well as detailed discussion of their relation to the LCEs the reader is referred, for example, to [9, 46, 54, 44] [92, pp. 304–305] for the KS entropy and to [79, 46, 47, 66, 44] for the information dimension.

In particular, Pesin [106] showed that under suitable smoothness conditions the relation between the KS entropy h and the LCEs is given by

$$h = \int_{\mathcal{M}} \left[\sum_{\chi_i(\mathbf{x})>0} \chi_i(\mathbf{x}) \right] d\mu,$$

where the sum is extended over all positive LCEs and the integral is defined over a specified region $\mathcal{M}$ of the phase space $\mathcal{S}$.

Kaplan and Yorke [79] introduced a quantity, which they called the *Lyapunov dimension*

$$D_L = j + \frac{\sum_{i=1}^{j} \chi_i}{|\chi_{j+1}|}, \tag{64}$$

where j is the largest integer for which $\chi_1 + \chi_2 + \ldots + \chi_j \geq 0$. The *Kaplan–Yorke conjecture* states that the information dimension D_1 is equal to the Lyapunov dimension D_L, i.e.,

$$D_1 = D_L, \tag{65}$$

for a typical system, and thus, it can be used for the determination of the fractal dimension of strange attractors. The meaning of the word "typical" is that it is not hard to construct examples where (65) is violated (see, e.g., [47]). But the claim is that these examples are pathological in that the slightest arbitrary change of the system restores the applicability of (65) and that such violation has "zero

probability" of occurring in practice. The validity of the Kaplan–Yorke conjecture has been proved in some cases [146, 87] although a general proof has not been achieved yet. We note that in the case of a $2N$D conservative system D_L is equal to the dimension of the whole space, i.e., $D_L = 2N$, because $j = 2N$ in (64) since $\sum_{i=1}^{2N} \chi_i = 0$ according to (47).

So, it becomes evident that developing an efficient algorithm for the numerical evaluation of few or of all LCEs is of great importance for the study of dynamical systems. In this section we present the different methods developed over the years for the computation of the spectrum of LCEs, focusing on the method suggested by Benettin et al. [14], the so-called *standard method*.

6.1 The Standard Method for Computing LCEs

The basis for the computation of few or even of all LCEs is Theorem 3, which states that the computation of a p-LCE from (44), considering a random choice of p ($1 < p \leq 2N$) linearly independent initial deviation vectors, leads to the evaluation of the p-mLCE $\chi_1^{(p)}$, which is equal to the sum of the p largest 1-LCEs (46).

In order to evaluate the p-mLCE of an orbit with initial condition $\mathbf{x}(0)$, one has to follow simultaneously the time evolution of the orbit itself and of p linearly independent deviation vectors with initial conditions $\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0)$ (using the variational equations (8) or the equations of the tangent map (11)). Then, the p-mLCE is computed as the limit for $t \to \infty$ of the quantity

$$\begin{aligned} X^{(p)}(t) &= \frac{1}{t} \ln \frac{\mathrm{vol}_p \left(d_{\mathbf{x}(0)}\Phi^t \, \mathbf{w}_1(0), d_{\mathbf{x}(0)}\Phi^t \, \mathbf{w}_2(0), \cdots, d_{\mathbf{x}(0)}\Phi^t \, \mathbf{w}_p(0) \right)}{\mathrm{vol}_p \left(\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0) \right)} \\ &= \frac{1}{t} \ln \frac{\| \mathbf{w}_1(t) \wedge \mathbf{w}_2(t) \wedge \cdots \wedge \mathbf{w}_p(t) \|}{\| \mathbf{w}_1(0) \wedge \mathbf{w}_2(0) \wedge \cdots \wedge \mathbf{w}_p(0) \|} = \frac{1}{t} \ln \frac{\left\| \bigwedge_{i=1}^{p} \mathbf{w}_i(t) \right\|}{\left\| \bigwedge_{i=1}^{p} \mathbf{w}_i(0) \right\|}, \end{aligned} \tag{66}$$

which is also called the *finite time p-mLCE*. So we have

$$\chi_1^{(p)} = \chi_1 + \chi_2 + \cdots + \chi_p = \lim_{t \to \infty} X^{(p)}(t). \tag{67}$$

We recall that the quantity $\mathrm{vol}_p \left(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p \right)$ appearing in the above definition is the volume of the p-parallelogram having as edges the vectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_p$ (see (106) and (105) in Appendix).

The direct numerical implementation of (66) and (67) faces one additional difficulty apart from the fast growth of the norm of deviation vectors discussed in Sect. 5.1. This difficulty is due to the fact that when at least two vectors are involved (e.g., for the computation of $\chi_1^{(2)}$), the angles between their directions become too small for numerical computations.

This difficulty can be overcome on the basis of the following simple remark: an invertible linear map, as $d_{\mathbf{x}(0)}\Phi^t$, maps a linear p-dimensional subspace onto

a linear subspace of the same dimension, and the coefficient of expansion of any p-dimensional volume under the action of any such linear map (for example, $\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(t)\right\| / \left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(0)\right\|$ in our case) does not depend on the initial volume [14]. Since the numerical value of $\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(0)\right\|$ does not depend on the choice of the orthonormal basis of the space (see Appendix for more details), in order to show the validity of this remark we will consider an appropriate basis which will facilitate our calculations.

In particular, let us consider an orthonormal basis $\left\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p\right\}$ of the p-dimensional space $E^p \subseteq \mathcal{T}_{\mathbf{x}(0)}\mathcal{S}$ spanned by $\left\{\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0)\right\}$. This basis can be extended to an orthonormal basis of the whole $2N$-dimensional space $\left\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p, \hat{\mathbf{e}}_{p+1}, \ldots, \hat{\mathbf{e}}_{2N}\right\}$ and $E^p \subseteq \mathcal{T}_{\mathbf{x}(0)}\mathcal{S}$ can be written as the direct sum of E^p and of the $(2N-p)$-dimensional subspace E' spanned by $\left\{\hat{\mathbf{e}}_{p+1}, \ldots, \hat{\mathbf{e}}_{2N}\right\}$

$$\mathcal{T}_{\mathbf{x}(0)}\mathcal{S} = E^p \bigoplus E'.$$

Consider also the $2N \times p$ matrix $\mathbf{W}(0)$ having as columns the coordinates of vectors $\mathbf{w}_i(0)$, $i = 1, 2, \ldots, p$ with respect to the complete orthonormal basis $\hat{\mathbf{e}}_j$, $j = 1, 2, \ldots, 2N$, in analogy to (102). Since $\mathbf{w}_i(0) \in E^p$ this matrix has the form

$$\mathbf{W}(0) = \begin{bmatrix} \widetilde{\mathbf{W}}(0) \\ \mathbf{0}_{(2N-p)\times p} \end{bmatrix},$$

where $\widetilde{\mathbf{W}}(0)$ is a square $p \times p$ matrix and $\mathbf{0}_{(2N-p)\times p}$ is the $(2N-p) \times p$ matrix with all its elements equal to zero. Then, according to (105) and (106) the volume of the initial p-parallelogram is

$$\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(0)\right\| = \left|\det \widetilde{\mathbf{W}}(0)\right|, \tag{68}$$

since $\det \widetilde{\mathbf{W}}^{\mathrm{T}}(0) = \det \widetilde{\mathbf{W}}(0)$ for the square matrix $\widetilde{\mathbf{W}}(0)$.

Each deviation vector is evolved according to (7) and it can be computed through (9) or (12), with $\mathbf{Y}(t)$ being the $2N \times 2N$ matrix representing the action of $d_{\mathbf{x}(0)}\boldsymbol{\Phi}^t$. By doing a similar choice for the basis of the $\mathcal{T}_{\boldsymbol{\Phi}^t(\mathbf{x}(0))}\mathcal{S}$ space, (102) gives for the evolved vectors

$$\left[\mathbf{w}_1(t)\ \mathbf{w}_2(t) \cdots \mathbf{w}_p(t)\right] = \left[\hat{\mathbf{e}}_1\ \hat{\mathbf{e}}_2 \cdots \hat{\mathbf{e}}_p\right] \cdot \mathbf{Y}(t) \cdot \mathbf{W}(0) = \left[\hat{\mathbf{e}}_1\ \hat{\mathbf{e}}_2 \cdots \hat{\mathbf{e}}_p\right] \cdot \mathbf{W}(t).$$

Writing $\mathbf{Y}(t)$ as

$$\mathbf{Y}(t) = \left[\mathbf{Y}_1(t)\ \mathbf{Y}_2(t)\right],$$

where $\mathbf{Y}_1(t)$ is the $2N \times p$ matrix formed from the first p columns of $\mathbf{Y}(t)$ and $\mathbf{Y}_2(t)$ is the $2N \times (2N-p)$ matrix formed from the last $2N-p$ columns of $\mathbf{Y}(t)$, $\mathbf{W}(t)$ assumes the following form:

$$\mathbf{W}(t) = \mathbf{Y}_1(t) \cdot \widetilde{\mathbf{W}}(0).$$

Then from (105) we get

$$\begin{aligned}\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(t)\right\| &= \sqrt{\det\left(\widetilde{\mathbf{W}}^{\mathrm{T}}(0)\cdot\mathbf{Y}_1^{\mathrm{T}}(t)\cdot\mathbf{Y}_1(t)\cdot\widetilde{\mathbf{W}}(0)\right)}\\ &= \sqrt{\det\widetilde{\mathbf{W}}^{\mathrm{T}}(0)\det\left(\mathbf{Y}_1^{\mathrm{T}}(t)\cdot\mathbf{Y}_1(t)\right)\det\widetilde{\mathbf{W}}(0)}\\ &= |\det\widetilde{\mathbf{W}}(0)|\sqrt{\det\left(\mathbf{Y}_1^{\mathrm{T}}(t)\cdot\mathbf{Y}_1(t)\right)}. \end{aligned} \tag{69}$$

Thus, from (68) and (69) we conclude that the coefficient of expansion

$$\frac{\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(t)\right\|}{\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(0)\right\|} = \sqrt{\det\left(\mathbf{Y}_1^{\mathrm{T}}(t)\cdot\mathbf{Y}_1(t)\right)}$$

does not depend on the initial volume but it is an intrinsic quantity of the subspaces defined by the properties of $d_{\mathbf{x}(0)}\Phi^t$. Note that in the particular case of $p = 2N$ the coefficient of expansion is equal to $|\det\mathbf{Y}(t)|$ in accordance to (43). An alternative way of expressing this property is that, for two sets of linearly independent vectors $\{\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0)\}$ and $\{\mathbf{f}_1(0), \mathbf{f}_2(0), \ldots, \mathbf{f}_p(0)\}$ spanning the same p-dimensional subspace of $\mathcal{T}_{\mathbf{x}(0)}\mathcal{S}$, the relation

$$\frac{\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(t)\right\|}{\left\|\bigwedge_{i=1}^{p}\mathbf{w}_i(0)\right\|} = \frac{\left\|\bigwedge_{i=1}^{p}\mathbf{f}_i(t)\right\|}{\left\|\bigwedge_{i=1}^{p}\mathbf{f}_i(0)\right\|} \tag{70}$$

holds [119].

Let us now describe the method for the actual computation of the p-mLCE. Similarly to the computation of the mLCE we fix a small time interval τ and define quantity $X^{(p)}(t)$ (66) as

$$X^{(p)}(k\tau) = \frac{1}{k\tau}\sum_{i=1}^{k}\ln\frac{\|\bigwedge_{j=1}^{p} d_{\mathbf{x}(0)}\Phi^{i\tau}\,\mathbf{w}_j(0)\|}{\|\bigwedge_{j=1}^{p} d_{\mathbf{x}(0)}\Phi^{(i-1)\tau}\,\mathbf{w}_j(0)\|} = \frac{1}{k\tau}\sum_{i=1}^{k}\ln\gamma_i^{(p)}, \tag{71}$$

where $\gamma_i^{(p)}$, $i = 1, 2, \ldots$, is the coefficient of expansion of a p-dimensional volume from $t = (i-1)\tau$ to $t = i\tau$. According to (70) $\gamma_i^{(p)}$ can be computed as the coefficient of expansion of the p-parallelogram defined by any p vectors spanning the same p-dimensional space. A suitable choice for this set is to consider an orthonormal set of vectors $\{\hat{\mathbf{w}}_1((i-1)\tau), \hat{\mathbf{w}}_2((i-1)\tau), \ldots, \hat{\mathbf{w}}_p((i-1)\tau)\}$ giving to (71) the simplified form

$$X^{(p)}(k\tau) = \frac{1}{k\tau}\sum_{i=1}^{k}\ln\gamma_i^{(p)} = \frac{1}{k\tau}\sum_{i=1}^{k}\ln\left\|\bigwedge_{j=1}^{p} d_{\mathbf{x}((i-1)\tau)}\Phi^{\tau}\,\hat{\mathbf{w}}_j((i-1)\tau)\right\|. \tag{72}$$

Thus, from (67) and (72) we get

$$\chi_1^{(p)} = \chi_1 + \chi_2 + \cdots + \chi_p = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \gamma_i^{(p)} \tag{73}$$

for the computation of the p-mLCE. This equation is valid for $1 \leq p \leq 2N$ since in the extreme case of $p = 1$ it is simply reduced to (59) with $\alpha_i \equiv \gamma_i^{(1)}$. In order to estimate the values of χ_i, $i = 1, 2, \ldots, p$, which is our actual goal, we compute from (73) all the $\chi_1^{(p)}$ quantities and evaluate the LCEs from

$$\chi_i = \chi_1^{(i)} - \chi_1^{(i-1)}, \quad i = 2, 3, \ldots, p, \tag{74}$$

with $\chi_1^{(1)} \equiv \chi_1$ [119].

Benettin et al. [14] noted that the p largest 1-LCEs can be evaluated at once by computing the evolution of just p deviation vectors for a particular choice of the orthonormalization procedure, namely performing the Gram-Schmidt orthonormalization method.

Let us discuss the Gram-Schmidt orthonormalization method in some detail. Let $\mathbf{w}_j(i\tau)$, $j = 1, 2, \ldots, p$ be the evolved deviation vectors $\hat{\mathbf{w}}_j((i-1)\tau)$ from time $t = (i-1)\tau$ to $t = i\tau$. From this set of linearly independent vectors we construct a new set of orthonormal vectors $\hat{\mathbf{w}}_j(i\tau)$ from equations

$$\left.\begin{aligned}
&\mathbf{u}_1(i\tau) = \mathbf{w}_1(i\tau), \quad \gamma_{1i} = \|\mathbf{u}_1(i\tau)\|, \quad \hat{\mathbf{w}}_1(i\tau) = \frac{\mathbf{u}_1(i\tau)}{\gamma_{1i}},\\
&\mathbf{u}_2(i\tau) = \mathbf{w}_2(i\tau) - \langle \mathbf{w}_2(i\tau), \hat{\mathbf{w}}_1(i\tau)\rangle \hat{\mathbf{w}}_1(i\tau),\\
&\qquad \gamma_{2i} = \|\mathbf{u}_2(i\tau)\|, \quad \hat{\mathbf{w}}_2(i\tau) = \frac{\mathbf{u}_2(i\tau)}{\gamma_{2i}},\\
&\mathbf{u}_3(i\tau) = \mathbf{w}_3(i\tau) - \langle \mathbf{w}_3(i\tau), \hat{\mathbf{w}}_1(i\tau)\rangle \hat{\mathbf{w}}_1(i\tau) - \langle \mathbf{w}_3(i\tau), \hat{\mathbf{w}}_2(i\tau)\rangle \hat{\mathbf{w}}_2(i\tau),\\
&\qquad \gamma_{3i} = \|\mathbf{u}_3(i\tau)\|, \quad \hat{\mathbf{w}}_3(i\tau) = \frac{\mathbf{u}_3(i\tau)}{\gamma_{3i}},\\
&\qquad\qquad\qquad \vdots
\end{aligned}\right\} \tag{75}$$

which are repeated up to the computation of $\hat{\mathbf{w}}_p(i\tau)$. We remark that $\langle \mathbf{w}, \mathbf{u}\rangle$ denotes the usual inner product of vectors $\mathbf{w}$, $\mathbf{u}$. The general form of the above equations, which is the core of the Gram-Schmidt orthonormalization method, is

$$\begin{aligned}
\mathbf{u}_k(i\tau) =& \mathbf{w}_k(i\tau) - \sum_{j=1}^{k-1} \langle \mathbf{w}_k(i\tau), \hat{\mathbf{w}}_j(i\tau)\rangle \hat{\mathbf{w}}_j(i\tau),\\
\gamma_{ki} =& \|\mathbf{u}_k(i\tau)\|, \quad \hat{\mathbf{w}}_k(i\tau) = \frac{\mathbf{u}_k(i\tau)}{\gamma_{ki}},
\end{aligned} \tag{76}$$

for $1 \leq k \leq p$.

As we will show in Sect. 6.3 the volume of the p-parallelogram having as edges the vectors $d_{\mathbf{x}((i-1)\tau)}\Phi^{\tau}\,\hat{\mathbf{w}}_j((i-1)\tau) = \mathbf{w}_j(i\tau)$, $j = 1, 2, \ldots, p$ is equal to the volume of the p-parallelogram having as edges the vectors $\mathbf{u}_j(i\tau)$, i.e.,

$$\left\| \bigwedge_{j=1}^{p} d_{\mathbf{x}((i-1)\tau)}\Phi^{\tau}\,\hat{\mathbf{w}}_j((i-1)\tau) \right\| = \left\| \bigwedge_{j=1}^{p} \mathbf{u}_j(i\tau) \right\|. \tag{77}$$

Since vectors $\mathbf{u}_j(i\tau)$ are normal to each other, the volume of their p-parallelogram is equal to the product of their norms. This leads to

$$\gamma_i^{(p)} = \left\| \bigwedge_{j=1}^{p} \mathbf{u}_j(i\tau) \right\| = \prod_{j=1}^{p} \gamma_{ji}. \tag{78}$$

Then, (73) takes the form

$$\chi_1^{(p)} = \chi_1 + \chi_2 + \cdots + \chi_p = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \left(\prod_{j=1}^{p} \gamma_{ji} \right).$$

Using now (74) we are able to evaluate the 1-LCE χ_p as

$$\chi_p = \chi_1^{(p)} - \chi_1^{(p-1)} = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \frac{\prod_{j=1}^{p} \gamma_{ji}}{\prod_{j=1}^{p-1} \gamma_{ji}} = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \gamma_{pi}.$$

In conclusion we see that the value of the 1-LCE χ_p with $1 < p \leq 2N$ can be computed as the limiting value, for $t \to \infty$, of the quantity

$$X_p(k\tau) = \frac{1}{k\tau} \sum_{i=1}^{k} \ln \gamma_{pi},$$

i.e.,

$$\chi_p = \lim_{k\to\infty} X_p(k\tau) = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \gamma_{pi}, \tag{79}$$

where γ_{ji}, $j = 1, 2, \ldots, p$, $i = 1, 2, \ldots$ are quantities evaluated during the successive orthonormalization procedures ((75) and (76)). Note that for $p = 1$ (79) is actually (59) with $\alpha_i \equiv \gamma_{1i}$.

6.2 The Numerical Algorithm for the Standard Method

In practice, in order to compute the p largest 1-LCEs with $1 < p \leq 2N$ we follow the evolution of p initially orthonormal deviation vectors $\hat{\mathbf{w}}_j(0) = \mathbf{w}_j(0)$ and every $t = \tau$ time units we replace the evolved vectors $\mathbf{w}_j(k\tau)$, $j = 1, 2, \ldots, p$, $k = 1, 2, \ldots$ by a new set of orthonormal vectors produced by the Gram–Schmidt orthonormalization method (76). During the orthonormalization procedure the quantities γ_{jk} are computed and $\chi_1, \chi_2, \ldots, \chi_p$ are estimated from (79). This algorithm is described in pseudo-code in Table 2 and can be used for the computation of few or even all 1-LCEs. A Fortran code of this algorithm can be found in [144], while [117] contains a similar code developed for the computer algebra platform "Mathematica" (Wolfram Research Inc.).

Let us illustrate the implementation of this algorithm in the particular case of the computation of the two largest LCEs χ_1 and χ_2. As shown in Fig. 8 we start our computation with two orthonormal deviation vectors $\mathbf{w}_1(0)$ and $\mathbf{w}_2(0)$ which are evolved to $\mathbf{w}_1(\tau)$, $\mathbf{w}_2(\tau)$ at $t = \tau$. Then according to the the Gram-Schmidt orthonormalization method (75) we define vectors $\mathbf{u}_1(\tau)$ and $\mathbf{u}_2(\tau)$. In particular,

Table 2 The standard method. The algorithm for the computation of the p largest LCEs $\chi_1, \chi_2, \ldots, \chi_p$ as limits for $t \to \infty$ of quantities $X_1(t), X_2(t), \ldots, X_p(t)$ (71), according to (79). The program computes the evolution of $X_1(t), X_2(t), \ldots, X_p(t)$ with respect to time t up to a given upper value of time $t = T_M$ or until any of the quantities $X_1(t), X_2(t), \ldots, X_p(t)$ attain a very small value, smaller than a low threshold value X_m

Input:	1. Hamilton equations of motion (2) and variational equations (8), or equations of the map (4) and of the tangent map (11). 2. Number of desired LCEs p. 3. Initial condition for the orbit $\mathbf{x}(0)$. 4. Initial *orthonormal* deviation vectors $\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0)$. 5. Renormalization time τ. 6. Maximal time: T_M and minimum allowed value of $X_1(t), X_2(t), \ldots, X_p(t)$: X_m.
Step 1	**Set** the stopping flag, SF $\leftarrow 0$, and the counter, $k \leftarrow 1$.
Step 2	**While** (SF $= 0$) **Do** **Evolve** the orbit and the deviation vectors from time $t = (k-1)\tau$ to $t = k\tau$, i. e. **Compute** $\mathbf{x}(k\tau)$ and $\mathbf{w}_1(k\tau), \mathbf{w}_2(k\tau), \ldots, \mathbf{w}_p(k\tau)$.
Step 3	Perform the **Gram-Schmidt orthonormalization** procedure according to (76): **Do** for $j = 1$ to p **Compute** current vectors $\mathbf{u}_j(k\tau)$ and values of γ_{jk}. **Compute** and **Store** current values of $X_j(k\tau) = \frac{1}{k\tau} \sum_{i=1}^{k} \ln \gamma_{ji}$. **Set** $\mathbf{w}_j(k\tau) \leftarrow \mathbf{u}_j(k\tau)/\gamma_{jk}$. **End Do**
Step 4	**Set** the counter $k \leftarrow k + 1$.

Step 5	**If** $[(k\tau > T_M)$ or (Any of $X_j((k-1)\tau) < X_m$, $j = 1, 2, \ldots, p)]$ **Then**
	Set SF $\leftarrow 1$.
	End If
	End While

Step 6	**Report** the time evolution of $X_1(t), X_2(t), \ldots, X_p(t)$.

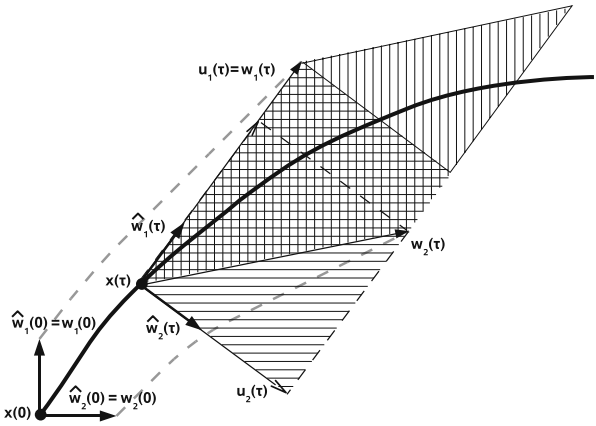

Fig. 8 Numerical scheme for the computation of the two largest LCEs χ_1, χ_2 according to the standard method. The orthonormal deviation vectors $\mathbf{w}_1(0)$, $\mathbf{w}_2(0)$ are evolved according to the variational equations (8) (continuous time) or the equations of the tangent map (11) (discrete time) for $t = \tau$ time units. The evolved vectors $\mathbf{w}_1(\tau)$, $\mathbf{w}_2(\tau)$, are replaced by a set of orthonormal vectors $\hat{\mathbf{w}}_1(\tau)$, $\hat{\mathbf{w}}_2(\tau)$, which span the same 2-dimensional vector space, according to the Gram–Schmidt orthonormalization method (76). Then these vectors are again evolved and the same procedure is iteratively applied. For each successive time interval $[(i-1)\tau, i\tau]$, $i = 1, 2, \ldots$, the quantities $\gamma_{1i} = \|\mathbf{u}_1(i\tau)\|$, $\gamma_{2i} = \|\mathbf{u}_2(i\tau)\|$ are computed and χ_1, χ_2 are estimated from (79)

$\mathbf{u}_1(\tau)$ coincides with $\mathbf{w}_1(\tau)$ while, $\mathbf{u}_2(\tau)$ is the component of vector $\mathbf{w}_2(\tau)$ in the direction perpendicular to vector $\mathbf{u}_1(\tau)$. The norms of these two vectors define the quantities $\gamma_{11} = \|\mathbf{u}_1(\tau)\|$, $\gamma_{21} = \|\mathbf{u}_2(\tau)\|$ needed for the estimation of χ_1, χ_2 from (79). Then vectors $\hat{\mathbf{w}}_1(\tau)$ and $\hat{\mathbf{w}}_2(\tau)$ are defined as unitary vectors in the directions of $\mathbf{u}_1(\tau)$ and $\mathbf{u}_2(\tau)$, respectively. Since the unitary vectors $\hat{\mathbf{w}}_1(\tau)$, $\hat{\mathbf{w}}_2(\tau)$ are normal by construction they constitute the initial set of orthonormal vectors for the next iteration of the algorithm. From Fig. 8 we easily see that the parallelograms defined by vectors $\mathbf{w}_1(\tau)$, $\mathbf{w}_2(\tau)$ and by vectors $\mathbf{u}_1(\tau)$ and $\mathbf{u}_2(\tau)$ have the same area. This equality corresponds to the particular case $p = 2$, $i = 1$ of (77). Evidently, since vectors $\mathbf{u}_1(\tau)$, $\mathbf{u}_2(\tau)$ are perpendicular to each other, we have $\mathrm{vol}_2(\mathbf{u}_1(\tau), \mathbf{u}_2(\tau)) = \gamma_{11}\gamma_{21}$ in accordance to (78).

6.3 Connection Between the Standard Method and the QR Decomposition

Let us rewrite (75) of the Gram-Schmidt orthonormalization procedure, by solving them with respect to $\mathbf{w}_j(i\tau)$, $j = 1, 2, \ldots, p$, with $1 < p \leq 2N$

$$\begin{aligned}
\mathbf{w}_1(i\tau) &= \gamma_{1i}\hat{\mathbf{w}}_1(i\tau) \\
\mathbf{w}_2(i\tau) &= \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_2(i\tau)\rangle \hat{\mathbf{w}}_1(i\tau) + \gamma_{2i}\hat{\mathbf{w}}_2(i\tau) \\
\mathbf{w}_3(i\tau) &= \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_3(i\tau)\rangle \hat{\mathbf{w}}_1(i\tau) + \langle \hat{\mathbf{w}}_2(i\tau), \mathbf{w}_3(i\tau)\rangle \hat{\mathbf{w}}_2(i\tau) + \gamma_{3i}\hat{\mathbf{w}}_3(i\tau) \\
&\vdots
\end{aligned} \tag{80}$$

and get the general form

$$\mathbf{w}_k(i\tau) = \sum_{j=1}^{k-1} \langle \hat{\mathbf{w}}_j(i\tau), \mathbf{w}_k(i\tau) \rangle \hat{\mathbf{w}}_j(i\tau) + \gamma_{ki} \hat{\mathbf{w}}_k(i\tau), \quad k = 1, 2, \ldots, p.$$

This set of equations can be rewritten in matrix form as follows:

$$\left[\mathbf{w}_1(i\tau)\ \mathbf{w}_2(i\tau) \cdots \mathbf{w}_p(i\tau) \right] = \left[\hat{\mathbf{w}}_1(i\tau)\ \hat{\mathbf{w}}_2(i\tau) \cdots \hat{\mathbf{w}}_p(i\tau) \right] \cdot$$
$$\cdot \begin{bmatrix} \gamma_{1i} & \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_2(i\tau) \rangle & \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_3(i\tau) \rangle & \cdots & \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_p(i\tau) \rangle \\ 0 & \gamma_{2i} & \langle \hat{\mathbf{w}}_2(i\tau), \mathbf{w}_3(i\tau) \rangle & \cdots & \langle \hat{\mathbf{w}}_2(i\tau), \mathbf{w}_p(i\tau) \rangle \\ 0 & 0 & \gamma_{3i} & \cdots & \langle \hat{\mathbf{w}}_3(i\tau), \mathbf{w}_p(i\tau) \rangle \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & \gamma_{pi} \end{bmatrix}.$$

So the $2N \times p$ matrix $\mathbf{W}(i\tau) = \left[\mathbf{w}_1(i\tau)\ \mathbf{w}_2(i\tau) \cdots \mathbf{w}_p(i\tau) \right]$, having as columns the linearly independent deviation vectors $\mathbf{w}_j(i\tau)$, $j = 1, 2, \ldots, p$ is written as a product of the $2N \times p$ matrix $\mathbf{Q} = \left[\hat{\mathbf{w}}_1(i\tau)\ \hat{\mathbf{w}}_2(i\tau) \cdots \hat{\mathbf{w}}_p(i\tau) \right]$, having as columns the coordinates of the orthonormal vectors $\hat{\mathbf{w}}_j(i\tau)$, $j = 1, 2, \ldots, p$ and satisfying $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_p$, and of an upper triangular $p \times p$ matrix $\mathbf{R}(i\tau)$ with positive diagonal elements

$$\mathbf{R}_{jj}(i\tau) = \gamma_{ji}, \quad j = 1, 2, \ldots, p, \quad i = 1, 2, \ldots.$$

From (80) we easily see that $\langle \hat{\mathbf{w}}_j(i\tau), \mathbf{w}_j(i\tau) \rangle = \gamma_{ji}$ and so matrix $\mathbf{R}(i\tau)$ can be also expressed as

$$\mathbf{R}(i\tau) = \begin{bmatrix} \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_1(i\tau) \rangle & \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_2(i\tau) \rangle & \cdots & \langle \hat{\mathbf{w}}_1(i\tau), \mathbf{w}_p(i\tau) \rangle \\ 0 & \langle \hat{\mathbf{w}}_2(i\tau), \mathbf{w}_2(i\tau) \rangle & \cdots & \langle \hat{\mathbf{w}}_2(i\tau), \mathbf{w}_p(i\tau) \rangle \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \langle \hat{\mathbf{w}}_p(i\tau), \mathbf{w}_p(i\tau) \rangle \end{bmatrix}.$$

The above procedure is the so-called QR decomposition of a matrix. In practice, we proved by actually constructing the $\mathbf{Q}$ and $\mathbf{R}$ matrices via the Gram-Schmidt orthonormalization method, the following theorem.

Theorem 4. *Let $\boldsymbol{A}$ be an $n \times m$ ($n \geq m$) matrix with linearly independent columns. Then $\boldsymbol{A}$ can be uniquely factorized as*

$$\boldsymbol{A} = \boldsymbol{Q} \cdot \boldsymbol{R},$$

where $\boldsymbol{Q}$ is an $n \times m$ matrix with orthogonal columns, satisfying $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I}_m$ and $\boldsymbol{R}$ is an $m \times m$ invertible upper triangular matrix with positive diagonal entries.

Although we presented the QR decomposition through the Gram–Schmidt orthonormalization procedure this decomposition can also be achieved by others, computationally more efficient techniques like for example the Householder transformation [62] [107, Sect. 2.10].

Observing that the quantities γ_{ji}, $j = 1, 2 \ldots, p$, $i = 1, 2 \ldots$, needed for the evaluation of the LCEs through (79) are the diagonal elements of $\mathbf{R}(i\tau)$ we can implement a variant of the standard method for the computation on the LCEs, which is based on the QR decomposition procedure [44, 62, 36, 40]. Similarly to the procedure followed in Sect. 6.2, in order to compute the p ($1 < p \leq 2N$) largest LCEs we follow the evolution of p initially orthonormal deviation vectors $\hat{\mathbf{w}}_j(0) = \mathbf{w}_j(0)$, $j = 1, 2 \ldots, p$, which can be considered as columns of a $2N \times p$ matrix $\mathbf{Q}(0)$. Every $t = \tau$ time units the matrix $\mathbf{W}(i\tau)$, $i = 1, 2, \ldots$, having as columns the deviation vectors

$$d_{\mathbf{x}((i-1)\tau)}\Phi^{\tau}\,\hat{\mathbf{w}}_j((i-1)\tau) = \mathbf{w}_j(i\tau), \quad j = 1, 2, \ldots, p,$$

i.e., the columns of $\mathbf{Q}((i-1)\tau)$ evolved in time interval $[(i-1)\tau, i\tau]$ by the action of $d_{\mathbf{x}((i-1)\tau)}\Phi^{\tau}$, undergoes the QR decomposition procedure

$$\mathbf{W}(i\tau) = \mathbf{Q}(i\tau) \cdot \mathbf{R}(i\tau) \tag{81}$$

and the new $\mathbf{Q}(i\tau)$ is again evolved for the next time interval $[i\tau, (i+1)\tau]$, and so on and so forth. Then the LCEs are estimated from the values of the diagonal elements of matrix $\mathbf{R}(i\tau)$ as

$$\chi_p = \lim_{k\to\infty} \frac{1}{k\tau} \sum_{i=1}^{k} \ln \mathbf{R}_{pp}(i\tau). \tag{82}$$

The corresponding algorithm is presented in pseudo-code in Table 3. From the above-presented analysis it becomes evident that the standard method developed by Shimada and Nagashima [119] and Benettin et al. [14] for the computation of the LCEs is practically a QR decomposition procedure performed by the Gram–Schmidt orthonormalization method, although the authors of these papers formally do not refer to the QR decomposition. We note that both the standard method and the QR decomposition technique presented here can be used for the computation of part ($p < 2N$) or of the whole ($p = 2N$) spectrum of LCEs.

As a final remark on the QR decomposition technique let us show the validity of (77) by considering the QR decomposition of matrix $\mathbf{W}(i\tau)$ (81). According to (105) and (106) we have

Table 3 Discrete QR decomposition. The algorithm for the computation of the p largest LCEs $\chi_1, \chi_2, \ldots, \chi_p$ according to the QR decomposition method. The program computes the evolution of $X_1(t), X_2(t), \ldots, X_p(t)$ with respect to time t up to a given upper value of time $t = T_M$ or until any of the these quantities becomes smaller than a low threshold value X_m

Input:	1. Hamilton equations of motion (2) and variational equations (8), or equations of the map (4) and of the tangent map (11). 2. Number of desired LCEs p. 3. Initial condition for the orbit $\mathbf{x}(0)$. 4. Initial matrix $\mathbf{Q}(0)$ having as columns *orthonormal* deviation vectors $\mathbf{w}_1(0), \mathbf{w}_2(0), \ldots, \mathbf{w}_p(0)$. 5. Time interval τ between successive QR decompositions. 6. Maximal time: T_M and minimum allowed value of $X_1(t)$, $X_2(t), \ldots, X_p(t)$: X_m.
Step 1	**Set** the stopping flag, SF $\leftarrow 0$, and the counter, $k \leftarrow 1$.
Step 2	**While** (SF $= 0$) **Do** **Evolve** the orbit and the matrix $\mathbf{Q}((k-1)\tau)$ from time $t = (k-1)\tau$ to $t = k\tau$, i. e. **Compute** $\mathbf{x}(k\tau)$ and $\mathbf{W}(i\tau)$.
Step 3	Perform the **QR decomposition** of $\mathbf{W}(i\tau)$ according to (81): **Compute** $\mathbf{Q}(k\tau)$ and $\mathbf{R}(k\tau)$. **Compute** and **Store** current values of $X_j(k\tau) = \frac{1}{k\tau}\sum_{i=1}^{k} \ln \mathbf{R}_{jj}(i\tau)$, $j = 1, 2 \ldots, p$.
Step 4	**Set** the counter $k \leftarrow k + 1$.

Step 5	**If** [$(k\tau > T_M)$ or (Any of $X_j((k-1)\tau) < X_m$, $j = 1, 2, \ldots, p$)] **Then**
	Set SF $\leftarrow 1$.
	End If
	End While

Step 6	**Report** the time evolution of $X_1(t), X_2(t), \ldots, X_p(t)$.

$$
\begin{aligned}
\left\| \bigwedge_{j=1}^{p} \mathbf{w}_j(i\tau) \right\| &= \sqrt{\det\left(\mathbf{W}^{\mathrm{T}}(i\tau)\cdot\mathbf{W}(i\tau)\right)} \\
&= \sqrt{\det\left(\mathbf{R}^{\mathrm{T}}(i\tau)\cdot\mathbf{Q}^{\mathrm{T}}(i\tau)\cdot\mathbf{Q}(i\tau)\cdot\mathbf{R}(i\tau)\right)} \\
&= \sqrt{\det\mathbf{R}^{\mathrm{T}}(i\tau)\det\mathbf{R}(i\tau)} = |\det\mathbf{R}(i\tau)| \\
&= \prod_{j=1}^{p}\gamma_{ji} = \prod_{j=1}^{p}\left\|\mathbf{u}_j(i\tau)\right\| = \left\| \bigwedge_{j=1}^{p} \mathbf{u}_j(i\tau) \right\|,
\end{aligned}
$$

where the identities $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_p$ and $\det\mathbf{R}(i\tau) = \prod_{j=1}^{p}\gamma_{ji}$ have been used.

6.4 Other Methods for Computing LCEs

Over the years several methods have been proposed and applied for computing the numerical values of the LCEs. The standard method we discussed so far is the first and probably the simplest method to address this problem. As we showed in Sect. 6.3 the standard method, which requires successive applications of the

Gram-Schmidt orthonormalization procedure, is practically equivalent to the QR decomposition technique.

The reorthonormalization of deviation vectors plays an indispensable role for computing the LCEs and the corresponding methods can be distinguished in discrete and continuous methods. The *discrete methods* iteratively approximate the LCEs in a finite number of (discrete) time steps and therefore apply to both continuous and discrete dynamical systems [62, 36, 40]. The standard method and its QR decomposition version are discrete methods. A method is called *continuous* when all relevant quantities are obtained as solutions of certain ordinary differential equations, which maintain orthonormality of deviation vectors continuously. Therefore such methods can only be formulated for continuous dynamical systems and not for maps. The use of continuous orthonormalization for the numerical computation of LCEs was first proposed by Goldhirsch et al. [63] and afterward developed by several authors [67, 62, 36, 40, 26, 110, 109, 94, 38].

Discrete and continuous methods are based on appropriate decomposition of matrices performed usually by the QR decomposition or by the SVD procedure. The discrete QR decomposition which has already been presented in Sect. 6.3 is the most frequently used method and has proved to be quite efficient and reliable. The continuous QR decomposition and methods based on the SVD procedure are discussed in some detail at the end of the current section.

Variants of these techniques have been also proposed by several authors. Let us briefly refer to some of them. Rangarajan et al. [110] introduced a method for the computation of part or of the whole spectrum of LCEs for continuous dynamical systems, which does not require rescaling and renormalization of vectors. The key feature of their approach is the use of explicit group theoretical representations of orthogonal matrices, which leads to a set of coupled ordinary differential equations for the LCEs along with the various angles parameterizing the orthogonal matrices involved in the process. Ramasubramanian and Sriram [109] showed that the method is competitive with the standard method and the continuous QR decomposition.

Carbonell et al. [20] proposed a method for the evaluation of the whole spectrum of LCEs by approximating the differential equations describing the evolution of an orbit of a continuous dynamical system and their associated variational equations by two piecewise linear sets of ordinary differential equations. Then an SVD or a QR decomposition-based method is applied to these two new sets of equations, allowing us to obtain approximations of the LCEs of the original system. An advantage of this method is that it does not require the simultaneous integration of the two sets of piecewise linear equations.

Lu et al. [94] proposed a new continuous method for the computation of few or of all LCEs, which is related to the QR decomposition technique. According to their method one follows the evolution of orthogonal vectors, similarly to the QR method, but does not require them to be necessarily orthonormal. By relaxing the length requirement Lu et al. [94] established a set of recursive differential equations for the evolution of these vectors. Using symplectic Runge–Kutta integration schemes for the evolution of these vectors they succeeded in preserving automatically the

orthogonality between any two successive vectors. Normalization of vectors occurs whenever the magnitude of any vector exceeds given lower or upper bounds.

Chen et al. [24] proposed a simple discrete QR algorithm for the computation of the whole spectrum of LCEs of a continuous dynamical system. Their method is based on a suitable approximation of the solution of variational equations by assuming that the Jacobian matrix remains constant over small integration time steps. Thus, the scheme requires the numerical solution of the $2N$ equations of motion but not the solution of the $(2N)^2$ variational equations since their solution is approximated by an explicit expression involving the computed orbit. This approach led to a computationally fast evaluation of the LCEs for various multidimensional dynamical systems studied in [24].

It is worth mentioning here a completely different approach, with respect to the above-mentioned techniques, which was developed at the early 1980s. In particular, Frøyland proposed in [60] an algorithm for the computation of LCEs, which he claimed to be quite efficient in the case of low-dimensional systems, and applied it to the Lorenz system [61]. The basic idea behind this algorithm is the implementation of appropriate differential equations describing the time evolution of volume elements around the orbits of the dynamical system, instead of defining these volumes through deviation vectors whose evolution is governed by the usual variational equations (8).

Apart from the actual numerical computation of the values of the LCEs, methods for the theoretical estimation of those values have been also developed. For example, Li and Chen [90] provided a theorem for the estimation of lower and upper bounds for the values of all LCEs in the case of discrete maps. These results were also generalized for the case of continues dynamical systems [91]. The validity of these estimates was demonstrated by a comparison between the estimated bounds and the numerically computed spectrum of LCEs of some specific dynamical systems [90, 91].

Finally, let us refer to a powerful analytical method which allows one to verify the existence of positive LCEs for a dynamical system, the so-called *cone technique*. The method was suggested by Wojtkowski [142] and has been extensively applied for the study of chaotic billiards [142, 143, 43, 97] and geodesic flows [41, 42, 19]. A concise description of the techniques can also be found in [7] [25, Sect. 3.13]. Considering the space $\mathbb{R}^n$ a *cone* $\mathcal{C}_\gamma$, with $\gamma > 0$, centered around $\mathbb{R}^{n-k}$ is

$$\mathcal{C}_\gamma = \left\{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^k \times \mathbb{R}^{n-k} : \|\mathbf{u}\| < \gamma \|\mathbf{v}\|\right\} \cup (\mathbf{0}, \mathbf{0}). \tag{83}$$

Note that $\{\mathbf{0}\} \times \mathbb{R}^{n-k} \subset \mathcal{C}_\gamma$ for every γ. In the particular case of $n = 3$, $k = 2$, $\mathcal{C}_\gamma$ corresponds to the usual 3-dimensional cone, while in the case of the plane ($n = 2$) a cone $\mathcal{C}_\gamma$ around an axis L is the set of vectors of $\mathbb{R}^2$ that make angle $\phi < \arctan \gamma$ with the line L. In the case of Hamiltonian systems (and symplectic maps) a cone can get the simple form $\delta\mathbf{q} \cdot \delta\mathbf{p} > 0$. Finding an invariant family of cones (83) in $\mathcal{T}_\mathbf{x}\mathcal{S}$, which are mapped strictly into themselves by $d_\mathbf{x}\Phi^t$, guarantees that the values of the $n - k$ largest LCEs are positive [142, 143]. We emphasize that the cone technique is not used for the explicit numerical computation of the LCEs, but for the analytical

proof of the existence of positive LCEs, providing at the same time some bounds for their actual values.

6.4.1 Continuous QR Decomposition Methods

The QR decomposition methods allow the computation of all or of the p ($1 < p < 2N$) largest LCEs. Let us discuss in more detail the developed procedure for both cases following mainly [62, 36, 94].

Computing the complete spectrum of LCEs

The basic idea of the method is to avoid directly solving the differential equation (10), by requiring $\mathbf{Y}(t) = \mathbf{Q}(t)\mathbf{R}(t)$ where $\mathbf{Q}(t)$ is orthogonal and $\mathbf{R}(t)$ is upper triangular with positive diagonal elements, according to Theorem 4. With this decomposition, one can write (10) into the form

$$\mathbf{Q}^{\mathrm{T}}\dot{\mathbf{Q}} + \dot{\mathbf{R}}\mathbf{R}^{-1} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q},$$

where, for convenience, we dropped out the explicit dependence of the matrices on time t, i.e., $\mathbf{Q}(t) \equiv \mathbf{Q}$. Since $\mathbf{Q}^{\mathrm{T}}\dot{\mathbf{Q}}$ is skew and $\dot{\mathbf{R}}\mathbf{R}^{-1}$ is upper triangular, one reads off the differential equations

$$\dot{\mathbf{Q}} = \mathbf{Q}\mathbf{S}, \tag{84}$$

where $\mathbf{S}$ is the skew-symmetric matrix

$$\mathbf{S} = \mathbf{Q}^{\mathrm{T}}\dot{\mathbf{Q}}$$

with elements

$$\mathbf{S}_{ij} = \begin{cases} (\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})_{ij} & i > j \\ 0 & i = j \\ -(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})_{ji} & i < j \end{cases}, \quad i, j = 1, 2, \ldots, 2N, \tag{85}$$

and

$$\frac{\dot{\mathbf{R}}_{pp}}{\mathbf{R}_{pp}} = (\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})_{pp}, \quad p, = 1, 2, \ldots, 2N \tag{86}$$

where $\mathbf{R}_{pp}$ are the diagonal elements of $\mathbf{R}$. As we have already seen in (82) the LCEs are related to the elements $\mathbf{R}_{pp}$, through

$$\chi_p = \lim_{t\to\infty} \frac{1}{t} \ln \mathbf{R}_{pp}(t).$$

Thus, in order to compute the spectrum of LCEs only (84) and (86) have to be solved simultaneously with the equations of motion (2). In practice, the knowledge of matrix $\mathbf{R}$ is not necessary for the actual computation of the LCEs. Noticing that

$$\frac{d}{dt}\left(\ln \mathbf{R}_{pp}\right) = \frac{\dot{\mathbf{R}}_{pp}}{\mathbf{R}_{pp}} = (\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})_{pp} = \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p, \tag{87}$$

where $\mathbf{q}_p$ is the pth column vector of $\mathbf{Q}$, we can compute the LCEs using

$$\chi_p = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt.$$

In practice, the LCEs can be estimated through a recursive formula. Let

$$X_p(k\tau) = \frac{1}{k\tau} \int_0^{k\tau} \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt.$$

Then we have

$$\begin{aligned} X_p\left((k+1)\tau\right) &= \frac{1}{(k+1)\tau} \int_0^{(k+1)\tau} \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt \\ &= \frac{1}{(k+1)\tau} \int_0^{k\tau} \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt + \frac{1}{(k+1)\tau} \int_{k\tau}^{(k+1)\tau} \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt. \end{aligned}$$

Replacing the first integral with $k\tau X_p(k\tau)$ we get

$$X_p\left((k+1)\tau\right) = \frac{k}{k+1} X_p(k\tau) + \frac{1}{(k+1)\tau} \int_{k\tau}^{(k+1)\tau} \mathbf{q}_p \cdot \mathbf{A}\mathbf{q}_p dt, \tag{88}$$

and

$$\chi_p = \lim_{k\to\infty} X_p(k\tau). \tag{89}$$

The basic difference between the discrete QR decomposition method presented in Sect. 6.3, and the continuous QR method presented here, is that in the first method the orthonormalization is performed numerically at discrete time steps, while the latter method seeks to maintain the orthogonality via solving differential equations that encode the orthogonality continuously.

Computation of the $p > 1$ largest LCEs

If we want to compute the p largest LCEs, with $1 < p < 2N$, we change (10) to

$$\dot{\mathbf{Y}}(t) = \mathbf{A}(t)\,\mathbf{Y}(t) \ , \quad \text{with } \mathbf{Y}(0)^{\mathrm{T}}\mathbf{Y}(0) = \mathbf{I}_p, \tag{90}$$

where $\mathbf{Y}(t)$ is in practice, the $2N \times p$ matrix having as columns the p deviation vectors $\mathbf{w}_1(t), \mathbf{w}_2(t), \ldots, \mathbf{w}_p(t)$. Applying Theorem 4 we get $\mathbf{Y}(t) = \mathbf{Q}(t)\mathbf{R}(t)$ where $\mathbf{Q}(t)$ is orthogonal so that the identity $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ holds but not the $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}$. Then from (90) we get

$$\dot{\mathbf{R}} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} - \mathbf{S}\right)\mathbf{R},$$

where $\mathbf{S} = \mathbf{Q}^{\mathrm{T}}\dot{\mathbf{Q}}$ is a $p \times p$ matrix whose elements are given by (85) for $i, j = 1, 2, \ldots, p$. Since $\mathbf{R}$ is invertible, from the relations

$$\dot{\mathbf{R}}\mathbf{R}^{-1} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} - \mathbf{S}$$

and

$$\dot{\mathbf{Q}} = \mathbf{A}\mathbf{Q} - \mathbf{Q}\dot{\mathbf{R}}\mathbf{R}^{-1}, \tag{91}$$

we obtain

$$\dot{\mathbf{Q}} = \left(\mathbf{A} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A} + \mathbf{Q}\mathbf{S}\mathbf{Q}^{\mathrm{T}}\right)\mathbf{Q},$$

or

$$\dot{\mathbf{Q}} = \mathbf{H}(\mathbf{Q}, t)\mathbf{Q}, \tag{92}$$

with

$$\mathbf{H}(\mathbf{Q}, t) = \mathbf{A} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A} + \mathbf{Q}\mathbf{S}\mathbf{Q}^{\mathrm{T}}.$$

Notice that the matrix $\mathbf{H}(\mathbf{Q}, t)$ in not necessarily skew-symmetric, and the term $\mathbf{Q}\mathbf{Q}^{\mathrm{T}}$ is responsible for lack of skew-symmetry in $\mathbf{H}$. Of course for $p = 2N$ (92) reduces to equation $\dot{\mathbf{Q}} = \mathbf{Q}\mathbf{S}$ (84). The evolution of the diagonal elements of $\mathbf{R}$ are again governed by (86), but for $p < 2N$, and so the p largest LCEs can be computed again from (87, 88, 89).

The main difference with respect to the case of the computation of the whole spectrum is the numerical difficulties arising in solving (92), since $\mathbf{H}$ is not skew-symmetric as was matrix $\mathbf{S}$ in (84). Due to this difference usual numerical integration techniques fail to preserve the orthogonality of matrix $\mathbf{Q}$.

A central observation of [36] is that the matrix $\mathbf{H}$ has a weak skew-symmetry property. The matrix $\mathbf{H}$ is called weak skew-symmetric if

$$\mathbf{Q}^{\mathrm{T}}\left(\mathbf{H}(\mathbf{Q}, t) + \mathbf{H}^{\mathrm{T}}(\mathbf{Q}, t)\right)\mathbf{Q} = \mathbf{0}, \quad \text{whenever } \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_p.$$

A matrix $\mathbf{H}$ is said to be strongly skew-symmetric if it is skew-symmetric, i.e., $\mathbf{H}^{\mathrm{T}} = -\mathbf{H}$. Christiansen and Rugh [26] proposed a method according to which, the numerically unstable equations (91) for the continuous orthonormalization could be

stabilized by the addition of an appropriate dissipation term. This idea was also used in [18], where it was shown that it is possible to reformulate (92) so that **H** becomes strongly skew-symmetric and thus, achieve a numerically stable algorithm for the computation of few LCEs.

6.4.2 Discrete and Continuous Methods Based on the SVD Procedure

An alternative way of evaluating the LCEs is obtained by applying the SVD procedure on the fundamental $2N \times 2N$ matrix $\mathbf{Y}(t)$, which defines the evolution of deviation vectors through (9) and (12) for continuous and discrete systems, respectively. According to the SVD algorithm a $2N \times p$ matrix ($p \leq 2N$) **B** can be written as the product of a $2N \times p$ column-orthogonal matrix **U**, a $p \times p$ diagonal matrix **F** with positive or zero elements σ_i, $i = 1, \ldots, p$ (the so-called *singular values*), and the transpose of a $p \times p$ orthogonal matrix **V**:

$$\mathbf{B} = \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^{\mathrm{T}}.$$

We note that matrices **U** and **V** are orthogonal so that

$$\mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} = \mathbf{V}^{\mathrm{T}} \cdot \mathbf{V} = \mathrm{I}_p. \tag{93}$$

For a more detailed description of the SVD method, as well as an algorithm for its implementation the reader is referred to [107, Sect. 2.6] and references therein. The SVD is unique up to permutations of corresponding columns, rows, and diagonal elements of matrices **U**, **V**, and **F** respectively. Advanced numerical techniques for the computation of the singular values of a product of many matrices can be found for example in [130, 101].

So, for the purposes of our study let

$$\mathbf{Y} = \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^{\mathrm{T}}, \tag{94}$$

where we dropped out as before, the explicit dependence of the matrices on time t. In those cases where all singular values are different, a unique decomposition can be achieved by the additional request of a strictly monotonically decreasing singular value spectrum, i.e., $\sigma_1(t) > \sigma_2(t) > \cdots > \sigma_{2N}(t)$. Multiplying (94) with the transpose

$$\mathbf{Y}^{\mathrm{T}} = \mathbf{V} \cdot \mathbf{F}^{\mathrm{T}} \cdot \mathbf{U}^{\mathrm{T}},$$

from the left we get

$$\mathbf{Y}^{\mathrm{T}} \cdot \mathbf{Y} = \mathbf{V} \cdot \mathbf{F}^{\mathrm{T}} \cdot \mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^{\mathrm{T}} = \mathbf{V} \cdot \mathrm{diag}(\sigma_i^2(t)) \cdot \mathbf{V}^{\mathrm{T}}, \tag{95}$$

where (93) has been used. From (95) we see that the eigenvalues of the diagonal matrix $\mathrm{diag}(\sigma_i^2(t))$, i.e., the squares of the singular values of $\mathbf{Y}(t)$, are equal to the

eigenvalues of the symmetric matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. Then from point 4 of the MET we conclude that the LCEs are related to the singular values of $\mathbf{Y}(t)$ through [62, 130]

$$\chi_p = \lim_{t\to\infty} \frac{1}{t} \ln \sigma_i(t), \quad p = 1, 2, \ldots, 2N,$$

which implies that the LCEs can be evaluated as the limits for $t \to \infty$ of the time rate of the logarithms of the singular values.

Theoretical aspects of the SVD technique, as well as a detailed study of its ability to approximate the spectrum of LCEs can be found in [101, 37, 38]. Continuous [67, 62, 39] and discrete [130] versions of the SVD algorithm have been applied for the computation of few or of all LCEs, although this approach is not widely used. A basic problem of these methods is that they fail to compute the spectrum of LCEs if it is degenerate, i.e., when two or more LCEs are equal or very close to each other, due to the appearance of ill-conditioned matrices.

7 Chaos Detection Techniques

A simple, qualitative way of studying the dynamics of a Hamiltonian system is by plotting the successive intersections of its orbits with a Poincaré surface of section (PSS) (e.g., [72] [92, pp. 17–20]). Similarly, in the case of symplectic maps one simply plots the phase space of the system. This qualitative method has been extensively applied to 2d maps and to 2D Hamiltonians, since in these systems the PSS is a 2-dimensional plane. In such systems one can visually distinguish between regular and chaotic orbits since the points of a regular orbit lie on a torus and form a smooth closed curve, while the points of a chaotic orbit appear randomly scattered. In 3D Hamiltonian systems (or 4d symplectic maps), however, the PSS (or the phase space) is 4-dimensional and the behavior of the orbits cannot be easily visualized. Things become even more difficult and deceiving for multidimensional systems. One way to overcome this problem is to project the PSS (or the phase space) to spaces with lower dimensions (see, e.g., [139, 140, 105]) although these projections are often very complicated and difficult to interpret. Thus, we need fast and accurate numerical tools to give us information about the regular or chaotic character of orbits, mainly when the dynamical system has many degrees of freedom.

The most commonly employed method for distinguishing between regular and chaotic behavior is the evaluation of the mLCE χ_1, because if $\chi_1 > 0$ the orbit is chaotic. The main problem of using the value of χ_1 as an indicator of chaoticity is that, in practice, the numerical computation may take a huge amount of time, in particular for orbits which stick to regular ones for a long time before showing their chaotic behavior. Since χ_1 is defined as the limit for $t \to \infty$ of the quantity $X_1(t)$ (54), the time needed for $X_1(t)$ to converge to its limiting value is not known a priori and may become extremely long. Nevertheless, we should keep in mind that the mLCE gives us more information than just characterizing an orbit as regular or

chaotic, since it also quantifies the notion of chaoticity by providing a characteristic timescale for the studied dynamical system, namely the Lyapunov time (51).

In order to address the problem of the fast and reliable determination of the regular or chaotic nature of orbits, several methods have been developed over the years with varying degrees of success. These methods can be divided in two major categories: Some are based on the study of the evolution of deviation vectors from a given orbit, like the computation of χ_1, while others rely on the analysis of the particular orbit itself.

Among other chaoticity detectors, belonging to the same category with the evaluation of the mLCE, are the fast Lyapunov indicator (FLI) [58, 59, 56, 89, 49, 69] and its variants [4, 5], the smaller alignment index (SALI) [122, 124, 125] and its generalization, the so-called generalized alignment index (GALI) [126, 127], the mean exponential growth of nearby orbits (MEGNO) [28, 29], the relative Lyapunov indicator (RLI) [115, 116], the average power law exponent (APLE) [95], as well as methods based on the study of spectra of quantities related to the deviation vectors like the stretching numbers [57, 93, 135, 138], the helicity angles (the angles of deviation vectors with a fixed direction) [32], the twist angles (the differences of two successive helicity angles) [33], or the study of the differences between such spectra [88, 136].

In the category of methods based on the analysis of a time series constructed by the coordinates of the orbit under study, one may list the frequency map analysis of Laskar [83, 86, 84, 85], the "0–1" test [64, 65], the method of the low-frequency spectral analysis [137, 81], the "patterns method" [120, 121], the recurrence plots technique [147, 148], and the information entropy index [100]. One could also refer to several ideas presented by various authors that could be used in order to distinguish between chaoticity and regularity, like the differences appearing for regular and chaotic orbits in the time evolutions of their correlation dimension [50], in the time averages of kinetic energies related to the virial theorem [74], and in the statistical properties of the series of time intervals between successive intersections of orbits with a PSS [80].

A systematic and detailed comparative study of the efficiency and reliability of the various chaos detection techniques has not been done yet, although comparisons between some of the existing methods have been performed sporadically in studies of particular dynamical systems [122, 125, 132, 133, 82, 95, 6].

Let us now focus our attention on the behavior of the FLI and of the GALI and on their connection to the LCEs. The FLI was introduced as an indicator of chaos in [58, 59] and after some minor modifications in its definition, it was used for the distinction between resonant and not resonant regular motion [56, 49]. The FLI is defined as

$$\mathrm{FLI}(t) = \sup_t \ln \|\mathbf{w}(t)\|,$$

where $\mathbf{w}(t)$ is a deviation vector from the studied orbit at point $\mathbf{x}(t)$, which initially had unit norm, i.e., $\|\mathbf{w}(0)\| = 1$. In practice, $\mathrm{FLI}(t)$ registers the maximum length

that an initially unitary deviation vector attains from the beginning of its evolution up to the current time t. Using the notation appearing in (59), the FLI can be computed as

$$\mathrm{FLI}(k\tau) = \sup_k \sum_{i=1}^{k} \ln \frac{D_i}{D_0} = \sup_k \sum_{i=1}^{k} \ln \alpha_i,$$

with the initial norm D_0 of the deviation vector being $D_0 = 1$.

According to (62) the norm of $\mathbf{w}(t)$ increases linearly in time in the case of regular orbits. On the other hand, in the case of chaotic orbits the norm of any deviation vector exhibits an exponential increase in time, with an exponent which approximates χ_1 for $t \to \infty$. Thus, the norm of a deviation vector reaches rapidly completely different values for regular and chaotic orbits, which actually differ by many orders of magnitude. This behavior allows FLI to discriminate between regular orbits, for which FLI has relatively small values, and chaotic orbits, for which FLI gets very large values.

The main difference of FLI with respect to the evaluation of the mLCE by (59) is that FLI registers the *current* value of the norm of the deviation vector and does not try to compute the limit value, for $t \to \infty$, of the mean of stretching numbers as χ_1 does. By dropping the time average requirement of the stretching numbers, FLI succeeds in determining the nature of orbits faster than the computation of the mLCE.

The generalized alignment index of order p (GALI_p) is determined through the evolution of $2 \leq p \leq 2N$ initially linearly independent deviation vectors $\mathbf{w}_i(0)$, $i = 1, 2, \ldots, p$ and so it is more related to the computation of many LCEs than to the computation of the mLCE. The evolved deviation vectors $\mathbf{w}_i(t)$ are normalized from time to time in order to avoid overflow problems, but their directions are left intact. Then, according to [126] GALI_p is defined to be the volume of the p-parallelogram having as edges the p unitary deviation vectors $\hat{\mathbf{w}}_i(t)$, $i = 1, 2, \ldots, p$

$$\mathrm{GALI}_p(t) = \|\hat{\mathbf{w}}_1(t) \wedge \hat{\mathbf{w}}_2(t) \wedge \cdots \wedge \hat{\mathbf{w}}_p(t)\|. \tag{96}$$

In [126] the value of GALI_p is computed according to (105), while in [2, 127] a more efficient numerical technique based on the SVD algorithm is applied. From the definition of GALI_p it becomes evident that if at least two of the deviation vectors become linearly dependent, the wedge product in (96) becomes zero and the GALI_p vanishes.

In the case of a chaotic orbit all deviation vectors tend to become linearly dependent, aligning in the direction which corresponds to the mLCE and GALI_p tends to zero exponentially following the law [126]:

$$\mathrm{GALI}_p(t) \sim e^{-[(\chi_1-\chi_2)+(\chi_1-\chi_3)+\cdots+(\chi_1-\chi_p)]t},$$

where $\chi_1, \chi_2, \ldots, \chi_p$ are the p largest LCEs. On the other hand, in the case of regular motion all deviation vectors tend to fall on the N-dimensional tangent space of the torus on which the motion lies. Thus, if we start with $p \leq N$ general deviation vectors they will remain linearly independent on the N-dimensional tangent space of the torus, since there is no particular reason for them to become linearly dependent. As a consequence GALI_p remains practically constant for $p \leq N$. On the other hand, GALI_p tends to be zero for $p > N$, since some deviation vectors will eventually become linearly dependent, following a particular power law which depends on the dimensionality N of the torus and the number p of deviation vectors. So, the generic behavior of GALI_p for regular orbits lying on N-dimensional tori is given by [126]:

$$\mathrm{GALI}_p(t) \sim \begin{cases} \text{constant} & \text{if } 2 \leq p \leq N \\ \frac{1}{t^{2(p-N)}} & \text{if } N < p \leq 2N \end{cases} . \tag{97}$$

The different behavior of GALI_p for regular orbits, where it remains different from zero or tends to zero following a power law, and for chaotic orbits, where it tends exponentially to zero, makes GALI_p an ideal indicator of chaoticity independent of the dimensions of the system [126, 127, 15]. GALI_p is a generalization of the SALI method [122, 124, 125] which is related to the evolution of only two deviation vectors. Actually $\mathrm{GALI}_2 \propto \mathrm{SALI}$. However, GALI_p provides significantly more detailed information on the local dynamics and allows for a faster and clearer distinction between order and chaos. It was shown recently [27, 127] that GALI_p can also be used for the determination of the dimensionality of the torus on which regular motion occurs.

As we discussed in Sect. 6.1 the alignment of all deviation vectors to the direction corresponding to the mLCE is a basic problem for the computation of many LCEs, which is overcome by successive orthonormalizations of the set of deviation vectors. The GALIs on the other hand, exploit exactly this "problem" in order to determine rapidly and with certainty the regular or chaotic nature of orbits.

It was shown in Sect. 4.1 that the values of all LCEs (and therefore the value of the mLCE) do not depend on the particular used norm. On the other hand, the quantitative results of all chaos detection techniques based on quantities related to the dynamics of the tangent space on a finite time, depend on the used norm, or on the coordinates of the studied system. For example, the actual values of the finite time mLCE $X_1(t)$ (54) will be different for different norms, or for different coordinates, although its limiting value for $t \to \infty$, i.e., the mLCE χ_1, will be always the same. Other chaos detection methods, like the FLI and the GALI, which depend on the current values of some norm-related quantities and not on their limiting values for $t \to \infty$ will attain different values for different norms and/or coordinate systems. Although the values of these indices will be different, one could expect that their qualitative behavior would be independent of the chosen norm and the used coordinates, since these indices depend on the geometrical properties of the deviation vectors. For example, the GALI quantifies the linear dependence or independence of deviation vectors, a property which obviously does not depend on the particular

used norm or coordinates. Indeed, some arguments explaining the independence of the behavior of the GALI method on the chosen coordinates can be found in [126]. Nevertheless, a systematic study focused on the influence of the used norm on the qualitative behavior of the various chaos indicators has not been performed yet, although it would be of great interest.

8 LCEs of Dissipative Systems and Time Series

The presentation of the LCEs in this report was mainly done in connection to conservative dynamical systems, i.e., autonomous Hamiltonian flows and symplectic maps. The restriction to conservative systems is not necessary since the theory of LCEs, as well as the techniques for their evaluation are valid for general dynamical systems like dissipative ones. In addition, within what is called time series analysis (see, e.g., [78]) it is of great interest to measure LCEs in order to understand the underlying dynamics that produces any time series of experimental data. For the completeness of our presentation we devote the last section of our report to a concise survey of results concerning the LCEs of dissipative systems and time series.

8.1 Dissipative Systems

In contrast to Hamiltonian systems and symplectic maps for which the conservation of the phase space volume is a fundamental constraint of the motion, a dissipative system is characterized by a decrease of the phase space volume with increasing time. This leads to the contraction of motion on a surface of lower dimensionality than the original phase space, which is called *attractor*. Thus any dissipative dynamical system will have at least one negative LCE, the sum of all its LCEs (which actually measures the contraction rate of the phase space volume through (43)) is negative and after some initial transient time the motion occurs on an attractor.

Any continuous n-dimensional dissipative dynamical system without a stationary point (which is often called a *fixed point*) has at least one LCE equal to zero [70] as we have already discussed in Sect. 4.5. For regular motion the attractor of dissipative flows represents a fixed point having all its LCEs negative, or a quasiperiodic orbit lying on a p-dimensional torus ($p < n$) having p zero LCEs while the rest $n - p$ exponents are negative. For dissipative flows in three or more dimensions there can also exist attractors having a very complicated geometrical structure which are called "strange."

Strange attractors have one or more positive LCEs implying that the motion on them is chaotic. The exponential expansion indicated by a positive LCE is incompatible with motion on a bounded attractor unless some sort of folding process merges separated orbits. Each positive exponent corresponds to a direction in which the system experiences the repeated stretching and folding that decorrelates nearby orbits on the attractor. A simple geometrical construction of a hypothetical strange attrac-

tor where orbits are bounded despite the fact that nearby orbits diverge exponentially can be found in [92, Sect. 1.5].

The numerical methods for the evaluation of the mLCE, of the p $(1 < p < n)$ largest LCEs and of the whole spectrum of them, presented in Sects. 5 and 6, can be applied also to dissipative systems. Actually, many of these techniques were initially used in studies of dissipative models [99, 119, 61, 62]. For a detailed description of the dynamical features of dissipative systems, as well as of the behavior of LCEs for such systems the reader is referred, for example, to [103, 44] [92, Sect. 1.5, Chaps. 7, and 8] and references therein.

8.2 Computing LCEs from a Time Series

A basic task in real physical experiments is the understanding of the dynamical properties of the studied system by the analysis of some observed time series of data. The knowledge of the LCEs of the system is one important step toward the fulfillment of this goal. Usually, we have no knowledge of the nonlinear equations that govern the time evolution of the system which produces the experimental data. This lack of information makes the computation of the spectrum of LCEs of the system a hard and challenging task.

The methods developed for the determination of the LCEs from a scalar time series have as starting point the technique of *phase space reconstruction* with *delay coordinates* [104, 134, 112] [78, Chaps. 3 and 9]. This technique is used for recreating a d-dimensional phase space to capture the behavior of the dynamical system which produces the observed scalar time series.

Assume that we have N_D measurements of a dynamical quantity x taken at times $t_n = t_0 + n\tau$, i.e., $x(n) \equiv x(t_0 + n\tau)$, $n = 0, 1, 2, \ldots, N_D - 1$. Then we produce $N_d = N_D - (d-1)T$ d-dimensional vectors $\mathbf{x}(t_n)$ from the x's as

$$\mathbf{x}(t_n) = \left[x(n)\ x(n+T) \ldots x(n+(d-1)T) \right]^{\mathrm{T}},$$

where T is the (integer) delay time. With this procedure we construct N_d points in a d-dimensional phase space, which can be treated as successive points of a hypothetical orbit. We assume that the evolution of $\mathbf{x}(t_n)$ to $\mathbf{x}(t_{n+1})$ is given by some map and we seek to evaluate the LCEs of this orbit.

The first algorithm to compute LCEs for a time series was introduced by Wolf et al. [144]. According to their method (which is also referred as the *direct method*), in order to compute the mLCE we first locate the nearest neighbor (in the Euclidean sense) $\mathbf{x}(t_k)$ to the initial point $\mathbf{x}(t_0)$ and define the corresponding deviation vector $\mathbf{w}(t_0) = \mathbf{x}(t_0) - \mathbf{x}(t_k)$ and its length $L(t_0) = \|\mathbf{w}(t_0)\|$. The points $\mathbf{x}(t_0)$ and $\mathbf{x}(t_k)$ are considered as initial conditions of two nearby orbits and are followed in time. Then the mLCE is evaluated by the method discussed in Sect. 5.2, which approximates deviation vectors by differences of nearby orbits. So, at some later time t_{m_1} (which is fixed a priori or determined by some predefined threshold violation of the vector's

length) the evolved deviation vector $\mathbf{w}'(t_{m_1}) = \mathbf{x}(t_{m_1}) - \mathbf{x}(t_{k+m_1})$ is normalized and its length $L'(t_{m_1}) = \|\mathbf{w}'(t_{m_1})\|$ is registered. The "normalization" of the evolved deviation vector is done by looking for a new data point, say $\mathbf{x}(t_l)$, whose distance $L(t_{m_1}) = \|\mathbf{x}(t_{m_1}) - \mathbf{x}(t_l)\|$ from the studied orbit is small and the corresponding deviation vector $\mathbf{w}(t_{m_1}) = \mathbf{x}(t_{m_1}) - \mathbf{x}(l)$ has the same direction with $\mathbf{w}'(t_{m_1})$. Of course with finite amount of data, one cannot hope to find a replacement point $\mathbf{x}(l)$ which falls exactly on the direction of $\mathbf{w}'(t_{m_1})$ but chooses a point that comes as close as possible. Assuming that such point is found the procedure is repeated and an estimation $X_1(t_{m_n})$ of the mLCE χ_1 is obtained by an equation analogous to (56):

$$X_1(t_{m_n}) = \frac{1}{t_{m_n} - t_0} \sum_{i=1}^{n} \ln \frac{L'_1(t_{m_i})}{L(t_{m_{i-1}})},$$

with $m_0 = 0$. A Fortran code of this algorithm with fixed time steps between replacements of deviation vectors is given in [144].

Generalizing this technique by evolving simultaneously $p > 1$ deviation vectors, i.e., following the evolution of the orbit under study, as well as of p nearby orbits, we can, in principle, evaluate the p-mLCE $\chi_1^{(p)}$ of the system, which is equal to the sum of the p largest 1-LCEs (see (67)). Then the values of χ_i $i = 1, 2, \ldots, p$ can be computed from (74). This procedure corresponds to a variant of the standard method for computing the LCEs, presented in [119] and discussed in Sect. 6.1, in that deviation vectors are defined as differences of neighboring orbits. The implementation of this approach requires the repeated replacement of the deviation vectors, i.e., the replacement of the p points close to the evolved orbit under consideration, when the lengths of the vectors exceed some threshold value. This replacement should be done in a way that the volume of the corresponding p-parallelogram is small, and in particular smaller than the replaced volume, and the new p vectors point more or less to the same direction like the old ones. This procedure is explained in detail in [144] for the particular case of the computation of $\chi_1^{(2)} = \chi_1 + \chi_2$, where a triplet of points is involved.

It is clear that in order to achieve a good replacement of the evolved p vectors, which will lead to a reliable estimation of the LCEs, the numerical data have to satisfy many conditions. Usually this is not feasible due to the limited number of data points. So the direct method of [144] does not yield very precise results for the LCEs. Another limitation of the method, which was pointed out in Wolf et al. [144], is that it should not be used for finding negative LCEs which correspond to shrinking directions, due to a cut off in small distances implied mainly by the level of noise of the experimental data. An additional disadvantage of the direct method is that many parameters which influence the estimated values of the LCEs like the embedding dimension d, the delay time T, the tolerances in direction angles during vector replacements and the evolution times between replacements have to be tuned properly in order to obtain reliable results.

A different approach for the computation of the whole spectrum of LCEs is based on the numerical determination of matrix $\mathbf{Y}_n$, $n = 1, 2, \ldots$, of (12), which defines

the evolution of deviation vectors in the reconstructed phase space. This method was introduced in [118] and was studied in more detail in [44, 45] (see also [78, Chap. 11]). According to this approach, often called the *tangent space method*, matrix $\mathbf{Y}_n$ is evaluated for each point of the studied orbit through local linear fits of the data. In particular, for every point $\mathbf{x}(t_n)$ of the orbit we find all its neighboring points, i.e., points whose distance from $\mathbf{x}(t_n)$ is less than a predefined small value ϵ. Each of these point define a deviation vector. Then we find the next iteration of all these points and see how these vectors evolve. Keeping only the evolved vectors having length less than ϵ we evaluate matrix $\mathbf{Y}_n$ through a least-square-error algorithm. By repeating this procedure for the whole length of the studied orbit we are able to evaluate at each point of the orbit matrix $\mathbf{Y}_n$ which defines the evolution of deviation vectors over one time step. Then by applying the QR decomposition version of the standard method, which was presented in Sect. 6.3, we estimate the values of the LCEs. The corresponding algorithm is included in the TISEAN software package of nonlinear time series analysis methods developed by Hegger et al. [71]. It is also worth mentioning that Brown et al. [17] improved the tangent space method by using higher order polynomials for the local fit.

If, on the other hand, we are interested only in the evaluation of the mLCE of a time series we can apply the algorithm proposed by Rosenstein et al. [111] and Kantz [77]. The method is based on the statistical study of the evolution of distances of neighboring orbits. This approach is in the same spirit of Wolf et al. [144] although being simpler since it compares distances and not directions. A basic difference with the direct method is that for each point of the reference orbit not one, but several neighboring orbits are evaluated leading to improved estimates of the mLCE with smaller statistical fluctuations even in the case of small data sets. This algorithm is also included in the TISEAN package [71], while its Fortran and C codes can be found in [78, Appendix B].

Acknowledgments The author is grateful to the referee (A. Giorgilli) whose constructive remarks and perceptive suggestions helped him improve significantly the content and the clarity of the paper. Comments from Ch. Antonopoulos, H. Christodoulidi, S. Flach, H. Kantz, D. Krimer, T. Manos and R. Pinto are deeply appreciated. The author would also like to thank G. Del Magno for the careful reading of the manuscript, for several suggestions, and for drawing his attention to the cone technique. This work was supported by the Marie Curie Intra-European Fellowship No MEIF-CT-2006-025678.

Appendix A: Exterior Algebra and Wedge Product: Some Basic Notions

We present here some basic results of the exterior algebra theory along with an introduction to the theory of wedge products following [1] and textbooks such as [128, 68, 129]. We also provide some simple illustrative examples of these results.

Let us consider an M-dimensional vector space V over the field of real numbers $\mathbb{R}$. The *exterior algebra* of V is denoted by $\Lambda(V)$ and its multiplication, known as

the *wedge product* or the *exterior product*, is written as $\wedge$. The wedge product is associative:

$$(\mathbf{u} \wedge \mathbf{v}) \wedge \mathbf{w} = \mathbf{u} \wedge (\mathbf{v} \wedge \mathbf{w}),$$

for $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and bilinear

$$\begin{aligned}(c_1\mathbf{u} + c_2\mathbf{v}) \wedge \mathbf{w} &= c_1(\mathbf{u} \wedge \mathbf{w}) + c_2(\mathbf{v} \wedge \mathbf{w}),\\ \mathbf{w} \wedge (c_1\mathbf{u} + c_2\mathbf{v}) &= c_1(\mathbf{w} \wedge \mathbf{u}) + c_2(\mathbf{w} \wedge \mathbf{v}),\end{aligned}$$

for $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $c_1, c_2 \in \mathbb{R}$. The wedge product is also alternating on V

$$\mathbf{u} \wedge \mathbf{u} = \mathbf{0},$$

for all vectors $\mathbf{u} \in V$. Thus we have that

$$\mathbf{u} \wedge \mathbf{v} = -\mathbf{v} \wedge \mathbf{u},$$

for all vectors $\mathbf{u}, \mathbf{v} \in V$ and

$$\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k = \mathbf{0}, \tag{98}$$

whenever $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in V$ are linearly dependent. Elements of the form $\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k$ with $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in V$ are called *k-vectors*. The subspace of $\Lambda(V)$ generated by all k-vectors is called the *k-th exterior power of* V and denoted by $\Lambda^k(V)$.

Let $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_M\}$ be an orthonormal basis of V, i.e., $\hat{\mathbf{e}}_i$, $i = 1, 2, \ldots, M$ are linearly independent vectors of unit magnitude and

$$\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j = \delta_{ij},$$

where "$\cdot$" denotes the inner product in V and

$$\delta_{ij} = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases}.$$

It can be easily seen that the set

$$\{\hat{\mathbf{e}}_{i_1} \wedge \hat{\mathbf{e}}_{i_2} \wedge \cdots \wedge \hat{\mathbf{e}}_{i_k} \mid 1 \leq i_1 < i_2 < \cdots < i_k \leq M\} \tag{99}$$

is a basis of $\Lambda^k(V)$ since any wedge product of the form $\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k$ can be written as a linear combination of the k-vectors of (99). This is true because every vector $\mathbf{u}_i$, $i = 1, 2, \ldots, k$ can be written as a linear combination of the basis vectors $\hat{\mathbf{e}}_i$, $i = 1, 2, \ldots, M$ and using the bilinearity of the wedge product this can be expanded to a linear combination of wedge products of those basis vectors. Any

wedge product in which the same basis vector appears more than once is zero, while any wedge product in which the basis vectors do not appear in the proper order can be reordered, changing the sign whenever two basis vectors change places. The dimension d_k of $\Lambda^k(V)$ is equal to the binomial coefficient:

$$d_k = \dim\Lambda^k(V) = \begin{pmatrix} M \\ k \end{pmatrix} = \frac{M!}{k!(M-k)!}.$$

Ordering the elements of basis (99) of $\Lambda^k(V)$ according to the standard *lexicographical order*

$$\boldsymbol{\omega}_i = \hat{\mathbf{e}}_{i_1} \wedge \hat{\mathbf{e}}_{i_2} \wedge \cdots \wedge \hat{\mathbf{e}}_{i_k}, \quad 1 \le i_1 < i_2 < \cdots < i_k \le M, \quad i = 1, 2, \cdots, d_k, \tag{100}$$

any k-vector $\bar{\mathbf{u}} \in \Lambda^k(V)$ can be represented as

$$\bar{\mathbf{u}} = \sum_{i=1}^{d_k} \bar{u}_i \boldsymbol{\omega}_i \ , \ \bar{u}_i \in \mathbb{R}. \tag{101}$$

A k-vector which can be written as the wedge product of k linear independent vectors of V is called *decomposable*. Of course, if the k vectors are linearly dependent we get the zero k-vector (98). Note that not all k-vectors are decomposable. For example, the 2-vector $\bar{\mathbf{u}} = \mathbf{e}_1 \wedge \mathbf{e}_2 + \mathbf{e}_3 \wedge \mathbf{e}_4 \in \Lambda^2(\mathbb{R}^4)$ is not decomposable as it cannot be written as $\mathbf{u}_1 \wedge \mathbf{u}_2$ with $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^4$.

Let us consider a decomposable k-vector $\bar{\mathbf{u}} = \mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k$. Then the coefficients $\bar{u}_i$ in (101) are the minors of matrix $\mathbf{U}$ having as columns the coordinates of vectors $\mathbf{u}_i, i = 1, 2, \ldots, k$ with respect to the orthonormal basis $\hat{\mathbf{e}}_i, i = 1, 2, \ldots, M$. In matrix form we have

$$\left[\, \mathbf{u}_1 \ \mathbf{u}_2 \cdots \mathbf{u}_k \,\right] = \left[\, \hat{\mathbf{e}}_1 \ \hat{\mathbf{e}}_2 \cdots \hat{\mathbf{e}}_M \,\right] \cdot \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & & \vdots \\ u_{M1} & u_{M2} & \cdots & u_{Mk} \end{bmatrix} = \left[\, \hat{\mathbf{e}}_1 \ \hat{\mathbf{e}}_2 \cdots \hat{\mathbf{e}}_M \,\right] \cdot \mathbf{U}, \tag{102}$$

where u_{ij}, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, k$ are real numbers. Then, the wedge product $\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k$ is written as

$$\bar{\mathbf{u}} = \mathbf{u_1} \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k = \sum_{i=1}^{d_k} \bar{u}_i \boldsymbol{\omega}_i = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le M} \begin{vmatrix} u_{i_1 1} & u_{i_1 2} & \cdots & u_{i_1 k} \\ u_{i_2 1} & u_{i_2 2} & \cdots & u_{i_2 k} \\ \vdots & \vdots & & \vdots \\ u_{i_k 1} & u_{i_k 2} & \cdots & u_{i_k k} \end{vmatrix} \hat{\mathbf{e}}_{i_1} \wedge \hat{\mathbf{e}}_{i_2} \wedge \cdots \wedge \hat{\mathbf{e}}_{i_k}, \tag{103}$$

where the sum is performed over all possible combinations of k indices out of the M total indices and $|\ |$ denotes the determinant. So, the coefficient of a particular k-vector $\hat{\mathbf{e}}_{i_1} \wedge \hat{\mathbf{e}}_{i_2} \wedge \cdots \wedge \hat{\mathbf{e}}_{i_k}$ is the determinant of the $k \times k$ submatrix of the $M \times k$ matrix of coefficients appearing in (102) formed by its $i_1, i_2, \ldots, i_k$ rows.

The inner product on V induces an *inner product* on each vector space $\Lambda^k(V)$ as follows: Considering two decomposable k-vectors

$$\bar{\mathbf{u}} = \mathbf{u_1} \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k \ \text{ and } \ \bar{\mathbf{v}} = \mathbf{v_1} \wedge \mathbf{v}_2 \wedge \cdots \wedge \mathbf{v}_k,$$

with $\mathbf{u}_i, \mathbf{v}_j \in V$, $i, j = 1, 2, \ldots, k$, the inner product of $\bar{\mathbf{u}}, \bar{\mathbf{v}} \in \Lambda^k(V)$ is defined by

$$\langle \bar{\mathbf{u}}, \bar{\mathbf{v}} \rangle_k \stackrel{\text{def}}{=} \begin{vmatrix} \mathbf{u}_1 \cdot \mathbf{v}_1 & \mathbf{u}_1 \cdot \mathbf{v}_2 & \cdots & \mathbf{u}_1 \cdot \mathbf{v}_k \\ \mathbf{u}_2 \cdot \mathbf{v}_1 & \mathbf{u}_2 \cdot \mathbf{v}_2 & \cdots & \mathbf{u}_2 \cdot \mathbf{v}_k \\ \vdots & \vdots & & \vdots \\ \mathbf{u}_k \cdot \mathbf{v}_1 & \mathbf{u}_k \cdot \mathbf{v}_2 & \cdots & \mathbf{u}_k \cdot \mathbf{v}_k \end{vmatrix} = \left| \mathbf{U}^{\mathrm{T}} \cdot \mathbf{V} \right|, \tag{104}$$

where $\mathbf{U}$, $\mathbf{V}$ are matrices having as columns the coefficients of vectors $\mathbf{u}_i$, $\mathbf{v}_i$, $i = 1, 2, \ldots, k$ with respect to the orthonormal $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_M\}$ (see (102)). Since every element of $\Lambda^k(V)$ is a sum of decomposable elements, this definition extends by bilinearity to any k-vector. Obviously for the basis (100) of $\Lambda^k(V)$ we have

$$\langle \boldsymbol{\omega}_i, \boldsymbol{\omega}_j \rangle_k = \delta_{ij} \ , \ i, j = 1, 2, \ldots, d_k,$$

implying that the basis is orthonormal. Inner product (104) defines a *norm* $\|\ \|$ for k-vectors by

$$\|\bar{\mathbf{u}}\| = \sqrt{\langle \bar{\mathbf{u}}, \bar{\mathbf{u}} \rangle_k} = \sqrt{\left| \mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} \right|}.$$

Thus, the norm of a decomposable k-vector (103) is given by

$$\|\bar{\mathbf{u}}\| = \|\mathbf{u_1} \wedge \mathbf{u}_2 \wedge \cdots \wedge \mathbf{u}_k\| = \sqrt{\left| \mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} \right|} = \left(\sum_{i=1}^{d_k} \bar{u}_i^2 \right)^{1/2} = \left\{ \sum_{1 \le i_1 < i_2 < \cdots < i_k \le M} \begin{vmatrix} u_{i_1 1} & u_{i_1 2} & \cdots & u_{i_1 k} \\ u_{i_2 1} & u_{i_2 2} & \cdots & u_{i_2 k} \\ \vdots & \vdots & & \vdots \\ u_{i_k 1} & u_{i_k 2} & \cdots & u_{i_k k} \end{vmatrix}^2 \right\}^{1/2}, \tag{105}$$

and it measures the volume $\mathrm{vol}(P_k)$ of the k-parallelogram P_k having as edges the k vectors $\mathbf{u_1}, \mathbf{u}_2, \cdots, \mathbf{u}_k$, since this volume is defined as (see, e.g., [75, p. 472])

$$\mathrm{vol}(P_k) = \sqrt{\left| \mathbf{U}^{\mathrm{T}} \cdot \mathbf{U} \right|}\ . \tag{106}$$

The use of a different orthonormal basis does not change the numerical value of $\mathrm{vol}(P_k)$. This can be easily seen as follows: Let $\hat{\mathbf{f}}_i$, $i = 1, 2, \cdots, M$ be a different orthonormal basis of V related to basis $\hat{\mathbf{e}}_i$ through

$$\left[\, \hat{\mathbf{e}}_1 \; \hat{\mathbf{e}}_2 \cdots \hat{\mathbf{e}}_M \,\right] = \left[\, \hat{\mathbf{f}}_1 \; \hat{\mathbf{f}}_2 \cdots \hat{\mathbf{f}}_M \,\right] \cdot \mathbf{A},$$

where $\mathbf{A}$ is an *orthogonal* $M \times M$ matrix, i.e., $\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}$. From (102) we get

$$\left[\, \mathbf{u}_1 \; \mathbf{u}_2 \cdots \mathbf{u}_k \,\right] = \left[\, \hat{\mathbf{f}}_1 \; \hat{\mathbf{f}}_2 \cdots \hat{\mathbf{f}}_M \,\right] \cdot \mathbf{A} \cdot \mathbf{U},$$

whence the volume $\mathrm{vol}'(P_k)$ with respect to the new basis $\hat{\mathbf{f}}_i$, $i = 1, 2, \cdots, M$ is given by

$$\mathrm{vol}'(P_k) = \sqrt{\left|(\mathbf{A} \cdot \mathbf{U})^{\mathrm{T}} \cdot \mathbf{A} \cdot \mathbf{U}\right|} = \sqrt{\left|\mathbf{U}^{\mathrm{T}} \cdot \mathbf{A}^{-1} \cdot \mathbf{A} \cdot \mathbf{U}\right|} = \sqrt{\left|\mathbf{U}^{\mathrm{T}} \cdot \mathbf{U}\right|} = \mathrm{vol}(P_k),$$

where the orthogonality of $\mathbf{A}$ was used. This result is not surprising since an orthogonal matrix corresponds to a rotation that leaves unchanged the norms of vectors and the angles between them.

Finally we note that the equality

$$\left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le M} \begin{vmatrix} u_{i_1 1} & u_{i_1 2} & \cdots & u_{i_1 k} \\ u_{i_2 1} & u_{i_2 2} & \cdots & u_{i_2 k} \\ \vdots & \vdots & & \vdots \\ u_{i_k 1} & u_{i_k 2} & \cdots & u_{i_k k} \end{vmatrix}^2$$

appearing in (105) is the so-called *Lagrange identity* (e.g., [68, p. 108], [16, p. 103]).

An Illustrative Example

In order to illustrate the content of the previous section we consider here a specific example. Let V be the vector space of $M = 4$-dimensional real vectors, i.e., $V = \mathbb{R}^4$ and

$$\hat{\mathbf{e}}_1 = (1, 0, 0, 0) \;,\; \hat{\mathbf{e}}_2 = (0, 1, 0, 0) \;,\; \hat{\mathbf{e}}_3 = (0, 0, 1, 0) \;,\; \hat{\mathbf{e}}_4 = (0, 0, 0, 1) \;, \quad (107)$$

the usual orthonormal basis of $\mathbb{R}^4$. Then the lexicographically ordered orthonormal basis (100) of the $d_2 = 6$-dimensional vector space $\Lambda^2(\mathbb{R}^4)$ is

$$\begin{aligned} &\boldsymbol{\omega}_1 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \;,\; \boldsymbol{\omega}_2 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_3 \;,\; \boldsymbol{\omega}_3 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_4 \;, \\ &\boldsymbol{\omega}_4 = \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_3 \;,\; \boldsymbol{\omega}_5 = \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_4 \;,\; \boldsymbol{\omega}_6 = \hat{\mathbf{e}}_3 \wedge \hat{\mathbf{e}}_4 \;. \end{aligned} \quad (108)$$

The $\Lambda^3(\mathbb{R}^3)$ vector space has dimension $d_3 = 4$ and the set

$$\mathbf{y}_1 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_3 \ , \ \mathbf{y}_2 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_4 \ ,$$
$$\mathbf{y}_3 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_3 \wedge \hat{\mathbf{e}}_4 \ , \ \mathbf{y}_4 = \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_3 \wedge \hat{\mathbf{e}}_4 \ ,$$

as an orthonormal basis, while the $d_4 = 1$-dimensional vector space $\Lambda^4(\mathbb{R}^4)$ is spanned by vector:

$$\mathbf{x}_1 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_3 \wedge \hat{\mathbf{e}}_4.$$

Let us now consider four linearly independent vectors $\mathbf{u_1}$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$ of $\mathbb{R}^4$ and the matrix

$$\mathbf{U} = [u_{ij}] = [\,\mathbf{u_1}\ \mathbf{u}_2\ \mathbf{u}_3\ \mathbf{u}_4\,] = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \end{bmatrix} \quad , \quad i, j = 1, 2, 3, 4,$$

having as columns the coordinates of these vectors with respect to the basis (107) of $\mathbb{R}^4$.

Considering basis (108) of $\Lambda^2(\mathbb{R}^4)$ the 2-vector $\mathbf{u}_1 \wedge \mathbf{u}_2$ is given by

$$\begin{aligned} \mathbf{u_1} \wedge \mathbf{u}_2 = &\begin{vmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{vmatrix} \boldsymbol{\omega}_1 + \begin{vmatrix} u_{11} & u_{12} \\ u_{31} & u_{32} \end{vmatrix} \boldsymbol{\omega}_2 + \begin{vmatrix} u_{11} & u_{12} \\ u_{41} & u_{42} \end{vmatrix} \boldsymbol{\omega}_3 + \\ &+ \begin{vmatrix} u_{21} & u_{22} \\ u_{31} & u_{32} \end{vmatrix} \boldsymbol{\omega}_4 + \begin{vmatrix} u_{21} & u_{22} \\ u_{41} & u_{42} \end{vmatrix} \boldsymbol{\omega}_5 + \begin{vmatrix} u_{31} & u_{32} \\ u_{41} & u_{42} \end{vmatrix} \boldsymbol{\omega}_6, \end{aligned}$$

according to (103). For the norm of this vector we get from (104) and (105):

$$\begin{aligned} \|\mathbf{u_1} \wedge \mathbf{u}_2\|^2 = &\begin{vmatrix} \|\mathbf{u}_1\|^2 & \mathbf{u}_1 \cdot \mathbf{u}_2 \\ \mathbf{u}_2 \cdot \mathbf{u}_1 & \|\mathbf{u}_2\|^2 \end{vmatrix} = \begin{vmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{vmatrix}^2 + \begin{vmatrix} u_{11} & u_{12} \\ u_{31} & u_{32} \end{vmatrix}^2 + \\ &+ \begin{vmatrix} u_{11} & u_{12} \\ u_{41} & u_{42} \end{vmatrix}^2 + \begin{vmatrix} u_{21} & u_{22} \\ u_{31} & u_{32} \end{vmatrix}^2 + \begin{vmatrix} u_{21} & u_{22} \\ u_{41} & u_{42} \end{vmatrix}^2 + \begin{vmatrix} u_{31} & u_{32} \\ u_{41} & u_{42} \end{vmatrix}^2, \end{aligned}$$

where $\|\ \|$ is used also for denoting the usual Euclidian norm of a vector.

In a similar way we conclude that the norm of the 3-vector produced by $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$

$$\begin{aligned} \mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3 = &\begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{vmatrix} \mathbf{y}_1 + \begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{41} & u_{42} & u_{43} \end{vmatrix} \mathbf{y}_2 + \\ &+ \begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{vmatrix} \mathbf{y}_3 + \begin{vmatrix} u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{vmatrix} \mathbf{y}_4 \end{aligned}$$

is

$$\|\mathbf{u_1} \wedge \mathbf{u}_2 \wedge \mathbf{u}_3\|^2 = \begin{vmatrix} \|\mathbf{u}_1\|^2 & \mathbf{u}_1 \cdot \mathbf{u}_2 & \mathbf{u}_1 \cdot \mathbf{u}_3 \\ \mathbf{u}_2 \cdot \mathbf{u}_1 & \|\mathbf{u}_2\|^2 & \mathbf{u}_2 \cdot \mathbf{u}_3 \\ \mathbf{u}_3 \cdot \mathbf{u}_1 & \mathbf{u}_3 \cdot \mathbf{u}_2 & \|\mathbf{u}_3\|^2 \end{vmatrix}$$

$$= \begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{vmatrix}^2 + \begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{41} & u_{42} & u_{43} \end{vmatrix}^2 + \begin{vmatrix} u_{11} & u_{12} & u_{13} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{vmatrix}^2 + \begin{vmatrix} u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \\ u_{41} & u_{42} & u_{43} \end{vmatrix}^2 ,$$

while the norm of the 4-vector produced by $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$

$$\mathbf{u_1} \wedge \mathbf{u}_2 \wedge \mathbf{u}_3 \wedge \mathbf{u}_4 = |\mathbf{U}|\, \mathbf{x}_1$$

is given by

$$\|\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3 \wedge \mathbf{u}_4\|^2 = \begin{vmatrix} \|\mathbf{u}_1\|^2 & \mathbf{u}_1 \cdot \mathbf{u}_2 & \mathbf{u}_1 \cdot \mathbf{u}_3 & \mathbf{u}_1 \cdot \mathbf{u}_4 \\ \mathbf{u}_2 \cdot \mathbf{u}_1 & \|\mathbf{u}_2\|^2 & \mathbf{u}_2 \cdot \mathbf{u}_3 & \mathbf{u}_2 \cdot \mathbf{u}_4 \\ \mathbf{u}_3 \cdot \mathbf{u}_1 & \mathbf{u}_3 \cdot \mathbf{u}_2 & \|\mathbf{u}_3\|^2 & \mathbf{u}_3 \cdot \mathbf{u}_4 \\ \mathbf{u}_4 \cdot \mathbf{u}_1 & \mathbf{u}_4 \cdot \mathbf{u}_2 & \mathbf{u}_4 \cdot \mathbf{u}_3 & \|\mathbf{u}_4\|^2 \end{vmatrix} = |\mathbf{U}|^2 .$$

References

1. Allen, L., Bridges, T.J.: Numerical exterior algebra and the compound matrix method. Numerische Mathematik **92**,197–232 (2002) 123
2. Antonopoulos, C., Bountis, T.: Detecting order and chaos by the linear dependence index (LDI) method. ROMAI J. **2**, 1–13 (2006) 118
3. Antonopoulos, C., Bountis, T., Skokos, Ch: Chaotic dynamics of N-degree of freedom Hamiltonian systems. Int. J. Bif. Chaos **16**, 1777–1793 (2006) 99
4. Bario, R.: Sensitivity tools vs. Poincaré sections. Chaos Solit. Fract. **25**, 711–726 (2005) 117
5. Bario, R.: Painting chaos: a gallery of sensitivity plots of classical problems. Int. J. Bif. Chaos **16**, 2777–2798 (2006) 117
6. Bario, R., Borczyk, W., Breiter, S.: Spurious structures in chaos indicators maps. Chaos Solit. Fract. (in press) (2009) 117
7. Barreira, L., Pesin, Y.: Smooth ergodic theory and nonuniformly hyperbolic dynamics. In: Hasselblatt, B., Katok, A. (eds.): Handbook of Dynamical Systems, vol. 1B, 57–263. Elsevier (2006)
8. Benettin, G., Strelcyn, J.-M.: Numerical experiments of the free motion of a point mass moving in a plane convex region: Stochastic transition and entropy. Phys. Rev. A **17**, 773–785 (1978) 92, 95
9. Benettin, G., Galgani, L.: Lyapunov characteristic exponents and stochasticity. In: Laval, G., Grésillon, D. (eds.): Intrinsic Stochasticity in Plasmas, 93–114, Edit. Phys. Orsay (1979) 75, 84, 86, 90, 99
10. Benettin, G., Galgani, L., Strelcyn, J.-M.: Kolmogorov entropy and numerical experiments. Phys. Rev. A **14**, 2338–2344 (1976) 74, 75, 92, 95, 96
11. Benettin, G., Galgani, L., Giorgilli, A., Strelcyn, J.-M.: Tous les nombres caractéristiques sont effectivement calculables. C. R. Acad. Sc. Paris Sér. A **286**, 431–433 (1978) 75, 84, 86, 89
12. Benettin, G., Froeschlé, C., Scheidecker, J.P.: Kolmogorov entropy of a dynamical system with an increasing number of degrees of freedom. Phys. Rev. A **19**, 2454–2460 (1979) 72, 75, 90, 98
13. Benettin, G., Galgani, L., Giorgilli, A., Strelcyn, J.-M.: Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; A method for computing all of them. Part 1: theory. Meccanica March: 9–20 (1980) 65, 67, 75, 76, 82, 84, 87, 89, 90

14. Benettin, G., Galgani, L., Giorgilli, A., Strelcyn, J.–M.: Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; A method for computing all of them. Part 2: Numerical application. Meccanica March: 21–30 (1980) 63, 65, 66, 67, 75, 76, 83, 86, 89, 92
15. Bountis, T., Manos, T., Christodoulidi, H.: Application of the GALI method to localization dynamics in nonlinear systems. J. Comp. Appl. Math. **227**, 17–26 (2009), nlin.CD/0806.3563 (2008) 119
16. Bourbaki, N.: Éléments de mathématique, Livre II: Algèbre, Chapitre 3. Hermann, Paris (1958) 127
17. Brown, R., Bryant, P., Abarbanel, H.D.I.: Computing the Lyapunov spectrum of a dynamical system from an observed time series. Phys. Rev. A **43**, 2787–2806 (1991) 123
18. Bridges, T.J., Reich, S.: Computing Lyapunov exponents on a Stiefel manifold. Physica D **156**, 219–238 (2001) 115
19. Burns, K., Donnay, V.: Embedded surfaces with ergodic geodesic flow. Int. J. Bif. Chaos **7**, 1509–1527 (1997) 111
20. Carbonell, F., Jimenez, J.C., Biscay, R.: A numerical method for the computation of the Lyapunov exponents of nonlinear ordinary differential equations. Appl. Math. Comput. **131**, 21–37 (2002) 110
21. Casartelli, M., Diana, E., Galgani, L., Scotti, A.: Numerical computations on a stochastic parameter related to the Kolmogorov entropy. Phys. Rev. A **13**, 1921–1925 (1976) 74
22. Casati, G., Ford, J.: Stochastic transition in the unequal-mass Toda lattice. Phys. Rev. A **12**, 1702–1709 (1975) 74
23. Casati, G., Chirikov, B.V., Ford, J.: Marginal local instability of quasi-periodic motion. Phys. Let. A **77**, 91–94 (1980) 96
24. Chen, Z.–M., Djidjeli, K., Price, W.G.: Computing Lyapunov exponents based on the solution expression of the variational system. Appl. Math. Comput. **174**, 982–996 (2006) 111
25. Chernov, N., Markarian, R.: Chaotic billiards. Mathematical Surveys and Monographs, Vol. 127. American Mathematical Society (2006)
26. Christiansen, F., Rugh, H.H.: Computing Lyapunov spectra with continuous Gram-Schmidt orthonormalization. Nonlinearity **10**, 1063–1072 (1997) 110, 114
27. Christodoulidi, H., Bountis, T.: Low-dimensional quasiperiodic motion in Hamiltonian systems. ROMAI J. **2**, 37–44 (2006) 119
28. Cincotta, P.M., Simó, C.: Simple tools to study global dynamics in non-axisymmetric galactic potentials-I. Astron. Astrophs. Supp. Ser. **147**, 205–228 (2000) 117
29. Cincotta, P.M., Giordano, C.M., Simó, C.: Phase space structure of multi-dimensional systems by means of the mean exponential growth factor of nearby orbits. Physica D **182**, 151–178 (2003) 117
30. Contopoulos, G.: Order and Chaos in Dynamical Astronomy. Springer, Berlin Heidelberg New York (2002) 66, 92
31. Contopoulos, G., Giorgilli, A.: Bifurcations and complex instability in a 4-dimensional symplectic mapping. Meccanica **23**, 19–28 (1988) 71
32. Contopoulos, G., Voglis, N.: Spectra of stretching numbers and helicity angles in dynamical systems. Cel. Mech. Dyn. Astron. **64**, 1–20 (1996) 117
33. Contopoulos, G., Voglis, N.: A fast method for distinguishing between ordered and chaotic orbits. Astron. Astrophs. **317**, 73–81 (1997) 117
34. Contopoulos, G., Galgani, L., Giorgilli, A.: On the number of isolating integrals in Hamiltonian systems. Phys. Rev. A **18**, 1183–1189 (1978) 75, 76, 92, 95, 99
35. Devaney, R.L.: An introduction to chaotic dynamical systems. 2nd Ed. Addison-Wesley Publishing Company, New York (1989) 64
36. Dieci, L., Van Vleck., E.S.: Computation of a few Lyapunov exponents for continuous and discrete dynamical systems. Appl. Num. Math. **17**, 275–291 (1995) 65, 108, 110, 112, 114
37. Dieci, L., Van Vleck, E.S.: Lyapunov spectral intervals: theory and computation. SIAM J. Numer. Anal. **40**, 516–542 (2002) 116

38. Dieci, L., Elia, C.: The singular value decomposition to approximate spectra of dynamical systems. Theoretical aspects. J. Diff. Eq. **230**, 502–531 (2006) 110, 116
39. Dieci, L., Lopez, L.: Smooth singular value decomposition on the symplectic group and Lyapunov exponents approximation. Calcolo **43**, 1–15 (2006) 116
40. Dieci, L., Russell, R.D., Van Vleck, E.S.: On the computation of Lyapunov exponents for continuous dynamical systems. SIAM J. Numer. Anal. **34**, 402–423 (1997) 108, 110
41. Donnay, V.: Geodesic flow on the two-sphere, Part I: Positive measure entropy. Erg. Theory Dyn. Syst. **8**, 531–553 (1988) 111
42. Donnay, V.: Geodesic flow on the two-sphere, Part II: Ergodicity. Lect. Notes Math. **1342**, 112–153 (1988) 111
43. Donnay, V., Liverani, C.: Potentials on the two-torus for which the Hamiltonian flow is ergodic. Commun. Math. Phys. **135**, 267–302 (1991) 111
44. Eckmann, J.-P., Ruelle, D.: Ergodic theory of chaos and strange attractors. Rev. Mod. Phys. **57**, 617–656 (1985) 65, 66, 76, 99, 108, 123
45. Eckmann, J.-P., Oliffson Kamphorst, S., Ruelle, D., Ciliberto, S.: Liapunov exponents from time series. Phys. Rev. A **34**, 4971–4979 (1986) 123
46. Farmer, J.D.: Chaotic attractors of an infinite-dimensional dynamical system. Physica D **4**, 366–393 (1982) 76, 99
47. Farmer, J.D., Ott, E., Yorke, J.A.: The dimension of chaotic attractors. Physica D **7**, 153–180 (1983) 99
48. Ford, J., Lunsford, G.H.: Stochastic behavior of resonant nearly linear oscillator systems in the limit of zero nonlinear coupling. Phys. Rev. A **1**, 59–70 (1970) 74
49. Fouchard, M., Lega, E., Froeschlé, Ch., Froeschlé, C.: On the relationship between the fast Lyapunov indicator and periodic orbits for continuous flows. Cel. Mech. Dyn. Astron. **83**, 205–222 (2002) 117
50. Freistetter, F.: Fractal dimensions as chaos indicators. Cel. Mech. Dyn. Astron. **78**, 211–225 (2000) 117
51. Froeschlé, C.: Numerical study of dynamical systems with three degrees of freedom II. Numerical displays of four-dimensional sections. Astron. Astrophs. **5**, 177–183 (1970) 74
52. Froeschlé, C.: Numerical study of a four-dimensional mapping. Astron. Astrophs. **16**, 172–189 (1972) 71, 74
53. Froeschlé, C.: The Lyapunov characteristic exponents—Applications to celestial mechanics. Cel. Mech. Dyn. Astron. **34**, 95–115 (1984) 76
54. Froeschlé, C.: The Lyapunov characteristic exponents and applications. J. de Méc. Théor. et Appl. Numéro spécial 101–132 (1984) 65, 66, 99
55. Froeschlé, C.: The Lyapunov characteristic exponents and applications to the dimension of the invariant manifolds and chaotic attractors. In: Szebehely VG (ed.) Stability of the Solar System and Its Minor Natural and Artificial Bodies, 265–282, D. Reidel Publishing Company (1985) 76
56. Froeschlé, C., Lega, E.: On the structure of symplectic mappings. The fast Lyapunov indicator: a very sensitive tool. Cel. Mech. Dyn. Astron. **78**, 167–195 (2000) 117
57. Froeschlé, C., Froeschlé, Ch., Lohinger, E.: Generalized Lyapunov characteristic indicators and corresponding Kolmogorov like entropy of the standard mapping. Cel. Mech. Dyn. Astron. **56**, 307–314 (1993) 94, 117
58. Froeschlé, C., Lega, E., Gonczi, R.: Fast Lyapunov indicators. Application to asteroidal motion. Cel. Mech. Dyn. Astron. **67**, 41–62 (1997) 117
59. Froeschlé, C., Gonczi, R., Lega, E.: The fast Lyapunov indicator: a simple tool to detect weak chaos. Application to the structure of the main asteroidal belt. Planet. Space Sci. **45**, 881–886 (1997) 117
60. Frøyland, J.: Lyapunov exponents for multidimensional orbits. Phys. Let. A **97**, 8–10 (1983) 111
61. Frøyland, J., Alfsen, K.H.: Lyapunov-exponent spectra for the Lorenz model. Phys. Rev. A **29**, 2928–2931 (1984) 111, 121

62. Geist, K., Parlitz, U., Lauterborn, W.: Comparison of different methods for computing Lyapunov exponents. Prog. Theor. Phys. **83**, 875–893 (1990) 65, 108, 110, 112, 116, 121
63. Goldhirsch, I., Sulem, P.-L., Orszag, S.A.: Stability and Lyapunov stability of dynamical systems: a differential approach and a numerical method. Physica D **27**, 311–337 (1987) 91, 97, 110
64. Gottwald, G.A., Melbourne. I.: A new test for chaos in deterministic systems. Proc. Roy. Soc. London A **460**, 603–611 (2004) 117
65. Gottwald, G.A., Melbourne, I.: Testing for chaos in deterministic systems with noise. Physica D **212**, 100–110 (2005) 117
66. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D **9**, 189–208 (1983) 99
67. Greene, J.M., Kim, J.-S.: The calculation of Lyapunov spectra. Physica D **24**, 213–225 (1987) 110, 116
68. Greub, W.: Multilinear Algebra. 2nd Ed. Springer, Berlin, Heidelberg, New York (1978) 123, 127
69. Guzzo, M., Lega, E., Froeschlé, C.: On the numerical detection of the effective stability of chaotic motions in quasi-integrable systems. Physica D **163**, 1–25 (2002) 117
70. Haken, H.: At least one Lyapunov exponent vanishes if the trajectory of an attractor does not contain a fixed point. Phys. Let. A **94**, 71–72 (1983) 90, 120
71. Hegger, R., Kantz, H., Schreiber, T.: Practical implementation of nonlinear time series methods: The TISEAN package. Chaos **9**, 413–435 (1999) 123
72. Hénon, M., Heiles, C.: The applicability of the third integral of motion: some numerical experiments. Astron. J. **69**, 73–79 (1964) 70, 74, 95
73. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985) 79
74. Howard, J.E. : Discrete virial theorem. Cel. Mech. Dyn. Astron. **92**, 219–241 (2005) 117
75. Hubbard, J.H., Hubbard, B.B.: Vector Calculus, Linear Algebra and Differential Forms: A Unified Approach. Prentice Hall, New Jersey (1999) 126
76. Johnson, B.A., Palmer, K.J., Sell, G.R.: Ergodic properties of linear dynamical systems. SIAM J. Math. Anal. **18**, 1–33 (1987) 65, 87
77. Kantz, H.: A robust method to estimate the maximal Lyapunov exponent of a time series. Phys. Let. A **185**, 77–87 (1994) 123
78. Kantz, H., Schreiber, T.: Nonlinear time series analysis. Cambridge University Press, Cambridge (1997) 120, 123
79. Kaplan, J.L., Yorke, J.A.: Chaotic behavior of multidimensional difference equations. In: Peitgen, H.-O., Walter, H.-O. (eds.): Functional Differential Equations and Approximations of Fixed Points, Lect. Notes Math. **730**, 204–227 (1979) 99
80. Karanis, G.I., Vozikis, Ch.L.: Fast detection of chaotic behavior in galactic potentials. Astron. Nachr. **329**, 403–412 (2008) 117
81. Kotoulas, T., Voyatzis, G.: Comparative study of the 2:3 and 3:4 resonant motion with Neptune: An application of symplectic mappings and low frequency analysis. Cel. Mech. Dyn. Astron. **88**, 343–163 (2004) 117
82. Kovács, B.: About the efficiency of Fast Lyapunov Indicator surfaces and Small Alignment Indicator surfaces. PADEU, **19**, 221–236 (2007) 117
83. Laskar, J.: The chaotic motion of the solar system: a numerical estimate of the size of the chaotic zones. Icarus **88**, 266–291 (1990) 117
84. Laskar, J.: Frequency analysis of multi-dimensional systems. Global dynamics and diffusion. Physica D **67**, 257–281 (1993) 117
85. Laskar, J.: Introduction to frequency map analysis. In: Simó, C. (ed.): Hamiltonian systems with three or more degrees of freedom, 134–150, Plenum Press, New York (1999) 117
86. Laskar, J., Froeschlé, C., Celletti, A.: The measure of chaos by the numerical analysis of the fundamental frequencies. Application to the standard map. Physica D **56**, 253–269 (1992) 117
87. Ledrappier, F., Young, L.-S.: Dimension formula for random transformations. Commun. Math. Phys. **117**, 529–548 (1988) 100
88. Lega, E., Froeschlé, C.: Comparison of convergence towards invariant distributions for rotation angles, twist angles and local Lyapunov characteristic numbers. Planet. Space Sci. **46**, 1525–1534 (1998) 117

89. Lega, E., Froeschlé, C.: On the relationship between fast Lyapunov indicator and periodic orbits for symplectic mappings. Cel. Mech. Dyn. Astron. **81**, 129–147 (2001) 117
90. Li, C., Chen, G.: Estimating the Lyapunov exponents of discrete systems. Chaos **14**, 343–346 (2004) 64, 111
91. Li, C., Xia, X.: On the bound of the Lyapunov exponents of continuous systems. Chaos **14**, 557–561 (2004) 111
92. Lichtenberg, A.J., Lieberman, M.A.: Regular and Chaotic Dynamics. Second Edition. Springer, Berlin, Heidelberg, New York (1992) 65, 66, 99, 121
93. Lohinger, E., Froeschlé, C., Dvorak R.: Generalized Lyapunov exponents indicators in Hamiltonian dynamics: an application to a double star system. Cel. Mech. Dyn. Astron. **56**, 315–322 (1993) 117
94. Lu, J., Yang, G., Oh, H., Luo, A.C.J.: Computing Lyapunov exponents of continuous dynamical systems: method of Lyapunov vectors. Chaos Sol. Fract. **23**, 1879–1892 (2005) 110, 112
95. Lukes-Gerakopoulos, G., Voglis, N., Efthymiopoulos, C.: The production of Tsallis entropy in the limit of weak chaos and a new indicator of chaoticity. Physica A **387**, 1907–1925 (2008) 117
96. Lyapunov, A.M. (1992) The General Problem of the Stability of Motion. Taylor and Francis, London (English translation from the French: Liapounoff A (1907) Problème général de la stabilité du mouvement. Annal. Fac. Sci. Toulouse **9**, 203–474. The French text was reprinted in Annals Math. Studies Vol. 17 Princeton Univ. Press (1947). The original was published in Russian by the Mathematical Society of Kharkov in 1892) 64, 73
97. Markarian, R.: Non-uniformly hyperbolic billiards. Annal. Fac. Sci. Toulouse **3**, 223–257 (1994) 111
98. Mathiesen, J., Cvitanović, P.: Lyapunov exponents. In: Cvitanović, P., Artuso, R., Mainieri, R., Tanner, G., Vattay, G. (eds.): Chaos: Classical and Quantum. Niels Bohr Institute, Copenhagen, http://chaosbook.org/version12/ (2008) 91
99. Nagashima, T., Shimada, I.: On the C-system-like property of the Lorenz system. Prog. Theor. Phys. **58**, 1318–1320 (1977) 75, 92, 95, 121
100. Núñez, J.A., Cincotta, P.M., Wachlin, F.C.: Information entropy. An indicator of chaos. Cel. Mech. Dyn. Astron. **64**, 43–53 (1996) 117
101. Oliveira, S., Stewart, D.E.: Exponential splittings of products of matrices and accurately computing singular values of long products. Lin. Algebra Appl. **309**, 175–190 (2000) 115, 116
102. Oseledec, V.I.: A multiplicative ergodic theorem. Ljapunov characteristic numbers for dynamical systems. Trans. Moscow Math. Soc. **19**, 197–231 (1968) 63, 65, 66, 67, 74, 76, 77, 79, 82, 84, 86
103. Ott, E.: Strange attractors and chaotic motions of dynamical systems. Rev. Mod. Phys. **53**, 655–671 (1981)
104. Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S.: Geometry from time series. Phys. Rev. Let. **45**, 712–716 (1980)
105. Paleari, S., Penati, S.: *Numerical Methods and Results in the FPU Problem*. Lect. Notes Phys. **728**, 239–282 (2008) 72, 116
106. Pesin, Ya. B.: Characteristic Lyapunov exponents and smooth ergodic theory. Russian Math. Surveys **32**, 55–114 (1977) 74, 99
107. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in Fortran. The Art of Scientific Computing. Cambridge University Press, Cambridge (2007) 115
108. Raghunathan, M.S.: A proof of Oseledec's multiplicative ergodic theorem. Isr. J. Math. **32**, 356–362 (1979) 65, 86
109. Ramasubramanian, K., Sriram, M.S.: A comparative study of computation of Lyapunov spectra with different algorithms. Physica D **139**, 72–86 (2000) 110
110. Rangarajan, G., Habib, S., Ryne, R.D.: Lyapunov exponents without rescaling and reorthogonalization. Phys. Rev. Let. **80**, 3747–3750 (1998) 110
111. Rosenstein, M.T., Collins, J.J., De Luca, C.J.: A practical method for calculating largest Lyapunov exponents from small data sets. Physica D **65**, 117–134 (1993) 123
112. Roux, J.-C., Simoyi, R.H., Swinney, H.L.: Observation of a strange attractor. Physica D **8**, 257–266 (1983)

113. Ruelle, D.: Analycity properties of the characteristic exponents of random matrix products. Adv. Math. **32**, 68–80 (1979) 86
114. Ruelle, D.: Ergodic theory of differentiable dynamical systems. IHES Publ. Math. **50**, 275–306 (1979) 65, 87
115. Sándor, Zs., Érdi, B., Efthymiopoulos, C.: The phase space structure around L4 in the restricted three-body problem. Cel. Mech. Dyn. Astron. **78**, 113–123 (2000) 117
116. Sándor Zs., Érdi, B., Széll, A., Funk, B.: The relative Lyapunov indicator: an efficient method of chaos detection. Cel. Mech. Dyn. Astron. **90**, 127–138 (2004) 117
117. Sandri, M.: Numerical calculation of Lyapunov exponents. Mathematica J. **6**, 78–84 (1996) 105
118. Sano, M., Sawada, Y.: Measurement of the Lyapunov spectrum from a chaotic time series. Phys. Rev. Let. **55**, 1082–1085 (1985) 76, 123
119. Shimada, I., Nagashima, T.: A numerical approach to ergodic problem of dissipative dynamical systems. Prog. Theor. Phys. **61**, 1605–1615 (1979) 67, 75, 84, 95, 102, 103, 108, 121, 122
120. Sideris, I.V.: Characterization of chaos: a new, fast and effective measure. In: Gottesman, S.T., Buchler, J.-R. (eds.): Nonlinear Dynamics in Astronomy and Astrophysics, Annals of the New York Academy of Science, 1045:79, The New York Academy of Sciences (2005) 117
121. Sideris, I.V.: Measure of orbital stickiness and chaos strength. Phys. Rev. E **73**, 066217 (2006) 117
122. Skokos Ch.: Alignment indices: a new, simple method for determining the ordered or chaotic nature of orbits. J. Phys. A **34**, 10029–10043 (2001) 98, 117, 119
123. Skokos, Ch., Contopoulos, G., Polymilis, C.: Structures in the phase space of a four dimensional symplectic map. Cel. Mech. Dyn. Astron. **65**, 223–251 (1997) 71
124. Skokos, Ch., Antonopoulos, Ch., Bountis, T.C., Vrahatis, M.N.: How does the smaller alignment index (SALI) distinguish order from chaos? Prog. Theor. Phys. Supp. **150**, 439–443 (2003) 117, 119
125. Skokos, Ch., Antonopoulos, Ch., Bountis, T.C., Vrahatis, M.N.: Detecting order and chaos in Hamiltonian systems by the SALI method. J. Phys. A **37**, 6269–6284 (2004) 117, 119
126. Skokos, Ch., Bountis, T.C., Antonopoulos, Ch.: Geometrical properties of local dynamics in Hamiltonian systems: The generalized alignment index (GALI) method. Physica D **231**, 30–54 (2007) 117, 118, 119, 120
127. Skokos, Ch., Bountis, T.C., Antonopoulos, Ch.: Detecting chaos, determining the dimensions of tori and predicting slow diffusion in Fermi-Pasta-Ulam lattices by the generalized alignment index method. Eur. Phys. J. Sp. Top. **165**, 5–14 (2008) 117, 118, 119
128. Spivak, M.: Calculus on Manifolds. Addison-Wesley, New York (1965) 123
129. Spivak, M.: Comprehensive Introduction to Differential Geometry, vol. 1. Publish or Perish Inc., Houston (1999) 123
130. Stewart, D.E.: A new algorithm for the SVD of a long product of matrices and the stability of products. Electr. Trans. Numer. Anal. **5**, 29–47 (1997) 115, 116
131. Stoddard, S.D., Ford, J.: Numerical experiments on the stochastic behavior of a Lennard–Jones gas system. Phys. Rev. A **8**, 1504–1512 (1973) 74
132. Süli, Á.: Speed and efficiency of chaos detection methods. In: Süli, Á., Freistetter, F., Pál, A. (eds.): Proceedings of the 4th Austrian Hungarian workshop on Celestial Mechanics, **18**, 179–189, Publications of the Astronomy Department of the Eötvös University (2006) 117
133. Süli, Á.: Motion indicators in the 2D standard map. PADEU **17**, 47–62 (2006) 117
134. Takens, F.: Detecting strange attractors in turbulence. Lect. Notes Math. **898**, 366–381 (1981)
135. Voglis, N., Contopoulos, G.: Invariant spectra of orbits in dynamical systems. J. Phys. A **27**, 4899–4909 (1994) 94, 117
136. Voglis, N., Contopoulos, G., Efthymiopoulos, C.: Method for distinguishing between ordered and chaotic orbits in four-dimensional maps. Phys. Rev. E **57**, 372–377 (1998) 117
137. Voyatzis, G., Ichtiaroglou, S.: On the spectral analysis of trajectories in near-integrable Hamiltonian systems. J. Phys. A **25**, 5931–5943 (1992) 117

138. Vozikis, Ch.L., Varvoglis, H., Tsiganis, K.: The power spectrum of geodesic divergences as an early detector of chaotic motion. Astron. Astrophs. **359**, 386–396 (2000) 117
139. Vrahatis, M.N., Bountis, T.C., Kollmann, M.: Periodic orbits and invariant surfaces of 4D nonlinear mappings. Int. J. Bif. Chaos **6**, 1425–1437 (1996) 116
140. Vrahatis, M.N., Isliker, H., Bountis, T.C.: Structure and breakdown of invariant tori in a 4-D mapping model of accelerator dynamics. Int. J. Bif. Chaos **7**, 2707–2722 (1997) 116
141. Walters, P.: A dynamical proof of the multiplicative ergodic theorem. Thans. Amer. Math. Soc. **335**, 245–257 (1993) 65, 87
142. Wojtkowski, M.: Invariant families of cones and Lyapunov exponents. Erg. Theory Dyn. Syst. **5**, 145–161 (1985) 111
143. Wojtkowski, M.: Principles for the design of billiards with nonvanishing Lyapunov exponents. Commun. Math. Phys. **105**, 391–414 (1986) 111
144. Wolf, A., Swift, J.B., Swinney, H.L., Vastano, J.A.: Determining Lyapunov exponents from a time series. Physica D **16**, 285–317 (1985) 76, 105, 121, 122, 123
145. Wu, X., Huang, T.-Y., Zhang, H.: Lyapunov indices with two nearby trajectories in a curved spacetime. Phys. Rev. E **74**, 083001 (2006) 95
146. Young, L.-S.: Dimension, entropy and Lyapunov exponents. Erg. Theory Dyn. Syst. **2**, 109–124 (1982) 100
147. Zou, Y., Pazó, D., Romano, M.C., Thiel, M., Kurths, J.: Distinguishing quasiperiodic dynamics from chaos in short-time series. Phys. Rev. E **76**, 016210 (2007) 117
148. Zou, Y., Thiel, M., Romano, M.C., Kurths, J.: Characterization of stickiness by means of recurrence. Chaos **17**, 043101 (2007) 117

Asteroid Dynamical Families

A. Cellino and A. dell'Oro

Abstract Asteroid dynamical families are extremely important for our understanding of the origin, evolution, and general physical properties of the asteroid population. First identified on the basis of their dynamical properties, families have been soon recognized as the products of well-defined physical processes, namely the disruption of single parent bodies as the consequence of energetic collisional events. The identification of dynamical families has opened important perspectives in all fields of research in asteroid science. The "paradigm" of interpretation of family data has been quickly evolving during the last decade and is now based on the evidence of a complex interplay of different physical and dynamical processes, some of which only recently have been fully recognized. In this chapter, we attempt to give a general and comprehensive review of the subject.

1 Introduction

Asteroid dynamical families are still a very important and fascinating subject in asteroid science, in spite of being a long debated topic that is now about 100 years old. It is difficult to find a line of research in asteroid science that does not lead sooner or later to face the enigmas posed by the families, as schematically shown in Fig. 1.

First discovered at the beginning of the twentieth century by Hirayama [1, 2], families were very soon interpreted as the likely products of collisional events taking place in the asteroid main belt. Any progress in the study of families, however, was long hampered by the difficulty in developing objective methods of family identification and by correspondingly huge discrepancies among the results obtained by different authors, as extensively discussed in [3].

A. Cellino (✉)
INAF-Ossevatorio Astronomico di Torino, Strada Osservatorio 20, 10025 Pino Torinese, Italy,
cellino@oato.inaf.it

A. dell'Oro
INAF-Ossevatorio Astronomico di Torino, Strada Osservatorio 20, 10025 Pino Torinese, Italy,
delloro@oato.inaf.it

Cellino, A., dell'Oro, A.: *Asteroid Dynamical Families*. Lect. Notes Phys. **790**, 137–193 (2010)
DOI 10.1007/978-3-642-04458-8_3

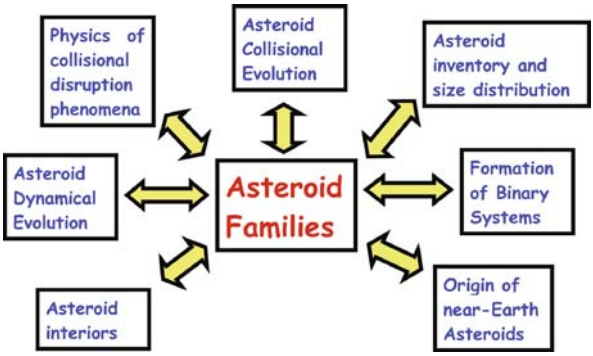

Fig. 1 Schematic view of the relations existing between family studies and the most important topics in asteroid science

Starting from the year 1990, new reliable methods of family identification led to the identification of at least 20 statistically robust asteroid families. This triggered a lot of activity in theoretical and observational studies of these groupings that led to important results and produced a general "paradigm" of interpretation. At the beginning of the new century, however, new theoretical and observational facts led to a deep change in this general paradigm of family interpretation. Currently, however, a new paradigm has not yet been completely established, some controversies are still open, and in general the interpretation of families is in a phase of general transition, in which most of the new developments are accepted, but the real extent to which some old results must be considered as fully incorrect is not yet completely clear.

In this chapter, we do not want to follow strictly a historical approach, but, starting from the most important pieces of evidence that have been accumulating with time, we try to give a comprehensive overview of the importance of dynamical families in the most general context of asteroid science. According to previous considerations, we pay particular attention to the important discontinuity that occurred in the general interpretation of family properties starting approximately since the year 2000, when the realization of the importance of new dynamical mechanisms that had not been previously taken into account has produced a big change of paradigm in the interpretation of family data. In particular, we focus on a number of subjects which are still debated, and we make a few predictions about possible developments in the future.

2 Families in the Twentieth Century

Figure 2 shows the observational evidence that was available to any attempt of identifying dynamical families in the asteroid main belt in the mid-1990s. This figure shows the plots of proper eccentricity and sinus of proper inclination versus proper semi-major axis, respectively. To understand the meaning of these plots, it is necessary first to understand what the proper orbital elements are. As is well known, the orbits of the bodies of our Solar System are not constant in time like in the ideal case of a two-body system, but they vary continuously due to the effect of mutual

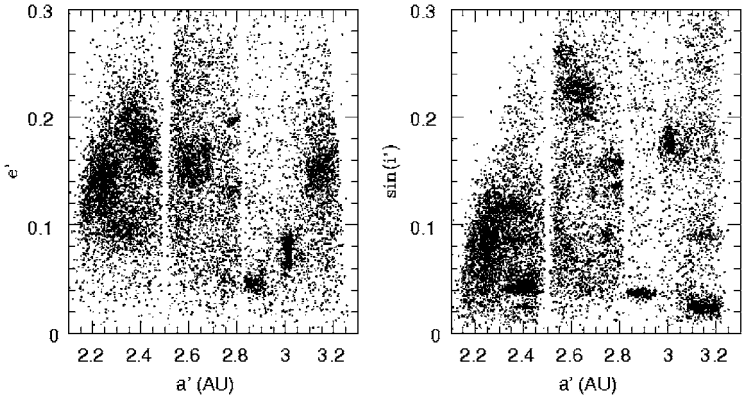

Fig. 2 Plots of proper eccentricity and proper inclination as a function of proper semi-major axis, using the data available in the early 1990s

gravitational perturbations between the many bodies present in the system. These effects may be very important for minor bodies like the asteroids, whose motion may be strongly perturbed by the major planets, and particularly by Jupiter. For this reason, the orbit of any given object at a certain epoch is described by a set of orbital elements, called *osculating* elements, which are not constant in time. If we want to make some quantitative analysis of the similarity of the orbits of different asteroids, therefore, we need to use, if possible, some quantities which may be more stable over time than the simple osculating elements. A classical definition states that what we call *proper* elements are quasi-integrals of the motion and that they are therefore nearly constant in time. Alternatively, one can say that proper elements are true integrals, but of a conveniently simplified dynamical system. In any case, proper elements are obtained as a result of the elimination of short- and long-term periodic perturbations from their instantaneous, osculating counterparts (the osculating elements) and thus represent a kind of "average" characteristics of motion, which normally varies very little over long timescales [4].

Since the early 1990s, the development of refined and fast techniques to compute proper elements [5] put at disposal of family searches increasingly larger databases of asteroid proper elements, much larger than those adopted in previous analyses. An example is given in Fig. 2. If one looks at this figure, it is easy to see that the main belt asteroids are not uniformly distributed in the space of orbital proper elements. Apart from the evident presence of forbidden zones that appear to be empty (like the vertical narrow strips known as "Kirkwood gaps"), which correspond to resonant orbits that are not stable [6]; it is clear that the distribution of the objects, in the populated regions of the proper elements space, is very irregular, and even a quick visual inspection is sufficient to find evidence of several more or less pronounced clusters of objects. According to the meaning of proper elements, these clusters represent groups of objects that have very similar orbits, even over long timescales. These clusters of objects sharing similar orbits are what we call *dynamical families*.

What is the interpretation of the existence of these families? There are not purely dynamical mechanisms that should be expected to be able to produce some very compact and sharp clusterings of orbits like some of those that are evident in Fig. 2, starting from some more homogeneous distribution of orbits. On the other hand, it is hard to imagine that asteroids were originally accreted in clusters, and they are still there after 4 billions of years. To understand the origin of families, we must consider what are the most important mechanisms that have determined the evolution of the asteroidal population since the epoch of its formation. In this respect, although there are still some uncertainties on the very early stages of the asteroids' history, in particular concerning the process that was responsible for the early depletion of over 99% of the solid matter originally located in this region of the Solar System [7], it is widely accepted that catastrophic collisions have been the major physical process that has governed the evolution of the asteroid population for most of the time passed since the early epochs of planetary accretion.

In particular, collisions may naturally explain the existence of dynamical families. The idea is that a family is a swarm of fragments created by the collisional disruptions of an original parent body. This is a nice example of a situation in which dynamical properties provide convincing evidence of the occurrence of very interesting physical processes. Asteroid families become, like the tilt of the Uranus' spin axis, the existence of our Moon, the presence of great impact basins on all atmosphereless bodies observed remotely and in situ, new witnesses of the complex collisional history of our Solar System.

The idea that families are collisional outcomes can be expressed in a more quantitative way. In particular, let us assume that a given body orbiting the Sun suffers a sudden velocity change due to some reason, like in the case of a fragment ejected from its parent body in a catastrophic collision event. As a consequence of this change of its velocity vector, the body will achieve a new orbit, described by a new set of orbital parameters. The relation between the velocity change experienced by the body and the variation of its orbital elements is well known. In particular, the conversion from velocities to orbital elements or vice versa is expressed by the so-called Gauss formulae, that can be written as follows, under the assumption that the velocity change is much smaller than the initial orbital velocity of the body:

$$\begin{cases} \delta a/a = \dfrac{2}{na(1-e^2)^{1/2}}\left[(1+e\cos f)\delta V_T + (e\sin f)\delta V_R\right] \\ \\ \delta e = \dfrac{(1-e^2)^{1/2}}{na}\left[\dfrac{e+2\cos f+e\cos^2 f}{1+e\cos f}\delta V_T + (\sin f)\delta V_R\right], \\ \\ \delta I = \dfrac{(1-e^2)^{1/2}}{na}\dfrac{\cos(\omega+f)}{1+e\cos f}\delta V_W \end{cases} \tag{1}$$

where n is the mean motion around the Sun, na is the mean orbital velocity, and δV_T, δV_R, and δV_W are the components of the velocity vector change (ejection velocity) along the direction of the motion, radial, and normal to the orbital plane,

respectively. The parameters f and ω are the a priori unknown true anomaly and argument of perihelion of the body at the epoch of the velocity change on its original orbit. If we consider the case of a fragment ejected from a parent body in a typical collisional event, we see that the condition that the velocity change is much smaller than the original orbital speed of the parent body is well satisfied. In fact, typical orbital velocities for main belt asteroids are of the order of 10 km/s, whereas the typical ejection velocities are generally more than one order of magnitude lower, according to current physical models of these events [8].

The Gauss equations (1) are fundamental in many respects in family studies, and we will refer often to them in this chapter. At this point we only note that they can be used to demonstrate that the collisional disruption of a parent body must necessarily be expected to produce a swarm of fragments with very similar orbits that, in the space of the orbital elements, should appear as a cluster of objects. Since the orbits are subject to perturbations and are subject to short- and long-timescale variations, the similarity of the original orbits will be kept more evident over longer times in the space of the proper elements, which are much less subject to time variations.

Having recognized that dynamical families are collisional outcomes, the first problem that was encountered in the early family studies was that of the *reliable identification* of these groupings. In other words, by looking at plots like those shown in Fig. 2, a fundamental question is which ones among the apparent clusters are real and correspond to true collisional processes, and which ones are local overdensities of objects in the proper elements space due only to chance and not to physical processes. This is the first problem that was faced by the family studies that were carried out in the twentieth century. The following sections are aimed at presenting in a schematic way what happened in the field of asteroid families starting since the last decade of the past century, when family studies experienced a moment of very intense development. In Sect. 3, we will then focus on what happened starting from the early years of the present century, when the importance of new dynamical mechanisms was realized, leading to new concepts and interpretations of available data.

2.1 Family Identification

We cannot make here a comprehensive summary of all the results produced by different authors in their identifications of dynamical families since the epoch of discovery of these groupings. Here we only recall the fact that the most notable property of the early results in this field was that different family searches wildly disagreed with each other, the number of identified families ranging from a few up to more than 100. This was due to big differences between the data sets of proper elements used by different authors, as well as by differences in the adopted identification criteria, which were often based on subjective analyses of the available data. A review of this topic can be found in [9].

Starting since 1990, a couple of new identification methods, based on reproducible and well-defined algorithms, were independently developed by two teams in Torino (Italy) and Nice (France). These two methods, named *HCM* (standing for hierarchical clustering method) and *WAM* (standing for wavelet analysis method), were completely independent, being based on a classical multivariate clustering analysis approach (HCM) and on a wavelet-based technique for local overdensity recognition (WAM), respectively. Both techniques had in common the idea of quantifying the distance between two points in the proper element space by introducing a suitable metric (a definition of distance) in that space. An identical standard metric was adopted, complemented by another alternative metric to be used as a check in order to test the stability of family identification upon the metric choice.

Due to the relation existing between differences in orbital elements and differences in fragment ejection velocity in an impact event (expressed by the Gauss equations (1) seen above), the distance in the space of proper elements was chosen to have the dimension of a velocity, expressed in m/s. Based on a number of considerations explained in a classical paper by Zappalà et al. [10], the adopted standard metric had the form

$$\partial v = na' \sqrt{5/4(\delta a'/a')^2 + 2(\delta e')^2 + 2(\delta \sin i')^2},$$

where ∂v is the distance between two points in the proper element space (corresponding to two orbits) expressed in m/s, and according to Gauss' equations, n is the mean motion and na' is correspondingly the mean orbital velocity of the first orbit.

Having defined a metric in the space of proper elements, the next step was to develop algorithms to identify clusters of objects that, from a statistical point of view, had zero probability to be due to chance. HCM and WAM differed in the way they identified object clusters, but apart from that both of them were based on the idea of eventually comparing the identified clusters with those resulting from a randomly generated synthetic population of objects. In other words, the idea was to compare the real clusters of objects present in a given region of the proper element space with those that may be produced by a random distribution of objects in the same region.

More in particular, it was thought that doing a purely random generation of objects in the same volume of the proper element space occupied by a given population of real objects could be misleading. The reason is that such a fully random population could be distributed in a too much different way with respect to the real objects present the same region of the proper element space, and any comparison between the real and the simulated objects might be questionable. For instance, in a given volume of the proper element space the real objects might be found to fill preferentially some regions of the volume, for a variety of reasons related to the overall history and dynamical properties of the asteroid main belt, whereas a fully random population would tend to fill the same volume in a homogeneous way. For this reason, in order to generate synthetic populations having something to do with the real objects, the synthetic objects were created imposing as an additional

constraint that the overall distribution of their three proper elements should separately fit the observed a', e', and i' distributions of the real objects.

In other words, the a', e', and i' histograms of the simulated populations cannot be distinguished from the analogous histograms of the real population, but the synthetic population does not contain any correlation between the a', e', and i' coordinates of each object. The synthetic populations generated in this way were called *quasi-random* populations. The general idea of the family identification algorithms was then to compare the clusters of the real and quasi-random population in a given region of the proper element space. The quasi-random population was used to identify the maximum local overdensity that can be randomly created among a population of objects distributed in some way in a given region of the proper element space. Families had to be clusters more compact and more populous than those produced by any quasi-random population, corresponding to groupings that could not be due to pure chance. Figure 3 shows, as an example, the comparison between the population of real objects present in an arbitrary volume of the proper element space and a corresponding quasi-random population built according to the above explanation. In particular, the figure shows the comparison in the a'– $\sin i'$ plane.

The HCM method is particularly suitable to explain in practice how the above approach can be actually implemented. Having introduced a metrics, the first step consists of computing all the mutual distances between each couple of objects of the considered sample. Having at disposal this distance matrix, an iterative procedure is performed, consisting of the following operations at each step:

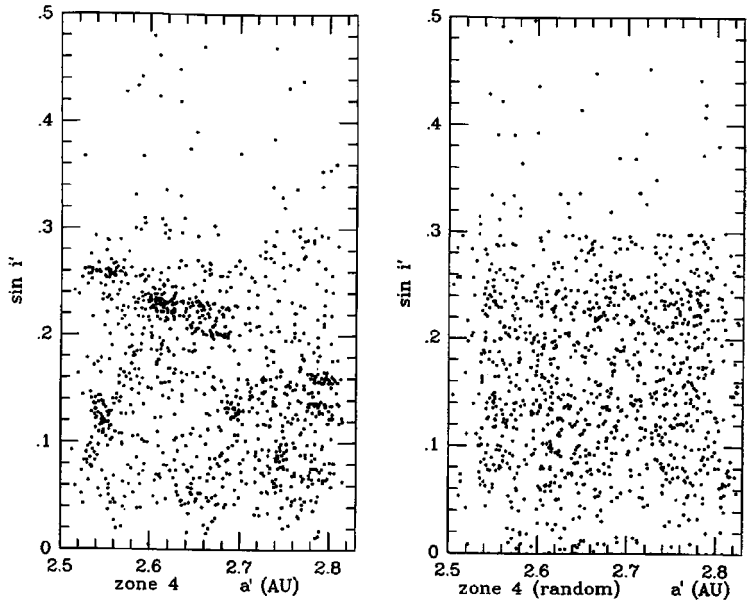

Fig. 3 Distribution in the (a', $\sin i'$) plane of the real asteroid population in a region of the main belt (*left*) and of the corresponding quasi-random population (*right*)

1. Identification of the two mutually closest objects (i, j) of the sample.
2. *Conglomeration of i and j*: this means that from this moment this couple counts as one single object $i \cup j$, and by definition the distance d of any object k from $i \cup j$ is assumed to be the minimum between $d(k, i)$ and $d(j, k)$.
3. The matrix of the distances is updated (it loses one row and one column) and go back to step (1).

The above procedure is iterated until there is only one remaining object, corresponding to the conglomeration of all the original objects of the considered sample.

In this way, a *dendrogram* can be built, which allows one to get immediately, being given some desired value of mutual distance, how many objects are clustered at that level of distance and also the identity of the members of the resulting groupings. In [10] it was introduced for the first time a kind of diagram that proved to be very useful to represent in a graphical way the results of the above analysis. One example of these diagrams, called *stalactite diagrams*, is shown in Fig. 4.

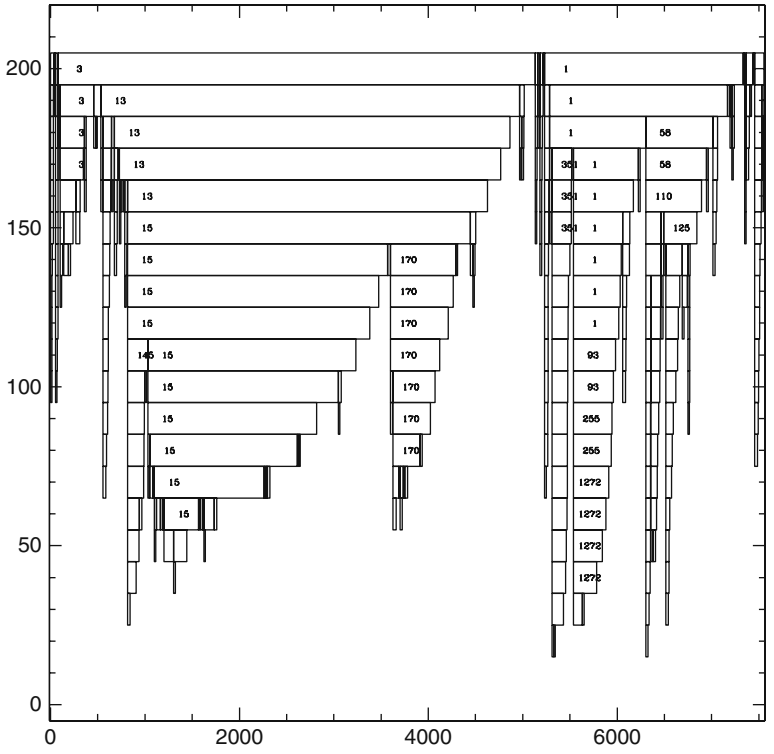

Fig. 4 Example of a *stalactite* diagram for a sample of main belt asteroids having proper semi-major axis between 2.501 and 2.825 AU. The vertical axis gives the mutual distance (in m/s) while the horizontal axis shows the number of objects that are included in separate clusters as a function of the distance level. The tips of the stalactites represent then very compact groupings formed by objects strictly similar in orbital proper elements

As one can see in this figure, as far as the considered level of distance increases, the objects tend to group together in increasingly bigger clusters, whereas only a few, compact groupings are found at small distance levels. Of course, the most compact and deepest groupings are the most interesting ones, since they correspond to very dense clusters of orbits. By producing stalactite diagrams for quasi-random populations in the same volume of the proper element space, it is possible to introduce some criteria for the identification of dynamical families, that is, groupings that cannot be due to pure chance. In several papers based on the HCM starting since 1990, the adopted criterion was, generally speaking, that families are either deeper than the deepest stalactites produced by the quasi-random population or they reach the same depth, but include much larger numbers of members. A minimum critical number of objects was introduced in this respect, and the reader should address to the original papers for a whole explanation. In addition, tests to check the statistical robustness of the resulting families for possible changes of the adopted metrics (distance function) and for variations of the proper elements of the objects correspondingly with the nominal accuracy of the proper elements computation were done.

This kind of analysis produced for the first time results that were based on a well-defined "objective" algorithm, did not depend on a visual inspection of the data, and were reproducible. This was a big step forward, and the result was the unambiguous identification of about 20 dynamical families, plus a number of other more uncertain groupings whose real interpretation was postponed to later times, when larger data sets of proper elements might become available. The last peer-reviewed paper in this series of family searches was published in 1995 [11]. It included the analysis of a sample of more than 12,000 proper elements, limited to objects having proper eccentricity and sinus of proper inclination both smaller than 0.3. For the first time, both HCM and WAM results were presented at the same time for the same sample of objects. This paper has been the reference for family studies for many years. The most important families identified in this analysis are listed in Table 1, while the families identified by the HCM (only) are shown in Fig. 5 (in the a'–e' plane) and 6 (in the plane a'–$\sin i'$).

2.2 Spectroscopic Confirmations

In the early 1990s, in a situation characterized by a big confusion in the field of family identification, the introduction of new methods like HCM and WAM could not be seen as a very important achievement in the absence of some convincing confirmation of their reliability. This kind of confirmation came soon, however, when Binzel and Xu carried out a spectroscopic survey of the Vesta family recently identified by the HCM [10].

A spectroscopic study of Vesta family members was at that time an ideal tool to test the supposedly collisional origin of this family. The reason is that Vesta had been for a long time a unique object among the asteroid population in terms of

Table 1 A comparison between the most prominent dynamical families identified by [11] in a joint analysis in which both the HCM and the WAM were applied to an identical sample of more than 12, 400 asteroids. Each family of the list is indicated using the name of its resulting least numbered member (that can be different for HCM and WAM). Only families having an intersection (numbers of objects in common) of 75% (upper block) or 50% (lower block) are listed. These families represent, therefore, the most reliable groupings identified in that analysis

Identified families		Number of members	
HCM	WAM	HCM	WAM
8 Flora	43 Ariadne	604	575
44 Nysa	135 Hertha	381	374
4 Vesta	4 Vesta	231	242
163 Erigone	163 Erigone	45	49
1 Ceres	83 Minerva	89	88
170 Maria	170 Maria	77	83
668 Dora	168 Dora	77	79
145 Adeona	145 Adeona	63	67
808 Merxia	808 Merzia	26	29
569 Misa	569 Misa	25	27
410 Chloris	410 Chloris	21	27
1644 Rafita	1644 Rafita	21	23
1128 Astrid	1128 Astrid	10	11
24 Themis	24 Themis	550	517
221 Eos	221 Eos	477	482
158 Koronis	158 Koronis	325	299
137 Meliboea	137 Meliboea	13	16
845 Naema	845 Naema	6	7
20 Massalia	20 Massalia	49	45
15 Eunomia	15 Eunomia	439	393
110 Lydia	110 Lydia	26	50
128 Nemesis	58 Concordia	20	38
1639 Bower	342 Endymion	10	15
10 Hygiea	10 Hygiea	103	175
490 Veritas	92 Undina	22	36
293 Brasilia	293 Brasilia	10	18

spectroscopic properties and corresponding mineralogic composition. At the beginning of the 1990s Vesta was still a fairly unique object characterized by a reflectance spectrum similar to that of terrestrial basalts, characterized by deep absorption features at wavelengths around 1 and 2 μm. As a consequence of these properties, Vesta was the prototype of a unique taxonomic class, named *V* after its name. The interpretation of its spectrum was that Vesta should likely be considered as a unique example of a fully differentiated asteroid, with the likely presence of a metallic core surrounded by an olivine mantle and a lighter basaltic crust. From the point of view of the studies of collisional evolution of the asteroid population, Vesta was considered to put some important constraint on the collision rate. In fact, any model of the collisional evolution process should have been constrained by the fact that the fragile basaltic crust of Vesta has remained intact until our days. On the other hand, the presence of a large, hemispheric-sized albedo spot, possibly due to the

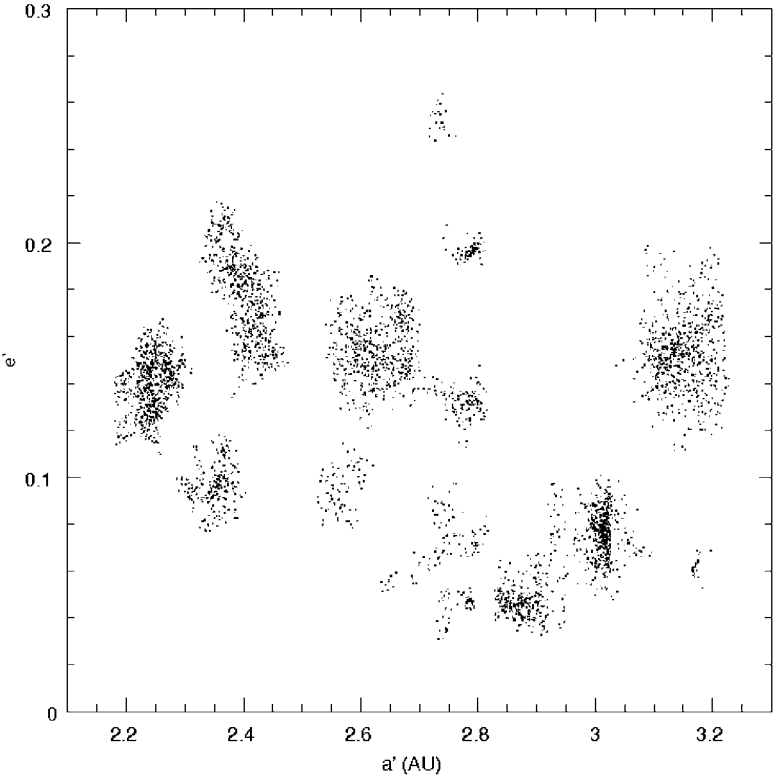

Fig. 5 Locations in the a'–e' plane of the families identified by the HCM according to [11]

presence of a large impact crater on the surface, had been discovered by means of polarimetric and photometric studies of this asteroid [12].

The presence of a dynamical family associated with Vesta was therefore not discouraged by the observational evidence available at that time, provided that the family could have been produced by an energetic cratering event, able to excavate a large crater on the surface, but without being able to break the object apart. At the same time, spectroscopy was the ideal tool to test the hypothesis of a dynamical family of Vesta, since the members of this supposed family would have been presumably sharing the unique spectral reflectance properties of Vesta itself, then should have been expected to belong to the V taxonomic class.

The results of the first spectroscopic investigation of the Vesta family published in 1993 [13] were a spectacular confirmation of the real collisional origin of the family identified by the new methods. Not only a sample of objects listed as Vesta members by an HCM analysis turned out to be V-type, but even a number of other, small objects not belonging to the family, but orbiting with orbital semi-major axes between that of Vesta and the inner border of the 3:1 mean motion resonance with Jupiter at 2.50 AU, corresponding to one of the major forbidden zones in the

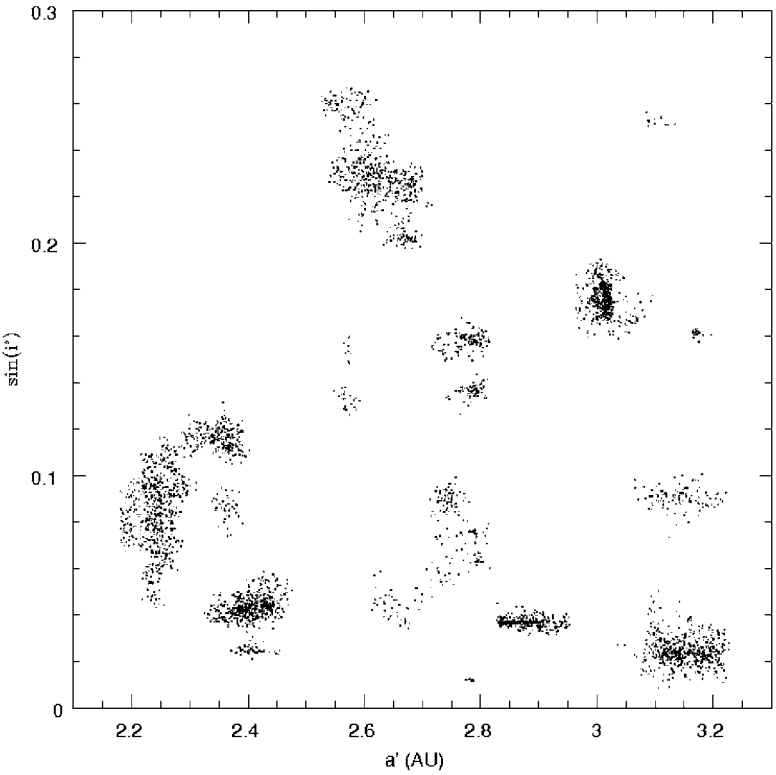

Fig. 6 Locations in the a'–$\sin i'$ plane of the families identified by the HCM according to [11]

asteroid belt (Kirkwood gaps), were found to have V-type spectra. These findings not only confirmed the existence of the Vesta family, but they also suggested that many fragments might have been ejected at large velocities, reaching a so large distance from the parent body that they cannot be recognized as family members. The fragments that might possibly have reached the 3:1 resonance, which is an effective "dynamical engine" to move objects from the asteroid main belt to the region of the terrestrial planets, as we will discuss in separate sections, might have become the parent bodies of HED achondrites, as well of the V-type near-Earth asteroids that were found around those years. A copy of the original plot showing the results of Binzel and Xu is shown in Fig. 7.

After the successful observations of the Vesta family, spectroscopic studies of families became very popular starting since the mid-1990s. The many studies published before 2002 are reviewed in the "Family Spectroscopy" chapter of the *Asteroids III* book [14]. One fundamental result of these campaigns was that families turn out to be quite homogeneous in spectral reflectance properties, and hence they are also likely homogeneous in composition. On one hand, this fact can be seen as a further proof that families identified by modern identification methods are real. In fact, no "impossible" assemblages of objects with spectra incompatible with the

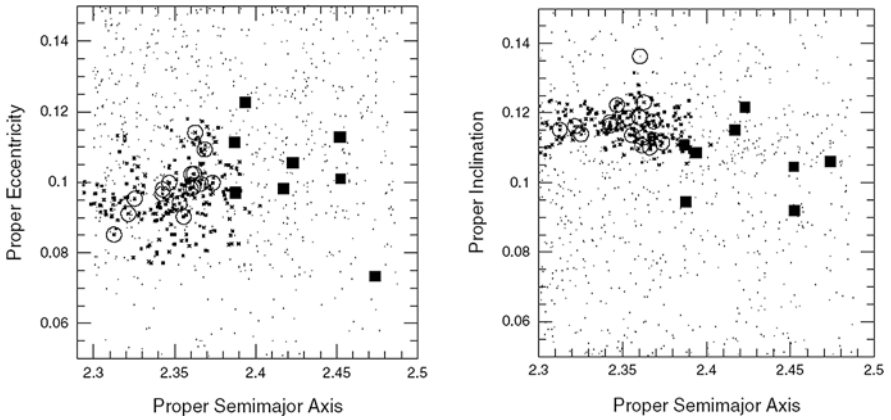

Fig. 7 Copy of the plots originally published by Binzel and Xu [13]. The *small dots in bold* are the supposed Vesta family members found at that time. Among them, those surrounded by a *circle* are objects found to exhibit a Vesta-like reflectance spectrum. The *full squares* indicate other Vesta-like objects not belonging to the Vesta membership list

hypothesis of a common origin have been found among families. Only in some cases some objects are found to have distinctly inconsistent spectra with respect to the majority of other members of the same family. In these cases, however, there are, in general, good reasons to suspect that the discrepant objects are random interlopers, not belonging to the family. As an example, Fig. 8 shows a sample of the brightest and largest members of the Dora family. This figure represents these objects as points in the a'-diameter and in the e'-diameter planes, where a' and e' are as usual the proper semi-major axis and eccentricity, respectively, and the diameter

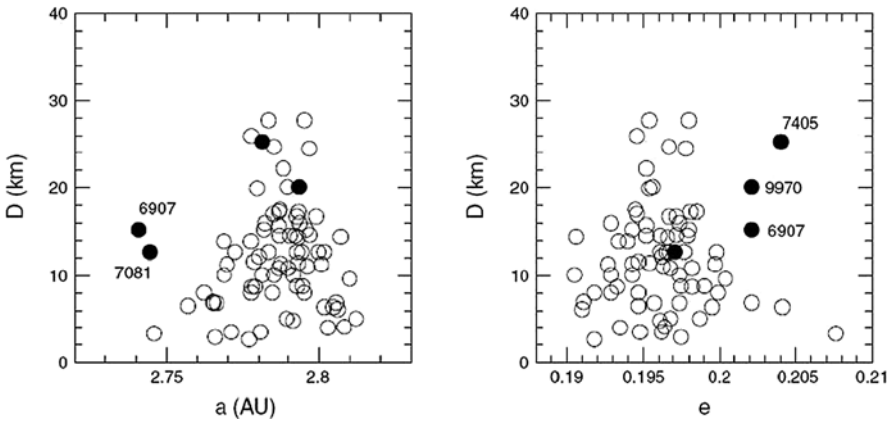

Fig. 8 a'–D (*left*) and e'–D (*right*) plots for some of the largest members of the Dora family. Filled symbols identify objects that have been found to have spectral reflectance properties hardly compatible with those of the majority of family members. Due also to their anomalous locations in these plots, these objects are thought to be likely interlopers in this family (see text)

in kilometers is derived from the known albedos and absolute magnitudes of the objects using the formula

$$\log D = 3.1236 - 0.2\,H - 0.5\log(p_V), \tag{2}$$

where D is the diameter in km, H is the absolute magnitude, and p_V is the albedo, and the value of 3.1236 constant is due to the definition of magnitude and the choice of expressing the sizes in kilometers.

Looking at Fig. 8, it is easy to see that the objects tend to occupy triangular domains in the a'–D and e'–D plots. In particular, smaller objects tend to have more spread proper elements. Thus, smaller objects are, in general, more dispersed in the proper elements space than larger ones. This is a fairly natural phenomenon if we interpret this in terms of ejection from an original parent body, since smaller objects may have been ejected at higher velocities and/or have experienced a more intense orbital evolution. We touch here some delicate point that will be more extensively discussed in Sect. 2.3. The filled circles in the figure represent objects that have anomalous spectra with respect to those that are normal for this family. If one looks at the plots, it is easy to say that these discrepant objects tend to occupy positions in the plots that are outside the triangular domains occupied by most family members. For this reason, we do believe that the discrepant objects are actually random interlopers, that only by chance share the same range of proper elements that characterizes the Dora family. The presence of random interlopers in the nominal membership lists is not unexpected. Due to the statistical criteria adopted to identify families, it is always possible that some objects that have nothing to do with a family are actually included in the member lists [15]. Another possibility is also just the opposite: in the case that the adopted criteria for family memberships are too conservative, it is possible to exclude from the member lists large numbers of actual family members. These facts are always to be taken into account, as we will see later, since it is always possible to infer erroneous conclusions on the family members inventory, simply looking at the nominal member lists.

The quoted example of Dora is fully representative of what is found in general among other families. Spectroscopy becomes in this way a very powerful tool not only for confirming the collisional origin of families but also to identify among the member lists some likely interlopers.

The fact that families turn out to be spectrally homogeneous has been seen as a fairly disappointing fact by some observers. The reason was that for some time, the Holy Graal of spectroscopic campaigns was to find evidence of some heterogeneous family, compatible with the scenario expected for the complete disruption of a fully differentiated object like Vesta, which means to find a number of M-type members (supposed to be metal-rich), in a larger cloud of V-type and A-type (the taxonomic class believed to be most likely diagnostic of an olivine composition). The fact that such a kind of family has not been found may put some further constraints to the models of collisional evolution of the main belt population. In particular, if at least some of the asteroids belonging to the M taxonomic class are really metallic in composition, we may conclude that Vesta is not a unique example of fully differentiated

asteroid, and other objects of this kind were produced in the early phases of the Solar System history (as shown in any case by the existence of metallic meteorites). However, if these differentiated objects were destroyed by collisions, the supposedly big families that were produced by these events are no longer recognizable today. According to some authors [16] this fact is surprising and can hardly be explained, if we do not conclude that these events took place very early, possibly in an era when a much larger total mass of planetesimals was still present in the region presently occupied by the asteroid main belt.

Apart from the discovery that families are likely homogeneous in surface composition, the spectroscopic campaigns led to other exciting discoveries. One of them is shown in Fig. 9.

This figure shows the results of comparative analysis of the reflectance spectra of members of the big Eos family, and those of a handful of objects that are currently located inside the 9:4 mean motion resonance with Jupiter. This resonance is known to be not among the most efficient resonant strips in the asteroid

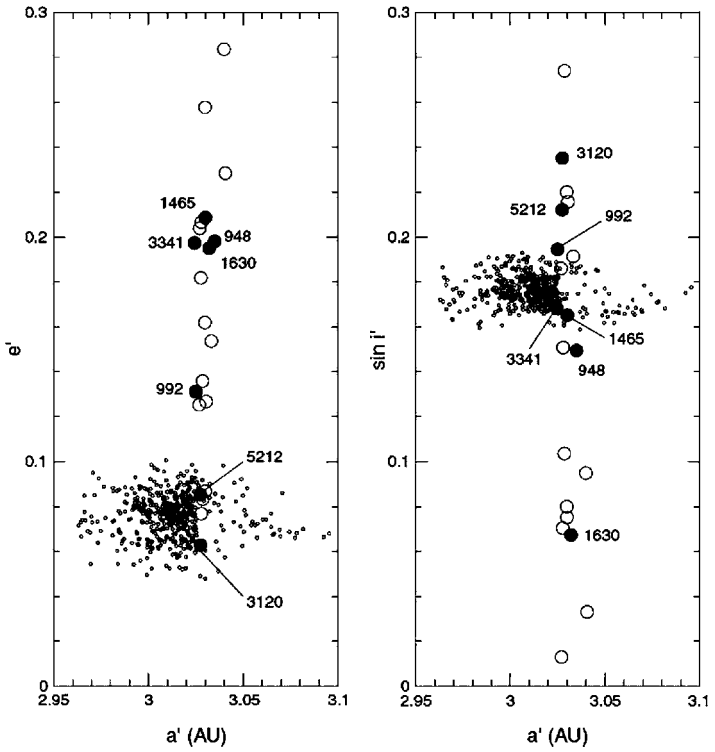

Fig. 9 These two plots in the a'–e' and a'–$\sin i'$ planes show the Eos family as it was known in the mid-1990s. Small symbols are the family members. *Open circles* represent a sample of asteroids not belonging to the family and located into the 9:4 mean motion resonance with Jupiter. *Filled circles* are a sample of the above objects that were observed spectroscopically [17] and were found to share the same unusual reflectance spectrum of the members of the Eos family

main belt, nevertheless dynamical studies show that objects injected into it start to undergo wide oscillations in eccentricity and inclination, and eventually are decoupled from the resonance as a consequence of planetary perturbations, to be either ejected from the solar system or with a much smaller probability, to be captured by Mars, eventually becoming near-Earth asteroids [17]. The resonant objects observed during the spectroscopic campaign did not belong to the Eos family, being clearly decoupled from it in the proper element space, as shown in Fig. 9. Their reflectance spectra, however, were unambiguously found to be similar to those of Eos family members. This result was made easier by the fact that the Eos family is composed of objects belonging to an unusual taxonomic class, named K, that can be clearly distinguished from more usual taxonomic classes. Since the observed objects in the 9:4 resonance were found to be K-type, it was natural to conclude that these objects came originally from the Eos family, and are seen now just at the beginning of a complex dynamical evolution that will eject them out of the asteroid main belt, a quite remarkable result.

The case of the Eos family, composed of uncommon K-type asteroids, is not unique. Actually, a surprisingly high number of families have been found to be composed of objects belonging to rare taxonomic classes, like F, L, K. Among them, the Polana family is a remarkable cluster of F-type objects located in a region of the inner belt where low-albedo objects like those belonging to the F class are quite rare. The Polana family is one of two overlapping families that were found to be clearly distinguishable only on the base of spectroscopic properties [18].

The interpretation of the evidence of a relative overabundance of families composed of objects belonging to rare taxonomic classes [14] has long been a puzzle, and the situation is still not clear. Some evidence of the presence of collisional heating seems to be present in some meteorites [19], but this subject has not been very deeply investigated so far.

What can certainly be said when mentioning spectroscopic observations of families is that this technique is becoming increasingly important as a powerful tool not only to analyze existing families and to look for interlopers, as mentioned above, but also to complement the family searches in the proper elements space, by adding a full new dimension to the problem. In fact, as we will see in Sect. 2.3, the number of objects for which there are currently computed proper elements is steadily growing, and we have now nearly 30 times the number of objects that were analyzed in the family searches carried out in the 1990s. This does not mean, however, that things are easier today. The situation is just the opposite in some respects. In particular, due to the huge number of objects in the current databases, it is extremely more difficult now to disentangle between different families in the regions of the proper element space where they tend to overlap. In these situations, only spectroscopy may be an effective tool to decide whether some objects belong to a family or to another.

2.3 Size Distributions

In the years between 1990 and 1995 the problem of family identification could be considered to have been convincingly solved, mainly after the first confirmations coming from spectroscopic studies. At that point, the attention started to focus on

the task of deriving from these families as much information as possible about the physical processes that had been responsible of their formation, namely collisions.

By researchers interested in the outcomes of catastrophic disruption processes, families can be seen as the results of experiments that are many orders of magnitude beyond what we can do in our laboratories in terms of energy in play. In the laboratory, it is possible to set up experiments of hypervelocity collisions in which targets having sizes of some centimeters can be disrupted, in order to analyze the outcomes of these events, including the size distribution of the fragments, the ejection velocity distributions, and the spin properties of the fragments. But there is no hope to have the possibility of disrupting bodies up to hundreds of kilometers in size, by means of collisions with projectiles with sizes between hundreds of meters and some kilometers, impacting the targets at typical speeds of 5 km/s. Nature performed for free this kind of experiments, when asteroid families were produced, because family-forming events exactly correspond to the kind of collisions just described above.

Since the physics of catastrophic collisions is obviously required to understand and develop models of the collisional history of the asteroid population, when families were finally identified in an unambiguous way, they quickly became a major source of information in this field and the objects of many researches.

One of the first topics of interest was the *size distribution* of these groupings. Figure 10 shows the cumulative size distributions of five of the most prominent asteroid families as they appeared to be about 10 years ago. The plots show the log of the numbers of objects larger than a given size D as a function of $\log D$. The log–log representation is due to the well-known fact that the size distributions of family (and non-family) asteroids are generally well fitted by power-laws, with the number of increasingly smaller objects increasing exponentially for decreasing sizes.

What turned out to be really remarkable in the early studies of family size distributions is the fact that these distributions, down to the values of size for which the family inventory is complete (objects of that size are all sufficiently bright to have been discovered) are really *steep*. To be more precise, family size distributions turned out to be much steeper than the theoretical slope of a collisionally relaxed population.

Some ancillary information is needed here. In the 1970s and 1980s, there was a lot of activity in the field of modeling the collisional history of the asteroid population. One of the first theoretical studies on this subject was done by Dohnanyi, who developed a model in which the disruption of a single object produced a swarm of fragments with a given size distribution [21, 22]. It was also assumed that this process is size-invariant, in the sense that the disruptions of bodies of very different sizes should behave in the same way, producing swarms of fragments whose size distributions scale with the size of the parent body. Under these assumptions, it was shown that a population of such objects, subject only to the evolutive process given by mutual collisions among its members, tends to quickly relax to a fixed cumulative size distribution described by a power-law with an exponent equal to -2.5.

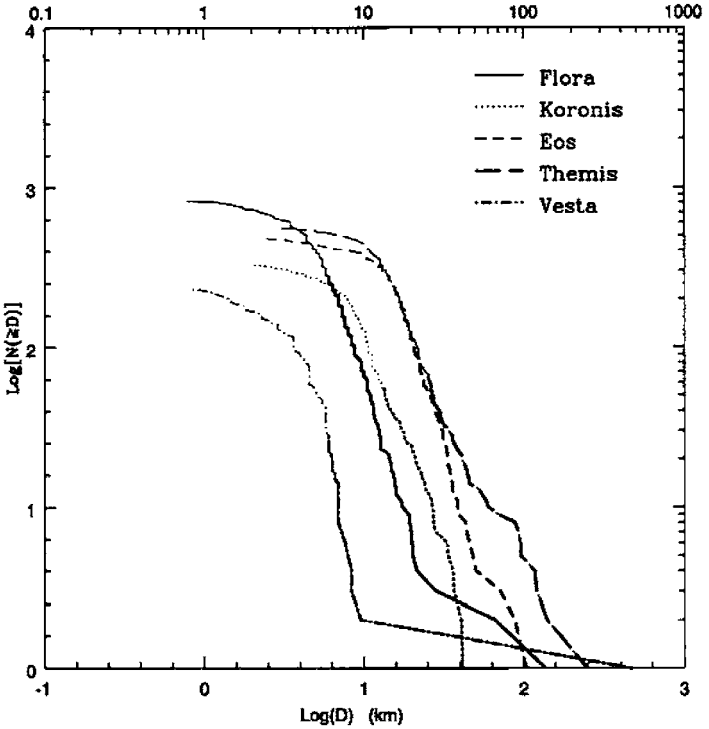

Fig. 10 Cumulative size distributions of the families of Flora, Eos, Themis, Koronis, and Vesta according to [20]

As opposite, starting since the early studies of the size distributions of families and non-family asteroids [23] it was found that the families have power-law exponents well beyond the -2.5 limit, in some cases even beyond -3, the limit value that gives an infinite reconstructed mass. For this reason, it was believed that at sizes smaller than the completeness limit these family size distributions had to relax to more moderate values, although it was not clear at all at which size this change of slope would usually take place.

At this point, historically a few years before the end of the twentieth century, two main problems were open: to discuss the consequences of what had been found concerning the steep size distributions of families and to understand how these size distributions could be so steep.

Let us start first with the latter problem. Let us make, as it was actually done in those years [24] two simple assumptions: (1) in a catastrophic collision the mass of the parent body is conserved (it is equal to the sum of the masses of its fragments); (2) the mass distribution of the fragments is represented by a bi-truncated power-law, in a range of masses between an upper limit, corresponding to the mass of the largest fragment (the largest remnant, as it is usually named), and a lower limit corresponding to the smallest produced fragment. The latter limit may well be very

low, corresponding to dust grains, but it is not zero. Now, by making the above two assumptions, and making a few computations, it is possible to derive some well-defined predictions for the size distributions of families. In particular, a well-defined relation can be found between the slope of the cumulative size distribution of the family's and the value of the ratio m_{LR}/m_{PB} between the mass of the family's largest remnant and the mass of the original parent body. Such a relation is shown in Fig. 11.

As can be seen, the slope of the power-law which describes the size distribution turns out to be increasingly steeper for decreasing values of the m_{LR}/m_{PB} mass ratio. This is not surprising in the framework of this model, because a more massive largest member means that a smaller amount of mass is left to be distributed among the other fragments. Unfortunately, this predicted trend is spectacularly contradicted by the behavior exhibited by real families. If we go back to Fig. 10, we may see that the steepest size distribution is that of families with a very big largest remnant, like in the case of Vesta, whereas a much shallower trend is exhibited by the family of Koronis, which is characterized by a large number of largest fragments having

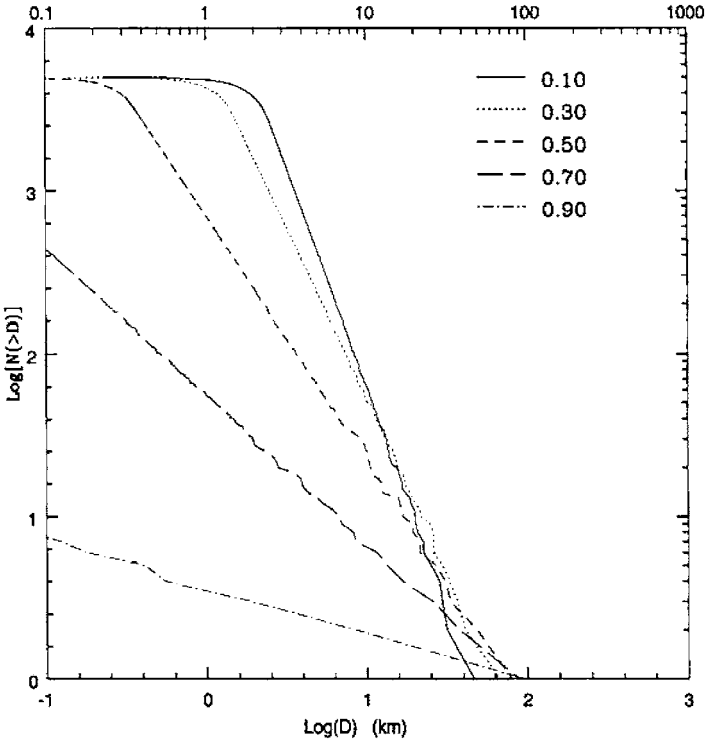

Fig. 11 Predicted size distributions of asteroid families characterized by different mass ratios between the largest remnant and the parent body (indicated in the figure) according to the purely mathematical model [24] (See text. Plot taken from [20])

approximately the same size. Then, something is wrong in the model expectations, and this must be explained in some way.

If one looks again at the size distribution of the Vesta family shown in Fig. 10, it is easy to see that the trend is characterized by a long straight segment linking the size of Vesta to that of the second largest member of the family. Starting from this object, the size distribution becomes a continuous curve composed of a large number of points, distributed according to a very steep power-law. This behavior is the key to understand the explanation that was given to the observed family size distributions in a classical paper in 1999 [20]. The argument is the following: any prediction based on an abstract mathematical form of the size distribution and on mass conservation must fail if it does not take into account some geometrical constraints imposed by the sizes of the individual fragments. In particular, different fragments cannot mutually overlap, and the space at disposal is limited by the finite volume of the parent body. Let us suppose for sake of simplicity that both the parent body and the fragments have all spherical shapes (an assumption which is not critical for the conclusions of this argument). Now, if the parent body has a diameter of, say, 100 km, and the largest remnant has a diameter of, say, 60 km, there is not any possibility for the second largest fragment to have a diameter larger than 40 km in the best possible case. In other words, the volume of each single fragment limits the space available to the formation of the others.

It is clear that we are implicitly making an oversimplification of the process of fragmentation, since real fragments are produced as a complicated effect of propagation of mechanical waves in the volume of the parent body. The latter is characterized by its own properties, including the presence of pre-existing material faults and cracks, and it is clear that the fragments are not obliged a priori to have predetermined shapes. Moreover, the final fragments may also be affected by effects of mutual reaccumulation. Bearing in mind these obvious objections, in order to avoid overinterpretation of the results, one can nevertheless write some simple numerical algorithm which simulates the formation of fragments taking into account the constraint of non-mutual overlapping and non-extrusion from the original parent body's volume. This exercise was first done in 1999 [20], and the results were striking. The resulting size distributions give an excellent fit of the observed size distributions of real families. Figure 12 is the same as Fig. 11, but this time it shows the predictions based on the geometric model just mentioned above. This time, the behavior is in good agreement with the trend shown by real families in Fig. 10. Families with a very big largest remnant, like Vesta, exhibit the steepest size distribution, whereas a more "democratic" family like Koronis, having a large number of largest fragments approximately equal in size, exhibits a much shallower slope.

Some fits of individual families, obtained by using the geometric model and looking for the values of the parent body size and the m_{LR}/m_{PB} mass ratio which produces a best fit to the data, are shown in Fig. 13. As can be seen, the fits appear quite good, even surprisingly good by taking into account the above-mentioned oversimplifications of this geometric model.

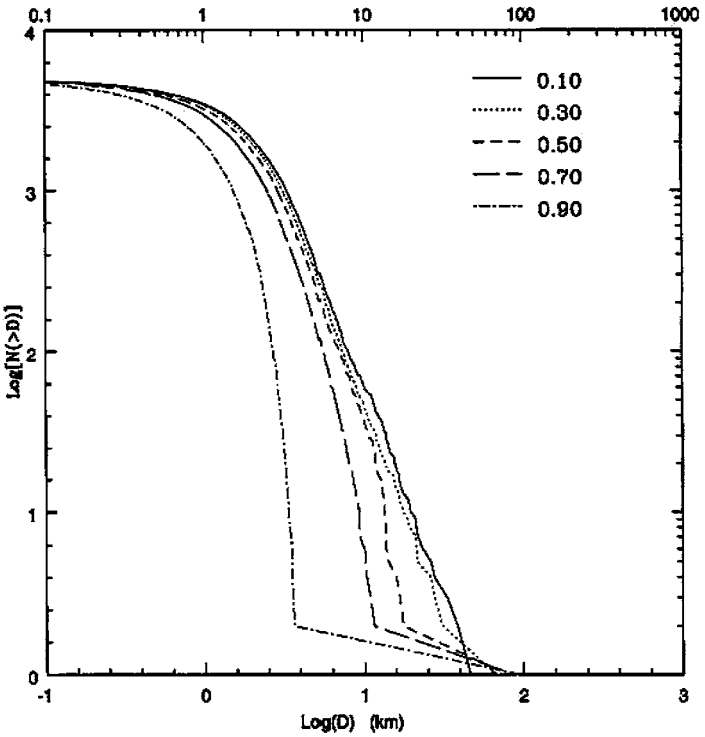

Fig. 12 Predicted size distributions of asteroid families characterized by different mass ratios between the largest remnant and the parent body (indicated in the figure) according to the geometric model [20]

After the publication of the paper presenting the geometric model [20], the steep slopes of the size distributions of asteroid families could be, at least qualitatively, explained. The geometric model not only gave some excellent fits of the size distributions exhibited by the major families known at that time, but also could be used to derive at least some indication concerning the a priori unknown values of the original parent bodies' sizes and the m_{LR}/m_{PB} mass ratios for these families. The general results of this exercise are summarized in Table 2. As can be seen, many families were formed, according to this kind of modeling, by the disruptions of objects up to 300–400 km in size. Moreover, some families are likely the outcomes of extremely energetic events, capable of producing largest remnants with masses only a few hundredths of the parent body.

We stress again that, due to obvious oversimplifications of the basic assumptions of the geometrical model, the results shown in Table 2 cannot be taken too literally and should be interpreted mostly in a statistical way than as an accurate fit of single families.

On the other hand, it is also worth to remind that since a long time it is known that some dust belts identified in the sky by thermal infrared surveys are associated with

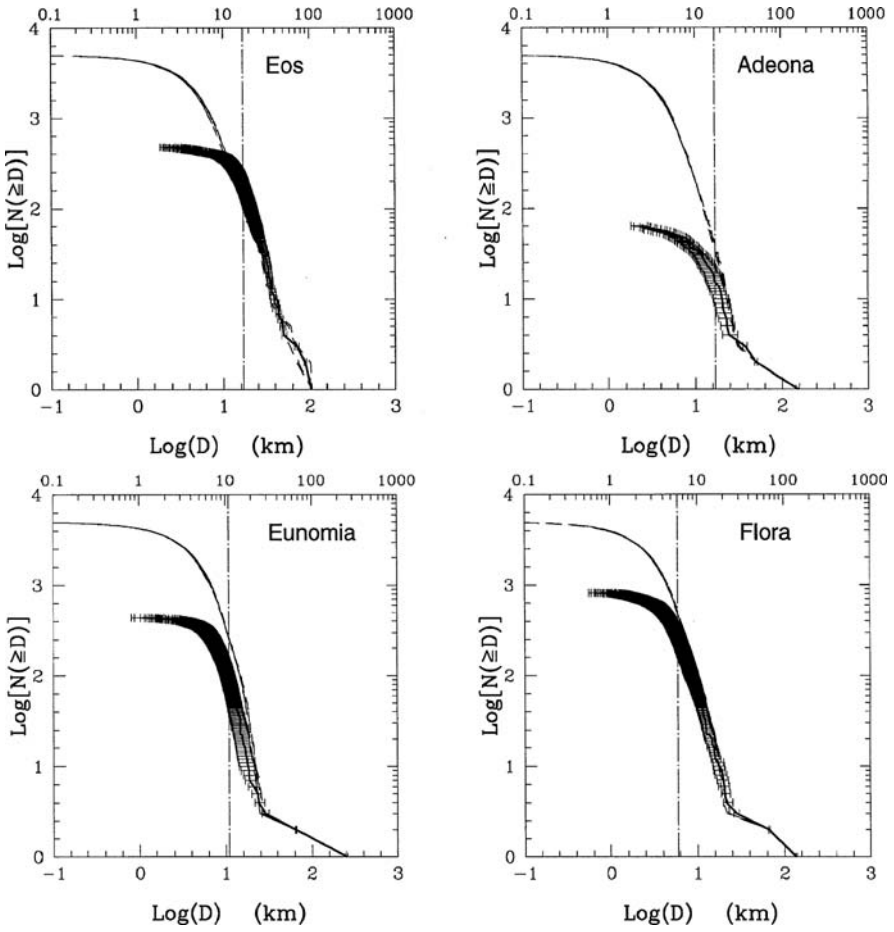

Fig. 13 Best fits of the size distributions of the families of Eos, Adeona, Eunomia, and Flora, obtained by applying the geometric model [20]. The thickness of the observed family size distributions is due to uncertainties in the sizes of the objects

some families. This is another reason to conclude that family-forming events, possibly followed by further second-generation object disruptions, may actually produce huge amounts of fragments down to very small sizes.

2.4 The Role of Families in the Asteroid Inventory

Figure 14 shows that, even qualitatively, asteroid families tend to become more evident if increasingly larger samples of asteroid proper elements are considered. According to this and other kinds of evidence discussed in Sect. 2.3, there was growing evidence in the 1990s that the size distributions of asteroid families were described by quite steep power-laws, much steeper than those found to describe

Table 2 The parent body size D_{PB} and the m_{LR}/m_{PB} mass ratio for some major asteroid families analyzed by [25]

Family	D_{PB} (km)	m_{LR}/m_{PB}
Adeona	189	0.51
Dora	88	0.03
Eos	218	0.11
Erigone	91	0.50
Eunomia	284	0.73
Flora	164	0.57
Gefion	74	0.06
Hygiea	481	0.61
Koronis	119	0.04
Maria	130	0.05
Massalia	151	0.90
Merxia	42	0.35
Themis	369	0.31
Vesta	468	0.95

the size distribution of non-family objects in different regions of the asteroid main belt. In particular, it turned out that non-family asteroids exhibited exponents of the size distribution shallower than the -2.5 theoretical value for a collisionally relaxed population, according to Dohnanyi's theory, as explained above. As opposite, families exhibited much steeper slopes than the Dohnanyi value. This fact had been already apparent based on a preliminary analysis of the database of asteroid sizes and albedos produced by the thermal IR observations of the IRAS satellite [23], as shown in Fig. 15.

Taken at face value, the above-mentioned results concerning the different size distribution of family and non-family asteroids have some important consequence on the inventory of the main belt population down to small sizes. In particular, if family size distributions are so much steeper than the size distribution of the population of non-family objects, it follows that at small sizes, below the limit of completeness of the observed population, family members should dominate the asteroid inventory.

Some care is needed, however, before drawing conclusions that might be shown to be erroneous. To better understand this delicate point, it may be useful to examine the plot shown in Fig. 16, which shows the size distribution of the Eunomia family according to data available around 1995.

The figure clearly shows that the size distribution of this, as of most families, is characterized by a trend corresponding to a steep power-law exponent down to some limit size value. At smaller sizes, the curve starts to become shallower, until the number of objects becomes constant, corresponding to the total number of known family members. The fact that the size distribution becomes shallower at small sizes should not be interpreted as an intrinsic property of the size distribution, because it is simply due to the fact that starting at some critical size level, the completeness of the sample is no longer full, or in other words we do not have yet observed all the really existing objects smaller than the completeness limit. The completeness size limit corresponds to the size value at which the cumulative distributions exhibit

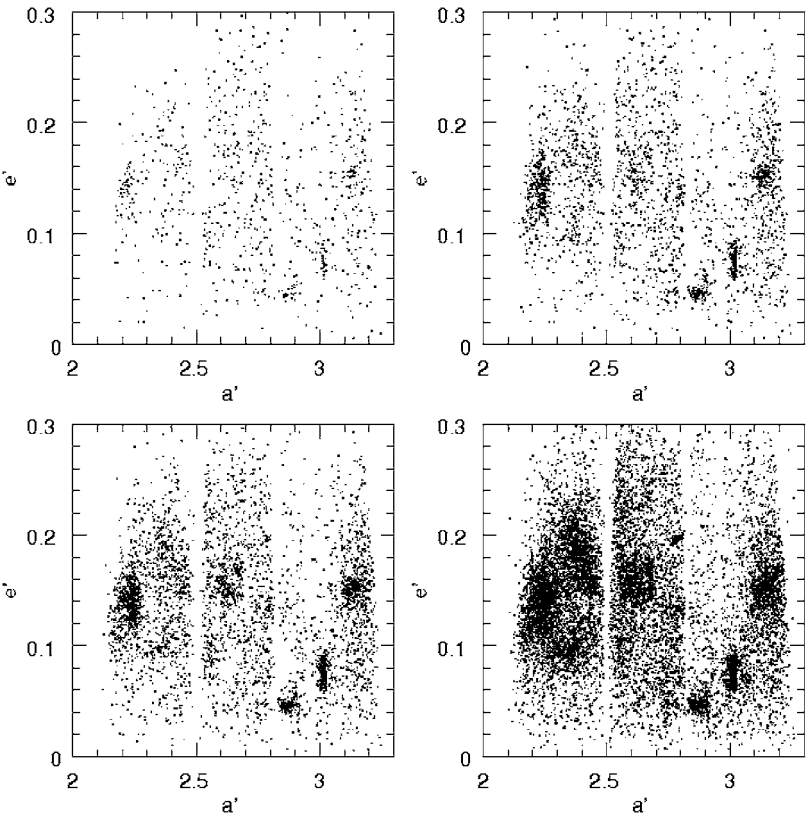

Fig. 14 Proper eccentricity versus proper semi-major axis plots for increasingly larger samples of asteroids. The plots show numbered asteroids up to $N = 1{,}000$ (*top left*), up to $N = 3{,}000$ (*top right*), up to $N = 5{,}000$ (*bottom left*), and $N = 12{,}000$ (*bottom right*). From the plots, it is apparent that asteroid families become progressively more evident when considering increasingly larger samples of objects

a change of slope and start to become shallow. In the figure, it turns out that the completeness limit for the Eunomia family as it was known a dozen of years ago is about 10 km.

To estimate the number of existing family members at sizes smaller than the completeness limit, some extrapolation is thus necessary. Such extrapolation, however, is a delicate affair.

In 1996 [26] this problem was analyzed using the following approach: the size distribution extrapolation was done down to a value of 1 km using two different methods. One method was a simple extrapolation of the observed size distribution above the completeness level. In this way, one gets some resulting n_1 number of objects larger than 1 km, as shown in Fig. 16. This led to an upper limit of the number of 1-km family members. Such kind of extrapolation may be questionable, however, because in many cases the size distribution above the completeness value is

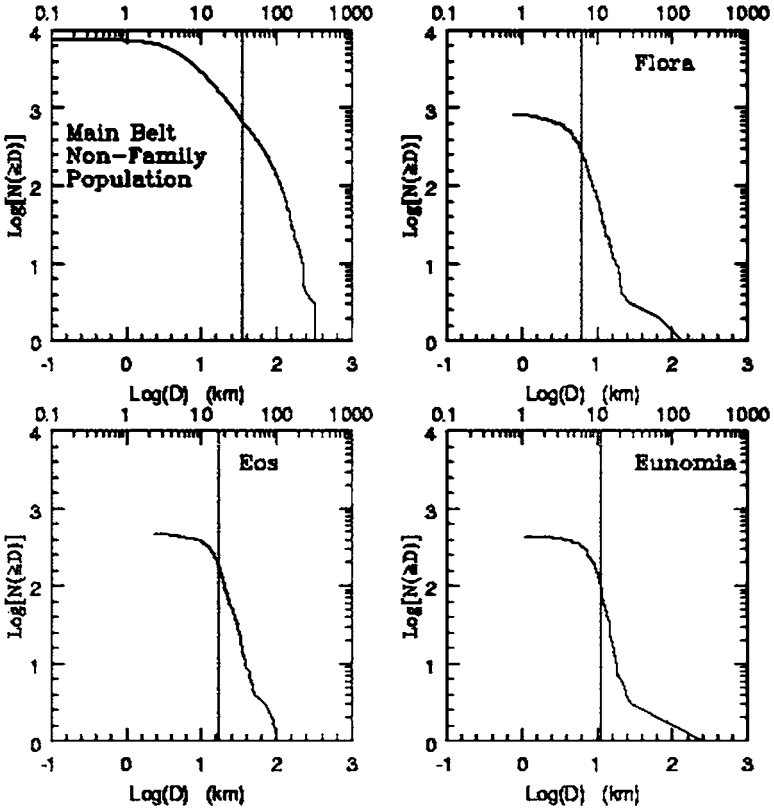

Fig. 15 Log–log plots of the size distribution of non-family main belt asteroids compared with that of some families as they were known in 1991 [23]

really very steep, so that a simple extrapolation of it may possibly lead to some likely overestimate of the number of objects at small sizes. When the size distribution exponent is above the -3 value, moreover, it is certain that the slope must relax to more moderate values at some size below the completeness limit, just because a simple extrapolation down to zero would give an infinite number of family members, corresponding to an infinite mass of the parent body. Unfortunately, on the other hand, there is not any a priori reason to believe that the size distribution should change at some known value of size, nor is it clear to which value of the exponent the size distribution should converge, if any.

The second extrapolation method adopted by [26] consisted of an extrapolation of the observed size distribution composed by two parts: the first one was a simple extrapolation of the observed size distribution, but limited only to the interval between the completeness size limit and a size of 5 km (chosen arbitrarily). Below 5 km, it was assumed that the size distribution followed a Dohnanyi law, characterized by a -2.5 power-law exponent. This led to an alternative value n_2 for the number

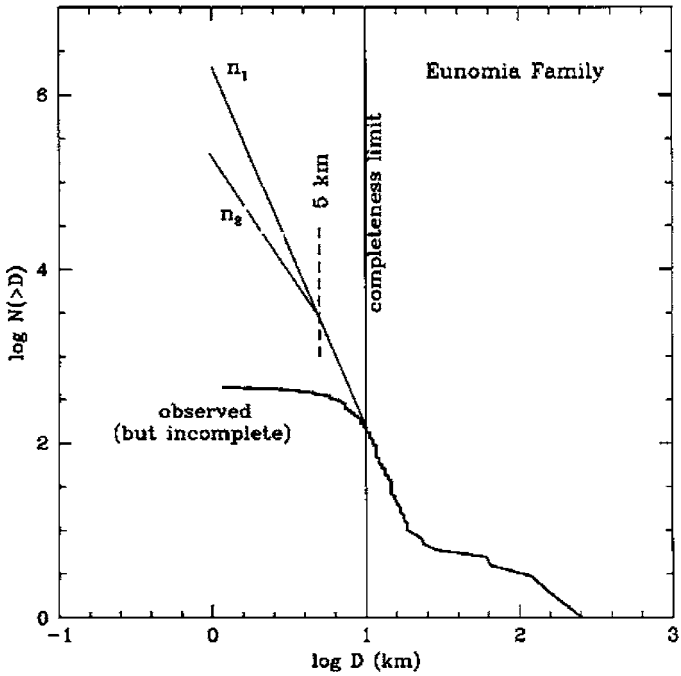

Fig. 16 The size distribution of the Eunomia family as it was known around the year 1995. The plot shows the size completeness limit at that epoch, as well as two possible extrapolations of the size distributions down to 1 km in size

of objects larger than 1 km. Figure 16 gives a graphical representation of the two methods.

The result of this analysis was that, although with large uncertainties, the contribution of asteroid families to the inventory of the main belt population is extremely important. While at a size of 10 km, family and non-family asteroids contribute approximately for a 50% each of the population, at smaller sizes the contribution of family members, mainly from a few very big ones like Themis, increases very much. The nominal value of the family contribution at a size of 1 km turned out to be 99%.

These results have never been universally accepted. Even in those same years, some analyses of the likely inventory and size distribution of the asteroid population, based on an assessment of the discovery efficiency of objects having different apparent magnitudes, concluded that the size distribution of main belt asteroids is not very steep at small sizes, and there is not any evidence of a likely domination of family members [27]. A more detailed discussion of the situation taking into account the observational evidence that is available today will be presented in Sect. 3.

A main belt population dominated by asteroid families would have several consequences, and some of them will be discussed in the following sections. Here, we only note that, among them, one would be an important effect on the intrinsic collision probability throughout the main belt. In particular, in a main belt dominated

by a few very populous families, the collision probability would be higher in the regions of the semi-major axis—eccentricity plane swept by the members of these families [28].

2.5 The Reconstruction of Family Velocity Fields

We have seen in Sect. 2 that a relation exists between the components of the ejection velocity of a fragment escaping from its parent body and the resulting difference between its orbital elements and those of the parent body. This relation is expressed by the Gauss equations (1). As a consequence, we have that in principle one could, having at disposal a family, try to infer the values of the original ejection velocity components of each family member by simply looking at its coordinates in the proper element space. The idea is that, if we assume that each object was ejected from the location of the current family barycenter, one could compute the components of its original ejection velocity from the parent body, by simply using the Gauss equations, knowing the differences in proper elements between the object and the family barycenter.

Of course, this would be in principle a very interesting result in many respects. In particular, a reliable reconstruction of the kinematical properties of an event of catastrophic disruption would provide very important constraints to the physical models of these events and possibly would shed some light on some structural properties of the family parent bodies.

However, the idea of reconstructing the original ejection velocity fields of family-forming events must face two fundamental problems. One, that will be discussed in Sect. 3, is related to the fact that it is not granted for sure that current family members have not experienced significant dynamical evolution since the time of their creation. In fact, if the proper elements of current family members have changed for any reason with respect to their original values, any attempt at deriving information on the original ejection velocity values starting from the present proper element values is intrinsically dangerous and might lead to misleading, or completely wrong, results. As quoted above, this problem will be more extensively discussed in Sect. 3, then for the moment let us forget it.

The second problem is that, if one looks at Gauss' equations, it is easy to see that the link between velocity components and proper element differences is not immediate, but it depends on the values of two a priori unknown parameters, namely the true anomaly f of the family parent body at the epoch of its disruption and the value of its argument of perihelion ω at the same epoch.

The problem that f and ω are unknown seems in principle a fundamental one. Any method of reconstruction of the original ejection velocity field of the fragments should be in principle able to produce some reliable estimate of these unknown angles, but it is hard to imagine how this could be achievable in practice. However, an analysis carried out in 1996 [29] showed that the problem is not hopeless. The basic idea is the following: if one tries to invert Gauss equations using the right

values of f and ω, the correct ejection velocity field will be obtained. The resulting field will be instead increasingly wrong as one chooses increasingly wrong values of the unknown angles. If the ejection velocity fields could be assumed to be completely random, with structures not showing any predictable properties, the inversion of the Gauss equations would not be possible in principle. But laboratory experiments tell us that the situation is different. Some common properties of the ejection velocity fields observed in laboratory hypervelocity collisions show that the ejection velocity fields generally present some characteristic properties. The most common and general property is that the fields turn out to be noticeably *axisymmetric*. The symmetry axis generally coincides with the diameter connecting the center of the disrupted object and the impact point.

According to the above considerations, one cannot pretend that the ejection velocity values of single fragments can be accurately predicted. However, based on the symmetry properties of the ejection velocity field, one can expect, in general, that the *distributions* of the velocity components V_T, V_R, and V_W are not simply random, but they satisfy some properties dictated by the general structure of the velocity field. This fact can be directly exploited to obtain an estimate of the unknown angles in Gauss equations. Focusing on the V_T, V_R velocity components, which are affected by the f angle only, some dimensionless parameters were built, which are functions of the unknown f angle in the Gauss equations. The following two were used in [29]:

$$Z = \frac{\sum_i V_{R_i}^2 - \sum_i V_{T_i}^2}{\sum_i V_{R_i}^2 + \sum_i V_{T_i}^2}$$

and

$$\alpha = \frac{\sum_i (V_{T_i} \cdot V_{R_i})}{\sqrt{\sum_i V_{R_i}^2 \cdot \sum_i V_{T_i}^2}}.$$

The above Z and α parameters can be used as indicators of the overall symmetry of the field and vary as a function of the assumed value of the unknown f angle. The dependence of Z and α upon f was tested in a number of numerical simulations in which synthetic ejection velocity fields were created, being characterized by a variety of possible structures (spherical fields, ellipsoidal fields, conic fields, etc.) and using different values of the "true" f angle. The result was that the hypothesis of axial symmetry of the resulting field could be translated into the requirement that the Z parameter reaches a minimum, or that α becomes equal to zero. These requirements were found to be sufficient to find a corresponding value of the f angle satisfactorily close to the "true" value used to build the simulations. In other words, the symmetry properties of the velocity field could be used to derive a fairly good estimate of the unknown f angle appearing in Gauss' equations. The same was found to be true also for the other unknown angle, ω, although in this case the

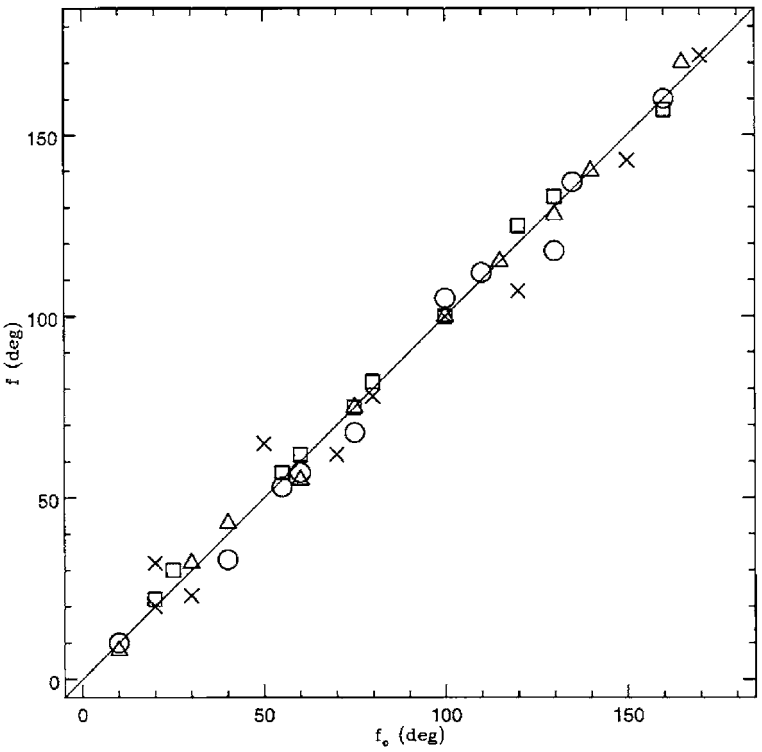

Fig. 17 Plot of the estimated angles f versus the "true" f_0 angles for a large set of simulations. Different symbols refer to conic (*crosses*), off-center spherical (*circles*), biaxial ellipsoidal (*triangles*), and asymmetric triaxial ellipsoidal (*squares*) fields, respectively. This figure is taken from the original paper [29]

uncertainty was larger. The results of this analysis are shown in Fig. 17 and 18, respectively.

The results of this analysis [29] were thus quite encouraging, and indicated that, in a large variety of simulated cases, the reconstructed fields obtained by applying this technique were on the average similar to the simulated fields, as shown in Fig. 19.

Based on this technique, it was possible to derive the overall structures of the fields of several families. A couple of results, referring to the families of Vesta and Maria, are shown in Figs. 20 and 21, respectively. In Fig. 20, both the apparent structure of the Vesta family in the proper element space is shown, as well as the corresponding structure in the space of ejection velocity components. Figure 21 only shows one projection of the Maria family structure in the velocity space. The overall kinematical structures of the two families look "reasonable" when compared with similar plots referring to the outcomes of experiments of catastrophic disruption in the laboratory. In both figures, the size of the symbols used to represent family members is directly proportional to the corresponding size of the object in

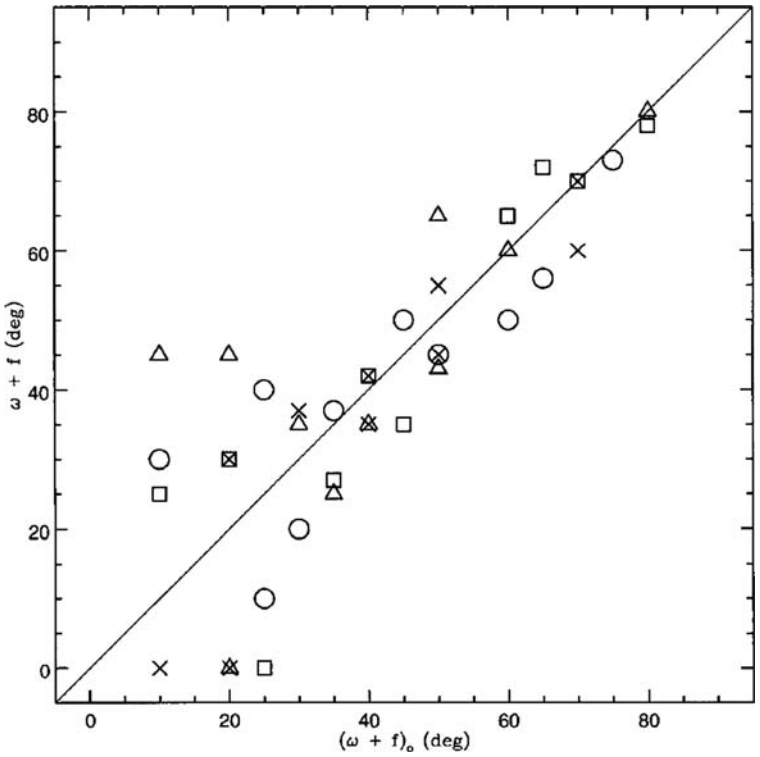

Fig. 18 The same as Fig. 17, but for the $\omega + f$ angle

kilometers. Since in this way the asteroid (4) Vesta would appear exceedingly big, it has not been included in Fig. 20, and its location is indicated by the intersection of two perpendicular lines in the plot. In Fig. 21 the plot also includes the locations of the borders of the strong 3:1 mean motion resonance with Jupiter in the velocity space. It is evident that the family is just hanging on "3:1 precipice," and this leads us to Sect. 2.6.

2.6 Families as Sources of Asteroid Showers on the Earth

The case of the Maria family, which appears to be located just on the border of one strong resonance, as shown in Fig. 21, is certainly not unique. Several important families are located on the border of one or more resonances. A list includes, in addition to Maria, the families of Themis, Eos, Koronis, Dora, and Gefion. The above list includes then several of the most important and populous families in the main belt. It is known that the most important mean motion resonances with Jupiter correspond to the well-known Kirkwood gaps in the asteroid main belt, namely some narrow strips corresponding to forbidden values of the orbital semi-major

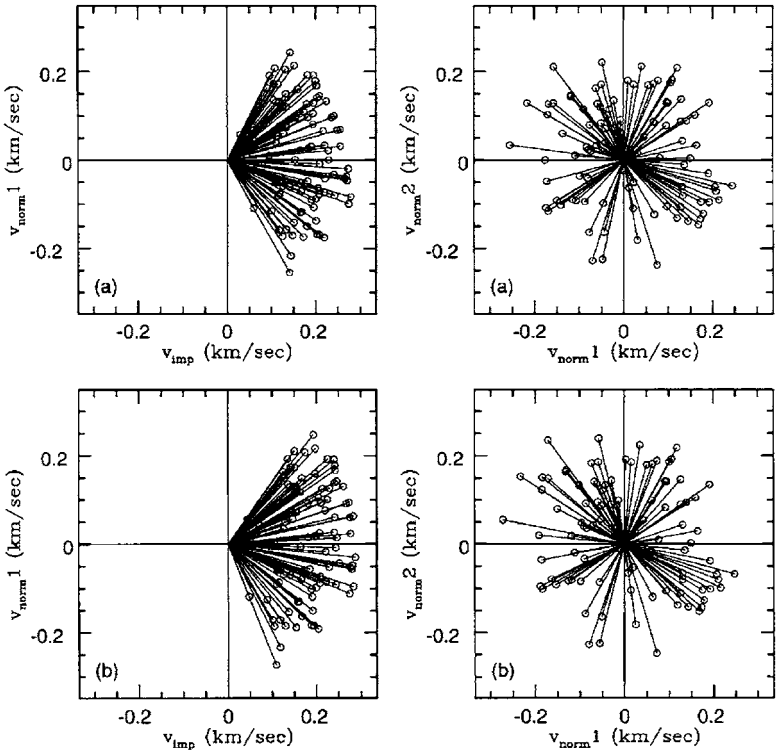

Fig. 19 Example of an application of the method of ejection velocity field reconstruction in a simulated case of a velocity field having an overall conic structure. Above, two projections of the simulated velocity field. *Bottom*, the same projections, but for the reconstructed field. Plot taken from [29]

axis. The same is true for the secular ν_6 secular resonance, as well as for a number of other resonances which are found to cross the proper element space in the region of the asteroid main belt. Figure 22, taken from [30], shows a visual representation of several of these resonances. The mean motion resonances with Jupiter, that produce the Kirkwood gaps, are also evident in other figures, for instance Fig. 2. The important fact is that all these resonances are associated with the notion of chaotic motion. In other words, any object whose orbital elements are such as to fall into one of these resonant zones, is subject to a chaotic dynamical evolution, which rapidly produces wide oscillations of the orbital elements, mainly eccentricity and inclination, possibly leading to close encounters with some major planet, producing big changes in orbital semi-major axis and consequent removal from the region of the asteroid belt.

The fact that many important families appear to be sharply cut by some neighboring resonance was interpreted in the 1990s as an indication that the original disruption events that produced these families had been sufficiently energetic as to eject many fragments into such resonances. These objects are no longer there,

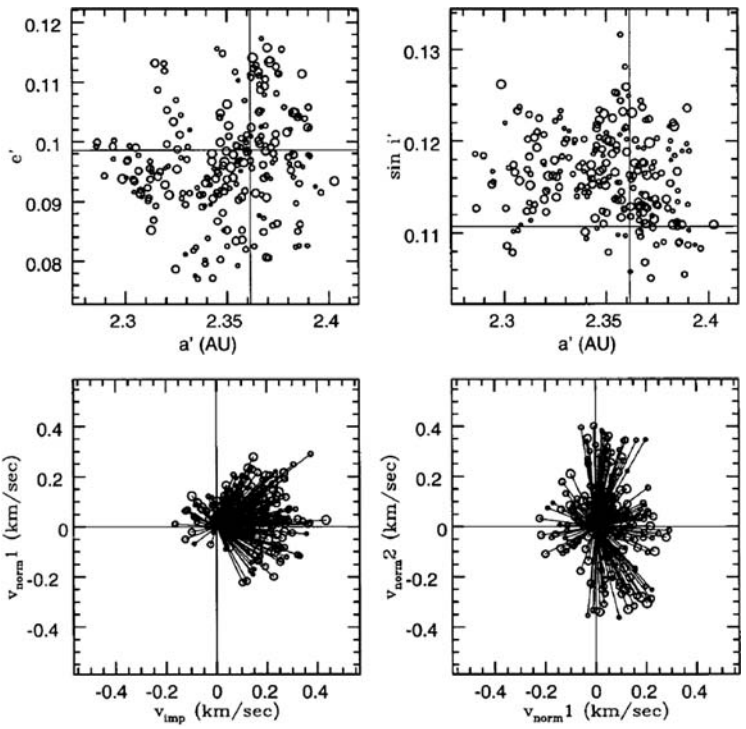

Fig. 20 The reconstruction of the ejection velocity field of the Vesta family according to [29]. The two plots on the *top* show the structure of the family in the proper elements space, whereas the two plots at *bottom* show the same, but in the space of the velocity ejection components. The sizes of the symbols are directly proportional to the corresponding sizes of the asteroids in kilometers. Since the very big (4) Vesta asteroid would be represented by an exceedingly large symbol in this plot, it is not represented in the plots, but the location of Vesta is indicated by the intersection of two *perpendicular lines* drawn throughout the plots

because they have experienced a chaotic dynamical evolution and have long been removed from the asteroid belt. In particular, many of them may have been moved to the region of the terrestrial planets, contributing to the inventory of near-Earth objects (NEOs).

It is known that the NEO population is composed by objects having short dynamical lifetimes and cannot exist for long times before being either disrupted or removed from the NEO region. For this reason, new objects must be steadily supplied by some mechanisms. In the 1990s, the discovery that many families are just on the border of some powerful resonances led to the natural idea that these families might be an important source of NEOs, and numerical simulations were performed to test this hypothesis.

In 1997 a fundamental paper [31] was published, presenting the results of an extensive analysis based on simulations of the orbital evolution of a large number of simulated family members injected into nearby resonances. The simulations were

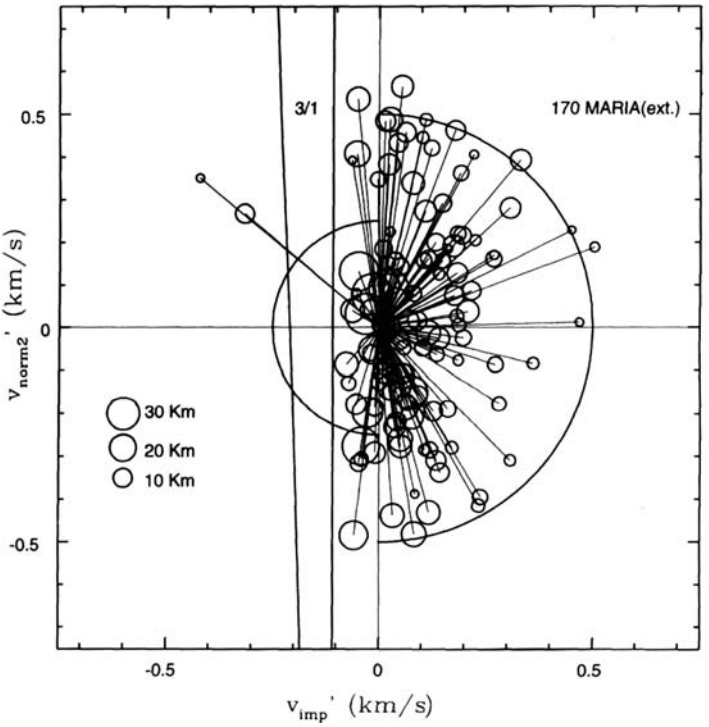

Fig. 21 The reconstruction of the Maria family according to [29]. Only one projection in the space of the components of ejection velocity is shown. The sizes of the symbols are proportional to the corresponding sizes of the Maria family members. Note that the sizes of the objects shown in this plot are not negligible, being in general of the order of 10 km or larger. The location in this plot of the borders of the 3:1 mean motion resonance with Jupiter is also shown. The two semi-circles have not any particular meaning, but for the fact of showing how an isotropic ejection velocity field (considered at two different values of velocity) appear to be cut and destroyed by the presence of the resonance, which is a strongly chaotic region in the space of orbital elements. Note also that the two objects in the plot located beyond the *left* border of the 3:1 resonance are likely not real family members

based on reasonable extrapolations of the possible original structures of a number of current families which are known to be cut by some resonance. Many of these family members were found to fall into these resonances, and their orbital evolution was numerically integrated to analyze their final fates.

The results of this analysis were striking. The orbital evolution of the objects injected into resonances were very fast. Objects injected into the 3:1 or ν_6 resonances were led to impact the Sun itself, as a consequence of orbital eccentricity being pumped up to the point that the perihelion distance becomes smaller than the Sun's size. It is worth to note that objects following such kind of evolution have also non-zero probability to impact the terrestrial planets during their evolution. Many other objects turned out to be quickly removed from the Solar System. This was mostly the case for objects injected into resonances in the outer region of the main

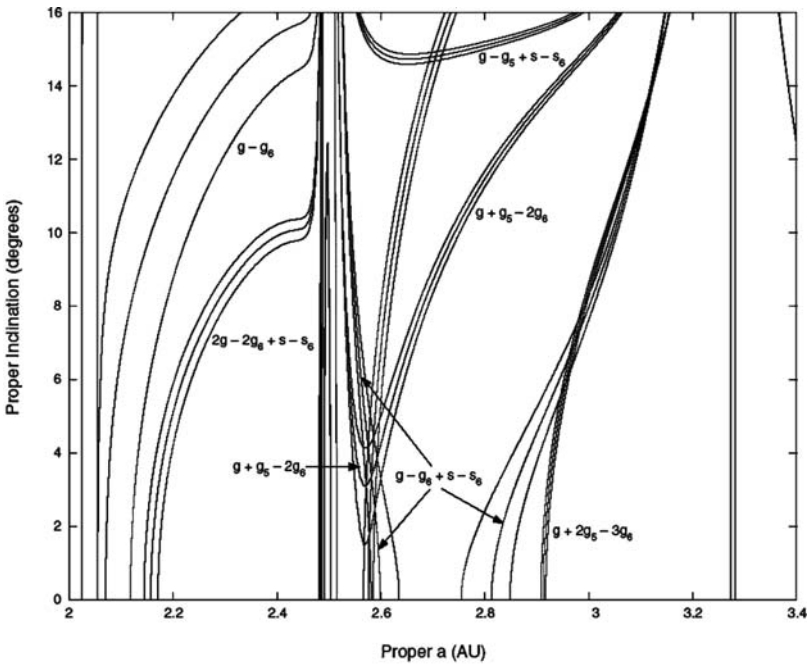

Fig. 22 The location of several of the most important resonances crossing the asteroid main belt. The plot is taken from [30]

belt, for which an eccentricity increase likely leads to close encounters with Jupiter. The typical timescales of the evolutions of objects injected into different resonances are shown in Fig. 23.

The simulation [31] showed clearly that the lifetimes of objects injected into resonances like the 3:1 or the ν_6 are extremely fast, even too fast, as we will discuss below. Taken at face value, and under the above-mentioned hypothesis that family members were immediately and directly injected into nearby resonances at the epoch of the disruption of the family parent body, the results show that family-forming events could produce real asteroid "showers" [32] which could affect the terrestrial planets. Table 3 summarizes for different families the number of 1-km family members that could be expected to have impacted the Earth as a consequence of family formation and the duration of these showers in Myr. The resulting number of potential impactors and the duration of the shower is a complicated function of the family structure, location, and efficiency of the involved resonance(s). In several cases it appeared that the expected showers could have been sufficiently energetic as to produce likely consequences on the evolution of the terrestrial biosphere, a quite interesting result per se.

It is important to note already at this stage that the dynamical evolutions of objects injected into resonances like 3:1 or ν_6 turned out to be unexpectedly fast at the epoch when these results were first obtained [31]. A couple of fundamental difficulties were that, to sustain the NEO population in a steady state, the required

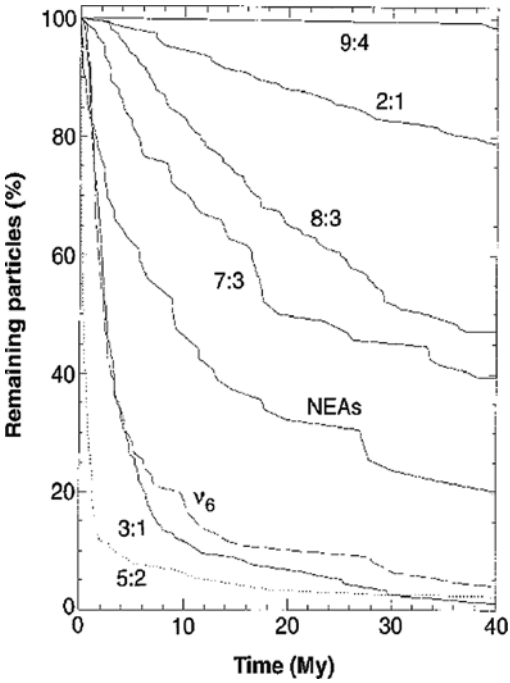

Fig. 23 This plot, taken from [31], shows the number of remaining objects (expressed in percent) as a function of time for samples of simulated family members injected into different resonances. The remaining objects are those that, as a function of time, are still existing, not having been removed from the Solar System and not having impacted the Sun (see text). For a comparison, the evolution of known near-Earth asteroids (NEAs) is also shown

flux of family members would be exceedingly high, if the evolution of these objects are so fast. In other words, to supply a steady state of NEOs able to explain the population that exists today, many family-forming events should be assumed to be necessary, if direct injection into resonance was the only one or the most important mechanism of NEO supply. There would be thus a problem of "missing families," since those that we identify today are not sufficient to justify a steady state NEO population over long timescales.

Table 3 Summary of predicted "asteroid showers" following the formation of different families according to [32]

Family	Resonance	$N_{impacts}$	Duration (Myr)
Flora	ν_6	4–11	30
Vesta	3:1	0–1	10
Eunomia	3:1	12–135	10
Eunomia	ν_6	0–4	15
Gefion	5:2	2–30	5
Dora	5:2	2–14	5
Koronis	5:2	0–2	5
Eos	9:4	2–10	140
Themis	2:1	3–7	90

The different columns give the family number, the resonance through which fragments are delivered to Earth, the range of predicted impacts by fragments 1 km in size, and the overall duration of the expected shower

Another great problem was that the resulting orbital evolutions, with typical lifetimes of 2 or 3 millions of years were exceedingly fast also when compared with the observed cosmic rays exposure ages exhibited by meteorites. Meteorite analyses show that these objects have been subject to irradiation from cosmic rays and solar wind over timescales much longer than the resulting dynamical lifetimes of their supposed progenitors, if we have to believe that direct injection into resonance following collisional events is the only one mechanism to supply NEOs and meteorites.

The solution of this paradox will be discussed in Sect. 3. Here, we only note that, in any case, whatever is the duration of the dynamical evolution of family members eventually injected into some resonant zone, it is in any case true that the events that produced big families must be expected to have produced large numbers of fragments that, possibly with a larger variety of possible timescales as it will be discussed in Sect. 3, may later have been delivered to the NEO region.

2.7 The Size–Ejection Velocity Relation in Families

The last logical step in the studies of the physical properties performed between 1990 and 2000 was an analysis of a possible size–ejection velocity relation among family members. We have seen above (see 2) that the sizes of the objects may be derived from knowledge of their absolute magnitudes and using an average albedo value for each family, as suggested by the overall homogeneity in surface composition of family members resulting from spectroscopic studies. On the other hand, the ejection velocity of a family member may also be derived from knowledge of the difference in its proper elements and those of the family barycenter, as we have also seen in Sect. 2.5.

In 1999, an extensive analysis of a size–velocity relation among family members was published [33]. The basic idea developed in that paper was that one may generalize to families a result found in laboratory experiments, namely that in a collisional event a fraction f_{KE} of the specific impact energy E/M is converted into kinetic energy of the fragments. Here, E indicates the impact energy, practically equal to the kinetic energy of the impacting body, while M is the mass of the impacted body. If one assumes that the resulting kinetic energy of ejection of any fragment is $1/2mv^2$, where m is the fragment's mass and v is its ejection velocity, one has that, assuming that a given fragment has a fraction A of the total amount of energy converted into kinetic energy of the fragments, it is possible to write

$$\frac{1}{2}\frac{m}{M}v^2 = A\, f_{KE}\frac{E}{M}.$$

By developing the above relation, one has that

$$\log(d/D) = -\frac{2}{3}\log v - K',$$

and

$$K' = -\frac{1}{3}\log(2Af_{KE}\frac{E}{M}),$$

where d is the size of the fragment and D is the size of the parent body. This relation should hold for all fragments. In other words, if one plots $\log(d/D)$ versus $\log(v)$ for the members of a family, it should be expected that the domain occupied by family members should be delimited by a straight line having a $-2/3$ angular coefficient, corresponding at each size to some permitted maximum value of kinetic energy. Since it is not reasonable to expect that a strict energy equipartition principle holds, the velocity of ejection of a fragment is not expected to be uniquely determined by its size. Instead, it may be expected that, at each size, family members should be distributed over an interval of possible velocities, up to a maximum limit depending on the size itself. These expectations were qualitatively confirmed as shown in Fig. 24.

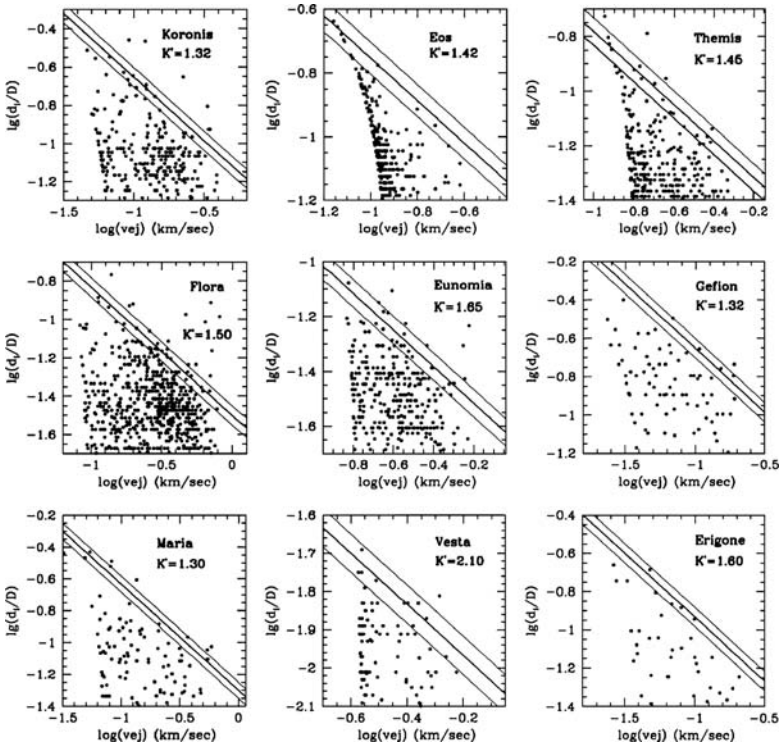

Fig. 24 Size–ejection velocity relation for some asteroid families, as published in [33]. The lines displayed in each plot have an angular coefficient equal to $-2/3$, the value predicted according to some simple physical considerations, considering a weak version of an energy equipartition principle (see text and quoted paper)

Although the plot shown in Fig. 24 looks fairly encouraging and in agreement with the expectations, some *caveats* are needed in order to avoid to overinterpret it. In particular, one basic assumption is implicit in this analysis, as well as in many other physical studies of families carried out in the same years: this implicit assumption, that we have already mentioned previously, is that the family members have not been dynamically evolving since the time of their formation. Under this hypothesis, the current proper elements correspond to those originally achieved at the epoch of the disruption from the parent body. We will see that this assumption has been found to be non-realistic when new dynamical effects, discussed in Sect. 3, have started to be taken into account.

The fact that something could be wrong in the physical studies of families carried out in the 1990s was already starting to emerge mainly for what concerns some apparent problems with the reconstruction of the ejection velocity fields of families derived by means of the methods described in Sect. 2.5, as it will be briefly mentioned in what follows.

2.8 Known Problems

We have seen in Sect. 2.5 that a method was developed in 1996 to carry out a reconstruction of the original ejection velocity fields of asteroid families. The method was able to estimate the values of the a priori unknown f and ω angles appearing in the Gauss equations (1). Numerical simulations were performed to show that the method was reasonably efficient and reliable.

When an analysis of the distribution of the f angle (the true anomaly of the parent body at the epoch of its disruption) was carried out, however, some unexpected feature became apparent. A priori, one should expect that, when analyzing several families, the resulting f angles should turn out to be uniformly distributed in the $(0^\circ, 180^\circ)$ range. The reason is that an analysis of the impact probabilities among asteroids predicts that the true anomaly of the impacted bodies should be distributed in a fairly homogeneous way, without any particular preference. A glance at Fig. 25, however, shows the distribution of the resulting f values turning out from a reconstruction of the ejection velocity fields of known families.

It is easy to see that the histogram of the resulting f values is strongly peaked on a value of about 90°. This fact, which cannot be easily explained, triggered a more detailed analysis [34] that, although published in 2004, is already described in this section, also because this analysis is strictly related to another problem in asteroid family studies that was well known since a long time. In particular, this is the problem of the apparent asymmetry of families. In [34] extensive numerical simulations were performed to derive what should be the statistically expected dispersions in orbital semi-major axis and eccentricity, resulting from a large number of collisions producing completely symmetric ejection velocity fields (spherical velocity fields) and occurring according to a uniform distribution of the values of the true anomaly angle f.

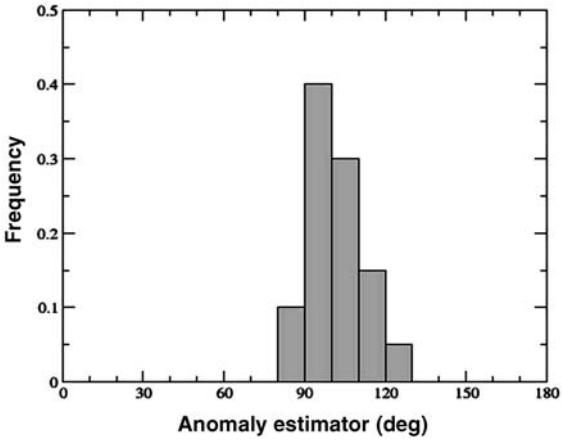

Fig. 25 Histogram of the f angles resulting from a reconstruction of the ejection velocity fields of known asteroid families using the method described in [29]

According to [34], if one calls a_0 the value of the semi-major axis of a family barycenter, Δa the range of semi-major axes of the members of the same family, and Δe and Δi the corresponding ranges of eccentricity and inclination, respectively, statistics predicts that for spherical velocity fields one should expect that for a large number of families one should find $\Delta e \simeq 0.77 \Delta a/a_0$ and $\Delta i \simeq 0.35 \Delta a/a_0$. If one looks at real families, however, one finds that Δe turns out to be about 1.2 times the predicted value, while ΔI turns out to be about twice the predicted value. As mentioned above, moreover, the distribution of the f angle should be expected to be fairly homogeneous, whereas this certainly not the case with real families.

In [34], some explanation of the above discrepancies was attempted. In particular, it was investigated whether some evolution of the orbital semi-major axis and eccentricity with respect to their original values achieved at the epoch of family formation could be responsible of the observed family asymmetries and non-homogeneous distribution of the f angle.

The result was that the observed asymmetries and f distributions could be explained if we assume that the current eccentricities of family members have been increased by a factor between 1.4 and 1.9, and at the same time the semi-major axes have been increased by a factor between 1.3 and 1.8, since the time of their formation. This analysis did not propose any particular mechanism to justify the above resulting orbital element increases, but it showed that such kind of diffusion would at the same time produce a distribution of f angles in complete agreement with the results of the reconstruction of real families. A very important conclusion of this study was that families that we see today should be on the average between 1.5 and 2 times more diffused in semi-major axis and eccentricity with respect to their original structures.

It is important to note that the above conclusions are not based on any assumption about the possible evolutive mechanisms of asteroid families, but are based on purely statistical arguments. The importance of these results will be more evident in Sect. 3.

2.9 Summary: The Twentieth Century Family "Paradigm"

To summarize the great body of results obtained in asteroid family studies in the years between 1990 and 2000, let us summarize now the "twentieth century family paradigm" as it appeared at the end of the above decade.

- Families exist and can be reliably identified.
- Families have a collisional origin.
- Family members dominate the asteroid inventory at small sizes.
- The original ejection velocity fields can be reconstructed from analysis of the current distribution of family members in the proper element space.
- Family members were ejected at high velocities, up to some hundreds of meters per second, and following some general size velocity relation.
- Families can be (or have been) important sources of NEOs through direct injection into neighboring resonances.
- The creation of big families triggers the collisional evolution of the whole asteroid population.
- The original parent bodies of asteroid families were not differentiated.

Some problems, like the real role played by family members in the overall asteroid inventory and the reasons of the observed structural asymmetries of many families and the anomalous distribution of the reconstructed f angles (see Sect. 2.8), were already apparent at this stage, but they were not yet considered so strong as to rule out the overall correctness of the above paradigm.

New facts, however, were going to be recognized in the immediately following years, leading to a general conceptual revolution whose full implications are not yet completely clear at the moment of writing this chapter.

3 Families in the Twenty-First Century

Since we have just explained above what we call in this chapter the "twentieth century" family paradigm, and we have mentioned several times that there has been in recent years a deep revision of common ideas about families, let us start this section by giving what seems to be the "twenty-first century" family paradigm, in order to directly introduce the changes that have taken place in recent years. The new paradigm is the following:

- Families exist and can be reliably identified.
- Families have a collisional origin.
- Family members do *not* dominate the asteroid inventory.
- Families did *not* eject collisional fragments at high velocities.
- Families have been strongly modified by evolutionary mechanisms.
- The original ejection velocity fields can *hardly* be reconstructed.

– Family ages can be evaluated.
– Family members are mostly re-accumulated.
– The original parent bodies of asteroid families were not differentiated.

As can be seen, some fundamental items in the family paradigm are still there (fortunately, the fact that families exist, can be identified and have a collisional origin has not been questioned!). On the other hand, several items in the above list directly contradict some ideas contributing to twentieth century paradigm. In particular, the most important change is that family members are now believed to have experienced important evolutionary processes, and this fact implies that several conclusions of the twentieth century paradigm, based on the implicit assumptions that family members are still directly reminiscent of the process of their formation, can no longer be accepted. As a consequence, the reconstruction of the ejection velocity fields of families as it was done in the 1990s [29] seems to be currently to the majority of researchers working in this field as a sterile exercise leading to misleading results.

In addition to this, it is now generally believed that families do not dominate the asteroid inventory even at small sizes, based on the results of some observing campaigns like the *Sloan Digital Sky Survey* (SDSS).

An important new item in the twenty-first century family paradigm is the idea that family ages can now be determined, something that was not considered to be possible in the framework of the older paradigm. Finally, there is the general idea that family members are mostly re-accumulated objects. This result comes from a number of numerical simulations based on refined hydrocodes used to study the fragmentation and fragment ejection process, followed by an N-body numerical integrator used to follow the trajectories and mutual interactions of the fragments immediately after their ejection.

In the following sections we will separately discuss the major facts that have led to the present family paradigm. Some problems that are apparently still unsolved will also be mentioned.

3.1 The Yarkovsky and YORP Effects

A big development in asteroid science that occurred in recent years has been the realization of the importance of the so-called Yarkovsky effect [35].

The physical process at the base of the Yarkovsky effect is the re-emission at thermal infrared wavelengths of the heat absorbed from the Sun. Two different version of the effect, named diurnal and seasonal Yarkovsky effect, respectively, exist. Since the seasonal effect turns out to be much weaker, we focus in what follows on the diurnal version of the effect. The mechanism is schematically shown in Fig. 26 and is briefly summarized in what follows. A rotating asteroid is exposed to incident sunlight. A minor fraction of the incident radiation is immediately scattered by the surface, but the rest is absorbed and delivers heat to the surface. The surface then irradiates the absorbed heat at thermal IR wavelengths. At this point, two processes

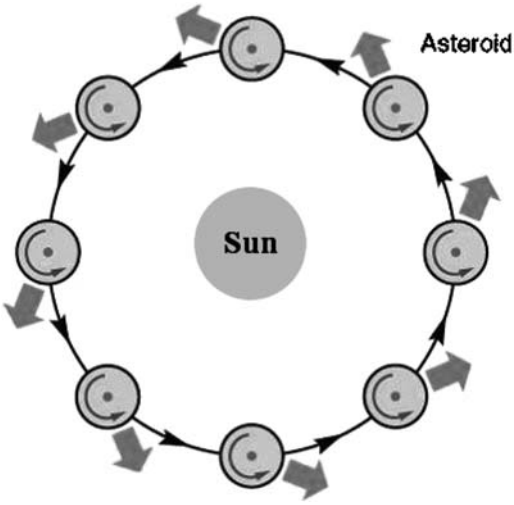

Fig. 26 Visual representation of the mechanism of the diurnal Yarkovsky effect. Plot derived from an original figure in [36]

take place, which determine the diurnal Yarkovsky effect. First, the asteroid surface is not an ideal medium, and some thermal inertia determines that the thermal flux is emitted not instantaneously with respect to the absorption of sunlight, but with some delay. Second, the object rotates around its spin axis, then the peak of the thermal emission is not directed toward the Sun, but along a direction that makes a small angle with respect to the direction of the star, due to the effect of rotation.

As a consequence, the irradiated thermal flux produces an impulse that can be either along the direction of the orbital motion or in the opposite direction, depending on the sense of rotation of the object. Consequently, the orbital motion of the body is either accelerated or decelerated, and its orbital semi-major axis changes accordingly. The net effect of the diurnal Yarkovsky effect is then a drift in orbital semi-major axis.

The efficiency of the effect is a function of many parameters. First, it depends on the obliquity angle of the asteroid, namely the angle between the plane of orbital motion and the direction of the polar axis of the object. Asteroids whose spin axis is directed toward the Sun do not experience a diurnal Yarkovsky effect. On the other hand, the effect is maximum when the obliquity angle is 90°. In addition, the effect is inversely proportional to the object's size and depends on the spin rate, thermal inertia, and the heliocentric distance. In particular, it turns out that the drift in orbital semi-major axis decreases approximately with the square of the orbital semi-major axis itself. The reason is, of course, that bodies orbiting at large heliocentric distances are more scarcely heated up by solar radiation.

The Yarkovsky effect is a nice example of a link between physical and dynamical mechanisms. The effectiveness of the effect has been estimated by several authors as a function of the different parameters mentioned above. According to [36], typical values of the drift in semi-major axis experienced by main belt asteroids of 1 km in size are around 10^{-4} AU per million of years, with an uncertainty of the order

of a factor of 2 or 3, depending on the value of the thermal inertia of the surface. For objects of 10 km in diameter the corresponding drift is ten times smaller. In the same paper, an estimate of the total drift in semi-major axis experienced by objects of different sizes during their expected collisional lifetimes is also given. The corresponding values are between 0.02 and 0.08 AU depending on the thermal inertia. For 10-km objects the corresponding interval is between 0.03 and 0.05 AU. A direct confirmation of the existence of a measurable Yarkovsky effect has also been obtained through radar ranging observations of the near-Earth asteroid (6489) Golevka [37].

The importance of the Yarkovsky effect for asteroid family studies is that it introduces an evolutionary mechanism that had not been taken into account in the twentieth century analyses. The idea is that newly born family members start to drift in semi-major axis due to the Yarkovsky effect. The families start then to diffuse in semi-major axis, mainly and more quickly at smaller sizes. Due to their semi-major axis drift, family members may be injected into resonant zones of the orbital element space. As a consequence, they experience chaotic changes in eccentricity and inclination, and they may be removed from their family, and start a complex dynamical evolution that may lead them to have close encounters with major planets, causing them to be injected in the region of the terrestrial planets or to be removed from the Solar System. Another possibility is also, for asteroids located in the inner region of the main belt, to steadily drift to smaller values of semi-major axis, until they become Mars-crossers. Close encounters with Mars lead them subsequently to become NEOs [38].

An example of the resulting evolution of a simulated Koronis family, based on numerical integrations of the orbits taking into account a model of the Yarkovsky effect, is shown in Fig. 27. The figure shows in the semi-major axis–eccentricity plane the time evolution of the simulated family. The simulated objects are indicated by segments showing their total orbital evolution at three epochs after the family formation. The members of the real Koronis family are shown as dots in the background. The plots show that the simulated objects progressively tend to mimic the distribution of the real asteroids, and in particular, it is possible to see that as objects cross a narrow resonance strip, which is found to cross the family, they start to increase their eccentricity and form the strange "tail" of family members exhibiting a larger eccentricity in the outer part of the family. Although the fit of the real family members is not really perfect, nevertheless the unusual structure of the family is qualitatively fit in a reasonable way, something that had not been possible to do in the pre-Yarkovsky era.

An indirect proof of the correctness of the Yarkovsky-based model is the fact that it explains why families are practically never found to include objects located beyond the borders of some powerful neighboring resonance, like the 3:1 or 5:2 mean motion resonances with Jupiter. The idea is that, if family members were originally ejected at high speeds, sufficient to reach these resonances and to inject bodies into them, then it would be strange that no objects are found today beyond the borders of these resonances. As opposite, by assuming that families were originally more compact and have been only subsequently spread in semi-major axis

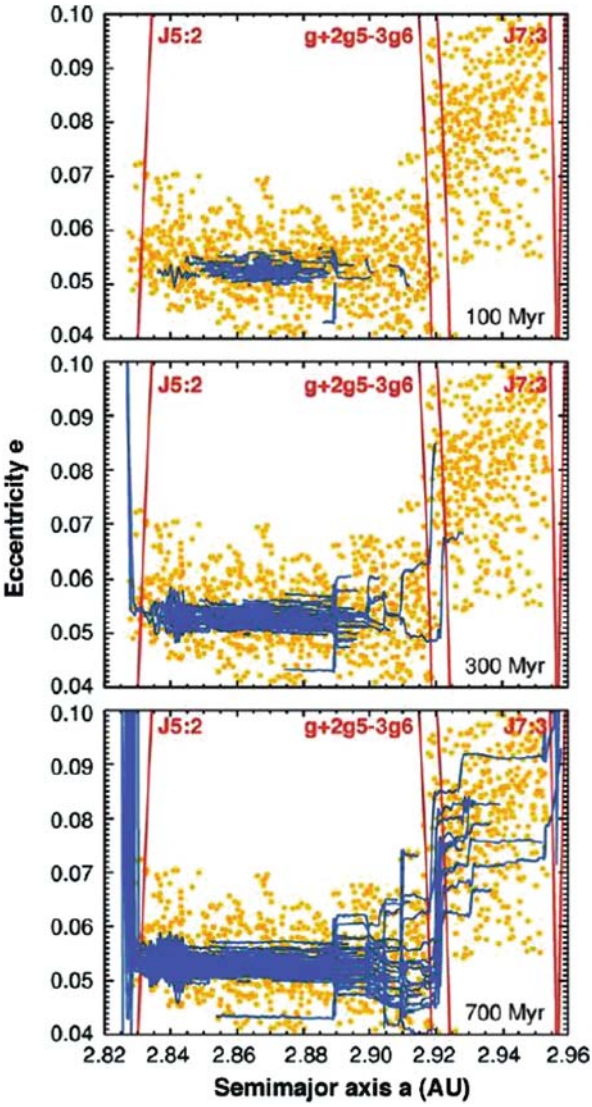

Fig. 27 Time evolution of a simulated Koronis family taking into account a model of the Yarkovsky effect in the numerical integration of the orbital motion of the simulated objects. The real Koronis members are shown as *dots* in the background. Plot taken from [36]

due to the Yarkovsky effect, one can explain why there are not resonance-crossing family members beyond very powerful resonances. If the objects underwent a slow Yarkovsky-driven orbital drift, it is reasonable to assume that, when reaching very powerful resonances like the 3:1 Kirkwood gap, they were quickly removed from the asteroid main belt and could not "reach the opposite shore of the river."

Another advantage of the Yarkovsky-based paradigm is that it naturally reconciles the extremely rapid dynamical evolutions of asteroids injected into the most important resonances [31] (see Sect. 2.6), with the much longer cosmic rays exposure ages exhibited by meteorites. The idea is that an immediate injection into

resonance of family members at the epoch of family creation would imply exceedingly short lifetimes for these objects, whereas a slower Yarkovsky drift in semi-major axis, eventually leading to injection into resonances, could reconcile the dynamical lifetimes of these objects with cosmic rays exposure ages of meteorites. At the same time, a slower process of delivery of objects into resonances, would also be more easily reconciled with a reasonable rate of NEO supply from the asteroid main belt.

After the realization of the importance of the Yarkovsky effect, numerical simulations have been performed to reproduce the observed structures of several families in the proper element space [39–41]. In these studies it was found that a better fit of the observed families may in several cases be obtained when one also includes in the model the role of the so-called Yarkovsky–O'Keefe–Radzievskii–Paddack (YORP) effect.

Like the Yarkovsky effect, the YORP effect is also due to a mechanism of thermal irradiation from the surface. What is important here, however, is the fact that due to the irregular shape of real objects, the thermal irradiation may well produce net torque effects which progressively modify the angular momentum of an object. In particular, since the moment of inertia remains constant as the object keeps its shape, what does change is the state of rotation. In particular, the YORP effect can modify both the spin period and the direction of the spin axis of an object [42].

Again, we deal here with a physical mechanism which depends in a complicate way upon many parameters, some of whom are poorly known. In particular, there is a dependence on the objects's shape, size, thermal conductivity, heliocentric distance, and spin axis orientation.

As for the shape, an object must have some "windmill" asymmetry for YORP to work, as shown in Fig. 28, taken from [36]; energy re-radiated from a fully symmetrical body (e.g., a sphere or an ellipsoid) produces no net YORP torque [42, 36]. These ideal shapes, however, are not encountered in the real world, so it may be

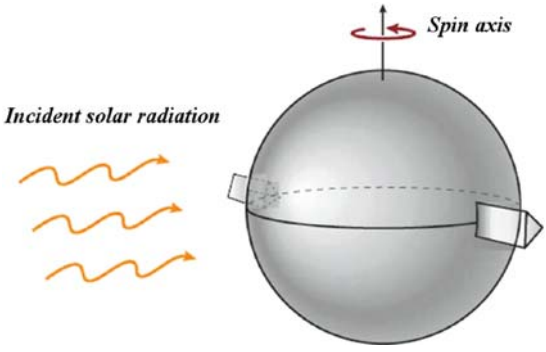

Fig. 28 Spin up of a simulated asteroid, ideally modeled as a sphere with two wedges attached to the equator. It is assumed that the asteroid is an ideal black body, so it absorbs all incident solar radiation and then re-emits it at infrared wavelengths as thermal radiation. Because the kicks produced by photons leaving the wedges are in different directions (note that the two wedges in the plot are *not* coplanar), a net torque is produced that, in the situation illustrated in this plot, with the object spinning as shown, causes the asteroid to spin up. Plot taken from [36]

expected that YORP torques always take place with real objects, although its actual effectiveness may vary much, depending on the exact shape and direction of the spin axis. Of course, the response of a body to the YORP torque is inversely proportional to its mass, then bigger objects are much less affected than small ones.

YORP can either spin up or spin down an object depending on its shape and rotation. YORP torque produces also a change of obliquity angle. The obliquity angle tends to reach an asymptotic value. In turn, however, when the obliquity angle increases sufficiently, the rotation rate may change, and possibly tumbling rotation occurs before a new stable rotation state is reached again, and so on, leading to the possible occurrence of YORP cycles [36].

It is now generally believed that including the YORP effect in the models of the evolution of the rotation state of main belt asteroids may be very important to explain the basic features of the distribution of measured rotation periods, in particular at small sizes, where there is abundance of both slow and fast rotators [43]. Moreover, another indirect proof of the role played by the YORP effect is also given by the discovery of an apparent bimodality in the distribution of the spin axis directions and spin rates of the members of the Koronis family [44]. More precisely, the observed bimodality of Koronis members should be due to the interplay of the YORP effect and a mechanism of spin–orbit resonance [45].

Of course, the YORP effect is important in affecting the effectiveness of the Yarkovsky evolution since it affects the rotation state, and the Yarkovsky effect depends on the direction of the spin and also on the spin rate. For instance, an ideally non-rotating body is not subject to the Yarkovsky effect. Similarly, objects rotating so rapidly as to become isothermal are neither affected by the Yarkovsky effect.

It is important to note that, when performing simulations of asteroids evolving under the effect of the Yarkovsky and YORP effects, it has to be taken into account that collisions also play a role in this game, since they may affect the rotation state by changing the angular momentum vector in such a way as to have a significant effect on the effectiveness of the thermal radiation mechanisms. The typical collision rates and collisional lifetimes for objects having different sizes thus become other non-negligible factors to be taken into account in simulations.

According to several studies, the inclusion of the YORP effect in numerical simulations seems to improve the fit of real families modeled by taking into account the Yarkovsky drift in semi-major axis. Some improved estimates of likely family ages come from these simulations, although the dynamical evolution of the objects masks the original ejection velocity fields and makes it difficult to evaluate ages of very old and/or small families [36]. Moreover, also some estimate of the initial ejection velocities of family members may be obtained as we will see below.

3.2 Ejection Velocities

We have seen that family studies performed in the 1990s were accepting as typical for the ejection velocity of small family members values up to 100 m/s or

even beyond. Such values resulted from interpreting the observed differences in semi-major axis between the smallest family members and the family barycenter in terms of ejection velocities according to Gauss' equations (1). This interpretation, according to more recent ideas, is wrong, because it does not take into account that the semi-major axis values of family members, mainly at smaller sizes, have been strongly affected by a Yarkovsky-driven drift.

Moreover, high values for the original ejection velocities of family members have always been hard to reconcile with the results of simulations of catastrophic disruption events [46, 47].

The importance of the initial ejection velocity values in family-forming events is that they represent the initial conditions to any simulation of the evolution of family members subject to the Yarkovsky effect. This leads to the possibility in principle to estimate the ages of asteroid families, by computing the rate of spreading in semi-major axis due to the Yarkovsky drift. Knowing the current width of families in semi-major axis, it becomes then possible to derive the time needed to reach the observed dispersions, starting from the initial conditions, namely the initial distributions of semi-major axis values of family members.

In the first analyses, the initial dispersion in semi-major axis of families was generally assumed to be very small, if not negligible, corresponding to very low values of the original ejection velocities, according to the results of hydrocode simulations. The most recent analyses, which include also the YORP effect, however, are more detailed and look for a simultaneous solution for the family age and the initial ejection velocities of family members.

According to a recent analysis of the Eos family [41], the original ejection velocity values for this family turn out to be of the order, on the average, of several tens of meters per second. The results indicate also that the original, post-impact width of the family in semi-major axis was about one half of what is observed today. By the way, such a result is in a very good agreement with the estimate of the average family spreading done by [34], based on the observed family asymmetry and distribution of the computed values of the f angle, as mentioned in Sect. 2.8.

A more systematic analysis of several families subject to the Yarkovsky and YORP effects has been recently published [48]. Typical ejection velocity values of a few tens of meters per second for family members having sizes of the order of 5 km were found. Although the authors claim that such values are in agreement with hydrocode results, taken at face value and assuming a simple dependence of the ejection velocity upon the inverse of size, the obtained values would imply ejection velocities of the order of 100 m/s for fragments having sizes of 1 km. Moreover, once again, the resulting initial spread of family members in semi-major axis turned out to be between 30 and 50% of the currently observed values, in good agreement with the independent above-mentioned estimates of the post-impact evolution of families based on the apparent asymmetries of the ejection velocity fields derivable by the current structures of families in the proper element space [34].

We think that the coincidence between the results of these completely independent results concerning the initial spreading in semi-major axis of asteroid families and the corresponding typical values of the ejection velocities of family members

are very remarkable and may suggest that the problem of evaluating these initial velocities seems to be close to a definitive solution.

This means also that the estimated values of the initial ejection velocities of family members are raising again somewhat, according to the most updated analyses, with respect to the assumptions made by the first Yarkovsky-based numerical integrations. This also means that the initial velocity values might have been not really so low for the smallest objects, and some revision of the hydrocode results might be necessary.

We note, finally, that the expected dispersion in eccentricity and inclination coming out from numerical simulations including the Yarkovsky and YORP effects turn out to be generally smaller by a factor of about 2 with respect to the observed values. We will come back to this point in Sect. 3.5

3.3 Inventory and Size Distributions

As mentioned in Sect. 2.3, the surprisingly steep slopes of family size distributions and their interpretation in terms of a dominance of small family members in the asteroid inventory have been long debated [25]. Recently, the results concerning the slope and shape of family size distributions based on the simple geometric model [20] mentioned in Sect. 2.3 has been generally confirmed by detailed hydrocode + N-body modeling, as shown in Fig. 29.

In the last years, the assumption that family size distributions are steeper than the size distributions of non-family asteroids down to small sizes has been adopted to develop a so-called statistical asteroid model (SAM), aimed at simulating the inventory and distributions of size, albedo, and orbital elements of main belt asteroids down to 1 km in size [50].

The dominance of family members at small sizes, however, has been questioned by several authors. From the point of view of the consequences of the previously neglected Yarkovsky effect, the idea is that the Yarkovsky drift should produce

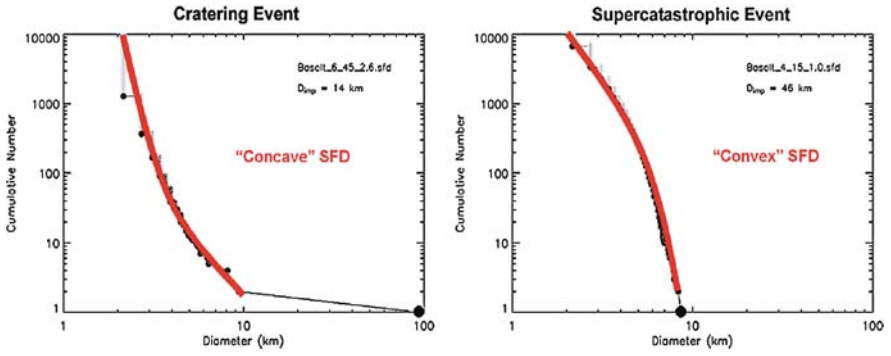

Fig. 29 A couple of size distribution frequencies (SFD) obtained by [49]. The obtained trends are in good agreement with previous results obtained by means of a much simpler model [20]

a rapid removal of the smallest family members. The final fate of these objects should be a complete removal from the main belt over fairly short timescales, due to resonance crossing occurring during their drift in semi-major axis. This seems confirmed by the fact that the SDSS [51] has found that the general size distribution of the asteroid main belt population is described by a power-law having an exponent much less steep than the value that would be predicted based on an extrapolation of the family size distributions observed beyond the limit of completeness.

In this respect, according to a study of SDSS results carried out a few years ago [52], it turns out that asteroid family size distributions at small diameters might be even less steep than the size distributions exhibited by non-family objects. Other authors have noted that such SDSS-based findings are not conclusive since SDSS data seem to be not always self-consistent [53]. In particular, it was noted that the SDSS-based inference about quite shallow slopes of family size distributions at small sizes would be in contradiction with another independent conclusion about the dominance of families in the overall asteroid inventory based on SDSS color data [54].

As a matter of fact, the fundamental problem seems to be that observations have not yet provided a conclusive evidence about the inventory and size distribution of the asteroid population. In particular, the results of different surveys are contradictory, and the interpretation of the data in terms of asteroid sizes is also not straightforward, as we will see in a moment.

On one hand, the SDSS data and the Subaru surveys [55, 56], both carried out from the ground at visible wavelengths, both found a quite shallow size distribution of the main belt population down to sizes smaller than 5 km. On the other hand, space-based surveys carried out at thermal infrared wavelengths find a much larger number of objects in the same size range [57]. The difference between ground-based and space-based surveys for objects 1 km in size turns out to be of the order of a factor between 2 and 3. This means that the results of thermal IR surveys are still compatible with predictions based on a predominance of family members [50], whereas ground-based surveys are not.

When deriving asteroid size distributions from sky surveys, one should always take into account that what is observed and recorded is a distribution of apparent magnitudes, not directly of sizes. The conversion from apparent magnitudes to sizes is done by converting apparent magnitudes into absolute magnitudes, and assuming some value of the albedo, according to Eqn. (2). Both the conversion to absolute magnitude and the assignment of an albedo value are important sources of errors.

Some recent studies have been based on the idea that the SDSS distribution of absolute magnitudes has a correct *slope*, and that a de-biased distribution of absolute magnitudes must be based on that slope value, but complementing it with the known number of asteroids having absolute magnitude $H < 12$. Having done so, the corresponding size distribution has been obtained by assigning to the objects an average albedo value equal to 0.092 [58]. According to this study, there should be about 1.2×10^6 main belt asteroids larger than 1 km. The above estimate is just in between the SDSS-based estimate of about 7×10^5 objects and the SAM estimate of 1.7×10^6.

As mentioned above, however, the derived size distributions are still quite uncertain. On one hand, the albedo is a parameter that may vary over an order of magnitude (between about 0.05 and 0.5), and is also dependent on the heliocentric distance, since darker objects are more abundant in the outer asteroid belt. Then, any albedo assignment based on an average value is intrinsically dangerous. On the other hand, another big problem has been becoming increasingly manifest in recent times, namely the problem of the reliability of the absolute magnitudes H. We remind that the absolute magnitude of an asteroid is an abstract parameter, having the meaning of the apparent magnitude that the object would exhibit if observed at unit distance from both the observer and the Sun and at zero phase angle (perfect Sun opposition). The values of H listed in asteroid catalogs are then derived from an extrapolation to zero phase angle of apparent magnitudes observed at (often, few) different epochs at corresponding phase angles different from zero. What seems currently to be a big problem is that the listed values of absolute magnitude seem to be very often extremely inaccurate, as shown in Fig. 30. The figure shows that the currently adopted H values may be wrong, in many cases dramatically wrong, due to the presence of both random and systematic errors, as a function of the magnitude itself. If this is the situation, any conclusion on the size distribution of the main belt population, and on the possible dominance of family members in the asteroid inventory at small sizes, seems premature. The problem is still open, and only new observations and a drastic correction of available catalogs of asteroid absolute magnitudes may lead to a real solution.

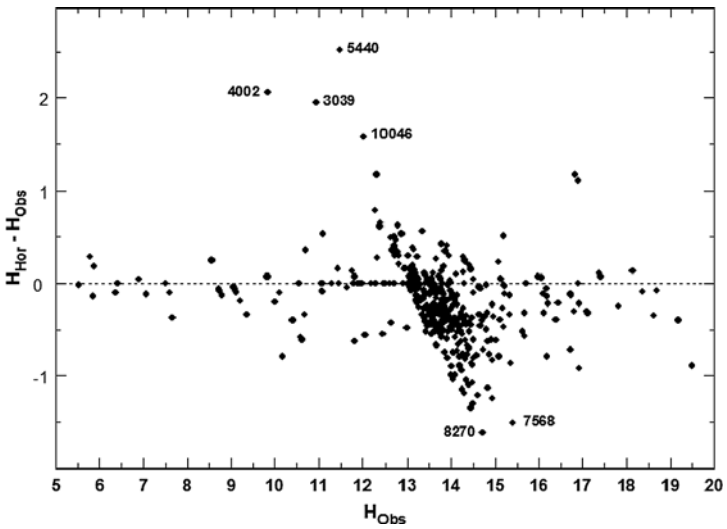

Fig. 30 Differences between the values of H absolute magnitudes of a sample of objects as they are listed in the JPL Horizon orbital element database and the H values derived from direct observations during a recent observing campaign (Plot taken from [25])

3.4 New Young Families and Implications

One big achievement of family studies in the last decade has been the discovery of a few very young families. In particular, the discovery of the Karin family [59]. This is a small clustering of objects within the much bigger Koronis family. It is likely the outcome of a second-generation collision involving an original member of the Koronis family, having a size around 30 km. The mass ratio between the parent body and the largest surviving fragment, (832) Karin, is about 0.15–0.2. The Karin family has been identified due to its peculiar structure in the proper element space, being characterized by a filament-like structure in the a'–e' plane. This kind of structure is expected, according to Gauss equations, in cases in which the true anomaly of the parent body, f, was very close to 0° to the epoch of the impact.

Numerical integrations back in time have directly shown that the Karin family should have an age around only 5.8 millions of years, and it is thus extremely young. It represents, therefore, an ideal grouping to test current ideas about the properties of asteroid families immediately after their formation. The cumulative size distribution of the family is shown in Fig. 31. As it can be seen, the size distribution is very steep and is fitted by a power-law having an exponent of -5.3 [60]. This family provides, therefore, a nice confirmation of the steep size distributions characterizing the outcomes of family-forming events.

According to [60] the ejections speeds of small fragments produced by the event were larger than those of larger fragments, in qualitative agreement with the general "twentieth century" idea of an original size–velocity relation for the members of asteroid families [33]. In particular it has been found that the ejection velocity shows a simple dependence on the inverse of size. It must be noted, however, that the

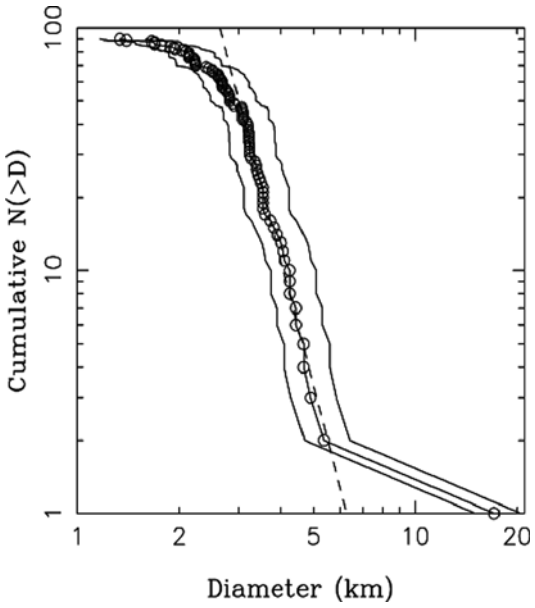

Fig. 31 The cumulative size distribution of the Karin family. Plot taken from [60]

smallest Karin members are still undiscovered, due to their apparent faintness. The mean ejection speeds of fragments above 3 km in diameter have been found to be of the order of 10 m/s, but the morphology of the observed ejection velocity field derived from the structure of the family in the proper element space has been found to be not easily reproducible using the approach adopted by [60]. It is also worth to note that, in spite of its youth, current models of the family evolution include some small Yarkovsky-driven evolution to improve the fit between the models and the appearance of the real family.

Another recent discovery has been that of the likely disruption of a parent body that produced the asteroid (298) Baptistina, an object previously included in the big Flora clan in the inner belt [11]. The Flora family has long been considered to be a puzzle, due to the fact that it is very big and dispersed, and there is the possibility that it might consist of the overlapping of separate groupings. According to [61], the collision that produced a family including (298) Baptistina may have occurred recently, about 100 millions of years ago. Such event, according to simulations, might have been responsible of an asteroid shower (see Sect. 2.6), which led to an increase of the lunar and terrestrial cratering rate during the last 100 Myr and was likely including the big impact occurred at the end of the Cretaceous about 65 Myr ago. This was the impact that, according to a consistent body of evidence, produced the Chicxulub crater in Yucatan and was likely responsible of the mass extinction event leading to the disappearance of dinosaurs.

3.5 Some Problems

The twenty-first century family paradigm is based on a convincing body of evidence, coming both from theory and observations. The inclusion of the Yarkovsky and YORP effects, in particular, constitutes certainly a big step forward in the interpretation of the properties of asteroid families.

Having clearly made the above statement, we cannot yet conclude that everything is now clear and that there are no pending problems. A list of problems affecting current ideas about families is the following (not in any particular order):

1. Up to which size the Yarkovsky effect is really effective?
2. Does YORP eventually strengthen or weaken Yarkovsky?
3. What is the explanation of the D versus proper elements relations observed for families?
4. How to put together consistently dynamical and physical effects having different and size-dependent timescales (resonance crossing, resonance-driven dynamical evolution, spin axis collisional realignment)?
5. How to explain why real families have e' and i' distributions which look often more dispersed than the results of Yarkovsky-based simulations?
6. The initial family structures are not directly known and are mostly derived from numerical simulations. Can they be estimated from the distributions of the largest members of families?

Item (1) in the above list is strictly related to item (6). The idea is that we know that the effectiveness of the Yarkovsky effect decreases with the size of the objects. In other words, the Yarkovsky-driven drift in semi-major axis is inversely proportional to the mass of the object. The net effect of the Yarkovsky drift is then that of mimicking a size–ejection velocity relation in orbital semi-major axis. On the one hand, this precludes the possibility to infer information on the original structure of families directly from a simple and direct inspection of the distribution of the current proper elements of family members at small sizes. On the other hand, however, it is still true that large family members, not appreciably affected by the Yarkovsky force, should be more directly reminiscent of their original ejection velocity values. The problem is to make a reliable assessment of a size limit beyond which we may assume that the Yarkovsky drift has been negligible. This depends on the age of the family and on the complicated dependence of the Yarkovsky effect itself upon many physical parameters, which are also subject to changes due to collisions, like the rotation period and the spin axis orientation. As a general comment, we note that the reconstruction of the ejection velocity fields of several families done in the pre-Yarkovsky era [29] included in general many family members that were fairly large, due to the effect that the family membership lists derived in the 1990s were limited to fairly small databases not including the proper elements of many small and faint objects that are available today. For this reason, it is not sure that all the old results are completely and systematically wrong, although it is clear that they should be deeply revised, taking also into account the indications coming from the uneven resulting distribution of the resulting f angles [34] (see Sect. 2.5).

The item (2) in the previous list expresses some uncertainty concerning the interplay between the Yarkovsky and YORP effects. The reason is that both effects depend on many parameters, and the resulting evolution determined by YORP is also related to complicated effects of spin–orbit resonance [45]. In this situation, it is not completely clear whether the YORP effect really makes Yarkovsky more effective, as it might look reasonable at a first glance, assuming that YORP simply tends to bring the obliquity angle to reach a value of 90°. The situation seems more complicated, and further analyses and modeling seem necessary.

The other items of the above problem list are different aspects of the following general problem: the effect of the Yarkovsky effect on the evolution of orbital eccentricity and inclination is eminently indirect. The effect itself has only a weak direct influence on the evolution of the eccentricity and no effect at all on the evolution of orbital inclination. Eccentricity and inclination change mostly because during its drift in semi-major axis under the effect of the Yarkovsky effect, an object may enter some resonant zone, and start wide chaotic oscillations in eccentricity and inclination. While being located in a resonant strip, the object still continues its drift in semi-major axis, and more or less quickly, depending on its mass, reaches the other border of the resonance, taking again a state of regular orbital motion. Of course, if it enters a wide and powerful resonance, like the 3:1 mean motion resonance with Jupiter, the chances of the object to be strongly perturbed and removed from the asteroid main belt before leaving the resonance are high, yet the main belt is crossed by many weaker resonances, that may be progressively reached by any single object

during its drift in semi-major axis, mainly if it is small and consequently its drift is more rapid. The residence time in any given resonance and the number of resonances crossed in a given time are then size dependent. At the same time, the object is also subject to a collisional evolution that may change its size and rotational state. The net result of such kind of complex evolution seems, therefore, quite complicated and hardly predictable. It is then not really surprising that Yarkovsky-based models do not produce in general very good fits of the structures of the families in eccentricity and inclination and of the resulting relations between size and the above orbital parameters.

The very surprising fact, however, is that the size–eccentricity and size–inclination relations appear to be fairly simple and regular, as shown in Fig. 32. In particular, in spite of complications related in several cases to the presence of nearby resonances, that may have some evident role in shaping the borders of the families in semi-major axis, it turns out that the dispersions of eccentricities and inclinations generally turn out be inversely proportional to the asteroid size, as it would be expected by assuming that a size–velocity relation holds. The problem then is: Can this apparent relation be simply explained by a mechanism based on a pure Yarkovsky-based evolution? Some further work seems still necessary to properly deal with this problem.

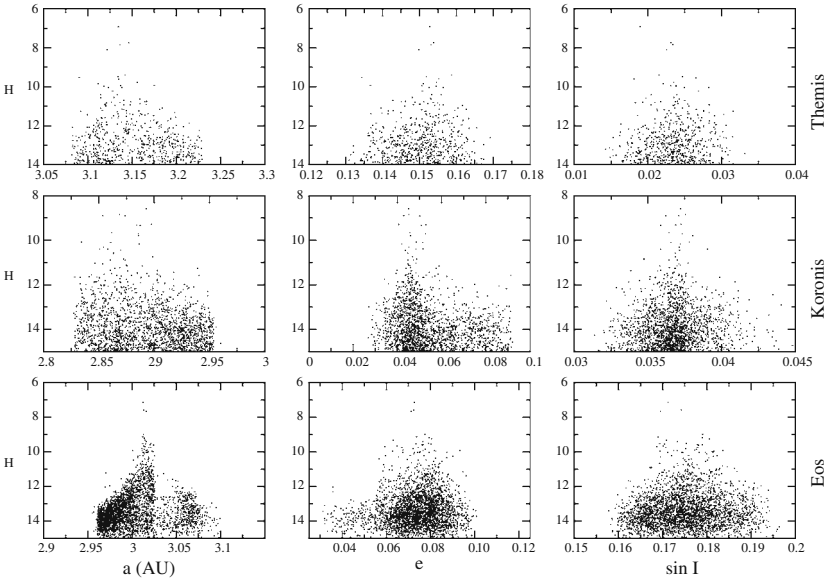

Fig. 32 The a', e', and $\sin i'$ versus size relations exhibited by the Themis, Koronis, and Eos families (from *top to bottom*). Plots based on still unpublished 2007 data

4 Discussion and Conclusions

In spite of tremendous improvements in our models, especially following the realization of the importance of thermal radiation mechanisms, a lot of work seems still necessary to achieve a really satisfactory comprehension of the whole family puzzle.

Based on the overall discussion made in the previous sections, it seems that a list of lines of research that should deserve a careful attention in planning future activities in this field includes the following:

- New, updated family lists are needed.
- More spectroscopic data are needed to identify interlopers and help in the assigning members to mutually overlapping groupings in the proper element space.
- Better estimates of Yarkovsky effectiveness seem desirable.
- A refined interpretation of the proper elements versus size plots are needed.
- A better assessment of what remains of the primordial structures of asteroid families in spite of the Yarkovsky evolution is needed. This mainly refers to the biggest objects belonging to family member lists.
- New photometric data are needed to test the possible existence of systematic trends in the orientations of the spin axes and on the rotation periods of the members of a same family, as in the case of the Koronis family discussed above [44]. These data will certainly be produced by the next generation of ground-based (Pan-STARRS) and space-based (Gaia) sky surveys.

For what concerns the need of new family lists, some attempts have been already done in recent years, but the situation is intrinsically difficult due to the effect of mutual family overlapping. This is an undesirable consequence of having at disposal, in the present situation, huge data sets of asteroid proper elements, containing much more small objects with respect to the databases adopted for family identification in the 1990s. We note also that the criterion adopted for establishing family membership is crucial, and a lot of care must be devoted to this problem. The reason is that a too liberal criterion may produce family member lists including large numbers of interlopers, or of actual members of other families. As opposite, a too restrictive criterion may produce member lists depleted of large numbers of actual family members, so increasing artificially the inventory of non-family objects. For the above reasons some improved methods of family identification must probably be developed, and at the same time the role played by ancillary spectroscopic data is going to become increasingly important.

Acknowledgments This chapter summarizes the work carried out by many scientists working on the field of asteroid families for many years. The figures in this chapter were taken by scientific articles written in part by ourselves, but mostly by a number of colleagues. These papers represent some important milestones in the development of asteroid family studies. Our long and fruitful collaborations with V. Zappalà, P. Paolicchi, C. Froeschlé, P. Tanga, and Ph. Bendjoya, have been decisive in building our own knowledge of the subject of asteroid families. Very important interactions with D. R. Davis, E. F. Tedesco, W. F. Bottke, A. Campo Bagatin, A. Milani, Z. Knežević, E. Asphaug, D. Durda, P. Michel, S. J. Bus, R. P. Binzel, F. Marzari, D. Lazzaro, and A. Doressoundiram are kindly acknowledged. Most of all, we would like to remember the decisive role played in this field by P. Farinella and F. Migliorini. We dedicate this chapter to the memory of these two friends.

References

1. Hirayama, K.: Proc. Phys. Math. Soc. Japan **9**, 354 (1918) 137
2. Hirayama, K.: Proc. Imp. Acad. Tokyo **9**, 482 (1933) 137
3. Carusi, A., Valsecchi, G.B.: Astron. Astrophys. **115**, 327 (1982) 137
4. Knežević, Z., Lemaître, A., Milani, A.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 603–612. University of Arizona Press, Tucson (2002) 139
5. Milani, A., Knežević, Z.: Celestial Mech. Dynamical Astron. **49**, 347 (1990) 139
6. Nesvorný, D., Ferraz Mello, S., Holman, M., Morbidelli, A.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 379–394. University of Arizona Press, Tucson (2002) 139
7. Wetherill, G.W.: Icarus, **100**, 307 (1992) 140
8. Michel, P., Tanga, P., Benz, W., Richardson, D.C.: Icarus **160**, 10 (2002) 141
9. Bendjoya, Ph., Zappalá, V.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 613–618. University of Arizona Press, Tucson (2002) 141
10. Zappalá, V., Cellino, A., Farinella, P., Knežević, Z.: Astron. J. **100**, 2030 (1990) 142, 144, 145
11. Zappalá, V., Bendjoya, Ph., Cellino, A., Farinella, P., Froeschlé, C.: Icarus **116**, 291 (1995) 145, 146, 147
12. Cellino, A., Zappalá, V., Di Martino, M., Farinella, P., Paolicchi, P.: Icarus **70**, 546 (1987) 147
13. Binzel, R.P., Xu, S.: Science **260**, 186 (1993) 147, 149
14. Cellino, A., Bus, S.J., Doressoundiram, A., Lazzaro, D. In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 633–643. University of Arizona Press, Tucson (2002) 148, 152
15. Migliorini, F., Zappalá, V., Vio, R., Cellino, A.: Icarus **118**, 271 (1995) 150
16. Davis, D.R., Farinella, P., Marzari, F.: Icarus **137**, 140 (1999) 151
17. Zappalá, V., Bendjoya, Ph., Cellino, A., Di Martino, M., Doressoundiram, A., Manara, A., Migliorini, F.: Icarus **145**, 4 (2000) 151, 152
18. Cellino, A., Zappalá, V., Doressoundiram, A., Di Martino, M., Bendjoya, Ph., Dotto, E., Migliorini, F.: Icarus **152**, 225 (2001) 152
19. Rubin, A.E.: Icarus **113**, 156 (1995) 152
20. Tanga, P., Cellino, A., Michel, P., Zappalá, V., Paolicchi, P., Dell'Oro, A.: Icarus **141**, 65 (1999) 154, 155, 156, 157, 158, 184
21. Dohnanyi, J.S.: J. Geophys. Res. **74**, 2531 (1969) 153
22. Dohnanyi, J.S.: Physical studies of minor planets. In: Gehrels, T. (ed.) NASA SP-267, pp. 263–295. Washington, DC (1971) 153
23. Cellino, A., Zappalá, V., Farinella, P.: Mon. Not. R. Astr. Soc. **253**, 561 (1991) 154, 159, 161
24. Petit J.M., Farinella, P.: Celest. Mech. Dynam. Astron. **57**, 1 (1983) 154, 155
25. Cellino, A., Dell'Oro, A., Tedesco, E.F.: Planet. Space Sci. **57**, 173 (2009) 159, 184, 186
26. Zappalá, V., Cellino, A.: Completing the inventory of the solar system. In: Rettig, T.W., Hahn, J.M. (eds.) ASP Conference Series, vol. 107, pp.29–44 (1996) 160, 161
27. Jedicke, R., Metcalfe, T.S.: Icarus **131**, 245 (1998) 162
28. Dell'Oro, A., Paolicchi, P., Cellino, A., Zappalá, V., Tanga, P., Michel, P.: Icarus **153**, 52 (2001) 163
29. Zappalá, V., Cellino, A., Dell'Oro, A., Migliorini, F., Paolicchi, P.: Icarus **124**, 156 (1996) 163, 164, 165, 167
30. Knežević, Z., Lemaître, A., Milani, A.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 603–612, University of Arizona Press, Tucson (2002) 167, 170
31. Gladman, B.J., Migliorini, F., Morbidelli, A., Zappalá, V., Michel, P., Cellino, A., Froeschlé, Ch., Levison H.F., Bailey M., Duncan M.: Science **277**, 197 (1997) 168, 170, 171, 180
32. Zappalá, V., Cellino, A., Gladman, B.J., Manely, S., Migliorini, F.: Icarus **134**, 176 (1998) 170, 171
33. Cellino, A., Michel, P., Tanga, P., Zappalá, V., Paolicchi, P., Dell'Oro, A.: Icarus **141**, 79 (1999) 172, 173, 187
34. Dell'Oro, A., Bigongiari, G., Paolicchi, P., Cellino, A.: Icarus **169**, 341 (2004) 174, 175, 183, 189
35. Bottke, W.F., Vokrouhlický, D., Rubincam, D.P., Brož, M.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 395–408. University of Arizona Press, Tucson (2002) 177

36. Bottke, W.F., Vokrouhlický, D., Rubincam D.P., Nesvorný, D.: Ann. Rev. Earth Planet. Sci. **34**, 157 (2006) 178, 180, 181, 182
37. Chesley, S.R., Ostro, S.J., Vokrouhlický, D., Capek, D., Giorgini, J.D., et al.: Science **302**, 1739 (2003) 179
38. Migliorini, F., Michel, P., Morbidelli, A., Nesvorný, D., Zappalá, V.: Science **281**, 2022 (1998) 179
39. Bottke, W.F., Vokrouhlický, D., Brož, M., Nesvorný, D., Morbidelli A.: Science **294**, 1693 (2001) 181
40. Carruba, V., Burns, J.A., Bottke, W.F., Nesvorný, D.: Icarus **162**, 308 (2003) 181
41. Vokrouhlický, D., Brož, M., Morbidelli, A., Bottke, W.F., Nesvorný, D., Laz-zaro, D., Rivkin, A.S.: Icarus **182**, 92 (2006) 181, 183
42. Rubincam, D.P.: Icarus **148**, 2 (2000) 181
43. Pravec, P., Harris, A.W., Michalowski, T.: In: Bottke, W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, pp. 113–122. University of Arizona Press, Tucson (2002) 182
44. Slivan, S.M.: Nature **419**, 49 (2002) 182, 191
45. Vokrouhlický, D., Nesvorný, D., Bottke, W.F.: Nature **425**, 147 (2003) 182, 189
46. Michel, P., Benz, W., Tanga, P., Richardson, D.C.: Science **294**, 1696 (2001) 183
47. Michel, P., Benz, W., Richardson, D.C.: Planet. Space Sci. **52**, 1109 (2004) 183
48. Vokrouhlický, D., Brož, M., Bottke, W.F., Nesvorný, D., Morbidelli, A.: Icarus **182**, 118 (2006) 183
49. Durda, D.D., Bottke, W.F., Nesvorný, D., Enke, B.L., Merline, W.J., Asphaug, E., Richardson, D.C.: Icarus **186**, 498 (2007) 184
50. Tedesco, E.F., Cellino, A., Zappalá, V.: Astron. J. **129**, 2869 (2005) 184, 185
51. Ivezic, Z., Tabachnik, S., Rafikov, R., Lupton, R.H., Quinn, T., Hammergren, M., et al.: Astron. J. **122**, 2749 (2001) 185
52. Morbidelli, A., Nesvorný, D., Bottke, W.F., Michel, P., Vokrouhlický, D., Tanga, P.: Icarus **162**, 328 (2003) 185
53. Cellino, A., Dell'Oro, A., Zappalá, V.: Planet. Space Sci. **52**, 1075 (2004) 185
54. Ivezic, Z., Lupton, R.H., Juric, M., Tabachnik, S., Quinn, T., Gunn, J.E., Knapp, G.R., Rockosi, C.M., Brinkmann, J.: Astron. J. **124**, 2943 (2002) 185
55. Yoshida, F., Nakamura, T.: Adv. Space Res. **33**, 1543 (2004) 185
56. Yoshida, F., Nakamura, T.: Planet. Space Sci. **55**, 1113 (2007) 185
57. Tedesco, E.F., Désert, F.X.: Astron. J. **123**, 2070 (2002) 185
58. Bottke, W.F., Durda, D.D., Nesvorný, D., Jedicke, R., Morbidelli, A., Vokrouhlický, D., Levison, H.: Icarus **175**, 111 (2005) 185
59. Nesvorný, D., Bottke, W.F., Levison, H., Dones, L.: Nature **417**, 720 (2002) 187
60. Nesvorný, D., Enke, B.L., Bottke, W.F., Durda, D.D., Asphaug, E., Richardson, D.C.: Icarus **183**, 296 (2006) 187, 188
61. Bottke, W.F., Vokrouhlický, D., Nesvorný, D.: Nature **449**, 48 (2007) 188

An Introduction to the Dynamics of Trojan Asteroids

P. Robutel and J. Souchay

Abstract The dynamics of Trojan asteroids constitutes one of the richest fields of celestial mechanics, as a real application of the three-body problem. It involves the L_4 and L_5 Lagrange points and the conditions of stability around these two points. In this chapter we propose to present the fundamentals of the dynamics of Trojan asteroids. After a brief historical overview, we come back to the definitions and characteristics of the collinear Lagrange points L_1, L_2, and L_3, as well as the triangular ones, L_4 and L_5. We show how observational data of Trojan asteroids have confirmed the existence of real bodies librating around these two last points. Then we focus on the linearization of the equations of motion around L_4 and L_5 from a general and purely theoretical point of view. In addition, we show how qualitative results can be extracted to describe the properties of Trojan asteroids. We complete our study by summarizing many previous and up-to-date investigations, which focus on their dynamical behavior.

1 Introduction

Dynamics of Trojans is directly linked with the notion of Lagrange points, traditionally called L points. These five points correspond to positions in which an object with a small mass, when subjected to the sole gravitational attraction of two other objects with a much larger mass, can theoretically be stationary relative to these objects in such a way that the geometrical configuration of the three objects remains permanently the same. If the mass of the small body is in fact negligible with respect to the two other masses (which corresponds to the so-called *Restricted three-body problem*), we can explain the presence of the stationary positions in the following way: the two large bodies undergoing a Keplerian motion, in the frame

P. Robutel (✉)
Observatoire de Paris/IMCCE, UMR8028 du CNRS, 77 Avenue Denfert-Rochereau, F-75014, Paris, robutel@imcce.fr

J. Souchay
Observatoire de Paris/SYRTE, UMR8630 du CNRS, 61 Avenue de l'Observatoire, F-75014, Paris, Jean.Souchay@obspm.fr

Robutel, P., Souchay, J.: *An Introduction to the Dynamics of Trojan Asteroids*. Lect. Notes Phys. **790**, 195–227 (2010)
DOI 10.1007/978-3-642-04458-8_4

of the two-body problem, we can simplify the problem by considering a circular motion. Thus we consider a rotating reference frame with the same period as the co-orbiting bodies.

Then the Lagrange points can be viewed as the positions where the combined gravitational attraction of the two large bodies on the third one and the centrifugal force are in balance, so that the third body is at rest in the rotating frame. Notice that the presence of these points of equilibrium still exists when we adopt an elliptical motion for the primaries, instead of a circular one. Among the five Lagrange points, three of them (L_1, L_2, and L_3) are collinear, along the line joining the two large bodies. The two remaining ones are symmetrical with respect to this line, in such a way that they form an equilateral triangle.

This chapter is devoted to the basic analytical development explaining the dynamical behavior of test particles around the L_4 and L_5 Lagrange equilateral positions and to the real case represented by the Trojans asteroids influenced by the combined gravitational torque of Jupiter and of the Sun. Our aim is principally to give the reader in the most detailed manner the theoretical foundations on which classical investigations of the dynamics of Trojan asteroids are carried out. Describing an exhaustive account of the recent and up-to-date developments is largely beyond the scope of this chapter. Despite this, we hope that it will be informative enough to allow any graduate student to easily acquaint themselves with the subject.

At first we will explain intuitively and without any calculation the physical meaning of the five Lagrange points. Then we will give a historical account of the subject, both on a theoretical and on an observational level. After that we will write the fundamental equation of the dynamics for the restricted three-body problem, thus showing the theoretical existence of the Lagrange points. A particular attention will be paid to the dynamical behavior of the equilibrium points L_4 and L_5 in the framework of the restricted three-body problem. The last section will present a brief discussion concerning the stability of (hypothetical) Trojan swarms harbored by the eight planets of the Solar System, including a substantial bibliography.

2 Intuitive Explanations of the Lagrange Points

In order to understand the meaning and characteristics of the Lagrange points, let us consider a planet P with mass M_P orbiting around the Sun S with mass M_S, on a circular orbit. According to the Kepler's third law, the closer a planet is to the Sun, the faster it will move around it, in terms of both angular velocity and amplitude of the velocity.

2.1 The L_1, L_2, and L_3 Lagrange Points

Following this principle, a body M with negligible mass orbiting on a circular orbit around the Sun at a distance smaller than SP will not be able to remain fixed with respect to the SP line. In fact this is not always true. If M is placed between the

Sun and the planet on the SP line, the gravity exerted by P pulls it in the opposite direction than that exerted by the Sun and cancels some of the attraction exerted by the Sun. Therefore, with a weaker pull toward the Sun, M will need less speed to maintain its orbit. The distance SM can be calculated so that the period of revolution of M will be exactly equal to the period of revolution of the planet. This distance corresponds to the Lagrange point L_1. Then, the three points S, M, and P remain aligned in that order.

We can explain the presence of L_2 with exactly the same kind of demonstration as for L_1: suppose now that M is aligned along SP with $SM > SP$. Then the period of revolution of M is a priori longer than the period of revolution of P, since its angular velocity is slower. In fact this is not always the case. The gravitational attraction of P on M is superimposed to that of the Sun. Therefore the central acceleration is bigger than the Keplerian motion and this allows M to move faster. If the distance PM is suitably chosen, the acceleration is such that the corresponding angular velocity is rigorously equal to that of P, and the three bodies S, P, and M remain aligned in that order.

The last aligned Lagrange point, L_3, is located at the opposite side of the Sun with respect to the planet P. Here, M is still subjected to the double attraction of S and P, as was the case in the L_2 configuration. This still causes an increase of orbital velocity with respect to a purely Keplerian motion, and if the distance SP is suitably chosen, the orbital period of M might become exactly identical to that of P.

Notice that if the mass of M is negligible with respect to the mass of P, as we have supposed above, then L_1 and L_2 are at approximately equal distances r_H from the secondary object. r_H corresponds to the radius of the Hill sphere, given by: $r_H \approx R(M_P/3M_S)^{1/3}$. In the case of the Sun–Earth system, the third mass should be placed at 1.5×10^6 km away from the Earth, and in the case of the Earth–Moon system it should be placed at 61,500 km away from the Moon.

The L_1, L_2, and L_3 Lagrange points are unstable, which means that if M slips off these positions, then it will softly drift away and irreparably leave the equilibrium.

2.2 The Lagrange Points L_4 and L_5

Still considering the planet P orbiting the Sun on a circular orbit, the two Lagrange points L_4 and L_5 lie at the same distance $SP = SM$ at 60° ahead of and behind P so that (S, P, M) are forming an equilateral triangle. This case is less easy to understand intuitively. In this situation, the ratio of the gravitational attraction exerted by the two massive bodies S and P on the third one M is the same as the mass ratio of the two bodies. As a consequence, the resultant force acts through the barycenter of the system. In addition, the fact that the three bodies lie on the avertices of an equilateral triangle ensures that the resultant acceleration is to the distance from the barycenter in the same ratio as that of the two massive bodies. This is exactly what is required to keep the body M in orbital equilibrium with the rest of the system.

Notice that the Lagrange points L_4 and L_5 correspond to neutral equilibrium points, which means that when M is gently pushed away from these positions of

equilibrium, it orbits around these positions without drifting farther and farther. This, in particular, is the start point of all the very interesting analytical developments of the next chapters.

3 A Few Historical Points

The three collinear Lagrange points L_1, L_2, and L_3 were first discovered by Euler (1707–1783) in 1765. He applied his calculations to the system Sun–Earth with the Moon as a test particle. He mentioned that if the Moon were four times more distant from the Earth than it is presently, then its motion would be such that it would be permanently in a full Moon configuration. In 1736 the mathematician Joseph Louis Lagrange was born in Torino (Piemont) and moved to Paris in 1787, where he remained until his death in 1813. In 1772, he worked actively on the three-body problem among other topics of celestial mechanics. Investigating the relative positions and velocities of the three bodies starting from the gravitational attraction, he found the famous equilibrium configurations where the three bodies are located at the vertices of an equilateral triangle[1].

A half-century later the French mathematician Joseph Liouville examined the position of equilibrium proposed by Euler concerning the Moon. In 1842, he demonstrated that this position was unstable. Notice that this had some philosophical impact. Indeed, Laplace in his *Exposition du Systême du monde* maintained that the Moon was created in order to shine on the Earth by night, and argued that the Moon would have been placed initially in the position of equilibrium mentioned above. By pointing the instability of the configuration, Liouville invalidated Laplace's argument.

One year later, in 1843, a fundamental property was found by Gascheau [26] when, while making specific studies about the equilateral configuration, he proved that for a circular motion of the three bodies, the positions of the three bodies at the vertex of an equilateral triangle were stable if their masses satisfied:

$$\frac{(m_0+m_1+m_2)^2}{m_0m_1+m_0m_2+m_1m_2} > 27.$$

In the case for which m_2 is negligible with respect to m_0 and m_1, this leads to $\mu(1-\mu) < 1/27$ with $\mu = \frac{m_1}{(m_0+m_1)}$. The corresponding value is: $\mu = \frac{1}{2}(1 - \sqrt{23/27}) \approx 0.0385$. Notice that we consider only the values of $\mu < \frac{1}{2}$ for which $m_1 < m_0$ (the opposite case being symmetrical). For instance we can immediately deduce from this law that in the case of the pair Sun–Jupiter ($\mu \approx 0.001$) and Earth–Moon ($\mu \approx 0.012$) the equilibrium is stable, whereas in the case of the pair Pluton–Charon ($\mu \approx 0.083$) it is unstable.

[1] This configurations lead to the two relative equilibria L_4 and L_5 when the mass of one of the bodies is negligible with respect to the masses of the others.

3.1 Physical Existence of the Lagrange Equilibria: Trojans satellites and other cases.

For more than one century the Lagrange points were only a subject of theoretical investigations, without any confirmation in the nature. In February 1906, the German astronomer Max Wolf put an end to this situation when discovering the first Trojan asteroid Achilles, with the number 588, at the L_4 Lagrange point. After a few months, the strangeness of its orbit was noticed, and soon after that other asteroids were discovered close to the L_4 and L_5 points of Jupiter, as Hector, the largest Trojan asteroid, was discovered in February 1907 by August Kopff, another German astronomer. With dimensions of 370 km $\times$ 200 km, Hector is particularly elongated with respect to other celestial objects of the same size. The theory which claims Hector should be a dual asteroid was suspected for a long time, and finely confirmed by a recent observation in July 2006 by the Keck 10 m telescope, with a resolution of 0.06 arcsec. All the asteroids found at the L_4 and L_5 points of Jupiter accepted names associated with Iliade. The L_4 group is named "group of Achilles" whereas the L_5 group is called group of Trojans, also called the "Patroclus group" to avoid confusion. In fact the Trojans group traditionally corresponds to the combination of the two above groups.

The drastic improvement of observational techniques led to the discovery of a large number of Trojans asteroids, more than one thousand in the two symmetric Lagrange points. As shown in Fig. 1, the Trojan asteroids can be found at very large distance from their respective Lagrange points, both for their projection on Jupiter's orbital plane (Fig. 1a) and on the plane perpendicular to the Sun–Jupiter line (Fig. 1b), for which the angular difference can reach $\pm 40^\circ$. Figure 2 shows the cumulative number of recorded L_4 and L_5 Trojans at a given year from 1900

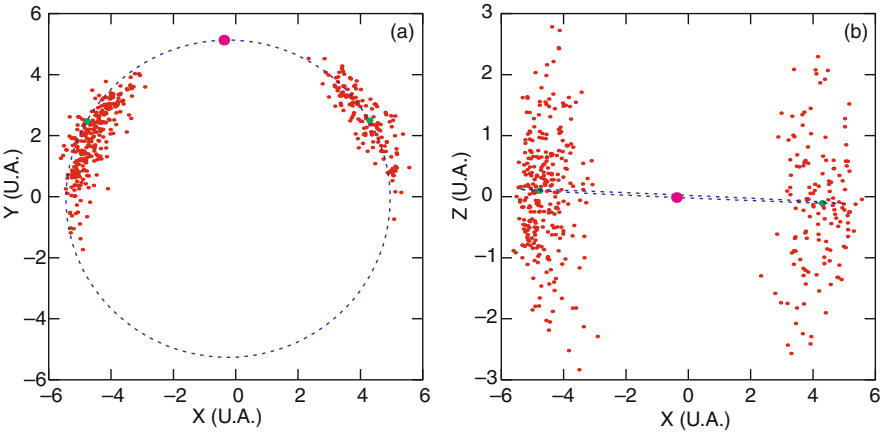

Fig. 1 Bi-dimensional positions of the first 400 recorded Trojan asteroids with respect to Jupiter (*small circle*). (**a**) Planar positions (projection on Jupiter's orbital plane). (**b**) Vertical positions (projections on the plane perpendicular to the Sun–Jupiter line)

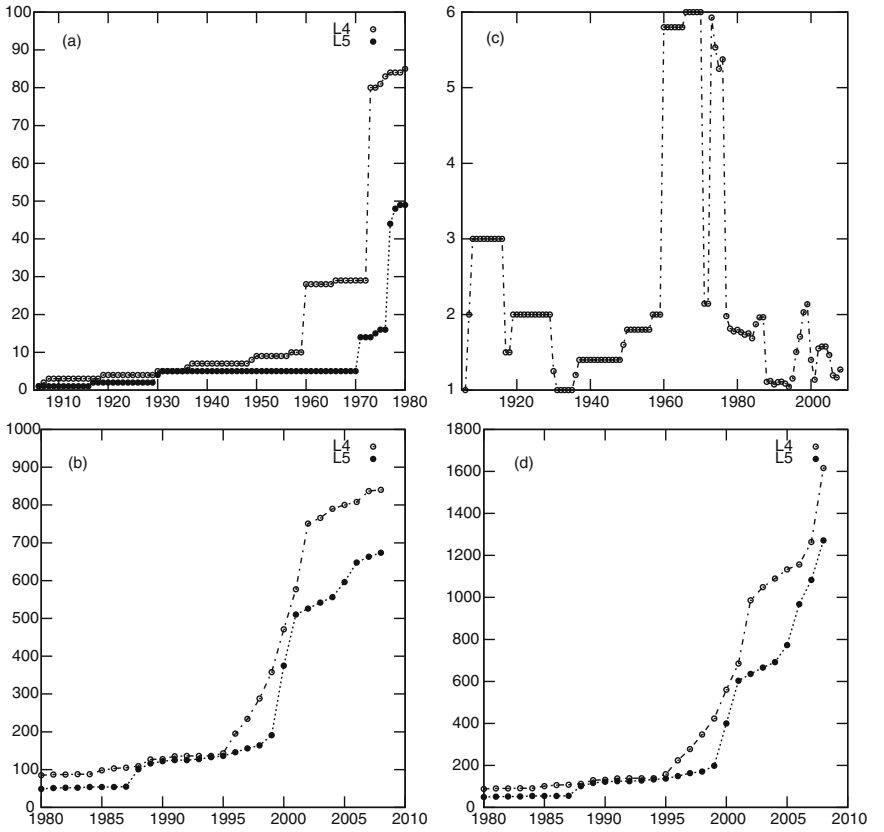

Fig. 2 Histograms representing the cumulative number of known Trojans asteroids at the L_4 and L_5 Lagrange points for a given year, between 1960 and 1980 (**a**) and between 1980 and 2008 (**b**). The corresponding histograms (**c**) represent the ratio of the numbered L_4 asteroids to the numbered L_5 ones. (**d**) Same as (**b**) for all observed Jovian Trojans (numbered as well as unnumbered)

to 1980 (Fig. 2a), from 1980 to 2008 (Fig. 2b), with the L_4/L_5 respective ratio (Fig. 2c). In contrast to Fig. 2b, c where only numbered Trojans[2] are taken into account, Fig. 2d gives the cumulative number are all observed bodies, numbered as well as unnumbered.

We can observe that the number of L_4 asteroids at a given date have always been larger than the corresponding number of L_5 ones. In some periods (for instance between 1960 and 1980), the abnormally large value of the ratio is obviously due to an observational bias (the L_4 zone being largely more explored than the L_5 one).

The Trojans can be, for instance, manually identified from their short trails compared to those of the main-belt asteroids. Close to opposition this proved to work fine, but with larger phase angles problems occur as explained by Lagerkvist

[2] A number is assigned to a given body after accurate orbital elements have been determined.

et al. [33]. A survey of for L_4 Trojans was made by Van Houten et al. [68] with the Palomar Schmidt telescope. They gave an estimate of around 700 objects down to absolute magnitude $H = 13$. This first survey was accompanied by additional ones during several apparitions in September 1973 for L_4, in March 1971, and October 1977 for L_5. They were, respectively called T1, T3 and T2. Lagerkvist et al. used the ESO Schmidt telescopes during apparitions in 1996, 1997, 1998 to study the L_4 point of Jupiter. For instance their first survey in 1996 covered a field of view of about 700 square degrees and they found 399 moving objects classified as Trojans. From this they concluded that about 1100 Trojans are present down to the absolute magnitude $H = 13$. These various systematic surveys carried out during limited time span explain the big jumps appearing in Fig. 2. On January 19, 2009, 1632 numbered Trojans were recorded at the L_4 point, 1277 at the L_5 one. Notice that a complete and up-to-date database of Jupiter Trojans discovered and confirmed, can be found at the following URL: http://www.cfa.harvard.edu/iau/lists/JupiterTrojans.html.

The equilateral triangle configuration has also been discovered in the Saturnian system. Saturn's moon Thetys is in relation with two small bodies located at the Saturn–Thetys' L_4 and L_5 Lagrange points, named respectively, Telesto and Calypso. The other Saturnian moon Dione is another example, with Helena at the L_4 point and Polydeuces at the L_5 one. These two bodies present large longitude variations with respect to the Saturn–Dione line, reaching more than 30° degrees in the case of Polydeuces.

In addition the Sun–Earth system and the Earth–Moon system are subject to some concentration of dust at their respective L_4 and L_5 points, called Kordylewski clouds in the second case (see [37]).

Mars itself possess four asteroids located at L_4 and L_5, which were discovered in 1990 (Eureka), 1998, 1999, and 2007. Neptune has six trojans discovered between 2001 and 2007.

At last the Earth companion Cruithne (3753) has a dynamical behavior similar to that of the Trojans. It alternates between two kinds of orbits due to close encounters with the Earth. When the asteroid is in the smallest and fastest orbit it gains orbital energy when close to the Earth, and then moves on the larger and slower orbit. A similar case of exchange of energy happens for the two satellites of Saturn Epimetheus and Janus.

3.2 Lagrange Collinear Points Artificial Population

The collinear Lagrange points L_1 and L_2 have recently been a big source of interest for people involved in present and future space missions. The Sun–Earth L_1 point is ideal for making observations of the Sun. Objects there are never shadowed by the Earth or by the Moon. The Sun–Earth L_2 point is a well-suited spot to carry out space-based observatories. One of the reasons is that a probe in the neighborhood of L_2 will always maintain the same orientation with respect to the Sun–Earth system,

and for this reason, shielding and calibration are much simpler. At the L_2 point, a spacecraft would not have to make constant orbits around the Earth, resulting in it passing in and out of the Earth's shadow and causing it to heat up and down, thus perturbing its observing mission. Free from that inconvenience and far away from the heat radiated by the Earth, the Sun–Earth L_2 point provides a very stable viewpoint.

Although the L_1 and L_2 points are nominally unstable, it is possible to find stable periodic orbits around these points, in the frame of the restricted three-body problem. These orbits can be ranged in three categories: vertical Lyapounor, horizontal Lyapounor and "halo" orbits. This classification is not obvious when the third body with negligible mass is undergoing the small gravitational perturbations exerted by other celestial bodies in the context of n-body problem, which is the case of the Solar System. Nevertheless, quasiperiodic orbits can be found following Lissajous curve trajectories in the N-body system. Although the orbits are not perfectly stable, a relatively small ballistic effort can allow a space probe to stay in a Lissajous orbit for a long period of time.

3.2.1 Recent Missions

- The ISEE-3 (International Sun Earth Explorer) was a probe launched on September 12, 1978 and sent directly on the L_1 Sun–Earth Lagrange point around which it described a halo to study the interactions between the Sun and the Earth, in particular the solar wind, the magnetosphere and the rays at high energy.
- The WIND probe describes a Lissajous orbit around the Sun–Earth L_1 point, after undergoing a swing-by around the Moon. It was launched on November 1, 1994 to study the solar wind.
- The SOHO (Solar Heliospheric Observatory) was launched on December 2, 1995 with the purpose of studying the Sun from the core to the corona. It reached directly the L_1 Sun–Earth Lagrange point after a direct transfer. Its orbit is a halo around the L_1 point.
- The ACE (Advanced Composition Explorer) launched on August 25, 1997.
- The WMAP (Wilkinson Microwave Anisotropy Probe) launched on June 30, 2001 reached the L_2 Sun–Earth Lagrange point after a 2 month travel and is now librating around it, taking huge data on the cosmic microwave background which represents the signature of the big bang. The probe lost only $1/10$ of its total fuel after reaching its location, and the remaining fuel will allow the probe to hang around the unstable L_2 point for nearly a century.

3.2.2 Future Missions

Because of the advantages mentioned above, the Sun–Earth L_2 Lagrange point is rapidly establishing itself as a prominent location for spacecrafts. For instance ESA has a number of missions that will make use of this very well-suited spot. As missions we can mention Herschel, Planck, Eddington, Gaia, the James Webb Space Telescope, and Darwin.

4 Restricted Three-Bodies Problem and the Lagrange's Equilibrium Points

4.1 n + 1 *Body Problem: General Setting*

Let us consider $n+1$ bodies noted $(P_0, P_1, \ldots, P_n)$ with respective masses $(m_0, m_1, \ldots, m_n)$ each of them undergoing the sole gravitational attraction exerted by the other ones. Let us consider the vector $\mathbf{u}_i$ such that

$\mathbf{u}_i = GP_i$ where G is the center of masse of the system. The fundamental equation of the dynamics writes

$$\ddot{\mathbf{u}}_i = \mathbf{G} \sum_{0 \le j \le n}^{j \neq i} m_j \frac{\mathbf{u}_j - \mathbf{u}_i}{\left\| \mathbf{u}_j - \mathbf{u}_i \right\|^3}. \tag{1}$$

Note that the equation is invariant by translation:

$$\mathbf{u}_i \longmapsto \mathbf{u}_i + \mathbf{v} \quad \mathbf{v} \in \mathbb{R}^3,$$

which is equivalent to the conservation of the linear momentum:

$$\dot{\mathbf{P}} = \sum_{0 \le i \le n} m_i \ddot{\mathbf{u}}_i = 0.$$

An usual and advantageous way to perform this reduction is to use coordinates r_i relative to one of the bodies, let us say P_0, so that we get the following relationships:

$$\begin{cases} \mathbf{r}_i = \mathbf{u}_i - \mathbf{u}_0, & 1 \le i \le n \\ \mathbf{r}_0 = \dfrac{m_0}{m}\mathbf{u}_0 + \cdots + \dfrac{m_n}{m}\mathbf{u}_n = 0, \; m = m_0 + \cdots + m_n \end{cases}. \tag{2}$$

Consequently, using relative coordinates (2) the (1) can be written as:

$$\begin{cases} \ddot{\mathbf{r}}_0 = & 0 \\ \ddot{\mathbf{r}}_i = -\mathbf{G}\dfrac{m_0 + m_i}{\left\| \mathbf{r}_i \right\|^3}\mathbf{r}_i + \sum_{1 \le j \le n}^{j \neq i} \mathbf{G} m_j \left(\dfrac{\mathbf{r}_j - \mathbf{r}_i}{\left\| \mathbf{r}_j - \mathbf{r}_i \right\|^3} - \dfrac{\mathbf{r}_j}{\left\| \mathbf{r}_j \right\|^3} \right) \end{cases}. \tag{3}$$

We are left with a $3n$ degrees of freedom system rather than $3(n+1)$. It is always possible to reduce the number of degrees of freedom of the differential system, by taking into account the conservation of angular momentum [7, 15] or by using the Jacobi reduction [58], or partial reduction [38], but in the following sections, we will not need these techniques.

4.2 Application to the Three-Body Problem: The Lagrance Points

In the case of the three-body problems (three bodies P_0, P_1, and P_2 with respective masses m_0, m_1, and m_2) (3) become:

$$\begin{cases} \ddot{\mathbf{r}}_0 = & 0 \\ \ddot{\mathbf{r}}_1 = -\mathbf{G}\dfrac{m_0+m_1}{\|\mathbf{r}_1\|^3}\mathbf{r}_1 + Gm_2\left(\dfrac{\mathbf{r}_2-\mathbf{r}_1}{\|\mathbf{r}_2-\mathbf{r}_1\|^3} - \dfrac{\mathbf{r}_2}{\|\mathbf{r}_2\|^3}\right) \\ \ddot{\mathbf{r}}_2 = -\mathbf{G}\dfrac{m_0+m_2}{\|\mathbf{r}_2\|^3}\mathbf{r}_2 + Gm_1\left(\dfrac{\mathbf{r}_1-\mathbf{r}_2}{\|\mathbf{r}_1-\mathbf{r}_2\|^3} - \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3}\right) \end{cases} \tag{4}$$

4.2.1 Collinear Solutions of Equilibrium L_1, L_2, and L_3

First of all we can investigate the existence of collinear positions of equilibrium, simply by setting $\mathbf{r}_2 = \alpha\mathbf{r}_1$. Notice that in that case α can a priori depend on t. Then $\mathbf{r}_1 - \mathbf{r}_2 = (1-\alpha)\mathbf{r}_1$. By substitution in (4) we find

$$\begin{cases} \ddot{\mathbf{r}}_1 = & -\mathbf{G}\left(m_0+m_1+m_2\left(\dfrac{\alpha}{|\,\alpha\,|^3} + \dfrac{1-\alpha}{|\,1-\alpha\,|^3}\right)\right)\dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} \\ \ddot{\mathbf{r}}_2 = -\mathbf{G}\left(m_0+m_2+m_1\dfrac{|\,\alpha\,|^3}{\alpha}\left(1-\dfrac{1-\alpha}{|\,1-\alpha\,|^3}\right)\right)\dfrac{\mathbf{r}_2}{\|\mathbf{r}_2\|^3} \end{cases} . \tag{5}$$

In addition, $\mathbf{r}_2 = \alpha\mathbf{r}_1$ leads to $\ddot{\mathbf{r}}_2 = \alpha\ddot{\mathbf{r}}_1$. This condition mixed with the equations above gives

$$\begin{aligned} m_0+m_1+m_2\left(\frac{\alpha}{|\,\alpha\,|^3} + \frac{1-\alpha}{|\,1-\alpha\,|^3}\right) = \\ \frac{1}{|\,\alpha\,|^3}\left(m_0+m_2+m_1\frac{|\,\alpha\,|^3}{\alpha}\left(1-\frac{1-\alpha}{|\,1-\alpha\,|^3}\right)\right). \end{aligned} \tag{6}$$

It can be shown that, when removing the absolute values, (6) leads to three different polynomial equations of degree 5 possessing only one real root each. These three real solutions of (6) verify $\alpha_3 < 0 < \alpha_1 < \alpha_2$. If the condition $\mathbf{r}_2 \wedge \dot{\mathbf{r}}_2 = \alpha_i^2\mathbf{r}_1 \wedge \dot{\mathbf{r}}_1$ is verified for $i = 1, 2, 3$ then the three points P_1, P_2, and P_3 are permanently aligned. The orbits of P_1 and P_2 around P_0 are two conics around P_0, coplanar and homothetic with the ratio α_i. The semi-major axes of these conics are aligned and they have the same eccentricity.

4.2.2 Equilateral Positions of Equilibrium L_4 and L_5

In order to understand the possibility of equilibrium in an equilateral triangle configuration, let us transform (4) in the following form:

$$\begin{cases} \ddot{\mathbf{r}}_0 = & 0 \\ \ddot{\mathbf{r}}_1 = -\mathbf{G}\dfrac{m_0+m_1+m_2}{\|\mathbf{r}_1\|^3}\mathbf{r}_1 + \mathbf{G}m_2\left(\dfrac{\mathbf{r}_2-\mathbf{r}_1}{\|\mathbf{r}_2-\mathbf{r}_1\|^3} - \dfrac{\mathbf{r}_2}{\|\mathbf{r}_2\|^3} + \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3}\right) \\ \ddot{\mathbf{r}}_2 = -\mathbf{G}\dfrac{m_0+m_1+m_2}{\|\mathbf{r}_2\|^3}\mathbf{r}_2 + \mathbf{G}m_1\left(\dfrac{\mathbf{r}_1-\mathbf{r}_2}{\|\mathbf{r}_1-\mathbf{r}_2\|^3} - \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} + \dfrac{\mathbf{r}_2}{\|\mathbf{r}_2\|^3}\right) \end{cases} . \quad (7)$$

Then if we set the following equalities $|\ \mathbf{r}_1\ | = |\ \mathbf{r}_2\ | = |\ \mathbf{r}_1 - \mathbf{r}_2\ |$, the second part of the right-hand side annulates and the equations above can be written in the simplified manner:

$$\begin{cases} \ddot{\mathbf{r}}_0 = & 0 \\ \ddot{\mathbf{r}}_1 = -\mathbf{G}(m_0+m_1+m_2)\dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} \\ \ddot{\mathbf{r}}_2 = -\mathbf{G}(m_0+m_1+m_2)\dfrac{\mathbf{r}_2}{\|\mathbf{r}_2\|^3} \end{cases} . \quad (8)$$

The conditions of equalities of the three distances above are satisfied at the condition that the three points P_0, P_1, and P_2 are located at the vertices of an equilateral triangle, whereas the two last correspond to the classical two-body problem. Therefore these conditions and these equations can be satisfied simultaneously if P_1 and P_2 are describing coplanar orbits around P_0, with the same semi-major axis and the same eccentricity, in fact if these orbits are the same but shifted by a 60° rotation angle around P_0. Moreover as the gravitational constant $\mu = G(m_0 + m_1 + m_2)$ is the same for the two Keplerian equations (8), the motions along the two orbits will be synchronous, with the same period in the case for which the two orbits are elliptic.

4.3 Equilateral Configurations in the Restricted-Three-Body Problem

In this section, we consider the restricted three-body problem, which means that one of the masses is zero. For instance we put $m_2 = 0$. Then the motion of P_1 with respect to P_0 becomes a two-body problem, and we are interested only in the motion of the third body P_2 with zero mass. As the motion of P_2 is the only subject of study, and for the sake of simplicity, we take $\mathbf{r} = \mathbf{r}_2$ Thus, the equations become:

$$\begin{cases} \ddot{\mathbf{r}}_1 = -\mathbf{G}\dfrac{(m_0+m_1)}{\|\mathbf{r}_1\|^3}\mathbf{r}_1 \\ \ddot{\mathbf{r}} = -\mathbf{G}\dfrac{(m_0+m_1)}{\|\mathbf{r}\|^3}\mathbf{r} + \mathbf{G}m_1\left(\dfrac{\mathbf{r}_1-\mathbf{r}}{\|\mathbf{r}_1-\mathbf{r}\|^3} - \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} + \dfrac{\mathbf{r}}{\|\mathbf{r}\|^3}\right) \end{cases} . \quad (9)$$

Denoting, respectively, by μ_0 and μ_1 the quantities $\mathbf{G}(m_0 + m_1)$ and $\mathbf{G}m_1$, the previous equation can be writen as:

$$\begin{cases} \ddot{\mathbf{r}}_1 = -\mu_0 \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} \\ \ddot{\mathbf{r}} = -\mu_0 \dfrac{\mathbf{r}}{\|\mathbf{r}\|^3} + \mu_1 \left(\dfrac{\mathbf{r}_1 - \mathbf{r}}{\|\mathbf{r}_1 - \mathbf{r}\|^3} - \dfrac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} + \dfrac{\mathbf{r}}{\|\mathbf{r}\|^3} \right) \end{cases}. \tag{10}$$

If we assume that each of these three bodies is located at a vertices of an equilateral triangle and that the primaries do not evolve on a straight line, we will show that this configuration is conserved at any instant if and only if this triangle lies in the plane of the motion of the two primaries.

Indeed, if the mutual distances between the three bodies are equal (but not necessarily constant), that is $\|\mathbf{r}\| = \|\mathbf{r}_1\| = \|\mathbf{r} - \mathbf{r}_1\| = \rho(t)$, the last factor of the second line of (10) vanishes. Consequently, the two vectors $\mathbf{r}$ and $\mathbf{r}_1$ satisfy the same differential equation:

$$\ddot{\mathbf{x}} = -\mu_0 \frac{\mathbf{x}}{\|\mathbf{x}(t)\|^3}. \tag{11}$$

It turns out that if the equilateral configuration is preserved, the motions of the bodies are Keplerian. Let us now assume that the motion of the primaries is bounded, that is to say elliptic (the following proofs are quite identical when the trajectories are parabolic or hyperbolic).

As $\|\mathbf{r}(t)\| = \|\mathbf{r}_1(t)\|$ for all t and knowing that

$$\|\mathbf{r}\| = a(1 - e\cos E) \quad \text{and} \quad \|\mathbf{r}_1\| = a_1(1 - e_1 \cos E_1), \tag{12}$$

we can make the following remarks: in the two ellipses we have the same minimum and maximum distances for P_0P_1 and P_0P_2 . Thus the two ellipses have the same semi-major axis and eccentricity: $a = a_1$, $e = e_1$. Consequently, we have $E(t) = E_1(t)$ for all t. In other words P_1 and P_2 describe exactly the same ellipse shifted with 60° and the eccentric anomalies E and E_1, as well as the true anomalies v and v_1 at any instant are the same.

It remains to show that the two ellipses lie in the same plane. In order to facilitate the demonstration, we can adopt the following reference frame: the plane (P_0, x, y) is the orbital plane of P_1 around P_0 and the axis (P_0, x) is chosen in such a way that it is directed toward the perihelion. Then we adopt the classical orbital parameters Ω, ω, v, i, and ϖ to represent the motion of P_2 in (P_0, x, y, z).

If S is the angle between $\mathbf{r}$ and $\mathbf{r}_1$ the relation $\cos S = 1/2$ must be satisfied, for the angle between the two points P_1 and P_2 is 60°. As $v_1 = v$, the coordinates of the two vectors can be written as:

$$\mathbf{r}_1 = \|\mathbf{r}_1\| \begin{pmatrix} \cos v \\ \sin v \\ 0 \end{pmatrix}, \tag{13}$$

$$\mathbf{r} = \|\mathbf{r}\| \begin{pmatrix} \cos(\omega + v)\cos\Omega - \sin(\omega + v)\sin\Omega\cos i \\ \cos(\omega + v)\sin\Omega + \sin(\omega + v)\cos\Omega\cos i \\ \sin i \sin(\omega + v) \end{pmatrix}, \tag{14}$$

which is equivalent to

$$\mathbf{r} = \|\mathbf{r}\| \begin{pmatrix} \cos(\Omega + \omega + v) + (1 - \cos i)\sin\Omega\sin(\omega + v) \\ \sin(\Omega + \omega + v) - (1 - \cos i)\cos\Omega\sin(\omega + v) \\ \sin i \sin(\omega + v) \end{pmatrix}. \tag{15}$$

A straightforward computation gives

$$\begin{aligned} \forall\, v:\ \frac{1}{2} = \cos S &= \frac{\mathbf{r}\cdot\mathbf{r}_1}{\|\mathbf{r}\|\,\|\mathbf{r}_1\|} \\ &= \cos(\Omega + \omega) + (\cos i - 1)\sin(\omega + v)\sin(v - \Omega). \end{aligned} \tag{16}$$

If $v = -\omega$ we get $\cos(\Omega + \omega) = \frac{1}{2}$ consequently, the condition (16), is equivalent to

$$\begin{cases} \cos(\Omega + \omega) = \frac{1}{2} \\ (\cos i - 1)\sin(\omega + v)\sin(v - \Omega) = 0 \ \ \forall\, v \end{cases} \quad \text{that is} \quad \begin{cases} \Omega + \omega = \pm\frac{\pi}{3} \\ i = 0 \end{cases}. \tag{17}$$

As a result, the two ellipses possess the same elliptic elements, excepted their arguments of the perihelia which is translated by $\pm\frac{\pi}{3}$.

5 Behavior of the Trajectories in a Neighborhood of Equilateral Points L_4 and L_5

In this section, we analyze in details the study of the motion of the massless particle P_2 in the vicinity of the Lagrange points L_4 and (or) L_5. For that purpose, we assume that the motion of the two primaries P_0 and P_1 is a bounded Keplerian motion such that $\mathbf{r}_1 \wedge \dot{\mathbf{r}}_1 \neq 0$. We have seen in Sect. 4 that the points L_1 to L_5 lie on the plane of the motion of the primaries, but in order to study their local stability, it is necessary to consider both planar and spatial (vertical) variations. To this aim, the motion of the primaries do not lying on a straight line, we define the normal unit vector $\mathbf{k}$ by

$$\mathbf{k} = \frac{\mathbf{r}_1 \wedge \dot{\mathbf{r}}_1}{\|\mathbf{r}_1 \wedge \dot{\mathbf{r}}_1\|} \quad \text{such that the basis } (\mathbf{r}_1, \dot{\mathbf{r}}_1, \mathbf{k}) \text{ is direct.}$$

Consequently, the vector $\mathbf{r} = \boldsymbol{P_0P_2}$ splits naturally in $\mathbf{r} = \mathbf{r}_p + \mathbf{z}$ where $\mathbf{r}_p$ is the orthogonal projection of $\mathbf{r}$ on the plane of the primaries (i.e., the plane generated by $(\mathbf{r}_1, \dot{\mathbf{r}}_1)$), and $\mathbf{z}$ is the projection of $\mathbf{r}$ on $\mathbf{k}$.

5.1 Linearization of the Equation of the Motion in a Neighborhood of the Equilateral Points

Let us assume that $\mathbf{s}_0(t)$ is a solution of the restricted three-body problem. In order to study the dynamics around this trajectory, the first step leads in the derivation of the variational equation (linearization of the equations of the motion along the solution $\mathbf{s}_0$). In other words, if $\mathbf{s}_0$ satisfy the equation $\ddot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ the variational equation along $\mathbf{s}_0$ is obtained by the linearization in the neighborhood of $\mathbf{s}_0$ of the equation: $\ddot{\mathbf{s}}_0 + \ddot{\boldsymbol{\eta}} = \mathbf{f}(\mathbf{s}_0 + \boldsymbol{\eta})$, that is

$$\ddot{\boldsymbol{\eta}} = D\mathbf{f}(\mathbf{s}_0(t))\boldsymbol{\eta}, \tag{18}$$

where $D\mathbf{f}(\mathbf{s}_0)$ is the differential of the function $\mathbf{f}$ (which can be identified to the Jacobian matrix $\mathbf{f}$) evaluated on the vector $\mathbf{s}_0$. Equation (18) is of great interest in the sense that its solutions drive the dynamics in a neighborhood of the trajectory $\mathbf{s}_0$. In particular, if $\boldsymbol{\eta}$ is small enough, the temporal evolution of $\boldsymbol{\eta}$ tells us what is the behavior of the neighboring solutions of $\mathbf{s}_0$, namely the solution $\mathbf{s}_0 + \boldsymbol{\eta}$. If $\boldsymbol{\eta}(t)$ remains always small, the solution $\mathbf{s}_0(t) + \boldsymbol{\eta}(t)$ evolves around $\mathbf{s}_0(t)$, and we will talk of linear stability. In the contrary if $\|\boldsymbol{\eta}\|$ goes to infinity with t, we will talk of instability, in the sense that the difference between the two neighbor solutions $\mathbf{s}_0(t)$ and $\mathbf{s}_0(t) + \boldsymbol{\eta}(t)$ diverge.

In order to derive the variational equations around the solution $\mathbf{s}_0$ of the equation:

$$\ddot{\mathbf{r}} = -\mu_0 \frac{\mathbf{r}}{\|\mathbf{r}\|^3} + \mu_1 \left(\frac{\mathbf{r}_1 - \mathbf{r}}{\|\mathbf{r}_1 - \mathbf{r}\|^3} - \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3} + \frac{\mathbf{r}}{\|\mathbf{r}\|^3} \right), \tag{19}$$

we first expend the right hand side of (19) at order one in $\boldsymbol{\eta}$, where $\mathbf{r} = \mathbf{s}_0 + \boldsymbol{\eta}$.

If $\mathbf{x}$ and $\mathbf{y}$ are two vectors of $\mathbb{R}^3$, we denote by $\mathbf{x} \cdot \mathbf{y}$ their usual scalar product and by $\mathbf{x}^2$ the scalar product of $\mathbf{x}$ by itself. Using this notations, the expansion of $(\mathbf{x} + \boldsymbol{\eta})/\|\mathbf{x} + \boldsymbol{\eta}\|^3$ at first order in $\boldsymbol{\eta}$ writes

$$\begin{aligned} \frac{\mathbf{x} + \boldsymbol{\eta}}{\|\mathbf{x} + \boldsymbol{\eta}\|^3} &= (\mathbf{x} + \boldsymbol{\eta})(\mathbf{x}^2 + 2\mathbf{x} \cdot \boldsymbol{\eta} + \boldsymbol{\eta}^2)^{-3/2} \\ &= \frac{\mathbf{x}}{\|\mathbf{x}\|^3} + \frac{\boldsymbol{\eta}}{\|\mathbf{x}\|^3} - 3\mathbf{x} \cdot \boldsymbol{\eta} \frac{\mathbf{x}}{\|\mathbf{x}\|^5} + O(\boldsymbol{\eta}^2). \end{aligned} \tag{20}$$

This expression leads to the expansions

$$\frac{\mathbf{s}_0 + \boldsymbol{\eta}}{\|\mathbf{s}_0 + \boldsymbol{\eta}\|^3} = \frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^3} + \frac{\boldsymbol{\eta}}{\|\mathbf{s}_0\|^3} - 3\mathbf{s}_0 \cdot \boldsymbol{\eta} \frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5} + O(\boldsymbol{\eta}^2) \tag{21}$$

and

$$\frac{\mathbf{r}_1-\mathbf{s}_0-\boldsymbol{\eta}}{\|\mathbf{r}_1-\mathbf{s}_0-\boldsymbol{\eta}\|^3}=\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^3}-\frac{\boldsymbol{\eta}}{\|\mathbf{r}_1-\mathbf{s}_0\|^3} +3(\mathbf{r}_1-\mathbf{s}_0)\cdot\boldsymbol{\eta}\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^5}+O(\boldsymbol{\eta}^2). \tag{22}$$

Replacing (21) and (22) in (19) we get

$$\begin{aligned}\ddot{\mathbf{s}}_0+\ddot{\boldsymbol{\eta}}=&-\mu_0\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^3}+\mu_1\left(\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^3}-\frac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3}+\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^3}\right)\\ &-\mu_0\left(\frac{\boldsymbol{\eta}}{\|\mathbf{s}_0\|^3}-3\mathbf{s}_0\cdot\boldsymbol{\eta}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5}\right)\\ &+\mu_1\left(\frac{\boldsymbol{\eta}}{\|\mathbf{s}_0\|^3}-3\mathbf{s}_0\cdot\boldsymbol{\eta}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5}-\frac{\boldsymbol{\eta}}{\|\mathbf{r}_1-\mathbf{s}_0\|^3}+3(\mathbf{r}_1-\mathbf{s}_0)\cdot\boldsymbol{\eta}\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^5}\right)\\ &+O(\boldsymbol{\eta}^2)\end{aligned} \tag{23}$$

and finally, the variational equation can be written as:

$$\begin{aligned}\ddot{\boldsymbol{\eta}}=&-(\mu_0-\mu_1)\left(\frac{\boldsymbol{\eta}}{\|\mathbf{s}_0\|^3}-3\mathbf{s}_0\cdot\boldsymbol{\eta}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5}\right)\\ &-\mu_1\left(\frac{\boldsymbol{\eta}}{\|\mathbf{r}_1-\mathbf{s}_0\|^3}-3(\mathbf{r}_1-\mathbf{s}_0)\cdot\boldsymbol{\eta}\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^5}\right).\end{aligned} \tag{24}$$

This second-order linear differential system which describes the infinitesimal variations around the solution $\mathbf{s}_0(t)$ is in general time dependent and consequently not integrable. We will see later that, when $\mathbf{s}_0(t)$ is an equilateral solution of the circular restricted three-body problem, this equations become independent of the time, and then can be solved.

Before going further, let us assume that the solution $\mathbf{s}_0(t)$ lies on the plane of the primaries. In this particular case, the vertical component (perpendicular to the primaries plan) of the variational equation (24) can be drastically simplified. Indeed, setting $\boldsymbol{\eta}=\mathbf{h}+\mathbf{z}=\mathbf{h}+z\mathbf{k}$, where $\mathbf{h}$ is the planar component of $\boldsymbol{\eta}$ and $\mathbf{z}$ its vertical one, the because $\mathbf{s}_0\cdot\mathbf{k}=\mathbf{r}_1\cdot\mathbf{k}=0$, (24) splits in two equations: the horizontal variational equation given by

$$\begin{aligned}\ddot{\mathbf{h}}=&-(\mu_0-\mu_1)\left(\frac{\mathbf{h}}{\|\mathbf{s}_0\|^3}-3\mathbf{s}_0\cdot\mathbf{h}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5}\right)\\ &-\mu_1\left(\frac{\mathbf{h}}{\|\mathbf{r}_1-\mathbf{s}_0\|^3}-3(\mathbf{r}_1-\mathbf{s}_0)\cdot\mathbf{h}\frac{\mathbf{r}_1-\mathbf{s}_0}{\|\mathbf{r}_1-\mathbf{s}_0\|^5}\right)\end{aligned} \tag{25}$$

and the vertical variational equation:

$$\ddot{z} = -\mu_0 \frac{z}{\|\mathbf{s}_0\|^3} + \mu_1 \left(\frac{1}{\|\mathbf{s}_0\|^3} - \frac{1}{\|\mathbf{r}_1 - \mathbf{s}_0\|^3} \right) z. \tag{26}$$

5.2 The Linear Vertical Variations

The linear equation (26) given the vertical infinitesimal variations of a planar solution of (19) is generally time dependent, but in some particular cases, its solution can be find easily. Let the trajectory $\mathbf{s}_0(t)$ be an equilateral solution restricted three-body problem, circular as well as elliptic. In both cases, the relation $\|\mathbf{s}_0\| = \|\mathbf{r}_1 - \mathbf{s}_0\| = \|\mathbf{r}_1\|$ holds, and the differential equation gives

$$\ddot{z} = -\frac{\mu_0}{\|\mathbf{r}_1\|^3} z. \tag{27}$$

If the motion of the primaries is circular, the vector $\|\mathbf{r}_1\|$ is constant. Thus, (27) is the equation of an harmonic oscillator whose eigenvalues are equal to $\pm i\sqrt{\mu_0/a_1^3} = \pm i n_1$. This implies linear stability in the vertical direction and imposes the solutions of the vertical variational equation to be $2\pi/n_1$ periodic. It worth mentioning that the "vertical frequency" and the orbital frequency are in 1:1 resonance.

When the planetary trajectory is an ellipse, $\|\mathbf{r}_1\|$ is no more constant and (27) is not autonomous, and its general solution is usually unknown. But here again, a very simple argument yields the explicit solution of (27). Indeed, the vector $\mathbf{r}_1$ satisfies the differential equation:

$$\ddot{\mathbf{r}}_1 = -\frac{\mu_0}{\|\mathbf{r}_1\|^3} \mathbf{r}_1, \tag{28}$$

which is obviously not a linear equation. But the motion of the primaries being given, $\|\mathbf{r}_1\|$ can be considered as a known time-dependent function. Therefore, x_1, y_1, and z (it is also true for x and y) are solution of the linear scalar differential equation:

$$\ddot{X} = -\frac{\mu_0}{\|\mathbf{r}_1(t)\|^3} X. \tag{29}$$

According to the theory of the linear differential equations, every solutions of (29) are a linear combination of two linearly independent particular solutions of (29). The determinant

$$\begin{vmatrix} x_1(t) & y_1(t) \\ \dot{x}_1(t) & \dot{y}_1(t) \end{vmatrix}$$

being different from zero for all time,[3], there exist two real numbers α and β, depending on the initial conditions $z(0)$ and $\dot{z}(0)$ such that

$$z(t) = \alpha x_1(t) + \beta y_1(t). \tag{30}$$

It turns out that the infinitesimal vertical variation is a periodic function which the frequency is equal to the mean motion n_1 of the primaries. As for the circular restricted three-body problem, the equilateral solutions are linearly stable in the vertical direction.

It worth mentioning that, using the same arguments, the solution (30) can also be written as

$$z(t) = \alpha' x(t) + \beta' y(t), \tag{31}$$

where x and y are the two co-ordinates of the vector $\mathbf{s}_0$ and (α', β') two real numbers depending only on $z(0)$ and $\dot{z}(0)$. Therefore, the trajectories associated to infinitesimal inclinations lie on a fixed plane. This implies that the precession frequency of the ascending node of an asteroid evolving in a neighborhood of L_4 or L_5 tends to zero with its inclination.

5.3 The Horizontal Variational Equation

We have seen in Sect. 5.1 that the variational equation (24) could be split in two independent systems of equation: one driving the vertical variations and the other one associated to the linearized motion in the plane of the primaries. The vertical equation derived from an equilateral solution was solved easily, even in the elliptic case. It will not be so straightforward for the horizontal equation (25). Let us first investigate the simplest situation of the circular restricted three-body problem.

5.3.1 The Case of the Circular Three-Body Problem

When the orbit of the trajectory $\mathbf{s}_0$ is circular, even if the quantities $\|\mathbf{s}_0\|$ and $\|\mathbf{r}_1 - \mathbf{s}_0\|$ are constant, (25) is not autonomous. But, as along this solution the mutual distances between the bodies remain constant, in a reference frame rotating with the two primaries the solution $\mathbf{s}_0$ becomes a fixed point. Consequently, in a suitable coordinates system the horizontal variational equation is a linear autonomous differential system. In order to derive this system, let us chose the most massive body P_0 as the origin of the mobile reference frame, its basis $(\mathbf{e}_1, \mathbf{e}_2)$ being defined by $\mathbf{e}_1 = P_0P_1$ and $\mathbf{e}_2 = R(\pi/2)\mathbf{e}_1$, where $R(\pi/2)$ is the rotation of angle $\pi/2$ in the plane of the primaries. More generally, the matrix of $R(\theta)$ (rotation of angle θ) is given by

[3] The motion being elliptic, the two vectors $\mathbf{r}_1$ and $\dot{\mathbf{r}}_1$ are always linearly independent.

$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

in the basis $(\mathbf{e}_1, \mathbf{e}_2)$. Let us notice that, as we consider in this section only planar motions, it is necessary to use a 3-dimensional co-ordinate system.

Let $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_1$ be the position of P and P_1 in the mobile reference frame. We have

$$\boldsymbol{\rho}_1 = \begin{pmatrix} a_1 \\ 0 \end{pmatrix}, \quad \mathbf{r}_1(t) = R(n_1 t)\boldsymbol{\rho}_1 \quad \text{and} \quad \mathbf{r}(t) = R(n_1 t)\boldsymbol{\rho}(t). \tag{32}$$

As

$$\frac{d^n}{d\theta^n} R(\theta) = R(\theta + n\pi/2),$$

the second derivative of the vector $\mathbf{r}$ with respect to t expressed in the mobile frame writes

$$\begin{aligned} \ddot{\mathbf{r}}(t) &= n_1^2 R(n_1 t + \pi)\boldsymbol{\rho} + 2n_1 R(n_1 t + \pi/2)\dot{\boldsymbol{\rho}} + R(n_1 t)\ddot{\boldsymbol{\rho}} \\ &= R(n_1 t)\left[\ddot{\boldsymbol{\rho}} + 2n_1 R(\pi/2)\dot{\boldsymbol{\rho}} - n_1^2 \boldsymbol{\rho}\right]. \end{aligned} \tag{33}$$

Using formulas (32), the right-hand side of (19) gives

$$\ddot{\mathbf{r}}(t) = R(n_1 t)\left[-\mu_0 \frac{\boldsymbol{\rho}}{\|\boldsymbol{\rho}\|^3} + \mu_1 \left(\frac{\boldsymbol{\rho}_1 - \boldsymbol{\rho}}{\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}\|^3} - \frac{\boldsymbol{\rho}_1}{\|\boldsymbol{\rho}_1\|^3} + \frac{\boldsymbol{\rho}}{\|\boldsymbol{\rho}\|^3} \right)\right] \tag{34}$$

and plugging the transformation (33) in the above equation, the equation of the motion of the massless body in the rotating frame is

$$\ddot{\boldsymbol{\rho}} + 2n_1 R(\pi/2)\dot{\boldsymbol{\rho}} - n_1^2 \boldsymbol{\rho} = -(\mu_0 - \mu_1)\frac{\boldsymbol{\rho}}{\|\boldsymbol{\rho}\|^3} + \mu_1 \left(\frac{\boldsymbol{\rho}_1 - \boldsymbol{\rho}}{\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}\|^3} - \frac{\boldsymbol{\rho}_1}{\|\boldsymbol{\rho}_1\|^3} \right). \tag{35}$$

In order to derive the horizontal variational equation in the mobile frame, two paths can be followed: First, we can linearize the previous equation in the neighborhood of $\boldsymbol{\rho}$ (which is a fixed point in the mobile frame). Second, in a more straightforward manner, the horizontal variational equation (25) is directly written in rotating coordinates by the mean of the transformation (32). The relations between the main quantities in fixed reference frame and in mobile frame are

$$\mathbf{s}_0(t) = R(\theta)\mathbf{r}_1(t) = R(\theta + n_1 t)\boldsymbol{\rho}_1, \tag{36}$$

where $\theta = \pi/3$ if we consider L_4 and $-\pi/3$ if it is L_5, and

$$\mathbf{r}_1(t) = R(n_1 t)\boldsymbol{\rho}_1, \quad \mathbf{h}(t) = R(n_1 t)\mathbf{u}(t), \tag{37}$$

where $\mathbf{u}$ if the infinitesimal horizontal variation in the rotating frame.

In the mobile frame the left-hand side of (25) becomes

$$\ddot{\mathbf{h}} = R(n_1 t)\left[\ddot{\mathbf{u}} + 2n_1 R(\pi/2)\dot{\mathbf{u}} - n_1^2 \mathbf{u}\right]. \tag{38}$$

Now, let us transform terms by terms on the right-hand side of (25). First of all, according to (36), (37), and because the motion of the primaries is circular, we have

$$\frac{\mathbf{h}}{\|\mathbf{r}_1 - \mathbf{s}_0\|^3} = \frac{\mathbf{h}}{\|\mathbf{s}_0\|^3} = \frac{\mathbf{h}}{a_1^3} = R(n_1 t)\frac{\mathbf{u}}{a_1^3}, \tag{39}$$

identically we have

$$\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5} = R(n_1 t)R(\theta)\frac{\boldsymbol{\rho}_1}{a_1^5} \tag{40}$$

and

$$\frac{\mathbf{r}_1 - \mathbf{s}_0}{\|\mathbf{r}_1 - \mathbf{s}_0\|^5} = R(n_1 t)\frac{\boldsymbol{\rho}_1 - R(\theta)\boldsymbol{\rho}_1}{a_1^5} = R(n_1 t)R(-\theta)\frac{\boldsymbol{\rho}_1}{a_1^5}, \tag{41}$$

the last equation being satisfied as long as $\theta = \pm\pi/3$. these formulas lead to the following expressions:

$$\mathbf{s}_0 \cdot \mathbf{h}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5} = \left(R(\theta)\boldsymbol{\rho}_1 \cdot \mathbf{u}\right) R(n_1 t)R(\theta)\frac{\boldsymbol{\rho}_1}{a_1^5} \tag{42}$$

and

$$(\mathbf{r}_1 - \mathbf{s}_0) \cdot \mathbf{h}\frac{\mathbf{r}_1 - \mathbf{s}_0}{\|\mathbf{r}_1 - \mathbf{s}_0\|^5} = \left(R(-\theta)\boldsymbol{\rho}_1 \cdot \mathbf{u}\right) R(n_1 t)R(-\theta)\frac{\boldsymbol{\rho}_1}{a_1^5}. \tag{43}$$

The two previous expressions are linear in $\mathbf{u}$ and thus can be written as $M\mathbf{u}$, where M is a 2×2 real matrix. Noticing that these two expressions can be written (if we forget the constant a_1^{-5} and the matrix $R(n_1 t)$ which will be factorized later) as $(\mathbf{x} \cdot \mathbf{u})\mathbf{x}$, the matrix $M = \mathbf{x}\mathbf{x}^T$ can be derived from the identities:

$$(\mathbf{x} \cdot \mathbf{u})\mathbf{x} = \mathbf{x}(\mathbf{x} \cdot \mathbf{u}) = \mathbf{x}(\mathbf{x}^T\mathbf{u}) = (\mathbf{x}\mathbf{x}^T)\mathbf{u} = M\mathbf{u} \tag{44}$$

Consequently, we get

$$\mathbf{s}_0 \cdot \mathbf{h}\frac{\mathbf{s}_0}{\|\mathbf{s}_0\|^5} = \frac{1}{4a_1^3} R(n_1 t) M_\epsilon \mathbf{u},$$
$$(\mathbf{r}_1 - \mathbf{s}_0) \cdot \mathbf{h}\frac{\mathbf{r}_1 - \mathbf{s}_0}{\|\mathbf{r}_1 - \mathbf{s}_0\|^5} = \frac{1}{4a_1^3} R(n_1 t) M_{-\epsilon} \mathbf{u} \quad \text{with} \tag{45}$$
$$M_\epsilon = 4\begin{pmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{pmatrix} = \begin{pmatrix} 1 & \epsilon\sqrt{3} \\ \epsilon\sqrt{3} & 3 \end{pmatrix} \text{ where } \epsilon = \frac{\theta}{|\theta|}.$$

Finally, using the expressions (36), (37), (38) and (42), (43), (44), (45), and after having eliminated the matrix $R(n_1 t)$ from the equations, the horizontal variational equation (25) written in the rotation frame leads to

$$\ddot{\mathbf{u}} + 2n_1 R(\pi/2)\dot{\mathbf{u}} - n_1^2 \mathbf{u} = -\frac{\mu_0}{a_1^3}\mathbf{u} + \frac{3\mu_0}{4a_1^3}\left(M_\epsilon + \mu\left(M_{-\epsilon} - M_\epsilon\right)\right)\mathbf{u}, \tag{46}$$

with $\mu = \mu_1/\mu_0 = m_1/(m_0 + m_1)$. According to the Kepler's third law, we have $n_1^2 = \mu_0/a_1^3$, and the terms $n_1^2\mathbf{u}$ and $\mu_0/a_1^3\mathbf{u}$ vanish. This leads to the second-order differential equation in $\mathbb{R}^2$:

$$\ddot{\mathbf{u}} + 2n_1 R(\pi/2)\dot{\mathbf{u}} - \frac{3}{4}n_1^2 M_{\epsilon,\mu}\mathbf{u} = 0 \quad \text{with}$$
$$M_{\epsilon,\mu} = \begin{pmatrix} 1 & \epsilon\sqrt{3}(1-2\mu) \\ \epsilon\sqrt{3}(1-2\mu) & 3. \end{pmatrix} \tag{47}$$

In order to solve this second differential system of second order in $\mathbb{R}^2$, a classical process is to bring it back to a differential system of order one in $\mathbb{R}^4$. Introducing $\mathbf{u}_1 = \mathbf{u}$ and $\mathbf{u}_2 = \dot{\mathbf{u}}$, we get

$$\begin{aligned} \dot{\mathbf{u}}_1 &= \mathbf{u}_2 \\ \dot{\mathbf{u}}_2 = \ddot{\mathbf{u}} &= -2n_1 R(\pi/2)\mathbf{u}_2 + \frac{3}{4}n_1^2 M_{\epsilon,\mu}\mathbf{u}_1, \end{aligned} \tag{48}$$

or

$$\dot{\mathbf{U}} = \frac{d}{dt}\begin{pmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix} 0 & I \\ \frac{3}{4}n_1^2 M_{\epsilon,\mu} & -2n_1 R(\pi/2) \end{pmatrix}\begin{pmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{pmatrix} = A\begin{pmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{pmatrix} = A\mathbf{U}. \tag{49}$$

The matrix of this linear system takes the form:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \alpha & \beta & 0 & \delta \\ \beta & \gamma & -\delta & 0 \end{pmatrix}, \tag{50}$$

where

$$\alpha = \frac{3}{4}n_1^2, \quad \beta = \frac{3}{4}\epsilon\sqrt{3}(1-2\mu)n_1^2, \quad \gamma = \frac{9}{4}n_1^2, \quad , \delta = 2n_1. \tag{51}$$

Now the horizontal variational equation is reduced to a linear and homogeneous differential system of order one in $\mathbb{R}^4$. In order to study its stability (that is the linear stability of the fixed points L_4 and L_5), we will use some classical results of the theory of the linear ordinary differential equations. For details concerning this theory, the reader will refer, for example, to [2, 56]. For the sake of simplicity, we assume that the matrix A is diagonalizable[4] (eventually in $\mathbb{C}$). In this case, their exists a basis, made of eigenvectors of A, such that the expression of A in this basis (denoted by D) is diagonal. The diagonal entries of D are the eigenvalues of A. If λ is an eigenvalue of A, i.e., a root of the characteristic polynomial of A given by $\mathcal{P}_A(\lambda) = det(\lambda I - A)$, we will denote by $\mathbf{e}_\lambda$ an eigenvector of A associated to λ. Obviously, this vector satisfies the relation $A\mathbf{e}_\lambda = \lambda\mathbf{e}_\lambda$. Using a decomposition in a basis of eigenvectors, the solution of (49) takes the form: $\mathbf{U}(t) = \sum_{1\le j\le 4} u_j(t)\mathbf{e}_{\lambda_j}$. By replacing this expression in (49), we get

$$\dot{\mathbf{U}} = A\mathbf{U} = \sum_{1\le j\le 4} \dot{u}_j\mathbf{e}_{\lambda_j} = \sum_{1\le j\le 4} u_j A\mathbf{e}_{\lambda_j} = \sum_{1\le j\le 4} u_j\lambda_j\mathbf{e}_{\lambda_j}. \tag{52}$$

By uniqueness of the decomposition we obtain for all j the equation $\dot{u}_j = \lambda_j u_j$ whose solutions are given by $u_j(t) = e^{\lambda_j t}c_j$ where c_j is an arbitrary constant number. Therefore, the general solution of (49) is

$$\mathbf{U}(t) = \sum_{1\le j\le 4} c_j e^{\lambda_j t}\mathbf{e}_{\lambda_j}. \tag{53}$$

If the eigenvalues are all real numbers, their eigenvectors have real co-ordinates and $\mathbf{U}(t)$ has real coefficients as long as the c_j are real. In this case, if one of the λ_j is strictly positive, the solution $\mathbf{U}(t)$ tends to infinity when $t \mapsto +\infty$. The fixed point is unstable. In the other cases, the solution remains bounded and the equilibrium is stable.

If an eigenvalue of A is complex, the solution given by (53) is no more valid, because the coefficients of $\mathbf{U}$ become complex numbers. But this difficulty can be overcome quite easily. Indeed, the coefficients of the characteristic polynomial $\mathcal{P}_A$ being real, if λ is one of its complex roots, the conjugated quantity $\bar{\lambda}$ is a root too. The corresponding eigenvectors, whose coefficients are complex themselves, satisfy the relationship: $\mathbf{e}_{\bar{\lambda}} = \overline{\mathbf{e}_\lambda}$. If we define the two real vectors $\mathbf{f}_\lambda$ and $\mathbf{g}_\lambda$ by $\mathbf{e}_\lambda = \mathbf{f}_\lambda + i\mathbf{g}_\lambda$, a straightforward computation shows that the projection of the vector $\mathbf{U}(t)$ on the subset spanned by the complex eigenvectors $\mathbf{e}_\lambda$ and $\mathbf{e}_{\bar{\lambda}}$ can be replaced by the real quantity:

[4] This hypothesis is not always satisfied, in this case Jordan's reduction is applied.

$$e^{\alpha t}\left[(a\cos(\beta t) - b\sin(\beta t))\mathbf{f}_\lambda + (b\cos(\beta t) + a\sin(\beta t))\mathbf{g}_\lambda\right], \tag{54}$$

where a and b are constant real numbers and α and β the real and imaginary parts of λ. This expression shows that in the plane spanned by the two vectors $\mathbf{f}_\lambda$ and $\mathbf{g}_\lambda$, the trajectories spirals toward the fixed point if $\alpha < 0$ and toward infinity when $\alpha > 0$. In the special case $\alpha = 0$ (pure imaginary eigenvalues), the trajectory rotates around the fixed point with a frequency equal to β. This case, called center, is stable.

Let us now come back to the Lagrange points. According to (50), the characteristic polynomial of A is equal to

$$\mathcal{P}_A(\lambda) = \lambda^4 + (\delta^2 - \alpha - \gamma)\lambda^2 + \alpha\gamma - \beta^2 = 0. \tag{55}$$

It turns out that if λ is one of its roots, $-\lambda$ is a root too. Consequently, if, for the sake of simplicity, we eliminate the degenerated cases where zero is a double or quadruple root of $\mathcal{P}_A$, we are left with four distinct dynamical situations.

(a) Hyperbolic fixed point: Saddle×Saddle (Fig. 3a).
The eigenvalues of A are equal to $(\alpha_1, -\alpha_1, \alpha_2, -\alpha_2)$ with $\alpha_j > 0$. In the two planes generated, respectively, by $(\mathbf{e}_{\alpha_1}, \mathbf{e}_{-\alpha_1})$ and $(\mathbf{e}_{\alpha_2}, \mathbf{e}_{-\alpha_2})$, the trajectories are hyperbolas defined by the parametric representation $(ae^{\alpha_j t}, be^{-\alpha_j t})$, unless $a = 0$ or $b = 0$. Let us notice that these two conditions correspond, respectively, to the contracting and to the expending directions. Because the product of the expansion factor e^{α_j} with the contraction factor $e^{-\alpha_j}$ is equal to one, the area of a given close domain is preserved by the flow of the system. Unless the initial condition is proportional to one of the two contracting eigenvectors $\mathbf{e}_{-\alpha_j}$, the solution tends to infinity with the time t. The fixed point is consequently unstable. Figure 3a shows the dynamics on this two planes.

(b) Hyperbolic fixed point: Saddle×Center (Fig. 3b)
This situation arises when one of the eigenvalues is real and another one pure imaginary. The four eigenvalues are thus given by $(\alpha, -\alpha, i\beta, -i\beta)$ with $\alpha, \beta > 0$. In the plane generated by $(\mathbf{e}_\alpha, \mathbf{e}_{-\alpha})$ the trajectories are almost always hyperbolas, as mentioned above, while in the plane $(\mathbf{f}_{i\beta}, \mathbf{g}_{i\beta})$ the trajectories are ellipses described periodically with a frequency equal to β. In this case too, the fixed point is unstable.

(c) Elliptic fixed point: Center×Center (Fig. 3c)
When the four roots are pure imaginary: $(i\beta_1, -i\beta_2, i\beta_1, -i\beta_2)$, the equilibrium is stable. The trajectories lie on a torus of dimension two in $\mathbb{R}^4$. If the frequencies β_1 and β_2 are commensurable (their ratio is a rational number), the motion is periodic. If the two frequencies are not commensurable, the motion is quasiperiodic, and the trajectory becomes dense in the torus.

(d) Hyperbolic fixed point: (stable) Focus× (unstable) Focus (Fig. 3d)
If $\lambda = \alpha + i\beta$, with $\alpha > 0$ and $\beta > 0$, the eigenvalues are $(\alpha + i\beta, \alpha - i\beta, -\alpha + i\beta, -\alpha - i\beta)$. The planes spanned by $(\mathbf{f}_\lambda, \mathbf{g}_\lambda)$ and $(\mathbf{f}_{-\lambda}, \mathbf{g}_{-\lambda})$ are invariant by the flow of the system. In the first plane the trajectories spiral outward to infinity, while in the second one the trajectories spiral inward.

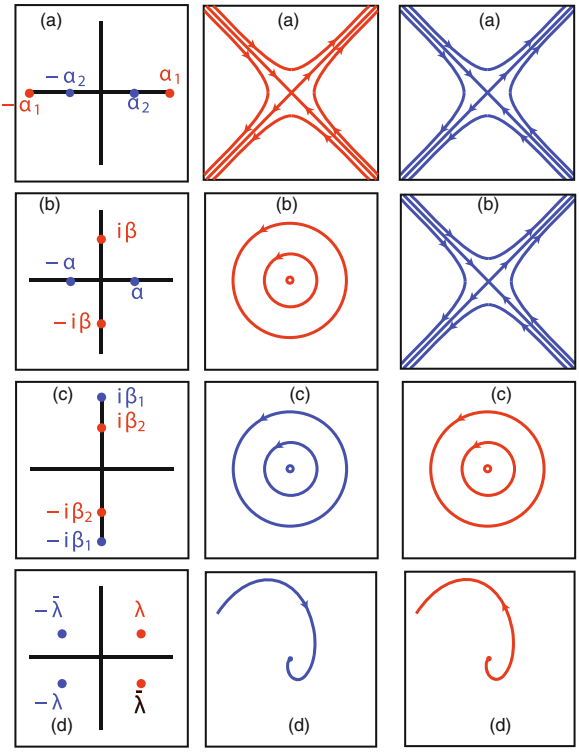

Fig. 3 Four different dynamical behaviors of the linear differential system in $\mathbb{R}^4$ associated to the matrix A. In the first column are plotted the four eigenvalues of A in the complex plan where the real axis (*horizontal line*) and the pure imaginary axis (*vertical line*) are represented. The two others columns represents the dynamics on two independent eigensubspaces. The lines labeled, respectively, by (a), (b), (c), and (d) correspond to the four situation described in the text. Among these four different kinds of dynamics, only the case (c) is stable

Unless the initial condition belongs to the contracting plane (the second one), the associated trajectory goes to infinity, and the equilibrium is unstable.

After these general considerations, let us return to the Trojans. According to (51) and (55) the characteristic polynomial of the matrix A is equal to

$$\mathcal{P}_A(\lambda) = \lambda^4 + n_1^2\lambda^2 + \frac{27}{4}\mu(1-\mu)n_1^4. \tag{56}$$

Its discriminant is given by

$$\Delta = (1 - 27\mu(1-\mu))n_1^4 = (1-a)n_1^4, \tag{57}$$

where $a = 27\mu(1-\mu)$. Consequently, the square of the roots of $\mathcal{P}_A$ satisfies

$$\lambda^2 = (-1 \pm \sqrt{1-a})\frac{n_1^2}{2} \tag{58}$$

The dynamical behavior of the equilibrium point depending on the location of its eigenvalues in the complex plane, let us first notice that $\lambda^2 \in \mathbb{R}$ if and only if $a \leq 1$, that is $\mu \in [0, \mu_c]$, where μ_c has been defined at the end of Sect. 3.2. a being, in this case always positive, we have: $\lambda^2 \leq 0$. Thus the roots of $\mathcal{P}_A$ are pure imaginary quantities given by $\lambda_j = \pm i\omega_j$ with

$$\begin{aligned} \omega_1 &= n_1\sqrt{(1-\sqrt{1-a})/2} = n_1\left(\sqrt{\frac{27}{4}\mu} + O(\sqrt{\mu})\right) \\ \omega_2 &= n_1\sqrt{(1+\sqrt{1-a})/2} = n_1\left(1 - \frac{27}{8}\mu + O(\mu)\right). \end{aligned} \tag{59}$$

We are left with an elliptic fixed point of the kind Centre×Center (see Fig. 3a) which insures the linear stability of the equilibrium. Let us notice that the two bounds of the interval of stability $[0, \mu_c]$ correspond to particular situations. When $\mu = 0$, we are in the presence of the well-known case of linearization along the Keplerian problem in the rotating frame, which is obviously degenerated. Indeed, on the opposite of the case $\mu \neq 0$ where 2 equilibrium points exist on the circle of radius a_1 centered at the most massive body, all the points of this circle are invariant. For this reason the problem is degenerated, imposing two eigenvalues to be equal to zero and to the other ones to be equal to $\pm i n_1$.

The upper bound of the interval, which limits the linear stability of the equilateral points is also of interest. Indeed, according to formulas (59),

$$\omega_1 = \omega_2 = \frac{n_1}{\sqrt{2}} \tag{60}$$

corresponds to a 1:1 resonance between these two frequencies. For $\mu = \mu_c$, a bifurcation arises and changes deeply the dynamical behavior of the fixed points. For $\mu \in]\mu_c, 1/2]$, the system becomes now unstable. In this interval, $1 - a$ is negative, and from (58), we deduce that the squares of the roots of the characteristic polynomial $\mathcal{P}_A$ become complex numbers. This corresponds to the unstable situation described in Fig. 3d. Thus, the four eigenvalues take the form $\lambda = \pm\alpha \pm i\beta$ with

$$\alpha = \frac{\sqrt{\sqrt{a}-1}}{2} n_1 \quad \text{and} \quad \beta = \frac{\sqrt{\sqrt{a}+1}}{2} n_1. \tag{61}$$

The quantity $-\alpha$ is associated to convergence speed toward the fixed point, α to the divergence speed from the fixed point, while β is the rotation frequency around the equilibrium point. Let us notice that the combination of these two motions, rotation

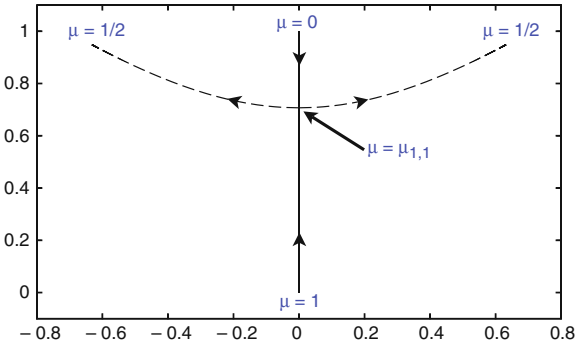

Fig. 4 Eigenvalues (divided by n_1) in the complex half-plane $\Im(z) \geq 0$. See the text for more details

plus convergence or divergence, leads the Trojan to spiral inward or outward (see Fig. 3d). Figure 4 summarizes the evolution of the eigenvalues of the triangular equilibrium points under the variation of the mass ratio μ from 0 to $1/2$. Due to the symmetry of the eigenvalues with respect to the real axis, λ and $\bar{\lambda}$ are both roots of the characteristic polynomial, only the two eigenvalues with positive imaginary part are represented. When $\mu = 0$ one of the eigenvalue starts at 0 and the other one starts at in_1; then, the first one increases (more precisely, its imaginary part), while the other one decreases until they collide for $\mu = \mu_c$. At this point, the stable behavior of the fixed points vanishes while the eigenvalues leave the imaginary axis. For $\mu > \mu_c$ these values evolve along two different branches symmetric with respect to the imaginary axis which end at

$$\pm\frac{\sqrt{\sqrt{27}-2}}{2\sqrt{2}} + i\frac{\sqrt{\sqrt{27}+2}}{2\sqrt{2}} \approx \pm 0.63208 + i\,0.94842,$$

when $\mu = 1/2$.

5.3.2 What Happens in the Elliptic Three-Body Problem

Let us assume that the motion of the planet harboring the Trojan is now elliptic. We have shown in Sect. 4 that the equilateral configurations still exist and that the elliptic elements of the test particle are the same as the elements of the planet, except the argument of the perihelion which is translated from $\pm\pi/3$. Consequently, if we use, as in the circular case, a reference frame in uniform rotation whose angular velocity is the planet mean one, the equilateral configurations will not be equilibrium points. In order to get fixed points, another way to proceed is to rewrite the equations of the motion in a rotating–pulsating reference frame related to the planet. In other words, we first chose a co-ordinate system in non-uniform rotation (the angle of rotation being the true anomaly of the planet) and rescale this system by a factor $1/\mathbf{r}_1^2(t)$. In this new co-ordinate system, the planet and the equilateral configurations are both

at rest. The derivation of the equation of the motion being straightforward but quite long, we prefer to refer the reader to classical books of celestial mechanics like [66]. The next step lies in the linearization of the differential system in the neighborhood of the fixed points $L4$ and $L5$. In contrast to the circular case, the linearized system is not autonomous, but depends periodically on the time (the period being equal to $2\pi/n_1$). Although the solutions of this kind of system cannot, in general, be expressed in a closed form, Floquet's theory (see for example [42, 56]) proves that a fundamental system of solutions $X(t)$ can be written as

$$X(t) = P(t)B,$$

where $P(t)$ is a matrix with periodic coefficients, and B a matrix which is independent of the time. As in the case of autonomous systems, the stability of the equilibrium can be deduced from the eigenvalues of the matrix B. This matrix being not explicitly known, two different approaches have been followed in order to understand the dynamical nature of the equilibria. The first one lies in numerical integrations of the linear system in order to compute the eigenvalues of B [12, 57], while the second one is based on an expansion of the system in power series of the planetary eccentricity, followed by the reduction of the equations to a normal form[44]. The results of these two approaches are presented in Fig. 5. In this figure, the X-axis is related to the mass of the two bodies by the parameter $\sigma = \sqrt{27\mu(1-\mu)}$, while the Y-axis corresponds to the eccentricity of the planet. For $e = 0$, the stability domain coincides with the one obtained in the circular problem, because $\sigma = 1$ is equivalent to $\mu = \mu_c$. For $e \neq 0$, the stability domain is split into two regions intersecting at the point of coordinates $(\sigma, e) = (\frac{\sqrt{3}}{2}, 0)$. This value of σ is reached for $\mu = (3 - 2\sqrt{2})/6 \approx 0.02859548$. On both sides of these points linear stability exist for non zero eccentricities. The right part of the stability domain shows that

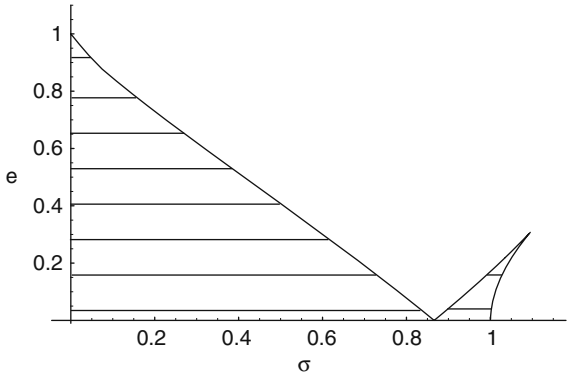

Fig. 5 Stability diagram of the equilateral solutions in the elliptic restricted three-body problem. The X-axis represents the quantity σ related to the mass ratio of the massive bodies by the relation $\sigma^2 = 27\mu(1-\mu)$. The eccentricity of the planet (of relative mass μ) is represented by the Y-axis. The fixed point is linearly stable in the hatched regions, while it is unstable outside

eccentricity of the orbits of the primaries have a stabilizing effect on the linearized problem, since it can be stable for $\mu > \mu_c$. Indeed, the coordinates of the extremal point of the right part of the stability domain are $(\mu, e) \approx (0.04698, 0.3143)$.

5.4 Beyond the Linear Stability

The linear stability of a fixed point is not a property that persists under perturbation. Practically, the stability properties of the equilateral solutions of the restricted three-body problem discussed above may disappear considering higher order expansion of the equation of the motion around the equilibrium point. When we consider the linearized system around an elliptic fixed point, every initial condition leads to a quasiperiodic trajectory. Consequently, a solution starting near the fixed point will always stay close to it. But this strong stability property does not, in general, persist when expansion of higher degrees is considered (even in Hamiltonian systems). Indeed, perturbations of the linear system may generate unstable directions along which the massless body escapes; and even if it is not the case, a lot of quasiperiodic trajectories mentioned above may not subsist, giving rise to diffusion phenomena. In Hamiltonian systems,[5] the behavior of quasiperiodic trajectories is at the core of KAM theory (see [3, 4, 52]). To give a very rough idea of this theory, we can say that starting with an integrable Hamiltonian system whose phase space is the union of invariant tori filled with quasiperiodic trajectories, under suitable conditions, a large amount of tori is preserved under a sufficiently small Hamiltonian perturbation. For Hamiltonian systems of 2 degrees of freedom, this theory allows to prove the stability in numerous situations. Indeed, for a Hamiltonian system of n degrees of freedom, the dimension of the phase space is $2n$. Due to the invariance of the Hamiltonian, a trajectory evolves on a $(2n-1)$-dimensional space, and therefore, the codimension of a torus is $n-1$. Consequently, in a 2-degrees of freedom Hamiltonian system, a KAM torus is a surface of codimension 1 which divides the phase space in two disconnected parts. In this case, the existence of KAM tori imposes a property of confinement which leads to stability for infinite time. The circular and planar restricted three-body problem corresponding to a 2-degrees of freedom Hamiltonian system, KAM theory can be applied to prove the stability in some particular domains of its phase space. This is the reason why KAM theory has been wildly employed in the aim to prove the stability of the equilateral solutions of the circular restricted three-body problem. According to [34], a direct application of KAM theory proves that invariant tori exist in the neighborhood of the points $L_{4,5}$ for almost all μ in the interval $[0, \mu_c]$. More recent studies [16, 43] show that the equilateral solutions are stable for all μ in $[0, \mu_c]$ except for two values: $\mu = (1 - \sqrt{213}/15)/2 \approx 0.013516$ and $\mu = (1 - \sqrt{1833}/45)/2 \approx 0.024294$, for which instabilities take place [39]. Regarding the spatial three-body problem or

[5] The restricted three-body problem, like the general n-body, can be written in Hamiltonian form (see [52]).

the planar and elliptic three-body problem, KAM theory does not allow to bound the trajectories. Indeed, for more than 2 degrees of freedom, the tori which are of codimension three or more do not separate the phase space anymore, leading to Arnold's diffusion phenomena [1]. Nevertheless, Nekhoroshev theorem makes possible to bound the diffusion on a finite, but exponentially long time [54]. In [6], the authors show that L_4 and L_5 are stable in the sense of Nekhoroshev for all but a few values of μ up to the Gascheau value μ_c. But this result, like the direct applications of KAM theory, does not give any information concerning the size of the stable neighborhood. Nevertheless, Nekhoroshev theorem leads to the concept of "effective stability" of the considered differential system, the idea being the following: a system is effectively stable if the time needed to observe significant changes is longer than the expected lifetime of the system itself (see [22]). This idea is particularly efficient in the case of Jovian Trojans. Indeed, numerous Trojans orbiting near L_4 or L_5 seem to be stable for billion years. This was first applied in [27], where the authors show, using Nekhoroshev-like estimations, that Jovian Trojans cannot escape from a bowl of radius R centered at $L_{4,5}$ before a period of time comparable to the age of the Solar System. This nice result, which is developed in the framework of the circular and planar restricted three-body problem is unfortunately valid only for R lower that 10 km (the closest observed Jovian Trojans orbit at more that 10^6 km from the Lagrangian points). By significant improvements of the estimates employed in [27] several authors [10, 28, 65] obtain stability radius big enough to include a few known Trojan asteroids. More sophisticated models and estimates of the stability radius have been developed recently [19, 22–24, 36], but in spite of all these efforts, this kind of methods seem far from being applicable to realistic models. Indeed, even if the elliptic and spatial restricted three-body problem provides a reasonable approximation of the motion, a lot of important dynamical phenomena arise when the gravitational perturbations of the four giant planets of the Solar system are taken into account (the phenomena already appear considering the asteroid–Sun–Jupiter–Saturn model [25, 60]). But these models possessing at least highest degrees of freedom are beyond the reach of the methods based on Nekhoroshev-like estimates.

6 Further Reading

This last section gathers some results dealing with the dynamics of Trojans in the solar system. It does not pretend to give a complete review of the subject and will not furnish an exhaustive bibliography. Its purpose is only to give some tracks for the interested reader. We have already mentioned in the previous section that the restricted three-body problem does not give a good approximation of the real trajectories of the Trojans of Jupiter and that the gravitational influence of the other giant planets, at least Saturn, must be taken into account. An important pioneer work was done by Erdi (see [20] and references therein), in the aim of increasing the precision

of the calculations of the Trojan orbits including the perturbations of the other planets. The most precise models are used at the present time for the calculation of the proper elements of the Jupiter Trojans. These quantities, which can be viewed as approximated integrals of the motion, i.e., as quantities varying in a very small and slow manner characterizing a given celestial object, have been introduced by J. G. Williams in 1969 [69]. In the case of an asteroid whose the motion is regular enough (in particular in the case of the lack of specific resonances), these proper elements are perfectly defined (although there are several ways to define them) and very stable in function of the time. On the opposite, for less regular orbits, the temporal variations of the proper elements can give a measurement of the irregularity of the motion (i.e. the orbital diffusion). One of the principal applications of the calculations of proper elements is the determination of the dynamical asteroid families. Asteroid families are clustering in the proper elements space, which are the result of the catastrophic disruption of a parent body after collision.

The calculation of the proper elements developed in the frame of the dynamical study of the asteroid belt (see [49]) has been applied to the Jupiter Trojans [5, 47, 48, 64] and has enabled to prove the existence of Trojan dynamical families (see [61]) as in the case of numerous ones in the asteroid belt (see Cellino in this volume). From the calculations of their proper elements, A. Milani [47] showed that some Trojans are subject to unstable behaviors, ranging from a bounded diffusion to the ejection from the swarm. In [35] the authors showed, starting from a numerical integration of the Jupiter Trojans for one billion years, that the swarms were not permanently stable. Some of the objects could escape from the swarm, their lifetime inside it depending on the distance to the Lagrange point. This phenomenon of slow erosion was confirmed a posteriori by [45, 55] which showed that unstable structures exist inside the swarms themselves. These unstable structures were studied exhaustively in [60, 59] where the authors showed that several Trojans known as unstable ones were evolving inside resonances. They also proved that diffusion phenomena along resonances could lead some objects orbiting deep inside the swarms to be ejected in a timescale of the order of 1 billion years. This phenomenon is related to the slow erosion suggested by [35].

Another important question concerns the high inclinations of some Trojans with respect to Jupiter's orbital plane. In some cases this inclination exceeds 40° as was shown in Fig. 1. Indeed, it is very unlikely that a population having a small initial inclination could reach more than 25° of inclination (see[40, 59]). Consequently, in order to reproduce the range of inclinations reached by the Trojans in the present Solar System, it seems necessary to have, in the initial population, Trojans with high inclinations. Moreover, if the dynamical mechanisms inside the Trojan swarms seem to be well understood now, except a few specific points, the question of the dissymmetry which seems to exist between the populations around the L_4 point and the L_5 one, as shown in Fig. 2, is not entirely solved. On this topic, an observational bias is no more considered as a suitable explanation. Moreover the cause of the dissymmetry does not seem to be linked to a dynamical effect as this was suggested by [18, 55, 60] for the Jupiter Trojans and by [17] for Neptune Trojans. The origin

of the dissymmetry might be found somewhere else, for instance in the formation of swarms inside a Solar System initially very dense and dominated by collisions between small bodies (see [13]).

If we have only very few clues about this last point, an effective scenario developed by Gomes, Levison, Morbidelli, and Tsiganis in 2005 seems to give a reasonable explanation to the formation and inclination problems. Indeed, it is shown for the first time in [29, 53, 67] that the planetary migration is compatible with the hypothesis that the Jupiter's Trojans are captured just after the crossing of the 1:2 mean motion resonance between Jupiter and Saturn. Moreover, these numerical simulations give a distribution of the Trojans inclination that agrees with the observed one.

In 1989, although no Trojan related to another planet than Jupiter was discovered, Innanen and Mikkola [31] presented a study of the dynamics of hypothetical Trojan swarms associated to the four giant planets using a numerical integration over 10 million years. They confirmed the existence of large regions of stability in the neighborhood of the Lagrange points L_4 and L_5 of Jupiter and showed that the same property held for Uranus and Neptune. On the opposite, their study showed the existence of strong instabilities around the equilateral configurations of Saturn.

These instabilities were confirmed and investigated more deeply by M. J. Holman and J. Wisdom [30]. The interesting dynamical situation of Saturn Trojans was detailed in other papers as [14, 40, 41]. According to these authors, the combination of the perturbations generated simultaneously by secular resonances and the quasi-resonance 2:5 between Jupiter and Saturn (called the "great inequality") leads to the instability as it is observed. The dynamical behavior of the Uranus and Neptune Trojans was also studied by Nesvorny and Dones [55], who predicted in which region of the sky they should be located in the case they really exist. It is inside such a region that the first Neptune Trojan, 2001QR322 was discovered. Some complementary studies can be found on the subject (see for instance [41, 32]).

Let us finish this section by some ideas about the Trojans of the tellurian planets. Numerous studies were devoted on their existence. We can refer for instance to [8, 9, 11, 21, 46] or [63, 62] for specific studies about Venus or Mars Trojans.

Except the planet Mars for which the four co-orbital objects discovered recently have a stable motion for a very long time, the possibility to discover stable Trojan swarms is fairly small. Indeed, according to the studies mentioned above, these regions are strongly unstable and do not harbor long-life Trojans. More precisely, according to [50, 51] the lifetime inside these regions should not exceed several millions years. But on the other hand, temporary populations of Trojans supplied by flux of asteroids visiting the inner Solar System might exist. But the conditions of observation of these potential bodies are hardly determined, in particular because of their small elongation with respect to the Sun.

Acknowledgments We are deeply grateful to Daniel Suchet and Rachelle Holman for having carefully reviewed the chapter. We are indebted to Alain Albouy for very interesting discussions concerning the historical aspect.

References

1. Arnold, V.I.: Instability of dynamical systems with several degrees of freedom. Sov. Math. Dokl. **5**, 581–585, (1964) 222
2. Arnold, V.I.: Ordinary Differential Equations. Springer-Verlag, New York (1992) 215
3. Arnold, V.I., Kozlov, V.V., Neistadt, A.I.: Mathematical Aspects of Classical and Celestial Mechanics. Encyclopaedia of Mathematical Sciences. Springer-Verlag, New York (2006) 221
4. Arnold, V.I.: Proof of A.N. Kolmogorov's theorem on the preservation of quasiperiodic motions under small perturbations of the Hamiltonian. Russ. Math. Surv. **18**(5), 9–36 (1963) 221
5. Beaugé, C., Roig, F.: A semianalytical model for the motion of the Trojan asteroids: Proper elements and families. Icarus **153**, 391–415 (2001) 223
6. Benettin, G., Fasso, F., Guzzo, M.: Nekhoroshev-stability of l4 and l5 in the spatial restricted three-body problem. Regul. Chaotic Dyn. **3**(3), 56–72 (1998) 222
7. Bennett, T.L.: On the reduction of the problem of n bodies. Messenger Math. **XXXIV**, 113–120 (1905) 203
8. Brasser, R., Innanen, K.A., Connors, M., Veillet, C., Wiegert, P., Mikkola, S., Chodas, P.W.: Transient co-orbital asteroids. Icarus **171**, 102–109 (September 2004) 224
9. Brasser, R., Lehto, H.J.: The role of secular resonances on Trojans of the terrestrial planets. MNRAS **334**, 241–247 (July 2002) 224
10. Celletti, A., Giorgilli, A.: On the stability of the Lagrangian points in the spatial restricted problem of three bodies. Celest. Mech. Dyn. Astron. **50**, 31–58 (1991) 222
11. Christou, A.A.: A numerical survey of transient co-orbitals of the terrestrial planets. Icarus **144**, 1–20 (March 2000) 224
12. Danby, J.M.A. Stability of the triangular points in the elliptic restricted problem of three bodies. Astron. Astrophys. **69**, 165 (1964) 220
13. de Elía, G.C., Brunini, A.: Collisional and dynamical evolution of the $L4$ Trojan asteroids. Astron. Astrophys. **475**, 375–389 (November 2007) 224
14. de La Barre, C.M., Kaula, W.M., Varadi, F.: A study of orbits near Saturn's triangular Lagrangian points. Icarus **121**, 88–113 (May 1996) 224
15. Deprit, A.: Elimination of the nodes in problems of N bodies. Celest. Mech. Dyn. Astron. **30**, 181–195, (June 1983) 203
16. Deprit, A., Deprit-Bartholome, A. Stability of the triangular Lagrangian points. Astron. J. **72**, 173–179, (March 1967) 221
17. Dvorak, R., Lhotka, Ch., Schwartz, R.: The dynamics of inclined Neptune Trojans. Celest. Mech. Dyn. Astron. **102**(1–3), 97–110 (2008) 223
18. Dvorak R., Schwarz, R.: On the stability regions of the Trojan asteroids. Celest. Mech. Dyn. Astron. **92**, 19–28 (2005) 223
19. Efthymiopoulos, C., Sándor, Z.: Optimized Nekhoroshev stability estimates for the Trojan asteroids with a symplectic mapping model of co-orbital motion. MNRAS **364**, 253–271 (November 2005) 222
20. Érdi, B.: The Trojan problem. Celest. Mech. Dyn. Astron. **65**, 149–167 (1997) 222
21. Evans, N.W., Tabachnik, S.A.: Asteroids in the inner solar system—II. Observable properties. MNRAS **319**, 80–94 (November 2000) 224
22. Gabern, F.: *On the dynamics of the Trojan asteroids*. PhD thesis, Departament de Matemàtica Aplicada i Anàlisi Universitat de Barcelona (2003) 222
23. Gabern, F., Jorba, À.: Generalizing the restricted three-body problem. The Bianular and Tricircular coherent problems. Astron. Astrophys. **420**, 751–762 (June 2004) 222
24. Gabern, F., Jorba, A., Locatelli, U.: On the construction of the Kolmogorov normal form for the Trojan asteroids. Nonlinearity **18**, 1705–1734 (2005) 222
25. Gabern, F., Jorba, A., Robutel, P.: On the accuracy of restricted three-body models for the Trojan motion. Discrete Contin. Dyn. Syst. **11**(4), 843–854 (2004) 222
26. Gascheau, G.: Examen d'une classe d'équations différentielles et application à un cas particulier du problème des trois corps. Compt. Rend. **16**(7), 393–394 (1843) 198

27. Giorgilli, A., Delshams, A., Fontich, E., Galgani, L., Simó, C.: Effective stability for a Hamiltonian system near an elliptic equilibrium point, with an application to the restricted three body problem. J. Diff. Eqns. **77**(1), 167–198 (1989) 222
28. Giorgilli, A., Skokos, C.: On the stability of the Trojan asteroids. Astron. Astrophys. **317**, 254–261 (January 1997) 222
29. Gomes, R.S., Levison, H.F., Tsiganis, K., Morbidelli, A.: Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Nature **435**, 466–469 (May 2005) 224
30. Holman, M.J., Wisdom, J.: Dynamical stability in the outer solar system and the delivery of short period comets. Astron. J. **105**, 1987–1999 (May 1993) 224
31. Innanen, K.A., Mikkola, S.: Studies on solar system dynamics. I—The stability of Saturnian Trojans. Astron. J. **97**, 900–908 (March 1989) 224
32. Kortenkamp, S.J., Malhotra, R., Michtchenko, T.: Survival of Trojan-type companions of Neptune during primordial planet migration. Icarus **167**, 347–359 (February 2004) 224
33. Lagerkvist, C.-I., Karlsson, O., Hahn, G., Mottola, S., Doppler, A., Gnädig, A., Carsenty, U.: The Uppsala-DLR Trojan Survey of L 4, the preceding Lagrangian cloud of Jupiter. Astronomische Nachrichten **323**, 475–483 (July 2002) 201
34. Leontovitch, A.M.: On the stability of the Lagrange's periodic solutions of the restricted three body problem. Dolk. Akad. Nouk USSR **43**, 525–528 (1962) 221
35. Levison, H.F., Shoemaker, E.M., Shoemaker, C.S.: The long-term dynamical stability of Jupiter's Trojan asteroids. Nature **385**, 42–44 (1997) 223
36. Lhotka, C., Efthymiopoulos, C., Dvorak, R.: Nekhoroshev stability at $L4$ or $L5$ in the elliptic-restricted three-body problem—Application to Trojan asteroids. MNRAS 384, 1165–1177 (March 2008) 222
37. Lissauer, J.J., Chambers, J.E.: Solar and planetary destabilization of the Earth Moon triangular Lagrangian points. Icarus **195**, 16–27 (May 2008) 201
38. Malige, F., Robutel, P., Laskar, J.: Partial reduction in the n-body planetary problem using the angular momentum integral. Celest. Mech. Dyn. Astron. **84**(3), 283–316 (2002) 203
39. Markeev, A.P.: Stability of the triangular Lagrangian solutions of the restricted three-body problem in the three-dimensional circular case. Sov. Astron. **15**, 682–668 (February 1972) 221
40. Marzari, F., Scholl, H.: The role of secular resonances in the history of Trojans. Icarus **146**, 232–239 (July 2000) 223, 224
41. Marzari, F., Tricarico, P., Scholl, H.: The MATROS project: Stability of Uranus and Neptune Trojans. The case of 2001 qr322. Astron. Astrophys. **410**, 725–734 (2003) 224
42. Meyer, K.R., Hall, G.R.: Introduction to Hamiltonian Dynamical Systems and the n-Body Problem. Springer-Verlag, New York (1992) 220
43. Meyer, K.R., Schmidt, D.S.: The stability of the Lagrange triangular point and a theorem of Arnold. J. Diff. Eqns. **62**(2), 222–236 (1986) 221
44. Meyer, K.R., Schmidt, D.S.: Elliptic relative equilibria in the n-body problem. J. Diff. Eqns. **214**, 256–298 (2005) 220
45. Michtchenko, T., Beaugé, C., Roig, F.: Planetary migration and the effects of mean motion resonances on Jupiter's Trojan asteroids. Astron. J. **122**, 3485–3491 (2001) 223
46. Mikkola, S., Innanen, K.: A numerical exploration of the evolution of Trojan type asteroidal orbits. Astron. J. **104**, 1641–1649 (October 1992) 224
47. Milani, A.: The Trojan asteroid belt: Proper elements, stability, chaos and families. Celest. Mech. Dyn. Astron. **57**, 59–94 (1993) 223
48. Milani, A.: The dynamics of the Trojan asteroids. In IAU Symp. 160: Asteroids, Comets, Meteors 1993, vol. 160, pp. 159–174 (1994) 223
49. Milani, A., Knežević, Z.: Secular perturbation theory and computation of asteroid proper elements. Celest. Mech. Dyn. Astron. **49**, 347–411 (1990) 223
50. Morais, M.H.M., Morbidelli, A.: The population of near-earth asteroids in coorbital motion with the earth. Icarus **160**, 1–9 (November 2002) 224
51. Morais, M.H.M., Morbidelli, A.: The population of near earth asteroids in coorbital motion with Venus. Icarus **185**(1), 29–38 (November 2006) 224

52. Morbidelli, A.: Modern Celestial Mechanics: Aspects of Solar System Dynamics. Taylor & Francis, London, ISBN 0415279399 (2002) 221
53. Morbidelli, A., Levison, H.F., Tsiganis, K., Gomes, R.S.: Chaotic capture of Jupiter's Trojan asteroids in the early Solar System. Nature **435**, 462–465 (May 2005) 224
54. Nekhoroshev, N.N.: An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. Russ. Math. Surv. **32**(6), 1–65 (1977) 222
55. Nesvorny, D., Dones, L.: How long-live are the hypothetical Trojan populations of Saturn, Uranus, and Neptune? Icarus **160**, 271–288 (2002) 223, 224
56. Perko, L.: Differential Equations and Dynamical Systems. Texts in Applied Mathematics. Springer-Verlag, New York (1991) 215, 220
57. Roberts, G.: Linear stability of the elliptic Lagrangian triangle solutions in the three-body problem. J. Diff. Eqns. **182**, 191–218 (2002) 220
58. Robutel, P.: Stability of the planetary three-body problem II: Kam theory and existence of quasiperiodic motions. Celest. Mech. Dyn. Astron. **62**, 219–261 (1995) 203
59. Robutel, P., Gabern, F.: The resonant structure of Jupiter's Trojan asteroids I: Long-term stability and diffusion. MNRAS **372**, 1463–1482 (2006) 223
60. Robutel, P., Gabern, F., Jorba, A.: The observed Trojans and the global dynamics around the Lagrangian points of the Sun–Jupiter system. Celest. Mech. Dyn. Astron. **92**, 53–69 (April 2005) 222, 223
61. Roig, F., Ribeiro, A.O., Gil-Hutton, R.: Taxonomy of asteroid families among the Jupiter Trojans: comparison between spectroscopic data and the Sloan Digital Sky Survey colors. Astron. Astrophys. **483**, 911–931 (June 2008) 223
62. Scholl, H., Marzari, F., Tricarico, P.: Dynamics of Mars Trojans. Icarus **175**, 397–408 (June 2005) 224
63. Scholl, H., Marzari, F., Tricarico, P.: The instability of Venus Trojans. Astron. J. **130**, 2912–2915 (December 2005) 224
64. Schubart, J., Bien, R.: On the computation of characteristic orbital elements for the Trojan group of asteroids. In Asteroids, Comets, Meteors II, pp. 153–156 (1986) 223
65. Skokos, C., Dokoumetzidis, A.: Effective stability of the Trojan asteroids. Astron. Astrophys. **367**, 729–736 (February 2001) 222
66. Szebehely, V.G.: Theory of Orbits: The Restricted Problem of Three Bodies. Academic Press, New-York (1967) 220
67. Tsiganis, K., Gomes, R.S., Morbidelli, A., Levison, H.F.: Origin of the orbital architecture of the giant planets of the Solar System. Nature **435**, 459–461 (May 2005) 224
68. van Houten, C.J., van Houten-Groeneveld, A., Gehrels, T.: Minor planets and related objects. V. The density of Trojans near the preceding Lagrangian point. Astron. J. **75**, 659–662 (June 1970) 201
69. Williams, J.G.: Secular Perturbations in the Solar System. PhD thesis, AA, University of California, Los Angeles (1969) 223

The Physics of Asteroids and Their Junction with Dynamics

M. Birlan and A. Nedelcu

Abstract The study of asteroid families is an important current topic. New insights obtained by dynamical considerations motivating the interest in completing the knowledge of the physics and the composition inside asteroid families. The discovery of young families offers a new frame of study of interaction between the family members, the family and the other solar system bodies, the consequence of a "hostile" medium for asteroid surface, the importance of cumulative, and long-term, non-gravitational effects. The last decade has shown that long-term dynamics of family objects can be explained by accounting for new physical effects such as Yarkovsky and Yarkovsky–O'Keefe–Radzievskii–Paddack effects. A review of these topic reveals the complexity and the importance of interdisciplinary research on these bodies.

1 Introduction

The asteroids are a population of objects in the solar system containing more than 404,923 objects.[1] Due to their large number and the location in their inner planetary system, this population represents a laboratory of study for celestial mechanics problems such as the dynamics of orbits, stability, chaos, and the long-term evolution of orbits.

New remote-sensing capabilities have opened the early history of individual asteroids and their parent bodies to sophisticated investigation. Based on the small size of the planetesimals and on meteorite chronologies, it is known that all significant chemical processes that affected these minor planets were essentially complete

M. Birlan (✉)
Institut de Mécanique Céleste et de Calculs des Éphémérides, 77 av Denfert-Rochereau, 75014 Paris Cedex, France, Mirel.Birlan@imcce.fr

A. Nedelcu
Astronomical Institute of the Romanian Academy, str Cu titul de Argint nr 5, Bucharest 4, Romania; Institut de Mécanique Céleste et de Calculs des Éphémérides, 77 av Denfert-Rochereau, 75014 Paris Cedex, France, nedelcu@astro.aira.ro

[1] Number of known objects of April 16, 2008, following ftp://ftp.lowell.edu/pub/elgb/astorb.html

Birlan, M., Nedelcu, A.: *The Physics of Asteroids and Their Junction with Dynamics*. Lect. Notes Phys. **790**, 229–250 (2010)
DOI 10.1007/978-3-642-04458-8_5

within the first 0.5% of solar system history. Asteroids represent the sole surviving in situ population of early inner solar system planetesimals, bodies from which the terrestrial planets subsequently accreted. Thus, one of the central questions of current asteroid physical studies concerns the geologic issues related to the original compositions of asteroidal parent bodies and the chemical and thermal processes that altered the original planetesimals [35].

The asteroid families are widely believed to be produced by large collisions over the solar system history. A short definition of the *syntagma* "asteroid family" is a cluster of objects which are genetically and dynamically linked, a result of a catastrophic event (collision of two bodies followed by the destruction of both target and impactor). As a corollary, clusters that are recognizable only because they occupy peculiar zones in the orbital elements space, which are isolated by the presence of secular and mean motion resonances, should not be termed as "family" [99].

Historically, the existence of families of asteroids was suggested by Hyraiama in 1918, who noticed "condensation here and there" in the distribution of the asteroids with respect their orbital elements, in particularly the mean motion n, the eccentricity e, and the inclination i [41]. Depicted by Hirayama as "curiously," "still curiously", or "remarkable coincidence," the distributions of some asteroids around the same value of orbital elements outline the major families of 158 Koronis, 221 Eos, and 24 Themis.

Asteroid families research has become a hot topic in the last decade [92]. This increasing interest will be developed in the following sections. The second section will treat briefly the dynamical aspects linked to families (both old and young ones). The overview of new, interesting physical aspects will be developed in the third section. The scientific aspects linked to the young families of asteroids will be developed in the fourth section of this article. Finally, some ideas and directions of research are proposed as conclusions.

2 Dynamical Considerations

2.1 Identification of Asteroid Families: Choice of Orbital Elements

The time variation of osculating elements of asteroids is due to the presence of several gravitational fields (major planets and other minor bodies). Jupiter's gravitational field is the most important for the evolution of the osculating elements of main-belt asteroids, but contributions of other planet gravitational fields should be also taken into account. This variation of osculating elements makes them inappropriate for the purposes of asteroid family identification [99].

The research of time-invariant orbital elements was developed in numerous articles [14, 1, 95, 22, 95, 46]. The appropriate semantics accepted by the researchers is "orbital proper elements" and designates the quasi-integrals of motion which are nearly constant in time [47].

Depending on the scientific approach, the computation of proper elements is based on analytical, semi-analytical, or synthetic methods [61–63, 52, 53, 47]. Knežević et al. [47] emphasized some particular cases (asteroids near resonances, Hildas, Trojans) of asteroids with proper elements on which different approaches could give different results. They also point out the importance of the accuracy, reliability, and the time-interval stability of the proper elements of minor bodies.

The importance of high-accuracy proper elements is crucial in the age determination of young families [73, 92]. These articles (along with others [71, 74, 21] use numerical integration in order to find the moment of nearly identical orbital elements of the family. However, orbital element integration backward in time, used as a direct method of age determination, is limited to families younger than 10 Myr [92]. Orbital element convergence of the family members becomes more precise when cumulative, non-gravitational, long-term forces [74] are taken into account.

2.2 Dynamics of the Families and the Dust Bands

The main belt is most probably a population dynamically relaxed which contained at least several times as much mass during planet formation as it does nowadays. Numerical simulations suggest that the asteroid belt was excited and depleted before the terrestrial planets completed their growth process [79]. In the assumption of giant planet migration proposed as the cause of the Late Heavy Bombardment (LHB), the asteroid belt was strongly perturbed [55]. The LHB largely erases the traces of the original distribution of objects in the region between Mars and Jupiter [65, 37]. Catastrophic collisions followed by the competition of superimposed gravitational influences of the Sun and planets "sculpted" the actual dynamical distribution of the main belt.

Main-belt collisions followed by disruption can liberate a wide range of fragments from micrometers to tens of kilometers. Two processes will differentiate the asteroid-sized fragments from the micrometer-sized ones. The large fragments will gravitationally evolve. Depending on their relative velocities, large fragments also will be spread somewhat in the interplanetary space. For low relative velocity fragments, the process of reaccretion(coalescence) could play an important role. Most of the asteroid-sized fragments which remain near the location of the parent body are identified nowadays as asteroid family members. Both family members and the small-size particle (sometimes defined as by-product of the collision) could give important insights on the main-belt evolution.

Numerical simulations of the collisional disruption of large asteroids were performed using sophisticated 3D codes [57, 56, 58, 60] and the gravitational interaction and evolution of the resulted fragments were traced. One of the simulation objectives was to deduce the formation process of big families such are Eunomia, Koronis, or Flora. Some major conclusions are drawn from the simulations, such as (i) all large family members must contain gravitationally re-accumulated fragments, (ii) the family distribution is composed of a large body and the rest of members follows a quasi-linear size–frequency distribution (SFD). New studies [29] suggest

that about 20 observed main-belt families are produced over the age of the solar system by catastrophic collisions of parent bodies larger than 100 km.

The small fragments produced by a disruptive collision have a different evolution. Models of interplanetary dust based on a variety of dynamical and physical processes (planetary perturbations, Poynting–Robertson drag, radiation pressure, electromagnetic forces, mutual collisions, sublimation, etc.) are used to explain the presence and evolution of fragments through a dust band. The origin of zodiacal dust bands[2] was related to the Eos, Koronis, and Themis large families of asteroids [27] and to newly identified young families of Karin and Veritas [73]. New studies of the evolution of dust trails into bands in the main-belt region suggest other very young families (e.g., Datura) reside at the origin of other zodiacal dust bands [94]. The budget of dust particles seems to be favorable to the younger families rather than the older ones. Morbidelli et al. [66] conclude there is collisional equilibrium for objects with diameter lower than 5 km inside the families of Eos, Themis, and Koronis, which limit their ability to produce dust particles.

The actual science of asteroid families is very well synthesized into the paradigm of Cellino et al. [23] in the frame of Yarkovsky and YORP non-gravitational forces. The post-Yarkovsky[3] paradigm allows plausible/reliable explanation for open problems such as [23]: (i) a better agreement for the size distribution objects in the main belt with some observational data; (ii) a better agreement between the observed structures of families and the hydrocode simulations; (iii) a natural explanation for the confinement of large families between powerful mean motion resonances.

Some asteroid families also contribute to the current population of near-Earth asteroids (NEAs). Large families as Themis and Eos were strongly depleted by the mean motion resonances 9:4 and 2:1 [64]. Numerical integrations [36] show that some objects injected in these resonances later achieve near-Earth-like orbits in only few million years. This is an indication that the NEA population and the impact rate to the terrestrial planets are related with collisional events in the main belt. From the analysis of terrestrial craters it was found a twofold increase of the impact flux from kilometer-sized bodies over the last 100 Myr. This apparent surge was produced by a catastrophic, family-forming impact in the inner region of the main belt 160 Myr ago [13]. The breakup of a 170 km parent body produced the current Baptistina family. This event was most likely the source of Chicxulub impactor that produced the Cretaceous/Tertiary (K/T) mass extinction event 65 Myr ago.

---

[2] The zodiacal dust bands were discovered by the satellite IRAS and could be defined as extended regions with strong emissions in the infrared region, slightly inclined to the ecliptic. The particles have a toroidal distribution located between Mars and Jupiter, but the ratio between the zodiacal dust produced by comets and that produced by asteroid collisions is not known.

[3] Yarkovsky effect is a thermal effect consisting of the absorption of solar radiation by a body and its subsequent anisotropic thermal reemission. The temperature differences on the surface, together with an irregular shape, produce a force and a torque. The strength of the reradiation force varies along the orbit as a result of thermal inertia. We can distinguish between a seasonal effect and a diurnal one. In the literature this thermal effect can be referenced as Yarkovsky and/or Yarkovsky/YORP [12].

Similar studies [78] were developed as a possible explanation of delivering meteorites from the ν_6 secular resonance to Earth-crossing orbits, and the authors underline the possibility of such "express delivery" scenario for explaining an L-chondrite meteorite falling in Sweden ≈470 Myr ago.

Investigation of the important link between the main belt and NEA populations requires a combined knowledge of their dynamical and physical evolutions and properties.

3 Physical Considerations

3.1 Physical Properties Inside a Family

In the case of asteroids, the physical properties of these bodies globally follow dynamical ones. Observations of physical properties were also enlarged to other wavelength regions, other than the visible region, from the ultraviolet to the infrared and the radio.

Two aspects will be developed during this subsection: insights in the asteroid population obtained by spectroscopy and constraints imposed by the spin of members of families of asteroids.

The visible spectroscopy of asteroids has become a dominant method of physical investigation during the last decade of the twentieth century. Results of large spectroscopic surveys [97, 20, 50], as well as spectral data of large families members, were published [28, 31, 32, 51].

Based on the visible spectroscopy, the family members share, globally, the same spectral behaviors [28, 31, 32, 51]. However, the articles that treat large families must deal with the spectra spanning a certain range of slopes, must speculate on the presence of interlopers in the observed sample, and must extrapolate the results obtained for a few dozen objects to the entire family.[4] Figures 1 and 2 present the visible spectra of families Flora and Eunomia available from the SMASSII [20] database,[5] with respect to the families determined by Zappala et al. [100].

One of the key parameters for statistical studies of the spectra of a family is the spectral slope, usually obtained using data from the spectral region 0.50–0.75 μm. The family members span a wide range of this spectral slope for each major family (e.g., Eos, Flora, Eunomia). Several scenarios were proposed in support of such variety, starting with a partially differentiated genitor of the family and finishing with subtle mechanisms of different surface alteration of family members by cosmic rays, solar wind, irradiation processes, etc., for which the generic term is *space weathering*.

---

[4] The families of Eos, Koronis, and Flora contain more than 300 objects each, thus these statistics concern roughly less than 10% of the largest bodies inside the family.

[5] Data are available online at http://smass.mit.edu

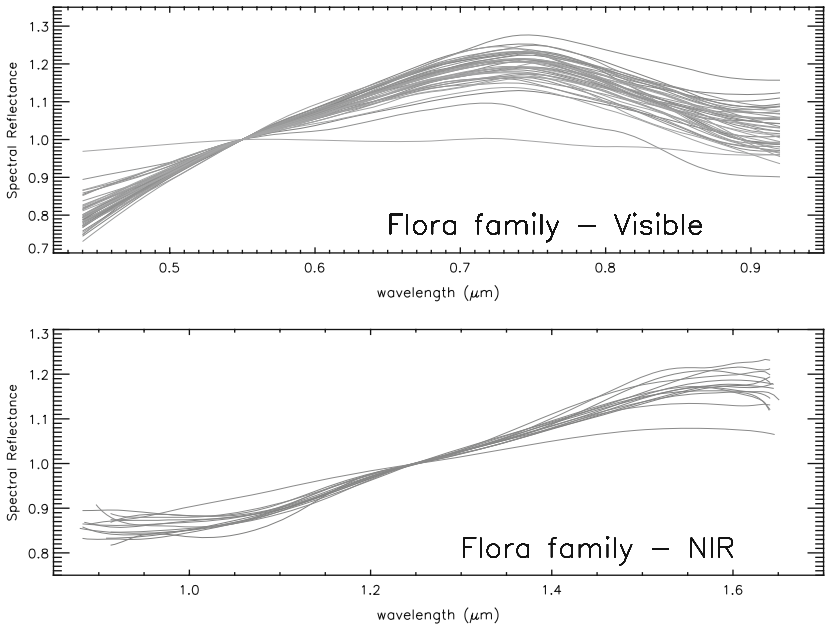

Fig. 1 Smoothed visible spectra of 45 asteroids belonging to Flora family extracted from SMASSII [20]. Two objects (asteroids 1324 and 2952) present distinct spectra. These two objects were not observed in S2OS3 [50]. These spectra were completed with spectra of 14 objects from SMASS-IR, obtained in the spectral region of 0.9–1.6 μm [18]. The spectra are normalized to 0.55 and 1.25 μm, respectively, and the family members were selected from Zappala et al. [100]. With some exceptions, these spectra reveal good similitude

Each spectroscopic study of asteroid families must also deal with the interlopers problem. For example, several authors identified objects whose spectra are quite different for the majority of the observed family members [28, 31, 32, 49]. Sometimes the percentage of objects with different spectra can reach 10% of the observed sample family. These objects are usually treated as objects non-genetically related to the family. Some explanations are based on the background objects sharing the same space as the family and the limitation of methods—the boundary limits [51] and the choice of metrics [75]—of family identification. The term of "clan" [30] was proposed as a designation, for groups for which unequivocal membership and/or separation from other background groups is not possible. This seems to be a good compromise because it answers the dilemma created by the presence of spectral characteristics for both primitive (B-,C-type asteroids) and evolved (E-, S-, A-types) among objects of the same family.

If we discuss our relative knowledge of spectral behaviors in the visible spectral range for some big asteroid families, the near-infrared (NIR) spectral properties are still poorly known. We can emphasize the efforts of Burbine and Binzel [18] who performed a spectral survey of about 181 objects in the frame of the MIT

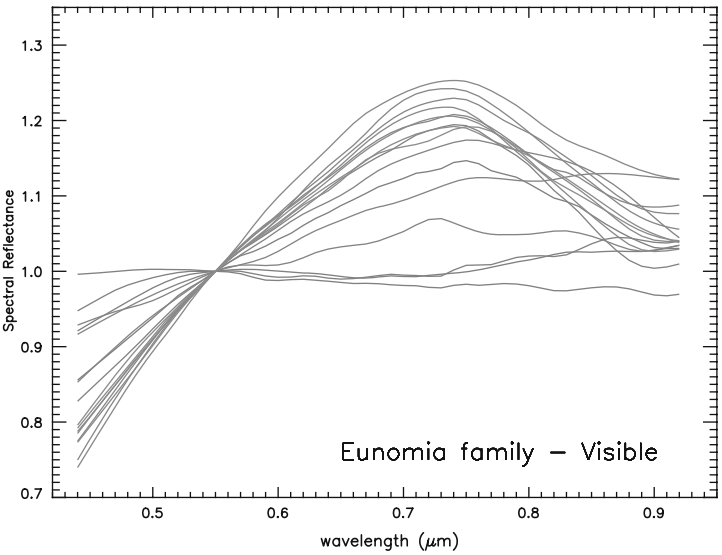

Fig. 2 Visible smoothed spectra of 16 asteroids belonging to the Eunomia family, extracted from SMASSII [20]. The spectra are normalized to 0.55 μm and the family members were selected from Zappala et al. [100]. The spectra span a wide range of slopes in the visible range. The asteroids 85, 141, and 1094 present flat no-features spectra, typical for C-F-X asteroids. A similar plot could be seen in Bus [19]

program *Small Main-Belt Asteroid Spectroscopic Survey*.[6] As long as the program was devoted to the main belt as a whole, the observed number of a certain family is limited (as it can be seen in Fig. 1); thus, general conclusions for families are difficult to be drawn.

Recently, a mineralogical analysis of 30 members of Eos family [69] was published using the spectral range 0.8–2.5 μm. The major conclusion is that the surface of the majority of their sample is dominated by forsteritic olivine, consistent with carbonaceous chondrites. One of the most plausible explanations is that Eos family (or at least their sample) might be composed by pieces of the mantle of a partial differentiated parent body.

These recent spectroscopic results bring forth the importance of a global visible+NIR spectral investigation extended to at least 2.5 μm. Indeed, this region contains a series of features (broadbands around 0.4, 1, and 2 μm, shallow absorption around 0.7 μm, etc.) that must be considered together when a mineralogical solution is computed.

Observations of rotation lightcurves for the family members is an important topic in deciphering the history of asteroid families. According to post-Yarkovsky paradigm [23], the family members should exhibit some preferential spin axis alignment. Laborious, long-term work on lightcurves for the Koronis family members

[6] The survey was performed in the spectral range 0.9–1.65 μm.

revealed a correlation between the lightcurve amplitude and the ecliptic longitude [5]; this correlation is a consequence of the alignment of spin axes [84]. This *Slivan state* deduced for the obliquity of spin axis was detailed later[7] [85]. This result could be corroborated to explain the rotation period distribution of 40 members of Koronis family [86]. This new result concludes a non-Maxwellian distribution of rotation rates inside this family strengthening the excesses in both slow- and fast-rotator objects.

The interpretation of such distribution of rotational period and spin axis orientation inside the Koronis family could be seen as a consequence of the YORP effect [91]. Thus, for the objects with prograde spin, the synodic periods are in the range 7.5–9.5 h and obliquities in the range of 42°–50°, while for the objects with retrograde rotation, the objects are slow ($P_{syn} \geq 15$ h) and fast ($P_{syn} \leq 5$ h) rotators, and the obliquity is in the range of 154°–169°. The non-random distribution of the orientations axis and of the synodic periods is considered as the consequence of the diurnal Yarkovsky effect. Moreover, the authors suggested that this thermal torque may be more important than collisions in changing the spin state of asteroids greater than 40 km in diameter.

3.2 Space Weathering and Asteroids Spectral Properties

Space weathering is defined by *Chapman* in 2004 [24] as being the observed phenomena caused by those processes (known or unknown) operating at or near the surface of an atmosphereless Solar System body that modify the remotely sensed properties of the body's surface from those of the unmodified, intrinsic, subsurface bulk of the body.

This definition stresses the difficulty of assessing a specific mineralogy of an atmosphereless body via remote observations. Indeed, accretion or erosion of particular materials, or modification of materials in situ by energetic impacts or irradiation will modify and contaminate the asteroid surface over a long period of time. The phenomenon of space weathering was first evidenced in the lunar soils. Laboratory analysis of the returned lunar soils revealed optical properties that differ from those of pristine lunar rocks.[8] These differences were attributed to several types of processes associated into the generic term of space weathering: regolith vitrification by high-speed micrometeorites, creation of grains (1–30 μm in size) of metallic iron (nanophase metallic iron), saturation of minerals by hydrogen implanted by the solar wind, and melted by micrometeoritic impacts.

---

[7] The article reveals a preferential spin axis alignment using the pole solution of 10 members of Koronis family, including the asteroid 243 Ida observed by Galileo spacecraft.

[8] The mature soil generally shows only the weak absorption features and red slope compared to the spectrum of the fragmental breccia.

In the case of asteroids, the phenomenon of space weathering is a very interesting subject for several reasons. One of the most important is linked to the OC paradigm.[9] From the 1970s, during several decades, the debate concerning the origin of ordinary chondrite meteorites proposed objects located in the inner part of main belt, or the extinct comets, or bodies which are delivered from the chaotic zones located inside secular resonances [67]. The association between some S-type[10] asteroids and OCs is still an open subject.

Another direction involving space weathering processes is that of explaining the spectral trend of Vesta or Vesta-like asteroids. The Vesta family is considered as the origin of most HED[11] meteorites. Their spectra exhibit features similar to a pristine, unaltered surface. While the maturation effect on asteroids surface from micrometeorite bombardment was estimated to be around two or three order of magnitude lower than on the Moon surface [33] the simple extrapolation of space weathering mechanisms explaining the Moon soils cannot be used. However, the long exposure of the surface to the solar wind must heavily alter its spectral properties, which is not the case for Vesta-like objects. New laboratory studies by irradiation of meteorites suggest that the pristine surface, such is that on Vesta asteroid, could be preserved in the presence of a remnant magnetic field of about 0.2 μT [87], acting like a shield against charged particles of the solar wind.[12] If this hypothesis seems to work for large, differentiated asteroids such as Vesta, the question of pristine materials on smaller ones, most probable fragments of Vesta crust and mantle, remains open.

Laboratory experiments might be also useful in simulating potential space weathering processes. These experiments allow the modification of the central wavelength of the 0.9–1.0 μm absorption band [68]. Such an alteration processes could allow different combinations of minerals which can simulate the central wavelength of the large band presented in asteroids spectra. Thus, space weathering is at the core of the debate of non-unique mineralogical interpretation of the surface of asteroids belonging to the same taxonomic class.

Experiments with pulse lasers [80], simulating micrometeorite impacts, conclude the modification of surface properties of samples and the production of nanophase iron deposits. The relevant timescale for the space weathering in this case was estimated to be of order of 100 Myr.

Space weathering becomes an interesting subject of study in the frame of catastrophic collisions in the main belt, namely (i) the young families discovered in the last decade and (ii) asteroids complexes (double, multiple, or binary asteroids). Indeed, the members of young families may exhibit surfaces younger (or rejuvenated) than the original parent body. This could be evidenced by spectroscopic mea-

---

[9] The ordinary chondrite (OC's) are by far the largest class of samples among meteorite falls; up to now there is no main-belt asteroid having spectral properties identical to that of OCs.

[10] We refer the reader to the articles of Gaffey et al. [34], Belton et al. [2, 3], Binzel et al. [6–8], Chapman [24].

[11] Group of Howardite, Eucrite, Diogenite meteorite classes.

[12] The presence of magnetic field of asteroids and its interaction with the solar wind was studied by Greenstadt [38], and Ip and Herbert [42].

surements and may give quantitative constraints on the timescale of space weathering alteration processes. In the case of binary or multiple asteroids, the spectroscopy of each component of the system could reveal differences in spectra. These differences could be interpreted in terms of homogeneity/heterogeneity of the original body, but also as a consequence of different ages of the surfaces.

Last but not least, observations "in situ" of spacecraft instruments reveal important clues about the asteroid surfaces. Spectral analysis (by NEAR-Shoemaker spacecraft) of (433) Eros craters shows albedo contrasts of order of factor of two [26], with fresh material on the rims and the crater walls. In the case of the asteroid (25143) Itokawa, recently visited by Hayabusa mission, spectroscopic observations reveal a single-scatter albedo 30–40% lower than that of Eros [45].

4 Young Families of Asteroids

Emerging directions in asteroid research which may bring forth the link between family members dynamics and their surface properties are being investigated. The most promising in this sense seem to be the study of young families [72–76]. Three families with ages between 1 and 10 Myr were identified during the last decade: the Iannini family (1–5 Myr old), Karin (5.75 ± 0.05 Myr old), and Veritas (8.3 ± 0.1 Myr old). Among them, the Karin cluster is located in a densely populated region of the large Koronis family,[13] while Iannini and Veritas are families with inclinated orbits ($12^\circ.15$ and $9^\circ.26$, respectively).

A new family was identified [76] around the asteroid (1270) Datura. This is a small family (only 7 members were identified) and it is considered as the result of a breakup of a main-belt asteroid approximately 450,000 years ago.

Using the osculating elements and a modified metric for the identification of young families [75], new clusters were proposed. This technique allows identification of three new clusters and to partially find again the families of Karin, Iannini, and Datura. These new clusters, each of them composed by three objects are (14627) Emilkowalski, 1992YC2,[14] and (21509) Lucascavin (Table 1 of [75]). The age estimation for the new clusters was less than 800,000 years for Emilkowalski members and less than 250,000 years for the other two clans. Recently [48, 90] efforts for characterizig these new families were published.

The synthesis of the young families currently proposed is given in Table 1.

It is important to mention the utilization of fine tuning induced by the Yarkovsky effect [15, 93] in the numerical integration backward in time for finding the origin of the catastrophic event at the origin of the young families.

---

[13] The largest body of the Karin family is the asteroid (832) Karin, identified previously also as member of the larger (and older) Koronis family.

[14] Identified also as 1989 AH5.

Table 1 The current knowledge of the young (less than 10 Myr old) families of asteroids. Family name, number of members, semi-major axis, eccentricity, and inclination of the largest member of the family (the osculating elements at April 16, 2008), and references are presented

Family name	Number of members	a AU	e	i °	References
Karin	90	2.864719	0.07861	1.00525	[74]
Iannini	49	2.642372	0.31233	11.09786	[96]
Veritas	259	3.168739	0.09886	9.2649	[73]
Datura	7	2.234749	0.20768	5.98964	[75]
Emilkowalski	3	2.598794	0.15047	17.73248	[75]
Lucascavin	3	2.280641	0.11288	5.98683	[75]
1992 YC2	3	2.622319	0.2188	1.62903	[75]

Two subjects must be mentioned, presented here as questions:
(1) What is the number of families that can be identified in the main belt? Can we talk about the completeness of the families in the main belt?
(2) How relevant are the physical parameters (and the parameters derived from spectroscopic measurements) in the general context of the new (young) families proposed by dynamicists?

4.1 Generalities on Karin Family

Older asteroid families ($\sim$1 Gyr) have been substantially eroded and dispersed, making difficult the accurate determination of the age or the nature of the family formation after the catastrophic impact. The younger families instead, experiencing little dynamical and collisional evolution after the breakup event, provide us with a valuable tool to understand disruptive asteroids collisions and even more subtle processes such as the dispersion of the asteroid families due to the Yarkovsky effect.

The announcement of the new family around the asteroid (832) Karin was made in 2002 [72]. This result proposed a cluster of 39 bodies on which the first two larger ones have comparable sizes ((832) Karin and (4507) 1990 FV). This new configuration stimulated the interests of scientists involved in collisional process within the main belt [59, 60].

The catastrophic disruption of the parent body asteroid was traced back in time by numerically integrating 13 numbered asteroids from the cluster of 39 asteroids in the (a_p, e_p, i_p space) [72]. It was found a remarkable agreement of the Ω and ω (nodal longitude and perihelion argument) for all the 13 asteroids 5.8 ± 0.2 Myr ago. Accordingly, at this time, they were following nearly identical orbits. Accounting for the undetected family members the diameter of the parent body was estimated at 24.5 ± 1 km. Later the age of Karin family was revised by numerically integrating a larger number of asteroids having osculating elements similar to those of the Karin cluster asteroids [74]. The output of the numerical integration was digitally filtered to suppress high frequencies retaining all the periods longer than $\sim$5 kyr.

The filtered signal was analyzed using the frequency-modified Fourier transform [83] to eliminate the terms corresponding to the secular planetary frequencies and to finally obtain the synthetic proper elements a_p, e_p, i_p, and the proper perihelion and nodal frequencies g and s. Applying HCM (hierarchical clustering method) on this proper elements set, 97 Karin cluster members, were identified. Seven of them were found to be interlopers since their nodal and perihelion longitude were not aligned with those of (832) Karin at t = −5.8 Myr. Among them was (4507) 1990 FV, previously considered as the second largest member of the cluster. However small (± 40°), the spread of Ω, $\bar{\omega}$ (nodal and perihelion longitudes) for the Karin cluster at t = −5.8 Myr is still too large to be a consequence of the breakup event itself that could account for ∼1° in both angles. A semi-major axis drift due to a non-gravitational effect was proposed as an explanation of this discrepancy and it was validated using a numerical integrator that explicitly accounted for the Yarkovsky effect [16]. The new integration[15] by improving the convergence of the proper elements provides a new estimate of the cluster age: 5.75 ± 0.05 Myr and, for the first time, the direct detection of the Yarkovsky effect for main-belt asteroids. With (4507) 1990 FV classified as an intruder, the size of the family parent body was revised to ∼20 km. Thus, the SFD of the Karin family becomes a classical one, containing a large body and a continuum of small members. Extrapolating the current semi-major axis drift rates, it was found that in ∼100 My the Yarkovsky effect will erase the genetic link between the cluster members making the family indistinguishable from the background asteroids for the HCM.

Hydrocode simulations which take into account the unobserved sub-kilometer fragments, which are believed to represent a large fraction of the parent body mass, obtain an estimate for the parent body of about 33 km in size [77]. The parent body of the Karin cluster was produced by the earlier collision that created the larger family of Koronis about 2–3 Gyr ago.

The discovery of young families (Karin being the most studied among them) offers an excellent opportunity for physical studies of the members that apparently suffered limited dynamical and collisional erosion. The spectroscopic investigation of the family members allows information on the structure and composition of the parent body. Thus, similar spectral features of the members are an indicator of the possible homogeneous composition of the parent body, while some differences in behavior or wavelength position of the spectral features could give some information about a possible differentiated structure of the parent body [88].

The processes of irradiation by cosmic and solar wind ions, the bombardment by interplanetary dust particles (micrometeorites), induce relevant surface modifications on atmosphereless bodies. Generally, the alteration affects the spectral properties of asteroids, induces progressive darkening, and reddening of solar reflectance

[15] The SWIFT code of Levison and Duncan [54] was principally used. Following the authors, several integration methods could be used via SWIFT: Wisdom–Holman Mapping (WHM or MVS), regularized mixed variable symplectic (RMVS), a fourth order T+U symplectic (TU4), and Burlisch–Stoer (BS). A particular package (SWIFT-RMVSY), which takes into account the Yarkovsky effect and the second-order symplectic integration scheme (MVS2) is also available and was used for backward integration for the young families.

spectra in the range 0.2–2.7 μm. The differences in the distribution of the spectral slopes of members inside the family can also be used for the evaluation of degree of space weathering . The precise dating of young family members (corresponding to the catastrophic collision) allows the link between remote spectroscopy and the laboratory data of irradiation experiments.

4.2 Physical Properties Inside the Karin Family

The dynamical considerations about this family should be completed by observations concerning the physics and the reflectance properties. Except for limited physical data for the asteroid (832) Karin, no physical properties were known for family members before the identification of the family.

4.2.1 Photometry

The most accessible member for observations is the asteroid (832) Karin. The first determination of its synodic period yielded 18.82 ± 0.01 h and the lightcurve displayed an amplitude of $0^m.32$ [4]. A new campaign of observations (performed during the opposition of 2003) revised these values to 18.35 ± 0.02 h for the synodic period and to $0^m.61 \pm 0^m.02$ for the composite lightcurve amplitude [98]. However, the authors mention a possible second period of 19.00 ± 0.03 h. The slope parameter was estimated of 0.19 ± 0.04.

Photometric observations were performed during the object's opposition in 2005, using the 1.2 m telescope at the Observatoire de Haute-Provence, France. The observations were obtained in February 9 and 11, 2005. Figure 3 presents these

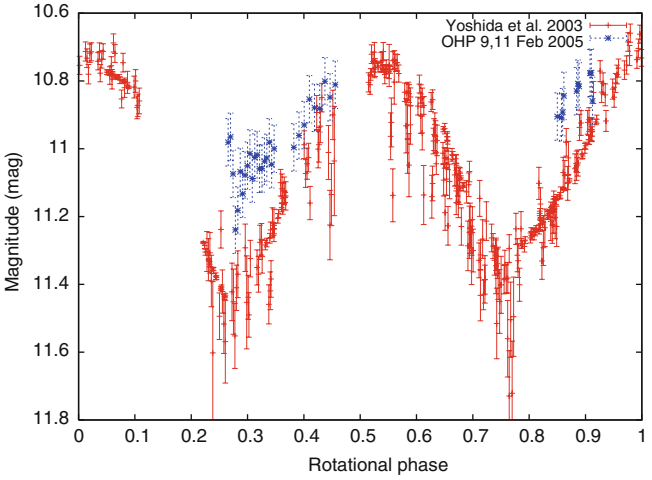

Fig. 3 Composite lightcurve of the asteroid (832) Karin. The observation points (with errorbars) were obtained on February 9 and 11, 2005, at Observatoire de Haute-Provence, France (*in blue*) and were superimposed by the composite lightcurve (*red color*) of Yoshida et al. [98]

observations (in blue) superimposed on the composite lightcurve obtained during the 2003 opposition [98]. These data are less dense but globally in accordance with the period previously proposed by Yoshida et al. [98].

B–V and U–B colors were reported in 1987 [4] while B–V, V–R, and V–I colors were reported in 2004 [98]. We underline the color variation over the rotational phase [98, 43] and the hypothesis concerning the inhomogeneity of Karin surface.

Both articles [4, 98] conclude for Karin having colors typical of S-type asteroids. In the assumption of an albedo of 0.2, Yoshida et al. [98] estimated the object size of an ellipsoid of 20.1×11.5 km.

Two other members of Karin family were observed recently [39]: (11728) Einer and (93690) 2000VE21. The rotational period was estimated to 12.92 ± 0.16 h, a lightcurve amplitude of $0^m.19$ for (11728) Einer, and no relevant discernable period for the asteroid (93690) 2000VE21.

As can be seen, efforts for characterizing this family in terms of rotational periods are still incipient, and at present these data do not allow a general conclusion.

4.2.2 Spectroscopy of the Karin Family

Spectroscopic results are more consistent for the Karin family, mainly due to a coordinated campaign [88] involving several observatories and telescopes. Visible spectroscopy was performed with NTT/EMMI (La Silla, Chile), CFHT/MOS (Mauna Kea, Hawaii), and TNG/Dolores (La Palma, Gran Canaria Island) for 24 members of Karin family, while 0.8–2.5 μm near-infrared spectroscopy was obtained using IRTF/SpeX (Mauna Kea, Hawaii) and TNG/NICS (La Palma, Gran Canaria Island) for six members. Observations with IRTF/SpeX were performed remotely using CODAM [9] infrastructure at the Paris Observatory.

By far, the Karin family is the most completely observed one in the visible range by spectroscopic techniques. Indeed we can find in the literature [88] the visible spectra for 26% of the family members.

The sixth-order polynomial function fitting the visible spectra are presented in Fig. 4. This representation of polynomial fit is preferred to the real data for qualitative considerations on the family as a whole. The global trend of the spectra is typical for a surface rich in silicates. Depending on Fe and Ca content in the olivine and pyroxene on the asteroid's surface, the maximum of spectra varies around 0.75 ± 0.02 μm. The wavelength variation for the maximum could be associated either with space weathering processes or with surface diversity (i.e., different mineralogies) among the family members.

Another variable used to describe the spectral trend is the slope parameter. For 23 objects the average slope is roughly $0.23 \pm 0.19\,\mu\mathrm{m}^{-1}$; the slope of one object (the asteroid (20089) 1994PA14) was estimated to have a value of 0.58 μm. A total of 40% of the family objects share the slope range similar to that obtained from the analysis of spectra of 300 ordinary chondrites[16] [49]. This result is consistent with

[16] The study of Lazzarin et al. [49] reveals that 95% of the OC slopes are below $0.208\,\mu\mathrm{m}^{-1}$; they associate this value as an indicator of detection of space weathering processes.

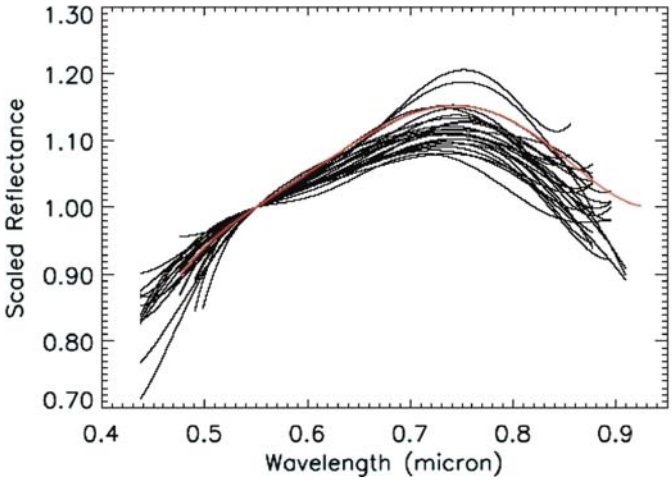

Fig. 4 Karin family spectra is depicted as sixth-order polynomial function fitting for 24 visible spectra and normalized to unity at 0.55 μm. The spectral trend is compatible with the presence of silicates for all the members and a slight difference in spectra is revealed. The *red line* represents the polynomial fit for the asteroid (832) Karin

the analysis of colors of Koronis members (especially objects from Karin family) of Sloan Digital Sky Survey Moving Object Catalog and the OC meteorites [44, 70]. These values may suggest that spectra of Karin family members are redder than the OCs and this trend could be associated to a low (but measurable) degree of space weathering.

How are the visible spectra of Karin family placed in the context of the Koronis family, at the origin of parent body of Karin clan? Figure 5 presents the spectral range of the Karin members (the domain is bordered by red color) in the context of other Koronis family members (green lines) obtained from the SMASS database. This comparison shows that the spectra of Karin members are less red that the ones of Koronis family.

NIR spectra (0.8–2.5 μm) of six members of the Karin family are also presented in the literature [11, 88, 17, 89, 25]. With one exception (the asteroid (832) Karin[17]) the data are very noisy and their interpretation is speculative. All the objects exhibit a detectable absorption band around 1 μm. Tentative interpretation (desirable to be improved in the future) of this spectral domain was done [88].

We can emphasize spectroscopy as a powerful tool to detect intruders inside a family. If we assume that Karin family members spectra are quite similar (i.e., coming from a relative homogeneous parent body), any asteroid having a spectral trend far from the majority of members is highly suspected to be an interloper.

[17] Karin spectra will be discussed in detail, the NIR counterpart of its composite spectrum is our basis for its mineralogical interpretation.

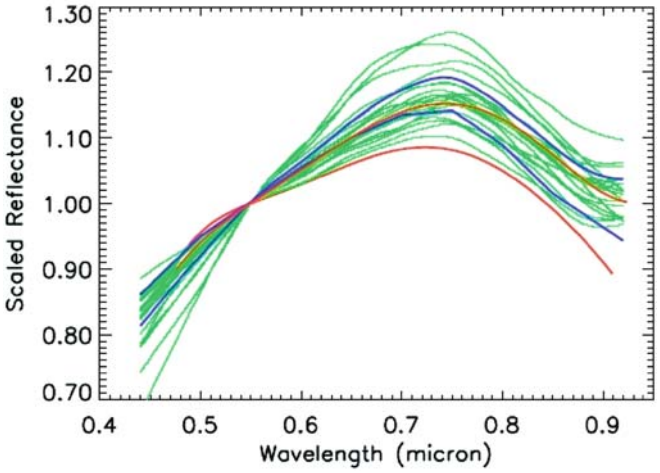

Fig. 5 The spectra of the Koronis family members (*green lines*) obtained from the SMASS database. In *red*, the domain representing 92% of the Karin members spectra. In *blue*, from *bottom to top*: the mean S-type spectrum and the mean S-type spectrum [20]. The slopes of Koronis members are greater than the ones of Karin young family. These values could be related to surfaces experiencing an important degree of space weathering for the Koronis members. The figure is reproduced from Vernazza et al. [87, 88]

The case of the asteroid (47640) 2000CA30 is interesting to be noted. The asteroid was observed on two nights, in November 4 and 5, 2003 using SpeX/IRTF. NIR spectrum shows a highly positive slope ($0.82 \pm 0.02\,\mu m^{-1}$) and no absorption feature (Fig. 6). The visible counterpart, obtained with Dolores/TNG used in LR-B mode on October 28, 2003, completes this figure. Considering the asteroid spectrum, there is a strong probability that (47640) 2000CA30 is a intruder. While low S/N spectra are recorded, new data are needed to confirm the asteroid spectrum.

This case of the asteroid (4507) 1991FV is puzzeling. Initially the asteroid was considered as member of Karin family [72]. However, its recently observed NIR spectrum is relatively close to the one of (832) Karin (Fig. 7).

4.2.3 Spectroscopy of (832) Karin

Soon after the publication of dynamical detection of the young family of Karin, efforts for observing spectroscopically the asteroid (832) Karin were undertaken. Near-infrared (0.8–2.5 μm) observations with CISCO/Subaru system were reported [81, 82]. The authors presented three spectra, identifying them on the composite lightcurve of (832) Karin obtained by Yoshida et al. [98] and discussed the differences among the observed spectra.[18] They noted a correlation of one of the spectrum

[18] The photometry was performed mainly during the 2003 opposition, and the spectral observations were carried out in September 2003.

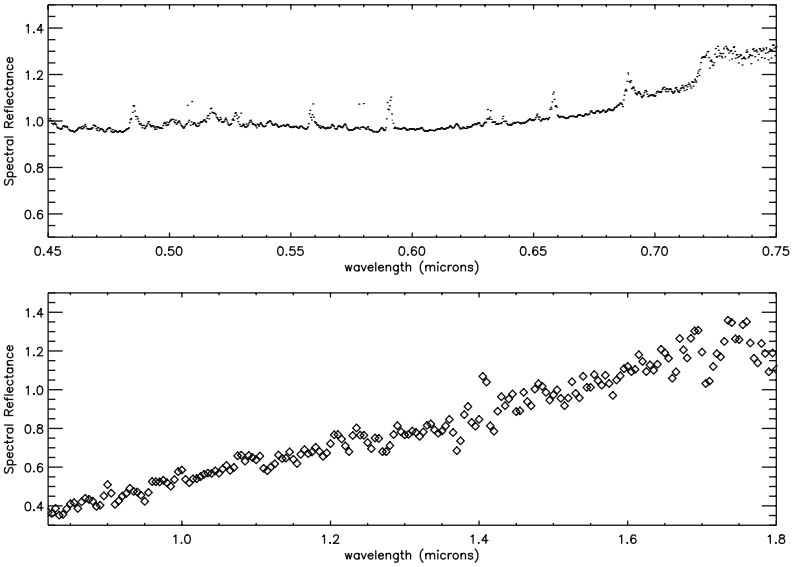

Fig. 6 Visible and NIR spectra of (47640) 2000CA30. The visible spectrum was obtained using Dolores/TNG used in LR-B mode. Data reduction was performed using the Hya64 standard star. NIR spectrum was obtained using SpeX/IRTF and reduced using the Landolt 93–101 standard star

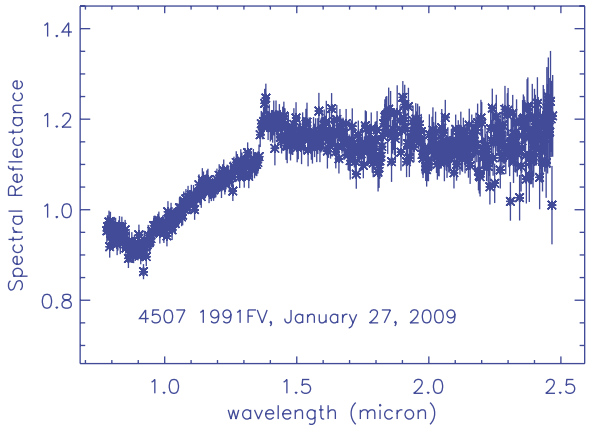

Fig. 7 NIR spectrum of the asteroid (4507) 1991 FV, initially considered as belonging to the Karin family. The observations were performed in January 27, 2009, using SpeX/IRTF and CODAM infrastructure. Data reduction was performed using Landolt 102-1081 standard star. The NIR spectrum (with errorbars and normalized to 1.25 μm) is similar to an S-type asteroid. The visible counterpart is presented by Vernazza et al. [87, 88]

with the color variation observed in Yoshida et al. [98]. The main conclusions were (i) Karin asteroid belongs to the S-type taxonomic class and (ii) because of spectral differences, Karin asteroid should have both mature and fresh surfaces, consequence of space weathering mechanisms.

This 0.8–2.5 μm spectral region was observed on November 4 and 5, 2003, using SpeX/IRTF [10]. These spectra, presented in Fig. 8, are in relatively good agreement with two of spectra published by Sasaki et al. [81, 82]. However, the absorption band around 1 μm exhibits a different depth, which cannot be explained by the error bars in the spectra. This may suggest that the surface spectral variation is real. However, this result shoud be reconsidered after some spectral anomalies reported on SpeX by Hardersen et al. [40].

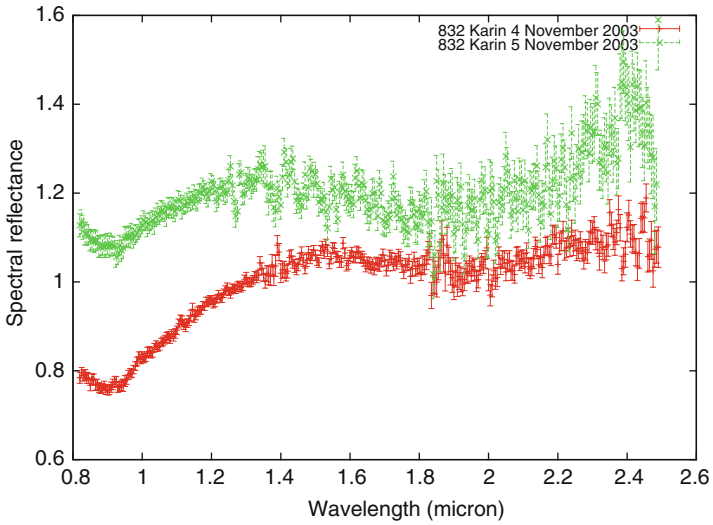

Fig. 8 NIR spectra obtained on November 4 and 5, 2003. The spectra are offset for clarity. The spectral trend is similar to a typical S-type asteroid. This graphic shows a relatively clear difference (which is not within the spectra error bars) in the region around 1 μm absorption band. This result should be reconsidered after some SpeX anomailes reported by Hardersen et al. [40]

New results were published recently [89, 25] based on observations in the visible and the near-infrared spectral regions. The new observations were performed between January and April 2006. These new results, obtained by two independent team of scientists, show no spectral variation in the Karins' spectra and sustain the hypothesis of a homogeneous superficial layer being at the origin of the reflected spectrum.

One of the questions that requires an answer is the following: What is the asteroid aspect during the 2003 opposition, compared to the 2006 one? This question could not have a trivial answer as long as the pole position of Karin is not determined yet. Thus, for this case our choice was to take into account all the possible pole solutions for the 2003 opposition and to see how they are placed during the 2006 opposition.

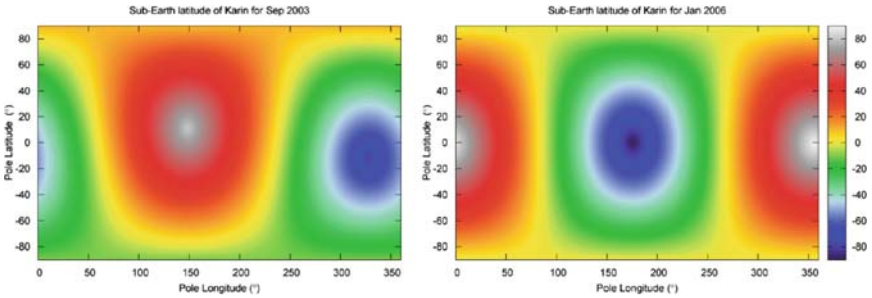

Fig. 9 Karin's sub-Earth latitudes for September 2003 and January 2006 as a function of all the possible pole solutions. The sub-Earth latitude is color coded; the domain yielding a near-equatorial (*yellow color*) aspect (sub-Earth latitude in the range [−15deg: +15deg]) is delimited by the *black lines*. All the possible aspect angles (sub-Earth latitudes) that Karin could take assuming an equatorial aspect during the September 2003 observations are clearly displayed on both figures. To avoid the redundancy of figures, we eliminate the representation from April 2006 while it is similar to that of January 2006

The result is presented in Fig. 9. We consider a near-equatorial aspect of the asteroid during the 2003 opposition and the comparison of sub-Earth latitudes for 2003 and 2006. From the region delimited by the black lines we can derive a high probability that Karin was at an equatorial aspect for both runs (Fig. 10). While the new spectral results [89, 25] cover all the rotational phase of the asteroid, the conclusion of a homogeneous surface is the most probable.

Quantitative results for space weathering were proposed [17] using both laboratory minerals and Karin reflectance spectra. Laboratory experiments on silicates by ion irradiation were modeled in terms of space weathering and accounting the Shku-

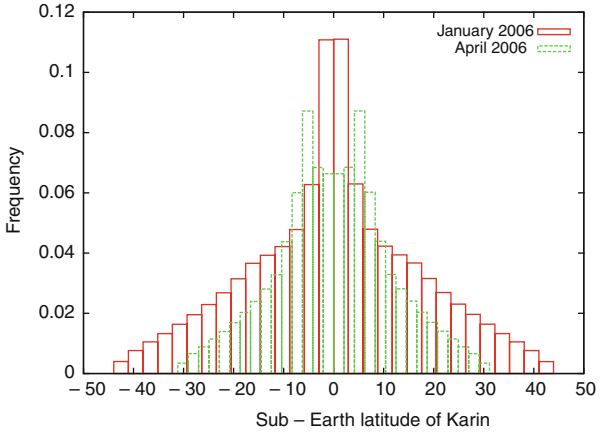

Fig. 10 Karin's sub-Earth latitudes in January and April 2006 versus their frequency. It appears that there is a quite high probability that Karin was close to the equatorial aspect during all observing runs

ratow law of diffusion. The proposed model (a mineralogical solution 58.5% olivine and 38% of orthopyroxene) was applied to the composite visible+NIR spectrum of (832) Karin. The results highlighted irradiation exposure time slightly lower than the dynamical age of the Karin family could be interpreted in terms of mechanisms allowing the renewal of surface with fresh materials.

5 Conclusions

The study of asteroid families is an important, current topic. New insights obtained by dynamical considerations (family identification research and procedures, discovery of young families, relation between families, and interplanetary dust bands) are also reflected by the increasing interest for completing the knowledge of the physics and the composition inside these families. In particular, the discovery of young families offers a new opportunity to study the interaction between the family members, the family and the other solar system bodies, the impact of a "hostile" medium for the asteroid surface, the importance of cumulative, and the long-term, non-gravitational effects. The last decade shows that long-term dynamics of family objects can be explained by accounting for new physical effects such as Yarkovsky and Yarkovsky–O'Keefe-Radzievskii–Paddack effects. Thus dynamics and physics of family members must be analyzed together while we must derive general properties for the whole family. This is what we call a necessary junction between dynamical concepts and the physical characterics of the family members.

Acknowledgments The chapter used results of observations acquired with IRTF, TNG, CFHT and NTT telescopes, and the CODAM remote facilities. The spectroscopic data of (4507) 1991FV were obtained and made available by the The MIT-UH-IRTF Joint Campaign for NEO Reconnaissance. The IRTF is operated by the University of Hawaii under Cooperative Agreement no. NCC 5-538 with the National Aeronautics and Space Administration, Office of Space Science, and Planetary Astronomy Program. The MIT component of this work is supported by the National Science Foundation under Grant No. 0506716. The authors thank Richard P. Binzel for comments.

References

1. Arnold, J.R.: Astron. J. **74**, 1235 (1969) 230
2. Belton, M.J.S., Veverka, J., Thomas, P., et al.: Science **257**, 1647 (1992) 237
3. Belton, M.J.S., Chapman, C.R., Veverka, J., et al.: Science **265**, 1543 (1994) 237
4. Binzel, R.P.: Icarus **72**, 135 (1987) 241, 242
5. Binzel, R.P.: In: Lagerkvist, C.-I., Rickman, H., Lindblad, B.A., Lindgren, M. (eds.) Asteroids, Comets, Meteors III, p. 15. University of Uppsala, Sweden (1990) 236
6. Binzel, R.P., Bus, S.J., Burbine, T.H., et al.: Science **273**, 956 (1996) 237
7. Binzel, R.P., Rivkin, A.S., Bus, S.J., et al.: Meteorit. Planet. Sci. **36**, 1167 (2001) 237
8. Binzel, R.P., Lupishko, D.F., DiMartino, M., et al.: In: Bottke, W.F., Jr., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, p. 183, University of Arizona Press, Tucson (2002) 237
9. Birlan, M., Barucci, A., Vernazza, P., et al.: New Astron. **9**, 343 (2004) 242

10. Birlan, M., Vernazza, P., Fulchignoni, M.: BAAS **36**, 1140 (2004) 246
11. Birlan, M.: Rom. Astron. J. **15**, 63 (2005) 243
12. Bottke, W.F., Vokrouhlický, D., Rubincam, D.P., Brož, M.: In: Bottke, W.F., Jr., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, p. 183, University of Arizona Press, Tucson (2002) 232
13. Bottke, W.F., Vokrouhlický, D., Nesvorný, D.: Nature **449**, 48 (2007) 232
14. Brouwer, D.: Astron. J. **56**, 9 (1951) 230
15. Brož, M., Vokrouhlický, D., Bottke, W.F., et al.: In: Lazzaro, D., Feraz-Melo, S., Fernández, J.A. (eds.) Asteroids, Comets, Meteors, Cambridge University Press, Cambridge (2005) 238
16. Brož, M.: PhD thesis, Charles University, Prague (2006) 240
17. Brunetto, R., Vernazza, P., Marchi, S., Birlan, M., et al.: Icarus **184**, 327 (2006) 243, 247
18. Burbine, T.H., Binzel, R.P.: Icarus **159**, 468 (2002) 234
19. Bus, S.J.: PhD thesis, Massachusetts Institute of Technology, Boston (1999) 235
20. Bus, S.J., Binzel, R.P.: Icarus **158**, 106 (2002) 233, 234, 235, 244
21. Carruba, V., Michtchenko, T.A., Roig, F., Ferraz-Mello, S., Nesvorný, D.: Astron. Astrophys. **441**, 819 (2005) 231
22. Carusi, A., Massaro, E.: Astron. Astrophys. Suppl. **34**, 81 (1978) 230
23. Cellino, A., Dell'Oro, A., Zappalà, V.: Planet. Sp. Sci. **52**, 1075 (2004) 232, 235
24. Chapman, C.: Annu. Rev. Earth Planet. Sci. **32**, 539 (2004) 236, 237
25. Chapman, C., Enke, B., Merline, W.J., et al.: Icarus **191**, 323 (2007) 243, 246, 247
26. Clark, B.E., Hapke, B., Pieters, C.L., Britt, D.: In: Bottke, W.F., Jr., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, p. 183, University of Arizona Press, Tucson (2002) 238
27. Dermott, S.F., Nicholson, P., Burns, J.A., Houck, J.R.: Nature **312**, 505 (1984) 232
28. Doressoundiram, A., Barucci, M.A., Fulchignoni, M., Florczak, M.: Icarus **131**, 15 (1998) 233, 234
29. Durda, D., Bottke, W.F., Jr., Nesvorný, D., Enke, B.L., Merline, W.J., Asphaug, E., Richardson, D.C.: Icarus **186**, 498 (2007) 231
30. Farinella, P., Davis, D.R., Cellino, A., Zappalà, V.: In: Harris, A.W., Bowell, E. (eds.) Asteroids, Comets, Meteors 91, p. 165 (1992) 234
31. Florczak, M., Barucci, M.A., Doressoundiram, A., et al.: Icarus **133**, 233 (1998) 233, 234
32. Florczak, M., Lazzaro, D., Mothé-Diniz, T., et al.: Astron. Astrophys. Suppl. Ser. **134**, 463 (1999) 233, 234
33. Gaffey, M.J.: LPSC **XIV**, 231 (1983) 237
34. Gaffey, M.J., Bell, J.F., Brown, R.H., et al.: Icarus **106**, 573 (1993) 237
35. Gaffey, M.J., Cloutis, E.A., Kelley, M.S., Reed, K.L.: In: Bottke, W.F., Jr., Cellino, A., Paolicchi, P., Binzel, R.P., (eds.), Asteroids III, p. 183, University of Arizona Press, Tucson (2002) 230
36. Gladman, B.J., Migliorini, F., Morbidelli, A. et al.: Science **277**(5323), 197 (1997) 232
37. Gomes, R., Levison, H.F., Tsiganis, K., Morbidelli, A.: Nature **345**, 466 (2005) 231
38. Greenstadt, E.W.: Icarus **14**, 374 (1971) 237
39. Hahn, G., Mottola, S., Sen, S.K., et al.: Bull. Astr. Soc. India **34**, 393 (2006) 242
40. Hardersen, P., Gaffey, M.J., Cloutis, E.A., et al.: Icarus **181**, 94 (2006) 246
41. Hirayama, K.: Astron. J. **31**, 743 (1918) 230
42. Ip, W.H., Herbert, F.: Moon Planets **28**, 43 (1983) 237
43. Ito, T., Yoshida, F.: Publ. Astron. Soc. Jpn. **59**, 269 (2007). 242
44. Jedicke, R., Nesvorný, D., Whiteley, R., et al.: Nature **429**, 275 (2004) 243
45. Kitazato, K., Clark, B.E., Abe, M., Abe, S., et al.: Icarus **194**, 137 (2008) 238
46. Kneževic, Z., Milani, A.: Astron Astrophys. **403**, 1165 (2003) 230
47. Kneževic, Z., Lemaître, A., Milani, A.: In: Bottke, W.F., Jr., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) Asteroids III, p. 603, University of Arizona Press, Tucson (2002) 230, 231
48. Krugly, Y.N., Gaftonyuk, N.M., Ïurech, J., et al.: The Solar System Bodies: From Optics to Geology Conference Kharkov, Ukraine, 26–29 May 2008 (abstract). 238
49. Lazzarin, M., Marchi, S., Magrin, S., Licandro, J.: MNRAS **359**, 1575 (2005) 234, 242
50. Lazzaro, D., Angeli, C.A., Carvano, J.M., et al.: Icarus **172**, 179 (2004) 233, 234
51. Lazzaro, D., Mothé-Diniz, T., Carvano, J.M., et al.: Icarus **142**, 445 (1999) 233, 234

52. Lemaître, A.: Cel. Mech. Dyn. Astron. **56**, 103 (1993) 231
53. Lemaître, A., Morbidelli, A.: Cel. Mech. Dyn. Astron **60**, 29 (1994) 231
54. Levison, H., Duncan, M.: Icarus **108**, 18 (1994) 240
55. Martin, H., Albarède, F., Clayes, P., Gargaud, M., et al.: EM&P **98**, 97 (2006) 231
56. Marzari, F., Farinella, P., Davis, D.R.: Icarus **142**, 63 (1999) 231
57. Michel, P., Benz, W., Tanga, P., Richardson, D.C.: Science **294**, 1696 (2001) 231
58. Michel, P., Benz, W., Tanga, P., Richardson, D.C.: Icarus **160**, 10 (2002) 231
59. Michel, P., Benz, W., Richardson, D.C.: Nature **421**, 608 (2003) 239
60. Michel, P., Benz, W., Richardson, D.C.: Icarus **168**, 420 (2004) 231, 239
61. Milani, A., Kneževiæ, Z.: Cel. Mech. Dyn. Astron. **49**, 347 (1990) 231
62. Milani, A., Kneževic, Z.: Icarus **98**, 211 (1992) 231
63. Milani, A., Kneževic, Z.: Icarus **107**, 219 (1994) 231
64. Morbidelli, A., Zappala, V., Moons, M., Cellino, A., Gonczi, R.: Icarus **118**, 132 (1995) 232
65. Morbidelli, A., Petit, J.-M., Gladman, B., Chambers, J.: M&PS **36**, 371 (2001) 231
66. Morbidelli, A., Nesvorný, D., Bottke, W.F., et al.: Icarus **162**, 328 (2003) 232
67. Morbidelli, A., Vokrouhlický, D.: Icarus **163**, 120 (2003) 237
68. Moroz, L.V., Fisenko, A.V., Semjonova, L.F., et al.: Icarus **122**, 366 (1996) 237
69. Mothé-Diniz, T., Carvano, J.M., Bus, S.J., et al.: Icarus **195**, 277 (2008) 235
70. Murchie, S.L., Pieters, C.L.: J. Geophys. Res. **101**, 2201 (1996) 243
71. Nesvorný, D., Morbidelli, A., Vokrouhlický, D., et al.: Icarus **157**, 155 (2002) 231
72. Nesvorný, D., Bottke, W.F., Dones, L., Levison, H.: Nature **417**, 720 (2002) 238, 239, 244
73. Nesvorný, D., Bottke, W.F., Levison, H.F., Dones, L.: Astrophys. J. **591**, 486 (2003) 231, 232, 238, 239
74. Nesvorný, D., Bottke, W.F.: Icarus **170**, 324 (2004) 231, 238, 239
75. Nesvorný, D., Vokrouhlický, D.: Astron. J. **132**, 1950 (2006) 234, 238, 239
76. Nesvorný, D., Vokrouhlický, D., Bottke, W.F.: Science **312**, 1490 (2006) 238
77. Nesvorný, D., Enke, B.L., Bottke, W.F., et al.: Icarus **183**, 349 (2006) 240
78. Nesvorný, D., Vokrouhlický, D., Bottke, W.F., et al.: Icarus **188**, 400 (2007) 233
79. Petit, J.-M., Morbidelli, A., Chambers, J.: Icarus **153**, 338 (2001) 231
80. Sasaki, S., Nakamura, K., Hamabe, Y.: Nature **410**, 555 (2001) 237
81. Sasaki, T., Sasaki, S., Watanabe, J., et al.: 35-th LPSC, abstract 1513 (2004) 244, 246
82. Sasaki, T., Sasaki, S., Watanabe, J. et al.: Astrophys. J. **615**, L161 (2004) 244, 246
83. Šidlichovský, M., Nesvorný, D.: CeMDA **65**, 137 (1996) 240
84. Slivan, S.M.: Nature **419**, 49 (2002) 236
85. Slivan, S.M., Binzel, R.P., Crespo da Silva, L.D., et al.: Icarus **162**, 285 (2003) 236
86. Slivan, S.M., Binzel, R.P., Boroumand, S.C., et al.: Icarus **195**, 226 (2008) 236
87. Vernazza, P., Brunetto, R., Strazzula, G., et al.: Astron. Astrophys. **451**, L43 (2006) 237, 244, 245
88. Vernazza, P., Birlan, M., Rossi, A., et al.: Astron. Astrophys. **460**, 945 (2006) 240, 242, 243, 244, 245
89. Vernazza, P., Rossi, A., Birlan, M., et al.: Icarus **191**, 330 (2007) 243, 246, 247
90. Vernazza, P., Binzel, R.P., Rossi, A., Birlan, M., et al.: In: It Asteroids, Comets, Meteors 2008 Conference, Baltimore 14–18 July 2008 (abstract) 238
91. Vokrouhlický, D., Nesvorný, D., Bottke, W.F.: Nature **425**, 147 (2003) 236
92. Vokrouhlický, D., Brož, M., Morbidelli, A., Bottke, W.F., Nesvorný, D., Laz-zaro, D., Rivkin, A.: Icarus **182**, 92 (2006) 230, 231
93. Vokrouhlický, D., Brož, M., Bottke, W.F., et al.: Icarus **182**, 118 (2006) 238
94. Vokrouhlický, D., Nesvorný, D., Bottke, W.F.: Astrophys. J. **672**, 696 (2008) 232
95. Williams, J.G.: In: Binzel, R.B., Gehrels, T., Matthews, M.S. (eds.), Asteroids II, p. 1034, University of Arizona Press, Tucson (1992) 230
96. Willman, M., Jedicke, R., Nesvorný, D., et al.: Icarus **195**, 663 (2008) 239
97. Xu, S., Binzel, R.P., Burbine, T.H., Bus, S.J.: Icarus **115**, 1 (1995). 233
98. Yoshida, F., Demawan, B., Ito, T., et al.: Publ. Astron. Soc. Jpn. **56**, 1105 (2004) 241, 242, 244, 246
99. Zappalà, V. Cellino, A.: Celest. Mech. Dynam. Astron. **54**, 207 (1992) 230
100. Zappalà, V., Bendjoya, Ph., Cellino, A., Farinella, P., Froeschle, C., Asteroid Dynamical Families. EAR-A-5-DDR-FAMILY-V4.1. NASA Planetary Data System (1997) 233, 234, 235

The Gaia Mission and the Asteroids

A Perspective from Space Astrometry and Photometry for Asteroids Studies and Science

Daniel Hestroffer, Aldo dell'Oro, Alberto Cellino, and Paolo Tanga

Abstract The Gaia space mission to be operated in early 2012 by the European Space Agency (ESA) will make a huge step in our knowledge of the Sun's neighbourhood, up to the Magellanic Clouds. Somewhat closer, Gaia will also provide major improvements in the science of asteroids, and more generally to our Solar System, either directly or indirectly. Gaia is a scanning survey telescope aimed to perform high-accuracy astrometry and photometry. More specifically, it will provide physical and dynamical characterisation of asteroids; a better knowledge of the solar system composition, formation and evolution; local test of the general relativity; and linking the dynamical reference frame to the kinematical ICRS. We develop here the general aspects of asteroid observations and the scientific harvest in perspective of what was achieved in the pre-Gaia era. In this lecture we focus on the determination of size of asteroids, shape and rotation, taxonomy, orbits and their improvements with historical highlight and also the dynamical model in general.

In memoriam of Jacques Henrard (1940–2008). We dedicate this chapter to this wonderful colleague from the FUNDP (Namur), who was—as long as possible—an

D. Hestroffer (✉)
IMCCE, CNRS, Observatoire de Paris, 77 avenue Denfert Rochereau, 75014 Paris, France, hestroffer@imcce.fr

A. dell'Oro
INAF-Osservatorio Astronomico di Torino, strada Osservatorio 20, 10025 Pino Torinese, Italy, delloro@oato.inaf.it

A. Cellino
INAF-Osservatorio Astronomico di Torino, strada Osservatorio 20, 10025 Pino Torinese, Italy;
IMCCE, CNRS, Observatoire de Paris, 77 avenue Denfert Rochereau, 75014 Paris, France, cellino@oato.inaf.it

P. Tanga
Cassiopée, CNRS, Observatoire de la Côte d'Azur, Mont Gros, 06304 Nice, France, Paolo.Tanga@oca.eu

Hestroffer, D. et al.: *The Gaia Mission and the Asteroids*. Lect. Notes Phys. **790**, 251–340 (2010)
DOI 10.1007/978-3-642-04458-8_6

assiduous participant to such winter schools of the CNRS and did always share his bright mind and high excitement in science and research.

Foreword

This chapter is a compulsion of several of this CNRS school courses given in Bad Hofgastein completed by some additional material. It is not intended to give a complete review of the Gaia capabilities for asteroids science or of the treatment of orbit determination and improvement since the beginning of orbit computation. Neither will it cover each of the different techniques used for any particular problem. We hope, however, that it gives an overview of the Gaia mission concept, astrometry and photometry of asteroids (and small bodies) in particular from space, and current developments in this research topic. Besides, this school being French in a German-speaking place, some French and German bibliography have sometimes been favoured or added.

1 Introduction

Before entering into the description of the Gaia mission observations and the discussion of the expected results for asteroids science, we will briefly remind the basic principles from the Hipparcos mission and then give an overview of the Gaia objectives (Sect. 2). We will present the Gaia satellite and instruments as well as its operational mode (Sect. 3) and expected scientific results for the Solar System (Sect. 4). In the next sections we will develop more specifically three aspects: the astrometric CCD signal yielding the fundamental astrometric position and marginal imaging capabilities (Sect. 5), the photometric measure yielding physical properties of asteroids (Sect. 6) and the dynamical model from the asteroids astrometry (Sect. 8). This provides a solid overview of what can be achieved in the domain of planetology and dynamical planetology of asteroids. The following sections are more general and not exclusively related to Gaia. There, we develop the general problem of orbit determination and improvement for the case of asteroids orbiting around the Sun (Sect. 9), we give a short description of both historical and modern methods. Finally, we treat the case of orbit determinations of binaries (Sect. 10) focusing mainly on resolved binaries. But since the problem of orbit reconstruction for extra-planetary systems appeared to be of importance for this school, it has been briefly addressed here through the problem of astrometric binaries (Sect. 10.4). Extra-solar planets is another topic actually addressed by Gaia, but this is out of the Solar System and out of the topic of the present lecture.

2 Gaia—The context

2.1 Before Gaia—The Hipparcos Legacy

Ten years have passed since the publication of the Hipparcos catalogue in 1997. This space mission did provide a scientific harvest in many fields of astronomy and even, indirectly, in Earth science [91]. The Hipparcos mission provided a homogeneous astrometric catalogue of stars with more sources and more precise than were the *Fundamental Katalog* series during the twentieth century. The acronym "High Precision Parallax Collecting Satellite"—in honour of the Greek astronomer Hipparchus—recalls that the basic output is, of course, the measure of the parallax of stars and their proper motions. In fact there were two programmes or instruments on board the satellite: Hipparcos and Tycho. Both provide astrometry and photometry of the celestial sources that were observed over the period 1989–1993, and two catalogues of stars were derived as well, the Hipparcos and Tycho catalogues. The Tycho data are based on the "sky mapper" which gives the detection of sources and triggers the main astrometric field observations for Hipparcos. It is hence less precise than Hipparcos but has more targets (about 2,500,000 stars in its second version Tycho2 in the year 2000) than Hipparcos (about 120,000) and provides two-dimensional astrometry as well as photometry in two bands close to the Johnson B and V. Additional treatment of the raw data has been undertaken [123]. Further details on the Hipparcos/Tycho missions and instrumentation can be found in, e.g. [59, 60] in the context of high-accuracy global astrometry, and [45, 49, 90, Vol. 1, Sect. 2.7] for all details on the solar system objects observations and catalogues. The Hipparcos and Tycho Solar System Annex files (`solar_ha`, `solar_hp`, `solar_t`) are available on the CDS database.[1] While Hipparcos and Tycho were designed to observe stars, they nevertheless could also provide data for solar system objects: 48 asteroids, 5 planetary satellites and 2 major planets.[2] There were different limiting factors depending on the programme instrumentation: for Hipparcos, magnitude brighter than $V \leq 12.4$ and size smaller than $\phi \leq 1$ arcsec and for Tycho, magnitude brighter than $V \leq 11.5$ and size smaller than $\phi \leq 4$ arcsec. But the more stringent one was that Hipparcos could only observe a very limited number of objects in its field of view (FOV), both for stars or asteroids.

The basic principle of the mission was to derive relative positions of targets observed simultaneously in two well-separated fields of view. The measure principle consists basically in observing the target while it crosses the field of view with a photometer and to record the photon flux as modulated by a periodic grid. In addition to the relative astrometry given by the grid, the photometers also recorded the total flux, providing the magnitudes in the broad $Hp \approx V_J + 0.3035\,(B - V)$ system for Hipparcos and in two (red and blue) bands for Tycho. The Tycho astrometric and

[1] URL: http://webviz.u-strasbg.fr/viz-bin/VizieR?-source=I/239/

[2] Pluto was not observed which spares us the trouble of having to name this object observed in the past, before last IAU resolution.

photometric data are less precise than their Hipparcos counterpart and concern only 6 asteroids compared to 48 asteroids observed within the Hipparcos mission.

The Hipparcos measure of position and magnitude of these selected solar system objects provided many scientific outcomes. We list a few, showing the diversity of the applications:

- From the observed positions of the satellites, and taking one model for their ephemerides, one can derive the (pseudo-)position of the system barycentre, and/or the centre of mass of the planet, as well. Such positions are model dependent since one uses the theory for the orbits of the satellites, but they are far more precise than direct observations of the planets themselves [78];
- The particular photometry derived by one of the data reduction consortium (FAST) comprises the classical apparent magnitude together with an additional one that is biased for non-point-like sources. This allowed to indirectly resolve the object and derive information on its size and light distribution [48];
- Hipparcos astrometry enabled the determination of the mass of (20) Massalia from a close encounter with the asteroid (44) Nysa [4];
- High-accuracy astrometry of asteroids enabled to improve their ephemerides [52] and also to detect small systematic effects due to the photocentre offset [46];
- High-accuracy astrometry of asteroids enables to link the dynamical inertial frame to the catalogue [107, 5, 22];
- The stellar astrometric catalogue, mostly Tycho2, has an indirect consequence on science for the Solar System. Such stars provide better astrometric reductions for modern CCD observations of solar system objects, as well as re-reduction for older plates (the problem is still the low density when compared to, e.g. UCAC2); they also provide much better predictions for stellar occultation path [28, 108, 30].

Compared to Gaia observation and science of solar system objects, Hipparcos was mostly a (nice) prototype or a precursor. They still have many common points:

- obviously space-based telescopes are exempt from many of all atmosphere-related problems (seeing, refraction, etc.) and provide better stability (mechanical thermal) and a better sky coverage (not limited to one hemisphere);
- scanning law with a slowly precessing spin axis (going down to solar elongations $\approx 45^\circ$);
- simultaneous observation of two fields of view for global astrometry;
- astrometric and photometric measurement, the colour photometry being mandatory for correction to the astrometry.

The main difference between the two missions arises from the use of CCD device for Gaia instead of a photo-multiplier channel with Hipparcos/Tycho, and to a lesser extent the capabilities of onboard data storage as well as data transmission to the ground segment. Albeit CCD observations were already in use in astronomy and astrometry in particular during the 1980s, Hipparcos could not benefit of this

technique: CCDs were still not validated for use in space (which requires a higher quality and robustness) and in any case the design of the satellite had to be set long before the beginning of the mission (often satellites are launched with outdated material). The consequences are dramatic in terms of number of celestial bodies observable and general astrometric and photometric accuracy: Gaia is not an Hipparcos-II. The limiting magnitude of Tycho was $V \leq 12$, but the catalogue is far from being complete[3] at that level; in comparison, the limiting magnitude of Gaia will be $V \leq 20$. The border of Hipparcos is at 200 kpc, while Gaia will reach a large part of the Milky Way up to the Magellanic Clouds.

In this respect also Gaia is different from the SIM mission [104, 105] (USA) which will provide very accurate parallax but for a very limited number of targets. The smaller Japanese mission Jasmine has more common points but, observing in the infrared, the scientific goals are orthogonal and complementary. Last, the two European satellites differ in their location in space, Hipparcos missed its geostationary orbit and was on an eccentric one, Gaia will be at the L2 Sun–Earth Lagrangian point. Gaia will also observe a large number of solar system objects (mainly asteroids) whose data will provide new insights and scientific results as developed in the next sections.

2.2 What Is Gaia?

The Gaia mission was designed to provide a renewed insight in the Galaxy structure, through a homogeneous set of very accurate measurements of stellar positions, motions and physical properties [73, 74]. However, the independence from any input catalogue grants that a very large number of non-stellar sources will be additionally observed. In fact, Gaia will automatically select observable sources with a criterium mainly based on a single parameter, the magnitude threshold ($V \leq 20$).

During the preliminary study of the mission, the community of planetologists realised that the observations of asteroids by Gaia may have a strong scientific impact, allowing a general improvement of our view of the Solar System of the same order as in the case for stars [72]. Of course, the reasons are similar and are built on the unprecedented astrometric accuracy of Gaia and on its spectro-photometric capabilities.

One will note, however, that the strongest quality of the Gaia data—besides accuracy—will reside in homogeneity. In fact, no other single survey has produced an equivalent wealth of data for 300,000 solar system objects, as is expected for Gaia [75]. As we will illustrate in the following, a complete characterisation of the small bodies of the Solar System will be possible.

[3] Hipparcos with its larger band filter can reach slightly fainter magnitudes, 12.5, but is anyway much less complete than Tycho.

We can also compare Gaia to other forthcoming deep surveys, such as the very important Pan-STARRS,[4] expected to map the whole observable sky three times per month at greater depth (V∼24, for the single observation at SNR=5). In this case, even more objects will be detected, and this will constitute a serious advantage to feed investigations of smaller or more distant and fainter asteroids. However, the astrometric and photometric accuracy—despite being optimised—will remain limited by ground-based conditions, and spectro-photometry will not be of the same level as in the case of Gaia. Other typical observational constraints for ground-based investigations also apply to Pan-STARRS, including the minimum Sun elongation that will be reached, severely limiting its capabilities for the investigation of peculiar asteroid categories, like Earth-crossers or the inner-Earth objects. Lastly, the sky coverage is limited to one hemisphere and *global astrometry* is hardly achieved with such systems (on the other hand, they will benefit the Gaia catalogue of stars).

For these reasons, and despite the limitation in brightness to V∼20, we believe that Gaia is rather unique and really has the potential to have a major impact in solar system science. In this lecture, we will try to focus on the kind of data that the mission will provide and on the corresponding data treatment that is being conceived in order to extract the relevant information. Hopefully, the observer will appreciate the techniques allowing to reach an exceptional accuracy, and the theoretician will find interesting new problems that must be solved to fully exploit the Gaia data scientific content.

3 The Gaia Mission

3.1 Launch and Orbit

The Gaia satellite will fit into the payload bay of a Soyuz fregat vector that will be launched from the European base of Kourou (French Guyana). After a ∼1-month travel, it will reach the L2 Lagrangian equilibrium point, that is situated 1.5 million km from Earth, opposite to the Sun. The satellite will then deploy the solar panels, fixed on a large, circular sunscreen (diameter: 10 m) that lies on a plane perpendicular to the spin axis (Fig. 1). The visibility of the satellite and the data link is reduced when compared to a geostationary orbit, but the environment is quieter. The L2 point is a dynamically unstable equilibrium location, requiring firing the satellite thrusters to apply trajectory corrections every ∼1 month. Gaia will thus be maintained on a Lissajous orbit around L2, allowing it to avoid eclipses of the Sun in the Earth shadow. The location thus appears as an optimal choice for constant sunlight exposure and for maximum thermal stability. The planned operational lifetime of the mission will be 5 years.

[4] http://pan-starrs.ifa.hawaii.edu/public/

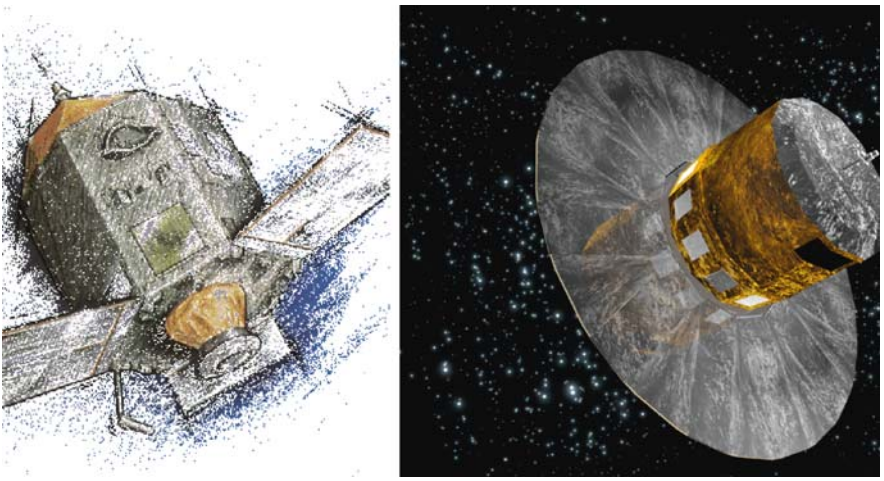

Fig. 1 The Hipparcos (*left*) and Gaia (*right*) satellites in a pictorial view. Hipparcos: one sees above the thruster and solar arrays one of the telescope baffle, for one of the observing direction. Gaia: the large, circular sunscreen will be deployed after launch and cargo to L2, it protects the instruments and permits their thermal stability. The telescopes, detectors and associated circuitry are situated inside the hexagonal or cylindrical housing (Copyright ESA)

3.2 The Spacecraft

The beating heart of the satellite is enclosed in an approximately cylindrical structure, including all the relevant parts: the thrusters openings and their fuel tanks, the electronic equipment, and, of course, the optical train. Since we are interested in the observations, we will just give some fundamental design principles concerning the latter component [87].

The optical bench structure is based on an octagonal toroid built in silicon carbide. This is a critical component, supporting all the optics and the focal plane. The measurement principle—being similar to that of Hipparcos—requires two different lines of sight, materialised by two Cassegrain telescopes. The primary mirrors are rectangular and measure 1.45×0.50 m^2. Five additional mirrors are required to fold the optical path, obtaining an equivalent focal length of 30 m. The light beams are combined and focused on a single focal plane, composed of a matrix of 106 CCDs. The extreme rigidity of the toroidal structure (actively monitored during the mission) ensures that the angle between the two telescopes (also called "basic angle") remains constant at 106.5°.

The CCD array constitutes a large, 1-G pixel camera that can be compared (in pixel number) to the Pan-STARRS cameras. However, it remains unrivalled in surface, since it extends over 0.93×0.46 m, by far the largest CCD array ever conceived.

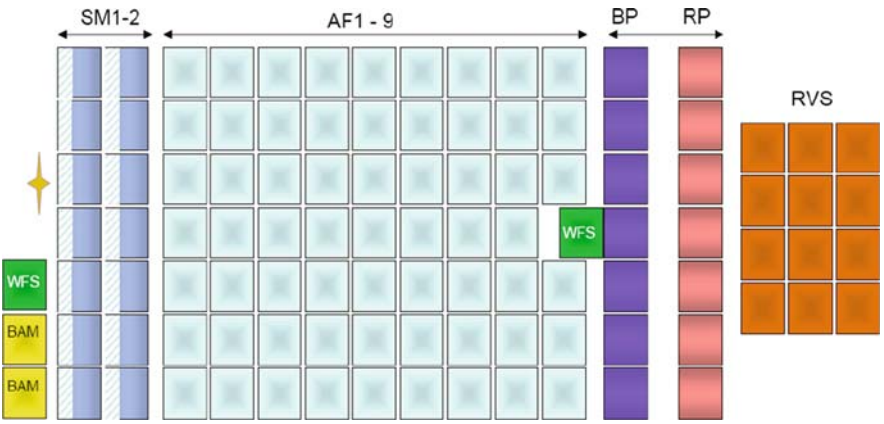

Fig. 2 The focal plane receives the light beams of both telescopes. While spinning, Gaia scans the sky in such a way that images of sources enter the focal plane from the *left* and cross it moving towards the *right*. The whole crossing takes about 1 min. Each CCD is crossed in 4.42 s. See the text for instrument details. The basic angle monitoring (BAM) and wavefront sensor (WFS) CCDs are used for monitoring tasks (Copyright EADS Astrium)

The resolution anisotropy due to the rectangular entrance pupil of the telescopes are matched by strongly elongated pixels, about three times larger along the direction in which the diffraction spot is more spread.

The general organisation of the focal plane is illustrated in Fig. 2. Different groups of CCDs are identified depending on their functions, since they correspond in all respects to different instruments. The main Gaia instrument is the astrometric field (AF), receiving unfiltered light, which is devoted to produce ultra-precise astrometry of the sources [65]. The other instruments are aimed at achieving spectral characterisation. The figure shows the red and blue photometers (RP and BP), which will be receiving light dispersed by a prism, and have optimised sensitivity in two different, contiguous portions of the visible spectrum. The resolution of RP and BP is rather low, each portion being spread on $\sim$30 pixels. The radial velocity spectrograph (RVS), on the other hand, provides a spectrum in a restricted wavelength range (847–874 nm) but with much higher resolution. It is aimed at sampling some significant spectral lines that can be diagnostic of stellar composition and can be used to derive radial velocities with a $\sim$ 1 km s^{-1} typical uncertainty. Due to star crowding (superposition of spectra) and SNR constraints, RVS will have a limit magnitude at $V = 17$. This instrument will provide no scientific information for asteroids. On the other hand, the measures for the bright asteroids will be used to calibrate the kinematic zero point of the RVS, as a complement to the data from IAU standard stars.

Since Gaia is continuously spinning with a rotation period of 6 h, the sources will drift on the focal plane, entering from the left in the scheme of Fig. 2 and travelling towards the right. The displacement will be compensated by a continuous drift of the photoelectrons on the CCD at the same speed; this technique (also used on fixed

ground-based telescopes) is known as "TDI mode" (from "time delay integration"). The resulting integration time (4.42 s) corresponds to the time interval required by a source to cross each CCD. Of course, this principle applies to all instruments. As a consequence, while an image of the source drifts on the AF CCDs, it will take the form of a dispersed spectra while moving on RP, BP and RVS.

An important part of the focal plane, the sky mapper (SM), requires some additional explanation. In fact, Gaia will neither record nor transmit to the ground reading values for all the pixels, due to constraints on the data volume. Conversely, only a reduced number of values ("samples"), representing the signal in the immediate surroundings ("window") of each source, will be processed and transmitted, and only for objects brighter than $V = 20$. These samples represent either the value of single pixels or of some binning of couples of pixels, depending on the star brightness. For stars with $V > 16$ (i.e. the largest fraction of sources) only six samples will be available, corresponding to the signal binned along six pixel rows in the direction perpendicular to the scan motion. These samples will be transmitted for each CCD in the AF field. Larger samples are required to accommodate the dispersion of RP, BP and RVS spectra.

As a consequence, the AF information on positions will be essentially one dimensional, being very precise along the scanning circle, but very approximate in the across-scan direction. One should also note that the windows are assigned by the onboard algorithm after SM detection. Beside a confirmation of detection that is expected from the first AF column, no other controls are executed all along the focal plane crossing, and the window follows the object on each CCD, assuming that it shifts at the nominal scanning speed. While this is true for stars, we will see that solar system objects will suffer measurement losses due to their apparent motion.

3.3 Observation Principles: The Scanning Law

The accuracy requirements of the mission can be reached only if the sky coverage is fairly uniform. To obtain this result, the direction of the spin axis of the probe cannot obviously stay fixed, but it must change, slowly but continuously, to change correspondingly the orientation of the scanning circle on the sky.

Two additional rotational motions are thus added to the four- revolutions per day satellite spin (Fig. 3).

The first one is a precession of the spin axis in 70 days, along a cone whose axis points towards the Sun. The change in orientation of the latter, provided by the orbital revolution around the Sun in 1 year, is the last additional rotation.

The overall motion, beside being derived from the scanning law optimisation, is also compatible with the need of keeping the system in thermal stability, since the incidence angle of the Sun light on the solar screens remains constant, and the enclosure of the scientific instruments is always consistently shadowed.

The scanning law that results from the combination of the three rotations permits 60–100 observations of any direction on the sky. Each one occurs with different

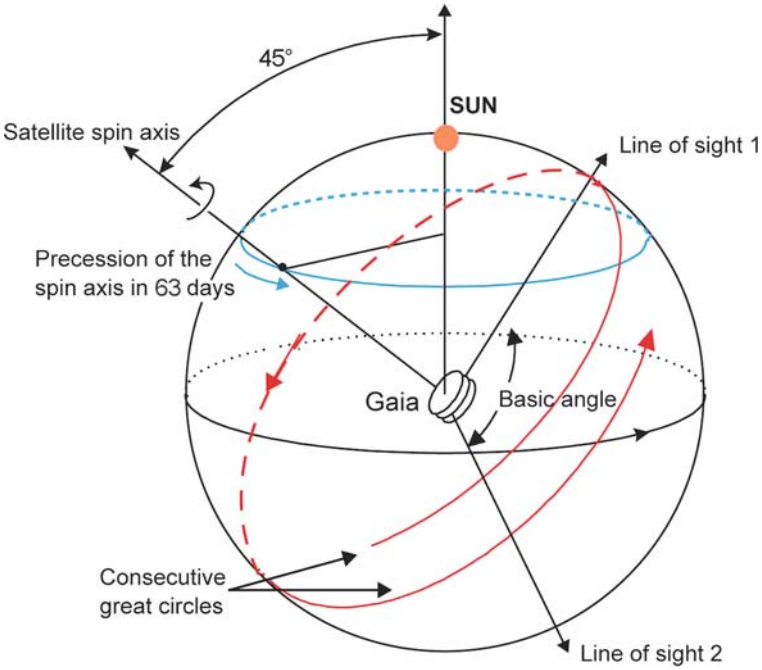

Fig. 3 The Gaia scanning law, composed of three rotations (spin, precession and orbital revolution) as explained in the text (Copyright ESA)

orientations of the scan circle, allowing to reconstruct the full two-dimensional position of fixed sources from positional measurements that are essentially one dimensional.

The availability of the observations in two simultaneous directions, and their multiplicity, has an important consequence: the star measurements contain both the information needed to map their position on the sky and that needed to reconstruct the orientation of the probe at any epoch. For this reason, the process of astrometric data reduction—the so-called Global iterative solution—is an inversion procedure allowing to retrieve at the same time the parameters that define star positions, their proper motion and the attitude of the probe. The applicability and performances of this strategy have been fully proved in the previous Hipparcos mission.

4 Solar System Science

While scanning the sky, sources corresponding to solar system objects (planets, dwarf planets, asteroids, comets, natural satellites, etc.) will enter the Gaia field of view and will be detected and recorded. The main difference with respect to stars will come from their motion. Their displacement on the sky has two main

consequences: they will not be re-observed at the same position; their motion will not be negligible even during a single transit.

These two basic statements are sufficient to dictate the need of a special data treatment for these objects. Several specific problems can thus be identified at first order, deserving an appropriate data reduction chain; we cite here:

– image smearing during integration time;
– signal shape due to resolved size and/or shape;
– de-centring relative to CCD windows or total loss during one transit;
– identification of new objects from detections at different epochs.

These issues are strictly related to the basic characteristics of the mission, but also other challenges are present when the science content to be extracted is considered. They will be discussed in the following sections.

A specific management activity for Solar System data reduction has been created in the frame of the DPAC (Data Processing and Analysis Consortium), as a part of the "Coordination Unit 4" devoted to process specific objects needing special treatment (double stars, exoplanet systems, extended objects, solar system objects).

Following the most recent mission specifications, one can identify the categories of objects that will be really observed by Gaia. In fact, all sources that will appear larger than $\sim$ 200 milli-arcs (mas) will probably be discarded and have no window assigned. This selection automatically excludes from observations the major planets, some large satellites (such as the Galilean satellites of Jupiter or Titan) and also the largest asteroids (or dwarf planets) when closer to Earth (Ceres, Pallas, Vesta, in particular).

On the other hand, small planetary satellites very close to major planets will be accessible, thanks to the low level of contamination from light scattered by the nearby planet. However, the vast majority of the observed objects will consist of asteroids of every category: mostly from the main belt, then small ones belonging to the near-Earth object population, and some additional tens among Jupiter Trojans, Centaurs and trans-Neptunians. Using the most recent survey data, we can estimate a population of $\sim$ 300,000 asteroids to be observable by Gaia, representing just 1/4000th of all the sources that Gaia will measure.

Each asteroid will be observed (by the AF) $\sim$ 60–80 times during the nominal mission operational lifetime of 5 years, although the number of detections can be much lower for near Earth-objects, sensibly depending on the geometry of the observation. On average, we can estimate that no less than 1 asteroid will enter the Gaia astrometric instrument when the viewing direction is close to the ecliptic.

Most asteroids will be known when Gaia will fly, so the discovery potential of Gaia remains low and does not constitute in itself a reason of special interest for the mission. One exception can probably be represented by specific object categories easily escaping most ground-based surveys, such as low-elongation inner-earth objects.

In fact, the geometry of the observations relatively to the positions of the Sun and the Earth can be easily estimated since most of the solar system objects orbit

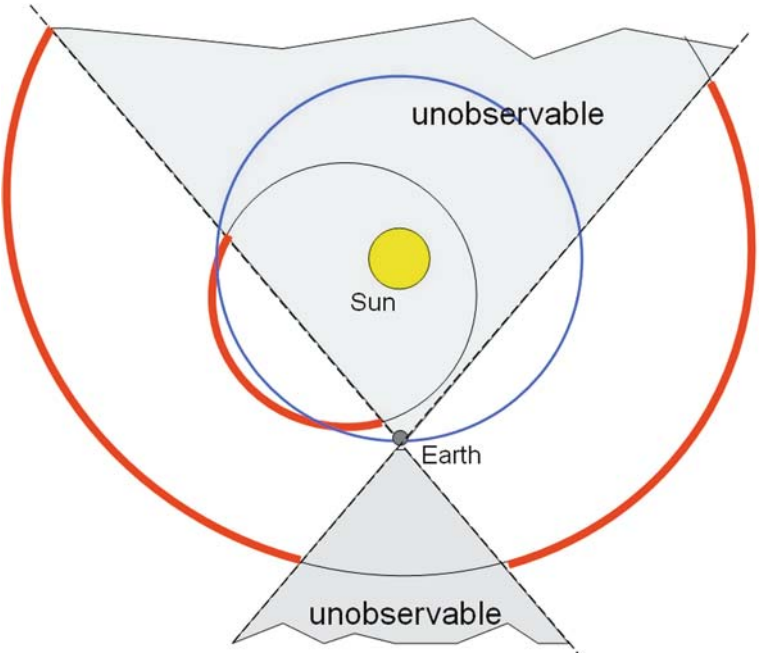

Fig. 4 At any given position along Earth orbit, the sky not accessible to the Gaia scan motion is delimited by two cones centred around opposition and conjunction with the Sun. When projected on the ecliptic, the unobservable region appears as the two *grey regions* in this picture. The observable portion of the orbit of a NEA and an MBA is enhanced

the Sun at very low ecliptic inclinations. It is thus straightforward to use the relevant angles to identify the regions of the ecliptic plane that can be visited by the Gaia scanning circle. The result is shown in Fig. 4 where the dashed sectors represent the viewing directions that are compatible with the scanning law. As one can see, Gaia will observe neither at the opposition nor towards conjunction, but mostly around quadrature.

More precisely, while quadrature represents the average direction, it would not be the most frequent one, since the intersection of the scan plane with the ecliptic spends most of the time preferentially close to the extremes of the accessible region.

The region at $\sim 45^\circ$ elongation from the Sun will thus be explored and could represent the most fruitful area for discovery purposes. Another poorly known population, not easily accessible from the ground, is the one represented by asteroid satellites. In fact, binaries larger than $\sim$120 mas will be seen as separate sources by Gaia, and today it is hard to estimate how many of them will be discovered.

However, we stress that the full physical and dynamical characterisation of known objects is the much more ambitious and rewarding goal that is expected from the Gaia mission. In fact, preliminary studies have shown that the precision of Gaia observations (both for photometry and astrometry) will be able to improve by more than two orders of magnitude the quality of asteroids orbits [110], to derive a mass

from mutual perturbations for ∼150 asteroids [75], to derive for most of them a shape and the rotational properties by lightcurve inversion [20], to measure general relativity PPN parameters [47], to directly measure asteroid sizes [25], to constrain non-gravitational accelerations acting on Earth-crossers [111]. We will detail in the following the most relevant of these issues.

As a result, we can already guess today that Gaia will open new perspectives for a better understanding of the Solar System—of asteroids in particular—portraying in a self-consistent way both dynamical and physical properties. Subsequent studies will take profit of the situation, for example, by rebuilding observational approaches that exploit better orbit and size and shape data (see, e.g., the asteroid occultation case [109]).

5 Analysis of the Astrometric Signals

5.1 Introduction

The CCD signal of a minor body transiting in the field of view of Gaia is characterised by some features that make it different from the ideal signal produced by a fixed star, and for this reason it requires a special treatment with respect to the standard processing pipeline adopted for stars.

First, it is important to understand what we mean in this lecture when we speak of "fixed stars". A fixed star is an ideal point-like source whose motion on the celestial sphere can be considered nil during the time taken by the source to make a transit across the Gaia field of view (FOV). In practical terms, the celestial objects that most closely fit the above definition are not stars, but distant quasars.

It is obvious that the property of being a non-moving source refers to the celestial sphere and not to the Gaia focal plane, since the Gaia field of view continuously changes while the satellite makes its scan of great circles on the celestial sphere (as a first approximation). However, as we will see in the following discussion, it may be said that the apparent motion of a fixed star in the Gaia FOV is at least approximately cancelled out by the adopted process of signal acquisition.

In general, a point-like source produces a photon flux distribution across the focal plane that, for a rectangular pupil like that of Gaia, is essentially the Fraunhofer diffraction pattern corrected for the aberrations introduced by the instrument optics itself. It is not our intention to discuss here all the details of the optical configuration of Gaia, but it is necessary to point out at least a few basic concepts. First, the response of the instrument to the distribution of photons produced by a point-like source is called point spread function (PSF). The PSF is necessarily the starting point to develop any description of the astrometric performances expected from Gaia. More in particular, the diffraction pattern produced on the focal plane by a point-like source *not moving across the field of view* is known as the *optical* PSF. By "diffraction pattern" we mean the bi-dimensional spatial distribution of the incident photons on the focal plane per unit area per unit of time. This includes the

aberrations introduced by the instrument optics, the transmittance of the instrument, the response of the CCD detectors, and it depends also on the spectral distribution of the source. In particular, if we call "quasi-monochromatic optical PSF" the optical PSF produced by a source for which the spectral distribution of its emitted radiation is limited to a very small interval of wavelengths, then we may define a polychromatic optical PSF as the average of different monochromatic optical PSFs, each one weighted according to the spectral distribution of the source.

The situation is more complex if we consider the effects related to the actual area of the focal plane (that in the case of Gaia is quite large). Distortions introduced by the instrument aberrations depend on the position of the source in the field of view. Sources in different positions in the field of view produce PSFs centred in different locations in the focal plane. Then, PSFs detected in different points of the focal plane, even if produced by the same kind of source and spectral distribution, show different features.

Moreover, due to the fact that in the adopted Gaia instrument configuration the signals are detected by an array of distinct CCDs, each having a well-defined quantum efficiency (QE) which is wavelength-dependent, it follows that the different spectral components of the PSF are detected with different efficiencies. For this reason, when we talk about the PSF, or in general the signal, of a generic source, it is more convenient and straightforward to refer to the number of photoelectrons produced by the CCD detectors per unit of time instead of the number of incident photons per time unit on the detector itself. In other terms, in the rest of this lecture we will call PSF the diffraction pattern including the effect of CCD quantum efficiency. It is worthwhile to point out that the QE of the CCDs in the Gaia astrometric focal plane is nominally the same for all of them and that no additional filters are placed in front of them.

As a matter of fact, however, the PSF of fixed stars is even not observable in the sense explained above, due to two fundamental reasons. First of all, the spatial resolution of the detector is finite because so is the dimension of the CCD pixels. As a consequence, the detector does not give the number of photoelectrons generated in any arbitrary point of the CCD per unit area and unit time, but it gives the total (integrated) number of photoelectrons collected by each pixel during the integration time. In other terms, the optical PSF has to be integrated within the rectangular area of the pixel and within the duration of the integration time. The second reason why the optical PSF is not directly observable is that, due to the complex motion of the satellite which allows it to scan the celestial sphere, any source moves within the field of view during the time in which it produces photoelectrons in the focal plane (source transit in the Gaia FOV). As a consequence, the PSF moves on the focal plane. Accordingly, the PSF is collected by all pixels and CCDs on which the PSF itself transits. Due to the effect of the apparent motion of the source in the FOV, the detected PSF should be smeared along the path that the image follows on the focal plane. In order to avoid this very undesirable effect, a transfer of charges from pixel to pixel towards the readout register is done at the same velocity and along the same direction of the apparent image motion. This technique of readout of the CCD is known as time delayed integration (TDI) mode. In TDI, the generated

photoelectrons "follow" the portions of the PSF that produced them. Ideally, if the motion of the charges followed exactly the same trajectory as that of the image in the focal plane and at the same continuous rate, the spatial distribution of the recorded photoelectrons would reproduce exactly the PSF. Unfortunately, this is not strictly true in practice, for the charge transfer is not a continuous process. What happens in practice is that the motion of the collected photoelectrons is produced by transferring all charges within one pixel into the adjacent one in the along-scan direction at regular time steps. The duration of the time step, indicated as the TDI period, is set in order to minimise the spread of the PSF due to the source motion. Nevertheless, the TDI mode cannot guarantee a perfect cancellation of the PSF smearing. The reason is that, even if the TDI period is kept constant and the charge transfer is done along the instantaneous scan direction, what happens is that the motion of the image is neither uniform nor aligned exactly in the along-scan direction. On one hand, this is simply due to the precession motion of the spin axis of the satellite, which is constantly changing the orientation of the along-scan direction. But even in the case that the motion of the image on the focal plane was perfectly uniform and exactly aligned with the along-scan direction, nevertheless the TDI mode could not correct completely the PSF smearing. The reason is that during each single TDI period the charges are not transferred, whereas the image moves and the signal is smeared. In this way, photons coming from the same point of the optical PSF may be collected by two different pixels. The final effect is that the PSF is broader and systematically displaced in the along-scan direction.

In principle, the TDI period should be equal to the along-scan dimension of the pixels divided by the image velocity. Non-uniformity of the along-scan motion introduces further distortion due to deceleration and acceleration of the image with respect to the TDI mean motion.

The PSF including the finite size of the pixels and the effect of the TDI mode is called *total* or *effective* PSF. But the real signal is affected also by other sources of noise and distortion. First of all, the signal is produced by a stochastic process, corresponding to the random sequence of photon arrivals and photoelectron generation. Then, the final number of charges collected in a given pixel is a random number, typically governed by the Poisson statistics. Moreover, the signal is distorted by the effect of cosmic rays on the CCD, affecting also the charge transfer efficiency (CTE) from one pixel to another. Some studies point out that some packets of charges can be entrapped and be released later without correlation with the regular TDI mode. All these effects introduce distortions of the final signal that are under investigation.

The treatment of the signals produced by solar system objects (SSOs) is even more difficult than in the "simple" case of fixed stars described above. SSOs are different from stars, in that they are characterised by a much faster apparent motion. When we say that SSOs move we intend that they display a motion with respect to the fixed stars. As a consequence, the trajectory of SSO images across the focal plane does not follow the same pattern of fixed stars in the same field of view and their motion is not properly corrected by the TDI mode.

SSOs have a residual velocity with respect to the apparent motion of a fixed star, and this residual velocity has components both on the along-scan and the

across-scan direction. This fact has the consequence that their final signals are spread. Moreover, while we may expect that for fixed stars the time interval between the transit on a given CCD and the adjacent one in the along-scan direction is constant, being about equal to the number of pixels in the along-scan direction in a CCD multiplied by the TDI period, this is no longer valid for a SSO. The transit of a SSO image on a CCD is either delayed or anticipated with respect to that of stars, depending on the residual velocity.

Velocity is not the only one peculiar features of SSOs to be taken into account. Another effect is that these objects may appear in general, or very often, as extended sources. This means that the final signal is not simply due to the PSF as for the case of point-like stars, but each point of a SSO image produces an independent PSF, and the final signal is the sum of all these PSFs. This means that the signal of a SSO depends on the size, the shape and the brightness distribution of its image.

For all the above reasons the analysis of the signals of SSOs requires a special treatment with respect to the simpler one used for fixed stars.

5.2 Signal Computation

Let x and y be the coordinates of a Cartesian system associated to the focal plane, having the x-axis set along the scan direction and directed as the movement of charges of the TDI mode. Let $I(x, y)$ be the photoelectron distribution produced by the optical image of a non-moving source, that is, the number of photoelectrons generated per unit area and unit time around the point of coordinates (x, y). The function $I(x, y)$ contains information about both the PSF photoelectron distribution $f(x, y)$ and the apparent brightness distribution of the observed object $g(x, y)$. From a mathematical point of view, I is the convolution of f and g:

$$I(x, y) = \iint g(x', y') f(x - x', y - y') \, \mathrm{d}x' \mathrm{d}y'. \tag{1}$$

For the sake of brevity multiple integrals signs will be omitted in the following. Let now τ be the TDI period. The signal, that is, the number of photoelectrons collected by a pixel whose centre has coordinates (α, β), during the interval of time τ, from a source image with centre located at coordinates (x_c, y_c), is given by:

$$S(\alpha, \beta) = \int I(x - x_c, y - y_c) \Pi\left(\frac{x - \alpha}{\Delta x}\right) \Pi\left(\frac{y - \beta}{\Delta y}\right) \Pi\left(\frac{t}{\tau}\right) \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t, \tag{2}$$

where Δx and Δy are the along-scan and across-scan dimensions of the pixels, t is the time and $\Pi(u)$ is the gate function equal to 1 for $-1/2 \leq u \leq 1/2$ and zero otherwise.

If now we assume that the image source moves with constant velocity, so that

$$x_c = \dot{x}t + x_0 \qquad y_c = \dot{y}t + y_0,$$

the signal becomes:

$$S(\alpha,\beta)=\int I(x-\dot{x}t-x_0,\,y-\dot{y}t-y_0)\Pi\left(\frac{x-\alpha}{\Delta x}\right)\Pi\left(\frac{y-\beta}{\Delta y}\right)\Pi\left(\frac{t}{\tau}\right)\mathrm{d}x\;\mathrm{d}y\,\mathrm{d}t. \tag{3}$$

Two comments are necessary at this point. For fixed stars it is assumed that $\Delta x = \dot{x}\tau$, but for SSOs this is not the case in general. In particular, if v_τ is the mean transfer velocity of the charges, so that $v_\tau = \Delta x/\tau$, the relevant parameter for SSOs is the residual along-scan velocity $\mu = \dot{x} - v_\tau$. Second, at every time interval of duration τ the charges in a pixel are moved to its immediate adjacent pixel. In this process $\alpha \rightarrow \alpha + \Delta x$, while at the same time $x_c \rightarrow x_c + (\mu + v_\tau)\tau = x_c + \Delta x + \mu\tau$ and $y_c \rightarrow y_c + \dot{y}\tau$. For fixed stars we have $\alpha - x_0 =$ constant, assuming an ideally perfect clocking, then in this case the total number of electrons recorded during the integration time T is simply T/τ times the number obtained over one pixel. This is not true for SSOs, and the computation of the integral has to be done for the entire integration time. For sake of simplicity the computation can be done under the assumption that the optical PSF is locally constant, or in other words that we may use the same PSF over one whole CCD.

In practical terms, the preferred option to attack the problem of signal computation has been so far based on a different, numerical ray-tracing approach. Let us then come back to consider the incoming photons from the source, before they start to interact with the optical system to produce the recorded image. We may imagine a coordinate system following the moving object and with origin in (x_c, y_c) and axes parallel to the along- and across-scan directions. In this reference system, the photons coming from the source would produce an image in an hypothetic detector, and the coordinates of a given point belonging to this (static) image would be x' and y'. Now, we can easily compute the corresponding coordinates (x, y) in the moving Gaia focal plane of the photons coming from the (x', y') point just defined above. These are:

$$\begin{aligned} x &= x' + \dot{x}t + x_0 \\ y &= y' + \dot{y}t + y_0. \end{aligned} \tag{4}$$

In the adopted method of image simulation, based on a Monte Carlo algorithm, all the quantities x', y' and t are randomly generated. The time t is uniformly generated between $-T/2$ and $T/2$ where T is the integration time, while x', y' are randomly generated according to a given surface luminosity distribution $g(x, y)$ of the object (chosen a priori), using a ray-tracing algorithm which also takes into account an assumed light scattering law characterising the object's surface. For the details of the numerical algorithm of generation of the sampling positions x' and y' see [26]. The number N of sampling points (x', y') depends on the magnitude of the object and on instrument characteristics including the CCD quantum efficiency. In particular, we have that $M = -2.5 \log N + C$, where M is the apparent magnitude and C is a suitable constant. At magnitude 12, corresponding to the saturation limit of the CCD in astrometric focal field, the number of collected photoelectrons are

about 1 million. The nominal magnitude limit of detection is $V \approx 20$ and corresponds to about 1000 photoelectrons [89]; see also Sect. 5.4.

Of course, the above optical image, or in other words the (x, y) distribution of incoming photons from the observed object, is not yet the final recorded signal. We have still to take into account the fact that each photon incident on the optical system suffers an angular deviation δx and δy, respectively, in the along-scan and across-scan directions due to the diffraction of the instrument aperture. The numerical model reproduces the instrument diffraction effect by generating δx and δy randomly, but according to a parent distribution given by the PSF of the system. In other words, the probability of a given deviation is proportional to the value of the PSF $f(\delta x, \delta y)$ for this particular deviation. In order to obtain the real arrival position of a single photon on the focal plane, we have then to add δx to x and δy to y. Obviously all the quantities x', y', δx and δy are expressed with the same unit.

As mentioned in the previous section, the spreading of the recorded photoelectrons in the along-scan direction on the focal plane due to the scanning motion of the satellite is reduced by the time delayed integration (TDI) readout mode. For this reason we introduce a moving coordinate system (X, Y) synchronised with the TDI charge transfer. Obviously $Y = y$ because no TDI correction is performed in the across-scan direction. In this way, the "effective" position of the photoelectron inside the CCD image turns out to be:

$$\begin{aligned} X &= \theta(x' + \delta x + \dot{x}t + x_0, t), \\ Y &= y' + \delta y + \dot{y}t + y_0, \end{aligned} \tag{5}$$

where θ is a special function accounting for the step-by-step TDI translation, and depending on time t explicitly, according to the phase of the TDI charge transfer. In practical terms, however, TDI blurring is included in the adopted PSF, so that the θ function reduces to a simple continuous translation (see [26] for details). In this case, we may simply write down the following:

$$\begin{aligned} X &= x' + \delta x + \dot{x}t + x_0, \\ Y &= y' + \delta y + \dot{y}t + y_0, \end{aligned} \tag{6}$$

where $\dot{x}$ and $\dot{y}$ represent the residual velocity of the object with respect to the TDI motion. For SSOs, this corresponds in practice to the apparent motion with respect to the fixed stars.

The final step of the numerical procedure is the allocation of the photoelectrons into the corresponding pixels. The Gaia CCD detectors have rectangular pixels with smaller side $\Delta X = \Delta x$ in the along-scan direction and larger side $\Delta Y = \Delta y$ in the across-scan direction. Each pixel is labelled with integer indexes i and j and corresponds to the rectangular region defined by the points (X, Y) such that

$$X_i \le X < X_i + \Delta X\,;\; X_i = i\Delta X \qquad Y_j \le Y < Y_j + \Delta Y\,;\; Y_j = j\Delta Y \tag{7}$$

in this way the origin of the coordinates system coincides with the left-hand, lowest corner of the pixel labelled as $i = 0$, $j = 0$.

In the astrometric focal plane of Gaia only a limited region of the CCD grid is actually readout. This window is, again, a rectangular $n \times m$ one, with a number n of pixels in the along-scan direction and m pixels in the across-scan direction. In this way $0 \leq i \leq n - 1$ and $0 \leq j \leq m - 1$. Following the sampling scheme proposed by [54], 12×12 pixels windows should be used for stars between 12 and 16 magnitude (in the G-band of Gaia) and 6×12 windows for stars between 16 and 20 magnitude. We indicate by N_{ij} the number of photoelectrons in the ith pixel of the window in the along-scan direction and the jth in the across-scan direction.

It is important to note that, in general, the recorded signal is binned, that is, the numbers of pixels in the across-scan direction are integrated (summed up), so that the final signal consists of the n numbers:

$$N_i = \sum_{j=0}^{m-1} N_{ij} \qquad 0 \leq i \leq n - 1. \tag{8}$$

In the following, we will assume that CCD window is always binned, and what we call "signal" is actually given by the set of discrete photoelectron counts just defined above.

5.3 CCD Processing

In general terms, we call "CCD processing" a sequence of numerical procedures aiming at extracting from the recorded signal N_i all possible information of interest. In particular, in the case of SSOs, four parameters are of great importance: the position of the object, its angular size, its velocity and its apparent magnitude. Basically, both for stars and SSOs, the adopted method of determination of all the relevant parameters is based on a best-fit procedure.

It is assumed that a mathematical model of the signal is at disposal, or in other terms that we are able to compute the numbers N_i for any set of values of the unknown parameters. This computation can be done analytically or throughout numerical codes. In addition to a signal model, one has to decide a criterion of comparison between the computed (C) and the observed (O) signal, in order to minimising or maximising some target function expressing the mathematical distance between C and O.

We note that, in principle, an alternative approach, based on a full reconstruction of the image, could also be adopted. A number of refined techniques of image reconstruction have been actually developed in the literature in several situations. On the other hand, it seems that such kind of approach can hardly be applied with success to poorly sampled signals covering only a few pixels, like those from faint, nearly point-like sources detected by Gaia. For this reason, we focus on the approach

based on the determination of a limited number of relevant unknown properties of the sources by means of a signal best-fit approach.

For sake of simplicity, in order to focus on a few fundamental concepts rather than on unnecessary technical details, let us limit now to a one-dimensional case. Let L be the model function used to fit the recorded signal. For a fixed star this reduces to the PSF integrated in the across-scan direction and convoluted with the Π function describing the shape of the pixel. This leads to derive the so-called line spread function (LSF). For fixed stars, this depends only on the position c in pixels of the star and on its apparent magnitude or, in other words, on the total number of collected photoelectrons. The signal is proportional to the flux and can be reproduced by translating the centre of the LSF into the star position. If we call L the LSF per unit of flux (that is, normalised to one collected photoelectron) the expected number of photoelectrons in the pixel i, $E(N_i)$, is given by:

$$E(N_i) = r^2 + b + NL(i - c), \tag{9}$$

where N is the total number of photoelectrons produced by the source and r is the RMS readout Poisson noise in electrons. The term b is the contribution of the background in electrons per sample. We assume that $r^2 + b$ is known with sufficient accuracy.

If N_i is the number of electrons really collected in the sample i, a method of estimation of the unknown parameters c and N, corresponding to the unknown photocentre position and apparent magnitude of the source, can be based on a maximum likelihood criterion. More precisely, c and N are varied in order to maximise the probability of collecting the observed counts N_i, assuming the signal model L. For Poisson statistics this probability is given by the likelihood function:

$$f(c, N) = \prod_{i=0}^{m-1} \frac{[r^2 + b + NL(i - c)]^{N_i}}{N_i!} \exp\left[-(r^2 + b + NL(i - c))\right]. \tag{10}$$

The numerical problem is then to find the values of c and N that maximise $f(c, N)$.

Without entering into too many details, let us take into account the case of a SSO. In this case, the signal model does not depend only on the parameters c and N. Assuming that the object has a known shape and brightness distribution, we have to introduce as additional parameters its size s in pixels and its velocity v. As we have already discussed above, the meaning of v is the difference between the velocity of the image on the focal plane and the mean TDI transfer velocity of the electrons. In this case the likelihood function becomes:

$$f(c, s, v, N) = \prod_{i=0}^{m-1} \frac{[r^2 + b + NL(i - c, s, v)]^{N_i}}{N_i!} \exp\left[-(r^2 + b + NL(i - c, s, v))\right]. \tag{11}$$

For fixed stars the TDI motion is set according to the image motion, so on the average the position c is at rest. In the case of a moving source, however, this is

no longer true. For this reason it is necessary to specify what we mean by c for a moving source. In particular, the meaning of c must be that of the position of the image at a particular epoch, for example, at the epoch of the signal readout.

We have just assumed that the brightness distribution of the source is known and the only parameter to be determined is its size s. This can be a critical point for some applications to SSOs and in particular to asteroids. In general, each asteroid has its own shape, spin axis orientation and rotation period. Moreover, when it is observed at some given epoch it is seen under some illumination condition, quantified by the value of the so-called phase angle. The phase angle is defined as the angle between the directions to the observer and to the Sun, as measured from the asteroid barycentre. We note that Gaia will never observe asteroids at zero phase angle, corresponding to object opposition from the Sun, but at phase angles larger than 10°, as a rule. This means that, due to the defect of illumination, the position of the "photocentre" of the collected signal will not be coincident with the position of the (sky-projected) barycentre of the object. This effect will have to be taken into account for the purposes of using Gaia astrometric measurements of asteroids to obtain refined orbital elements.

In addition to the above problems, the surface of any asteroid is characterised by its own reflectance properties that may well be different for different objects and also may vary, in principle, from point to point of the same object, determining an apparent brightness distribution which will vary depending on the observing circumstances. Finally, other problems may also arise in the cases of non-resolved binary systems and of objects characterised by the presence of some kind of cometary activity (and Gaia will certainly observe comets).

Of course, the above-mentioned effects make the CCD processing of SSOs a quite challenging task. As opposite to other kind of sources, the signal of any asteroid transiting in the field of view is inherently different with respect to the signals collected for the same asteroid during different transits, corresponding to different observing circumstances. All this makes SSOs completely different for what concerns the CCD processing and reduction of the collected astrometric signals, with respect to the case of fixed, point-like stars.

5.4 Accuracy Estimation

In order to estimate the accuracy of the measurement of the most important parameters, that for asteroids include apparent magnitudes, motions, positions and sizes, it is necessary to take into account all the disturbing factors that affect the generation of the recorded signal. The main sources of noise include photon statistics, CCD readout noise (RON), dark current, the noise introduced by the background, errors in the calibration of the PSF, perturbation of the electronic devices due to the satellite's environment, like the radiation damage of the CCDs produced by cosmic rays.

Some of these noise sources are still under investigation, including PSF calibration and CCD radiation damage, whereas for others an estimation is already available. For what concerns the RON, its total amount should be about 4 photoelectrons

(e^-). Including also the dark noise and other electronic disturbing factors, the total detection noise (TDN) should be of the order of 6–7 e^- for each CCD in the focal plane. This means that in the signal acquisition process the counted numbers N_i are affected by an uncertainty of this order.

Background contribution to the signal deserves a separate discussion. Sky brightness does not introduce an error in the photoelectrons readout, but rather an additional contribution of some number of photoelectrons to be added to the number of photoelectrons produced by the source. According to the *HST/WFPC2 Instrument Handbook*, the sky brightness at high ecliptic latitudes should be around 23.3 mag.arcs^{-2}, while it is around 22.1 mag.arcs^{-2} on the ecliptic plane, because of the zodiacal light [89]. According to the most updated configuration of the instrument [102], taking into account the size of the CCD pixels in the astrometric field (10×30 μm), the focal length of the telescope (35 m), the pupil aperture (1.45×0.50 m), and the CCD integration time (4.42 s), we expect a background contribution of about 2 e^-/pixel on the ecliptic and 0.6 e^-/pixel at high latitude. These two values should be considered as the extreme limits of background contribution per CCD.

The major source of noise is certainly due to photon statistics. The total number N of collected photons depends on the source's magnitude M, and $M = -2.5 \log N + C$, where C is a constant (neglecting the effect of the windowing cut-off). Visual magnitude $M = 0$ corresponds to an energy flux of 2.52×10^{-8} Wm^{-2}. Considering a mean wavelength of collected photons about 550 nm, this energy flux corresponds to a photon flux around 7×10^6 photons/cm^2/s. Taking into account an integration time of 4.42 s and an aperture of 1.45×0.50 m^2, the number of incoming photons should be around $1.6 \times 10^{11} 10^{-M/2.5}$ or $N_s \sim 10^{11} 10^{-M/2.5}$ including a factor $1/2$ due to CCD quantum efficiency. So we expect about 10^6 photoelectrons at magnitude 12, corresponding also to the saturation level of the CCDs, and about 10^3 at magnitude 20, that is, the nominal detection limit.

A quick estimation of the error in the position measurement can be done in the following way. Let us assume for simplicity the case of the signal from a nearly point-like, non-moving source. Its signal corresponds more or less to the instrument LSF. As we have explained above, the signal can be regarded as the distribution over a few pixel rows of the recorded photoelectrons. Then, the mean value of the photoelectrons' positions can be used as an estimator of the image position. As it is well known from statistics, the standard deviation of the mean is equal to the standard deviation of the distribution divided by the square root of the number of samples, that is, in this case, the number of the collected photoelectrons. The standard deviation of the LSF is about 2–3 pixels. Thus, at magnitude 12, which gives 10^6 electrons as seen above, we expect that the standard deviation of the mean is about $2–3 \times 10^{-3}$ pixels ($\sim$ 0.1–0.2 mas). In the same way, at magnitude 20, when we expect to collect more or less 10^3 electrons, the accuracy of the mean should be about 0.05–0.1 pixels ($\sim$ 3–6 mas).

We note that the above values for the effective uncertainty in the photocentre determination of SSOs are much higher than the final, end of mission uncertainty in the determination of the positions of single stars. The reason is that for SSOs, which

move and are seen in different observational circumstances in different transits on the Gaia focal plane, we cannot merge together the results of different detections, and we are forced to limit ourselves to derive separately the positions of the objects at different transits based on signals collected at these single transits. As opposite, the determination of the position of the stars will be performed by cross-matching the information derived from the measurements of tens of different transits. For this reason the final accuracy for stars at magnitude 12 is expected to be of the order of 10^{-4} pixel (see [89]), corresponding to about 0.004 mas.

The accuracy in the determination of the angular size of SSOs may be estimated by means of some numerical tests based on a simplified method of measurement. We have seen that the position of the object is related to the *mean* of the spatial distribution of the recorded photoelectrons. In the same way, the angular size of the source is related to the *standard deviation* of the same photoelectron distribution. Let us assume for simplicity that the shape of the object of which we want to measure the angular size is a perfect sphere with a uniform brightness distribution. Let us assume also that the object is observed at zero phase angle. In other words, we assume that the object appears as a flat, uniform disc in the sky. In order to avoid complications with the truncation of the signal, we assume also that the object's residual velocity is zero. This means that the final signal is well centred and it is not truncated. If the number of collected photons was infinite, then a deterministic relationship between the standard deviation σ of the signal and the diameter D of the disc would exist. So, by computing the expected standard deviation for each possible value of the disc's size, it should be possible to obtain the corresponding value of the size from the observed signal. In the real world, however, due to the fluctuations of the numbers N_i caused by photon statistics, to each value of the measured size of the disc may correspond different values of D. The distribution of these possible values is characterised by a mean and a dispersion around it. An example is given in Fig. 5, obtained by means of a numerical Monte Carlo simulation. The figure shows that a given value of σ can be produced by some interval of possible values of D. Using this kind of plot we can assess the expected accuracy in size determination.

In particular, the dispersions of the values of σ for a single value of the size D depend on the magnitude of the source, and it increases as the magnitude increases (or the flux decreases). Moreover, the relation between σ and D is not linear, but rather it can be written as:

$$\sigma^2 = \frac{D^2}{16} + \sigma_0^2,$$

where the first term is introduced by the angular distribution of the incoming photons and σ_0^2 is the variance introduced by the LSF. It follows that if we wish to estimate the value of D from the measured value of σ, the error $\mathrm{d}D$ is given by:

$$\mathrm{d}D = \frac{16\,\sigma\,\mathrm{d}\sigma}{D}, \tag{12}$$

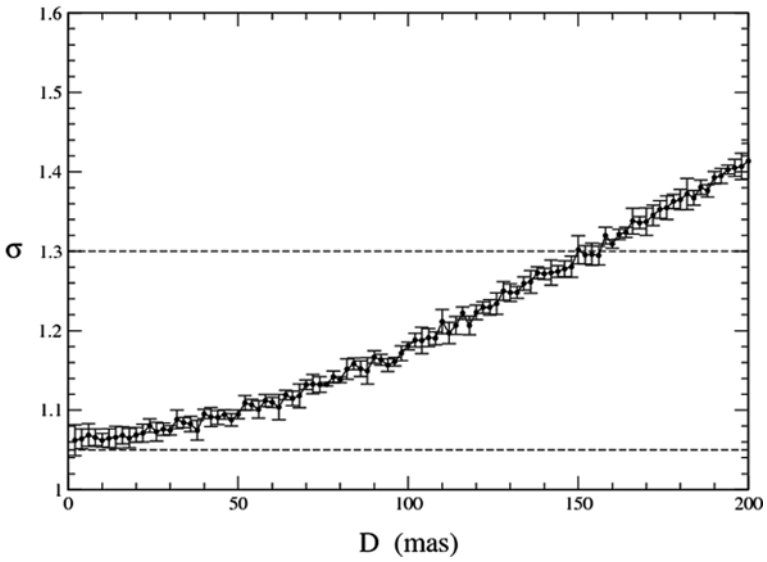

Fig. 5 Standard deviation σ (in pixels) of signals produced by an ideal spherical and homogeneously emitting object seen at zero phase angle, as a function of its apparent diameter D. The bar corresponds to the standard deviation of σ due to photon statistics

where $\mathrm{d}\sigma$ is the error, mainly due to the photon statistics, in the measurement of the standard deviation σ of the signal. So, for a given value of $\mathrm{d}\sigma$, the uncertainty of the size estimation increases as the size itself decreases. In other terms the slope $\mathrm{d}\sigma/\mathrm{d}D$ of the function $\sigma(D)$ decreases as D decreases, as it is clear from Fig. 5.

Since $\mathrm{d}\sigma$ is magnitude-dependent, it is possible to associate to any possible magnitude value a corresponding critical value of the size D for which the resulting relative error $\mathrm{d}D/D$ is equal to some given limit, like 10%, as an example. The result of this exercise is shown in Fig. 6. In the figure, the size limit corresponding to a relative size determination accuracy of 10% is plotted *versus* the magnitude. The domain of this plot below the 10% line corresponds to observational circumstances in which the image cannot be distinguished from that of a point-like source. At magnitude 12 it is possible to appreciate the apparent size for objects with a diameter of 20 mas, but at magnitude 20 this limit raises up to 150 mas. In conclusion, points below the 10% line in the figure correspond to observational circumstances for which the object is too far or is too faint for its angular size to be measured with an accuracy better or equal to 10%.

5.5 Size Measurement of Main Belt Asteroids

In order to assess the capabilities of Gaia in measuring the sizes of main-belt asteroids, we take advantage of existing simulations of observations of these objects by Gaia during its operational lifetime [72]. These simulations provide the list of

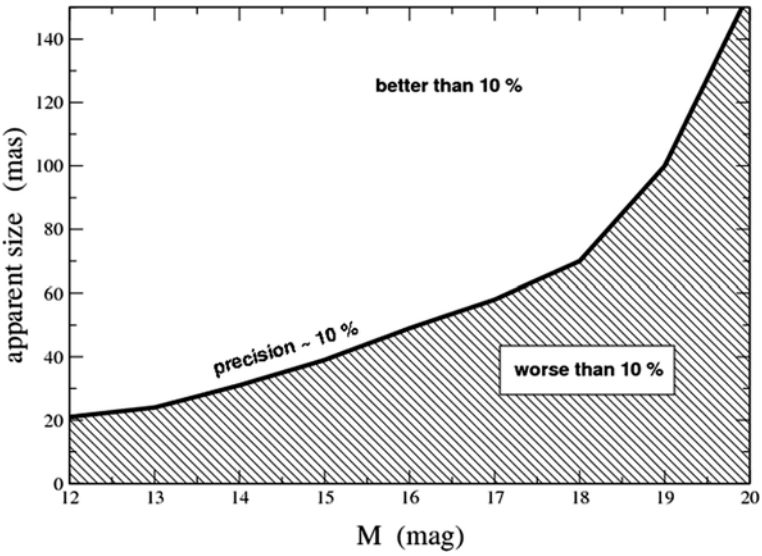

Fig. 6 The smallest size measurable with a precision of 10% plotted as a function of the object apparent magnitude

the transits of main-belt objects on the field of view of the instrument, specifying the distance r from the satellite and its apparent magnitude M. Among the full set of simulated observations, we selected only those of objects for which a value of diameter d is available (taken from the most recent issue of the IRAS catalogue of asteroid sizes and albedos [114]) and for which the apparent magnitude should range between 12 and 20. We assumed for sake of simplicity that the objects are spherical. The apparent size of the objects for each of the above observations will be obviously $D = d/r$. Having at disposal such a set of simulated observing circumstances for a large sample of objects (about 2000), it is possible to assess whether for each single detection the object's size can be determined with a precision better (*good observation*) or worse (*bad observation*) than 10% on the basis of the diagram in Fig. 6.

For any object, in general, some observations will be good and others will be bad according to the observing circumstances. Let S be the total number of observations of one single asteroid, and let s be the total number of *good* observations of the object, in the sense explained above. The resulting ratio $\rho = s/S$ can be taken as an evaluation of the efficiency of Gaia in measuring the size of this asteroid.

We show in Fig. 7 the result of this exercise. Each asteroid is plotted as a small cross in the efficiency – diameter plane. The solid line is the average value of the efficiency versus real diameter resulting from a running-box analysis. For asteroids having a size larger than 100 km the measurement efficiency is well above 50%, and almost all observations are good. Below 20 km no good observation is possible, because the objects are either too small or too faint or both (note the group of crosses

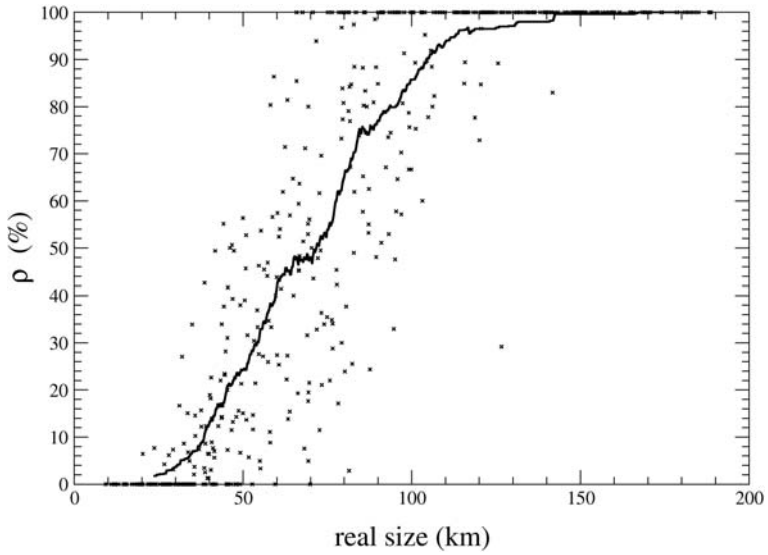

Fig. 7 Efficiency of Gaia in measuring the diameters of the main-belt asteroids

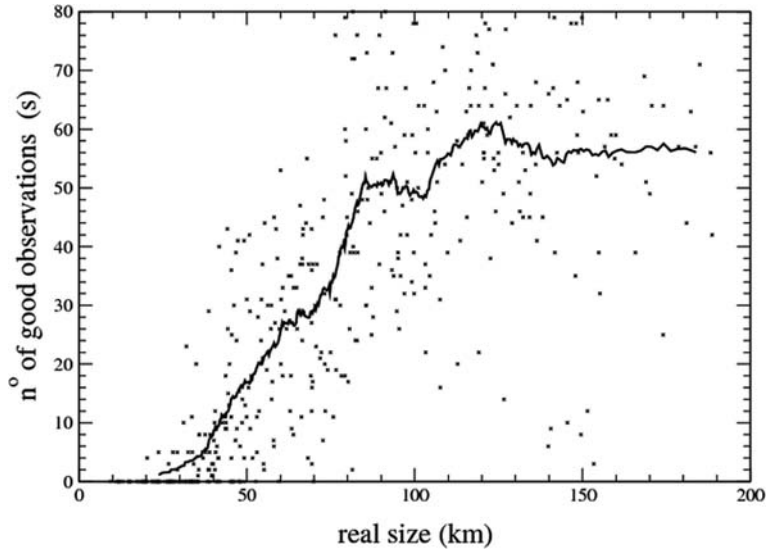

Fig. 8 Number of observations of main-belt asteroids of different sizes, allowing a size measurement with an accuracy of 10% or better

with $\rho = 0\%$). At the limiting size of 20–30 km, the measurement efficiency is only a few percent.

A slightly different plot is shown in Fig. 8, where the number of good observations s is plotted against the real diameter for each asteroid. Again, the solid line

is a running-box average. The crosses are rather scattered meaning that apart from the average scenario there are very different situations, depending on the individual orbital and physical properties of the objects. We note that in the case of very large asteroids the observation efficiency can decrease due to the paradoxical fact that in many cases they are too bright and they reach the CCD saturation limit of magnitude 12 (examples are the group of crosses in Fig. 8 at $D \simeq 150$ km and $s \leq 10$ corresponding to the asteroids 11 Parthenope, 18 Melpomene, 20 Massalia, 39 Laetitia, 89 Julia, 349 Dembowska).

5.6 Limits and Margins of Improvement

The results discussed in the previous sections have been obtained under some simplifying assumptions. In particular, we assumed that the objects had spherical shapes, and that the optical properties of the emitting surfaces were homogeneous. Or, in other words, we made the assumption that the surface albedo of the objects was homogeneous throughout the surface. Moreover, even if we did not mention this explicitly, the results shown in Figs. 5, 6, 7, and 8 were based on simulations in which some assumptions concerning the diffusive properties of the object surfaces were made. In fact, a well-defined light scattering law was assumed when running the ray-tracing part of the signal simulator. Without entering into details, the assumption was that of surfaces scattering the incident sunlight according to a composition of a Lambertian and a Lommer–Seeliger scattering law.

Now, we can ask ourselves whether the above simplifying assumptions are reasonable, and if there is some way to possibly improve the model. In this respect, the answer seems to be yes, and the way of improving the model is based on ancillary information that is expected to become available when the full set of recorded signals from each object at all transits collected during the Gaia operational lifetime will be available. In particular, the idea is that of taking profit of the analysis of the disc-integrated magnitude measurements performed at each object transit. The measurement of the apparent magnitude at each transit on the Gaia focal plane is a less complicated task with respect to the determination of the astrometric position and apparent size of the object, as seen in the previous section. Given the full set of measured apparent magnitudes of an object, it will be possible to derive from that a big deal of information concerning the rotational properties (spin rate and direction of the spin axis) and overall shape, assumed for simplicity to be that of a triaxial ellipsoid. The derivation of the above parameters is described in section 6 of this lecture. Having at disposal the object's pole direction and a more realistic triaxial shape, it will be possible to compute for each transit the corresponding observational circumstances in terms of apparent shape and orientation of the illuminated part of the body visible from Gaia at the epoch of the observation. In this way, a much improved object's model, with respect to that of a simple homogeneous sphere, will be adopted and used to derive refined estimates of the object's size and also of the offset between the position of the barycentre and that of the photocentre during the transits for which the object becomes resolvable.

As for the choice of the scattering law, the situation is intrinsically more difficult, yet not completely hopeless. In particular, it will be possible to take profit of the fact that at least for a few objects (433 Eros, 243 Ida, 951 Gaspra, 253 Mathilde, and probably in the near future 1 Ceres and 4 Vesta) we have at disposal data taken in situ by space probes. For these objects, we have very detailed information about size, shape, spin and surface properties. The idea is then that of using this ancillary information to possibly improve the adopted scattering laws in the reduction of Gaia data. In particular, the most correct scattering law, at least for the objects of the above list, should be the one producing a better agreement between the Gaia results and the known properties of the objects.

6 The Determination of Asteroid Physical Properties

6.1 Introduction

The determination of asteroid physical properties will be one of the fundamental contributions of Gaia to Planetary Science. As we have seen in Sect. 5, asteroid sizes will be directly measured for a number of objects that should be of the order of 1000, according to current signal simulations. This spectacular result, however, will be only one of a longer list that includes

- the measurement of about 150 asteroid masses;
- as just mentioned above, the direct measurement of about 1000 asteroid sizes;
- based on the measured masses and volumes, the determination of the average densities for about 100 objects belonging to practically all the known taxonomic classes;
- the determination of the rotational properties (spin period and polar axis orientation) and overall shapes for a number of the order of 10,000 objects;
- a new taxonomic classification based on reflectance spectra (including wavelengths in the blue region of the spectrum) obtained for several tens of thousands of objects;
- the measurement of the Yarkovsky acceleration for some tens of near-Earth objects.

To the items of the above list, we have to add, of course, the derivation of much improved orbits for a data set of about 300,000 objects, taking profit of the unprecedented astrometric accuracy of the Gaia mission. A detailed description of the expected performances of Gaia in the determination of orbits, masses and Yarkovsky acceleration is given in Sect. 8.

In what follows, we will focus on the remaining items of the above list, namely spin properties and taxonomy.

6.2 Inversion of Disk-Integrated Photometric Data

Historically, photometry has been one of the first observing techniques applied to obtain information on some physical properties of the asteroids.

Due to the fact that these objects have non-spherical shapes and their apparent brightness is due to solar photons scattered by the illuminated surface, there is a periodic modulation of the brightness due to the object's rotation. Photometric observations over a sufficiently long interval of time to cover a full rotation cycle produce then what is usually called a *lightcurve*, namely a plot of the apparent brightness as a function of time, which shows the periodic variation in magnitude due to rotation. Lightcurves provide thus in a straightforward way one of the many physical properties of an object, namely its rotation period.

The morphology of a lightcurve at a given epoch is the complex result of the shape of the object, the orientation of its rotation axis (the "asteroid pole") and the light scattering properties of the surface. The difference in magnitude between the maximum and the minimum brightness of the object during a rotational cycle is called *lightcurve amplitude*.

The fact that asteroids are moving objects on the celestial sphere has the consequence that the geometric configuration Sun–asteroid–observer changes for observations carried out at different epochs. The observing circumstances at a given epoch are characterised by different values of the heliocentric and geocentric distances of the illumination conditions, which are described by the so-called phase angle described in Sect. 5.3, and by two quantities that describe the orientation of the asteroid spin axis with respect to the direction to the observer and to the Sun. These two quantities are two angles: the *aspect angle*, normally indicated as ξ, and the *obliquity angle*. The aspect angle is the angle between the directions of the spin axis and the direction to the observer, measured at the object's barycentre. The obliquity angle is the angle between the plane containing the object, the observer and the spin axis of the object and the plane containing the asteroid-observer direction – which is perpendicular to the plane containing the observer, the object and the Sun. This angle is indicated as o in Fig. 9, which shows in a graphical way the meaning of the above angles. Finally, another angle ϕ describes the rotational phase of the object in its rotation around its axis. In practice, the values of all the above-mentioned angles reduce to the computation of the *sub-observer* (normally called *sub-Earth*) and *sub-Solar* points, namely the latitude and longitude coordinates of the two points which are the intersections of the body's surface with the vectors from the object's barycentre to the observer and to the Sun, respectively.

Due to the variations of the observing circumstances at different epochs, lightcurves taken at different oppositions of the same object will generally exhibit a variation in their morphologies, and in particular their amplitude. This is due to the fact that the aspect angle of an object is a function of the ecliptic longitude (or equivalently, the Right Ascension) at which it is observed at a given epoch. For instance, if we assume that the object has the shape of a triaxial ellipsoid with semi-axes $a > b > c$, spinning around the shortest axis c, the maximum lightcurve amplitude will be reached when the object is seen in equatorial view, when $\xi = 90^\circ$. This view

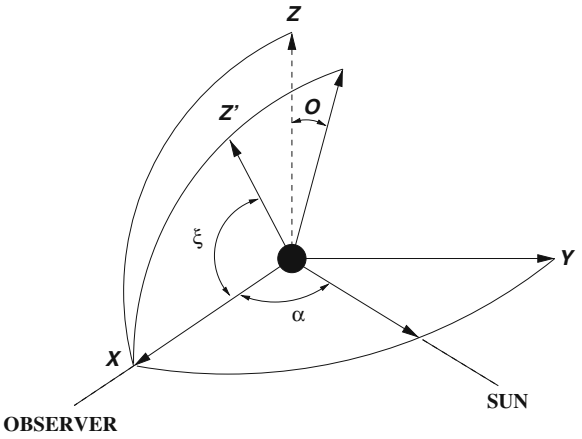

Fig. 9 Graphical explanations of the aspect (ξ) and the obliquity (o) angles. The angle α in this figure is the phase angle. Vector $\mathbf{Z}'$ is the direction of the positive spin

is in principle always reachable and corresponds to two well-defined values of the Right Ascension of the object (separated by an angle of 180°), which depend on the spin axis orientation. Just to make the things more clear, to observe an object at $\xi = 90^\circ$ means observing it when it reaches one of its two equinoxes.

We only remind that the choice that we make here of a triaxial ellipsoid shape with semi-axes $a > b > c$ is particularly suitable to illustrate the predictable magnitude variations. Triaxial ellipsoids seen at zero phase angle project in the sky an ellipse. The apparent surface S of this ellipse is given by the following relation:

$$S = \pi a^2 (A_1 \sin^2(\phi) + A_2)^{1/2}, \tag{13}$$

where ϕ is the rotation angle of the object, assumed to be zero at lightcurve maximum (when the projected ellipse has the major semi-axis equal to a). A_1 and A_2 are given by:

$$A_1 = \left(\frac{c}{a}\right)^2 \left[\left(\frac{b}{a}\right)^2 - 1\right] \sin^2(\xi) \tag{14}$$

and:

$$A_2 = \left(\frac{c}{a}\right)^2 \sin^2(\xi) + \left(\frac{b}{a}\right)^2 \cos^2(\xi). \tag{15}$$

It is easy to verify that, at the maximum and the minimum of luminosity, the projected areas of the ellipse are, respectively,

$$S_{max} = \pi a \cdot \sqrt{b^2 \cos^2(\xi) + c^2 \sin^2(\xi)} \tag{16}$$

and

$$S_{min} = \pi b \cdot \sqrt{a^2 \cos^2(\xi) + c^2 \sin^2(\xi)}\,. \tag{17}$$

We may assume in first approximation that the received flux of scattered sunlight coming from the asteroid will be simply proportional to the apparent projected surface S, so that, by neglecting any realistic effect of light scattering on the surface, we can simply write $m = -2.5\log(S) + c$, where c is a constant. As a consequence, it is easy to verify that if we call A the lightcurve amplitude, namely the difference of magnitude between the maximum and the minimum brightness, when we are in equatorial view ($\xi = 90^\circ$) we have that $A = -2.5\log(a/b)$. This means that observations of a triaxial ellipsoid asteroid taken in equatorial view, something that is always possible to obtain sooner or later, in principle provide also an estimate of the value of the a/b ratio.

The lightcurve amplitude progressively decreases as the aspect angle decreases from its maximum possible value of 90°. The minimum possible value of the aspect angle for a given object depends on the orientation of its spin axis. For an object whose spin axis lies on the ecliptic plane, a pole-on view becomes possible ($\xi = 0$), and in that geometric configuration the lightcurve amplitude in principle becomes zero and the object's magnitude does not change during the rotation. In that situation, the projected area will be $S = \pi ab$, which corresponds to the maximum possible value among all the possible projections of the triaxial ellipsoid, and then it corresponds also to the maximum possible brightness of the object. In general, however, the spin axis will be oriented in such a way as not to allow to reach a pole-on view.

Due to the fact that the lightcurve changes as a function of the aspect angle, having at disposal several lightcurves of an object obtained at different oppositions, it becomes in principle possible to derive the direction of its spin axis ("asteroid pole"). Different techniques have been developed for this purpose [68, 56]. Predictions concerning asteroid shapes and spin axis directions based on ground-based photometry have been found to be fairly accurate, according to the results of in situ investigations carried out by space probes [56].

One major advantage of observing from an orbiting platform like Gaia, with respect to traditional ground-based observations, is that from space it is easier to observe the objects even when they are far from opposition, since from space there is no strong observing constraint related to the diurnal and seasonal cycles of the Earth. In particular, from space the asteroids can be seen at small solar elongation angles, which are hardly achievable from the ground.

Each main-belt asteroid will be typically observed tens of times during the 5 years of the planned operational lifetime of Gaia. The simulations indicate that each object will be detected over a wide variety of ecliptic longitudes. Correspondingly, Gaia will make a good sampling of the interval of possible aspect angles of each object. The same variety of aspect angles, needed to derive the orientation of the spin axis, may be sampled from the ground only over much longer times. As a

consequence, Gaia will be very efficient in providing disk-integrated photometry data sufficient to derive the poles of the asteroids in a relatively short time, as well as the sidereal periods and the overall shapes.

However, it must be noted that there is a fundamental difference with respect to the situation usually encountered in traditional ground-based asteroid photometry. In fact, Gaia will detect each object only during very short transits when it crosses the field of view at epochs determined by the rotational and precessional motion of the satellite and by the orbital motion of the asteroid.

This means that Gaia photometry will not consist of full lightcurves, but only of a number of sparse, single photometric measurements lasting a few seconds. This would seem in principle a crucial limitation, but it is more than compensated by the high number of photometric measurements that will be recorded for each object (on the average, between 60 and 80 for main-belt asteroids) and by the good accuracy of Gaia photometry. In particular, the photometric accuracy will depend on the brightness of the target and it is expected to be better than 0.01 mag for single detections of objects down to approximately $V = 20.0$.

As we have seen above, the magnitudes of the objects detected at different epochs will depend on several parameters: the most important ones being the sidereal period, the shape and the orientation of the spin axis and the illumination circumstances, described by the phase angle. Additional variations may come in principle also from possible albedo variegation of the surfaces, but this is not expected to be very relevant for the majority of the objects. The possible existence of a non-negligible fraction of binary systems must also be taken into account, but for the moment we will not deal with this problem.

Gaia will obviously measure apparent magnitudes that will be immediately converted to magnitudes at unit distance from both the Sun and the satellite (reduced magnitudes). When reducing photometric data, the epochs of observations will also be corrected for light-time, being known the distance of the object at each detection.

Assuming now for the sake of simplicity to deal with an object having a triaxial ellipsoid shape, orbiting around the Sun along a typical main-belt asteroid orbit, it is possible to compute how the reduced magnitude is expected to vary as a function of time, depending on the coordinates of the pole. In particular, when making this exercise, it is convenient to work in terms of *differences* of reduced magnitude with respect to a reference observation (for instance, the first one in a series of different observations collected at different epochs). In this way, any potential error related to the constant appearing in the definition of magnitude ($m = -2.5 \log(\Phi) + c$, where m is the magnitude, Φ is the received–normalised–flux and c is a constant depending on the chosen units) is automatically removed.

If, for the sake of simplicity, we make also the assumption that the object is always observed at zero phase angle, it is easy to produce the plots shown in Figs. 10, 11, 12, 13, and 14, which have been computed assuming that the object has the same orbit of asteroid (39) Laetitia, a typical main-belt asteroid. Note that a choice for the object orbit must be done, because the aspect angle at any epoch depends both on the coordinates of the rotation pole and on those of the object itself, which is a function of its orbit.

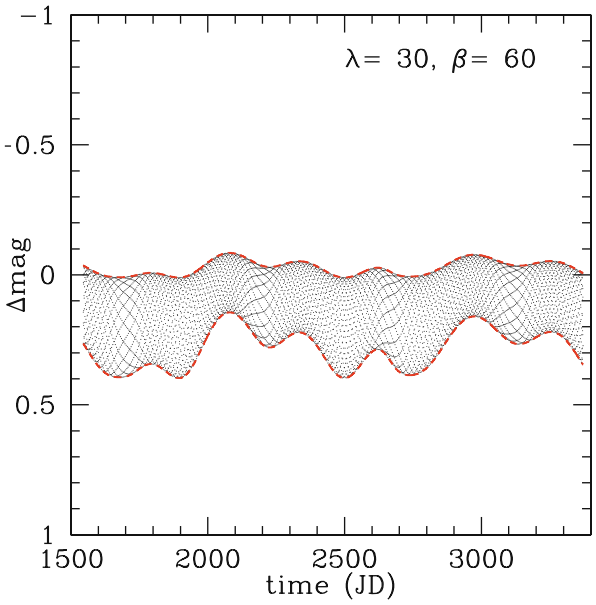

Fig. 10 The domain of possible values of magnitude (expressed as a difference with respect to the magnitude measured at a reference epoch) as a function of time (expressed in Julian Days) for a triaxial ellipsoid object having the same orbit of the asteroid (39) Laetitia. The axial ratios of the object are $b/a = 0.7$ and $c/a = 0.5$. The coordinates of the asteroid pole are given in the figure (λ being the ecliptic longitude and β the ecliptic latitude of the positive pole). The time interval covers 5 years, equal to the expected operational lifetime of Gaia

The plots shown in Figs. 10, 11, 12, 13, and 14 represent the domains in the time–Δm plane, Δm being the difference between the magnitude at time t and the magnitude at reference time $t = t_0$, occupied by the object during an interval of 5 years. In particular, the figures show how the domain of permitted Δm changes as a function of the object's shape (axial ratios) for a fixed pole of rotation or vice versa, as it changes by assuming a fixed shape, but varying the coordinates of the rotation pole. It must be noted that the above plots are used only to give a qualitative idea of the role played by shape and pole orientation in determining the possible range of photometric variation of an object, but they are not at all detailed representations of the real world.

In particular, the plots correspond to a very ideal situation in which the object has a perfect triaxial ellipsoid shape, and it is always seen at perfect Sun opposition. Thus, the effect on the apparent magnitude that arises from the fact of observing the object at different phase angles in different illumination conditions is not taken into account. Equally important, no realistic effect of light scattering is taken into account, and the magnitude is assumed to be purely due to the extent of visible illuminated area seen from the observer.

Of course, the real objects will only be sparsely observed by Gaia during the mission operational lifetime. As a consequence, what is more interesting is not the

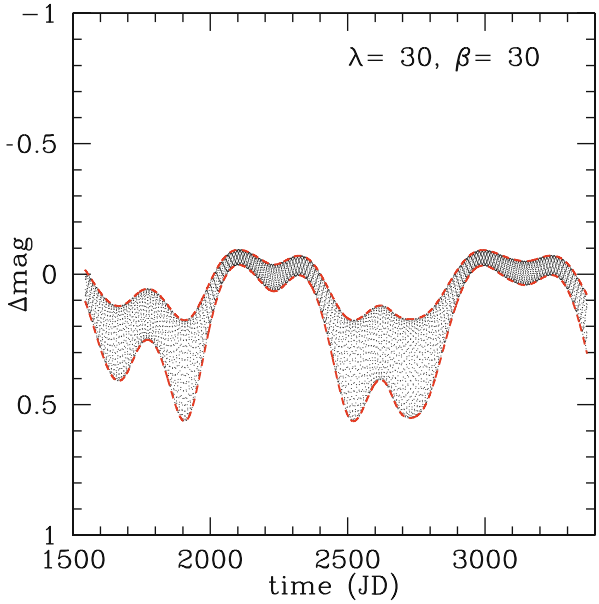

Fig. 11 The same as Fig. 10, but this time the ecliptic latitude of the asteroid's pole is $\beta = 30^\circ$. Together with Fig. 12, these three figures give some idea of the role played by the latitude of the pole in determining the photometric behaviour of a triaxial ellipsoid object

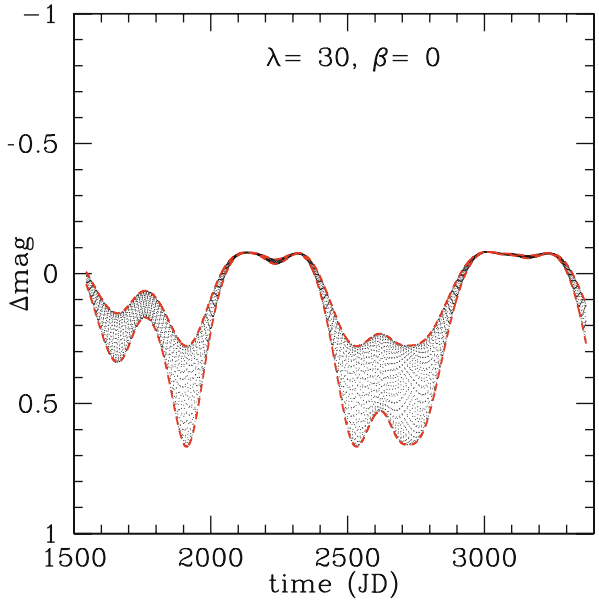

Fig. 12 The same as Fig. 10, but this time the ecliptic latitude of the asteroid's pole is $\beta = 0^\circ$

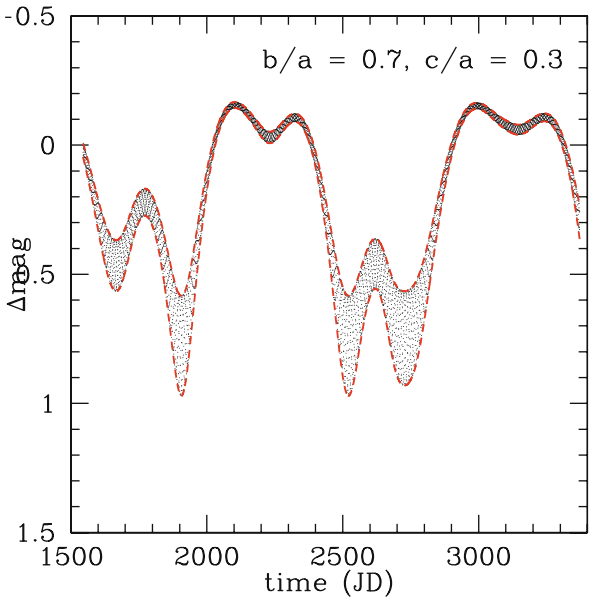

Fig. 13 The same as Fig. 10, but this time for an object having the same pole ($\lambda = 30^\circ$, $\beta = 60^\circ$), but a more elongated shape: $b/a = 0.7$, $c/a = 0.3$

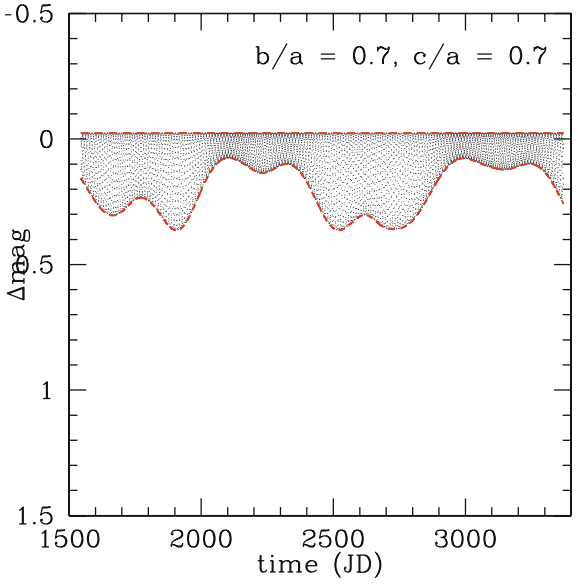

Fig. 14 The same as Fig. 13, but this time for an object having the same pole, but axial ratios: $b/a = c/a = 0.7$. The fact that $b = c$ has the effect of making constant the lightcurve maximum at all aspects

whole possible photometric domain where a given object can be seen during 5 years, but rather the way how the actual Gaia observations will sample the whole domain of photometric variation. To do this, one can take profit of detailed simulations of the mission, including observing circumstances of known asteroid detections, developed by a team including one of the authors of the present chapter (PT).

Figure 15 shows the results of such an exercise. In particular, it shows how Gaia would sample the photometric variation of an ideal triaxial ellipsoid object having the same orbit of the asteroid (311) Claudia; a spin period $P = 19.15$ h; axial ratios $b/a = 0.86$, $c/a = 0.71$; and pole ecliptic coordinates $\lambda_P = 49^\circ$, $\beta_P = 51^\circ$. As can be seen, the sampling of the photometric variation seems fairly good. Even better, it is interesting to do the same computation, but showing the results in an ecliptic longitude–magnitude plot, as done in Fig. 16. (expressing the coordinates in the equatorial reference frame, Right Ascension and Declination would be totally equivalent.) We recall that different ecliptic longitudes of the object correspond to different aspect angles. Figure 16 shows therefore that an object having a typical main-belt orbit, like (311) Claudia, will be observed in a fairly wide variety of aspect angles. This is encouraging when trying to develop methods to use these sparse photometric data to achieve an actual data inversion, leading to the determination of objects' rotation properties and overall shapes.

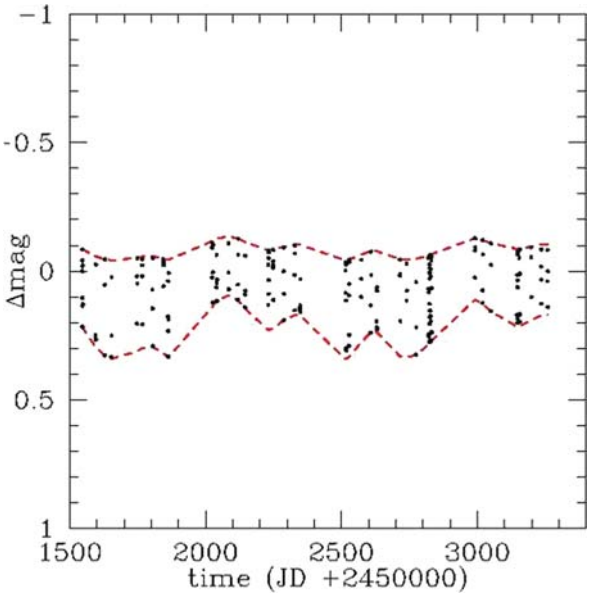

Fig. 15 Simulated observing circumstances of Gaia observations of a triaxial ellipsoid object having the same orbit of the main-belt asteroid (311) Claudia; rotation period $P = 19.15$ h; axial ratios $b/a = 0.86$, $c/a = 0.71$; and pole ecliptic coordinates $(49^\circ, 51^\circ)$. The predicted magnitudes as a function of time are plotted. The envelop represents the whole domain of photometric variation for such an object during the predicted operational lifetime of Gaia (5 years)

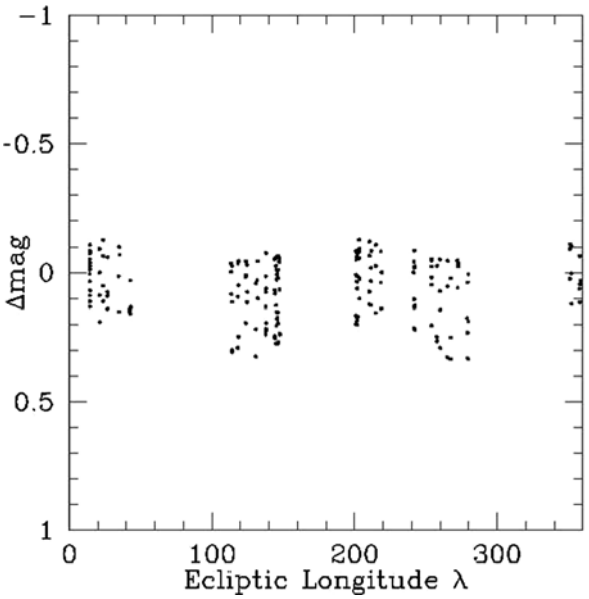

Fig. 16 The same as Fig. 15, but here the predicted magnitude is plotted against the ecliptic longitude of the object at the epochs of Gaia observations

Plots like the one shown in Fig. 16 are useful, because in principle a well-sampled longitude–magnitude plot may provide per se a lot of information. In particular, we have seen that the maximum lightcurve amplitude is achieved in equatorial view and directly provides an estimate of the b/a axial ratio, if we assume a triaxial ellipsoid shape. Moreover, the minimum lightcurve amplitude is reached at an ecliptic longitude corresponding to that of the asteroid pole. A minimum lightcurve amplitude close to zero means that the object may be seen nearly pole-on. Finally, a nearly flat maximum of the observed magnitude corresponds to an axial ratio b/c close to unity.

Of course, what is not known a priori when examining a longitude–magnitude plot like the one shown in Fig. 16 is the rotation period of the object. It is this parameter that determines the measured magnitude of the object at each single epoch of observation, which may be considered as a single snapshot of a continuous magnitude modulation superposed to the effects of the varying geometric observing circumstances.

In other words, the set of sparse photometric measurements that will be obtained by Gaia, as well as, it is important to note this, those from the ground-based telescopes of the next generation of sky surveys like Pan-STARRS, can be considered as single points of a complex, time-extended hyper-lightcurve. The goal of the analysis

of these data will be then that of being able to derive from them the main physical properties of the objects that are responsible for these observed hyper-lightcurves.

In particular, these are a list of parameters that should be determined by the photometric inversion:

- the asteroid's rotation period;
- the coordinates of the asteroid's pole, namely the intersection of the direct spin axis of the object with the celestial sphere. These two coordinates may be indifferently expressed in the equatorial or in the ecliptic J2000 system. Following the IAU convention [103] the ecliptic J2000 system should be preferred;
- some parameters describing the shape of the asteroid. In the case of a triaxial ellipsoid shape with semi-axes $a \geq b \geq c$, this reduces to two simple parameters, namely the axial ratios b/a and c/a;
- a parameter specifying the variation of the asteroid's brightness as a function of the phase angle. In practice, since Gaia will observe asteroids in a range of phase angles far from Sun opposition, and because the phase function is generally linear in the range 20–30°, the simplest approximation is to consider a linear magnitude–phase relation, described by one single slope parameter;
- a value specifying the rotational phase of the asteroid at a reference epoch, usually taken to be the epoch of the first recorded observation;

The above list corresponds therefore to a set of seven unknown parameters to be determined by the inversion, assuming that the objects have triaxial ellipsoid shapes. Of course, this particular shape choice is not the unique possibility that might be considered. Some explanation is then necessary and is given in the next section.

6.3 Notes on the Choice of a Triaxial Ellipsoid Shape

Asteroids are small rocky bodies whose shapes are normally determined by solid state forces rather than by self-gravitation. Asteroid images taken from short distances by space probes have generally revealed, mainly for the smallest visited objects, fairly irregular, "potato-like" shapes. Even objects of greater sizes, like (253) Mathilde, have revealed shapes strongly affected by the presence of giant concavities due to large impact craters. Moreover, in recent times there has been the development of methods of lightcurve and even sparse-data inversion that have produced complex reconstructed shapes and that have been in several cases confirmed, directly or indirectly, by other pieces of independent evidence [56].

Based on the above evidence, the choice of approximating objects by means of triaxial ellipsoid shapes looks a priori very questionable. On the other hand, there are several reasons why this choice has been done for at least a preliminary inversion of the future Gaia data. Most of these reasons are strictly related to the need of reducing as much as possible the number of free parameters to be determined by the photometry inversion algorithm that will be described in the next section.

Summarising, a list of reasons justifying the choice of a triaxial ellipsoid approximation is the following:

- In spite of being simple and depending on only two parameters (the b/a and c/a axial ratios), triaxial ellipsoid shapes are fairly flexible and allow to represent a fairly wide variety of shapes, from elongated "cigar-like" shapes, to flat discs, up to regular spheres.
- There are reasons to believe that, at least among the biggest asteroids, triaxial ellipsoid shapes might be a good approximation. The reason is that this is the equilibrium shape expected for re-accumulated objects having a large angular momentum. These objects, commonly called "rubble piles", are expected to exist, and correspond to the so-called LASPA (large-amplitude short-period asteroids) objects, identified in the 1980s [34, 35].
- Triaxial ellipsoid shapes have been successfully used by several authors in the past to compute asteroid rotation poles that have been confirmed also by other techniques [68].
- The shapes of impact fragments collected in laboratory experiments have been approximated in the past by triaxial ellipsoids [16]. The reason is that when dealing with a variety of irregular fragments produced by energetic impact experiments, circumscribed triaxial ellipsoids turned out to be simple and sufficiently accurate to be used for building a reasonable statistics of the shapes of these fragments.
- One fundamental advantage of triaxial ellipsoid shapes is that they admit *analytical solutions* of the problem of computation of visible and illuminated areas as seen from an observer in any geometric configuration. This property is decisive for the choice of this kind of shape to be implemented in the kind of inversion algorithm that will be described in the following section.
- As we will see, the choice of triaxial ellipsoid shapes is justified a posteriori by the fact that it can be proven to be able to produce successful inversion not only of simulated objects having a wide variety of shapes and light scattering properties but also of real objects previously observed during space missions in the past.

As a final remark, we note that using a model in which a triaxial ellipsoid shape is assumed to fit the objects simply means that the corresponding inversion of available data will look for the triaxial ellipsoid shape that best fits the observations. One should be aware that this triaxial ellipsoid can well not be adequate to represent the fine details of the real shape, but this is not equivalent to state that the resulting inversion must be necessarily bad, in particular for the period determination. The goodness of the solutions must be checked a posteriori by extensive simulations and by applications of the inversion algorithm to real data. We also note that it might be over-ambitious to try to use sparse photometric data to get, in addition to spin periods and poles, also very detailed reconstruction of the shapes. We remind, for instance, that even the most advanced methods of shape reconstruction based on lightcurve data cannot deal with shape concavities in a straightforward way [56].

For this reason, even an apparently modest triaxial ellipsoid approximation can well prove to be useful to derive global shape estimations, and this may be convenient if this can be made with only a modest investment of CPU power, as we will see below.

6.4 Photometry Inversion by Means of a Genetic Algorithm

Before describing the details of the adopted algorithm for the inversion of sparse photometric data, it should be useful to point out that this is a problem that hardly can be attacked by means of a pure "brute force" approach based on the power of modern computers. The reason is that, although the number of unknown parameters is only seven (the rotation period, two pole coordinates, two axial ratios, one magnitude–phase linear slope, and one initial rotation phase), the number of possible combination of these parameters, corresponding to single possible solutions, is large, mostly due to the required accuracy in the spin period determination.

In fact, if one considers for instance the case of an object having a rotation period of 6 h, the need to have the final photometric observations "in phase" within a not-so-small error of 0.8 in rotational phase, means that the spin period must be computed with an accuracy not worse that 10^{-5} h. A brute force approach based on computing the best solution by building of grid of possible cases in the space of the seven unknown parameters would then lead to the need of performing a huge number of iterations of the order of 10^{19}, as it is easy to verify taking into account the range of variability of the seven parameters (the period between 0 and tens of hours, the pole longitude between 0 and 360°, the pole latitude between −90 and +90°, the axial ratios between 0 and 1, the initial rotational phase between 0 and 0.5 (an ambiguity of 0.5 is permitted for this parameter), linear slope of the magnitude–phase relation between 0 and some tenths of magnitude per degree).

The above estimate means that the photometric inversion of each object would require an exceedingly big investment in CPU power. On the other hand, other approaches exist that may avoid the use of brute computing power and may permit to save a lot of computing time.

Among these approaches, a one based on the development of a so-called *genetic algorithm* has been chosen for the processing of Gaia photometric data. Genetic algorithms are adaptive algorithms developed to solve some classes of problems that are particularly suited to being attacked using this kind of approach. Some general ideas which are at the basis of the genetic algorithm approach are borrowed from natural sciences, and in particular from the processes of natural evolution of living species and survival of the fittest.

In particular, we can consider any possible solution of the problem of inversion of sparse photometric data as an individual "organism" characterised by its own "genome". The genome consists of a single value for each of the seven unknown parameters to be determined by the inversion. Any possible solution, therefore, is uniquely characterised by its own set of parameters (spin period, pole coordinates,

axial ratios, slope of the magnitude–phase relation and initial rotational phase), which can be seen as the "genes" or the "DNA" of the given solution.

Of course, different solutions can be more or less good. The range goes from completely bad solutions of the problem up to excellent solutions, which may be accepted as the result of photometry inversion. The goodness of any given solution is assessed on the basis of its corresponding residuals with respect to a set of real (or simulated) observations. Better solutions give a better fit of the observational data, corresponding to smaller residuals $(O - C)$ between the observed data and those computed according to the given set of parameters. In the application to the Gaia problem, the parameter that has been used so far to quantify the goodness of a given solution is the following:

$$\epsilon = \sqrt{\frac{\sum_i (O_i - C_i)^2}{N_{obs}}}\,, \tag{18}$$

where N_{obs} is the number of available observations and O_i and C_i are the observed and calculated values of the ith observation, respectively.

The idea of the genetic approach is then to find a good solution of the inversion problem by taking an initial set of tentative solutions, randomly generated, and then let them "evolve" by mixing their genomes during a series of successive generations, until the correct solution of the problem emerges spontaneously. In other words, starting from an initial set of "parent" solutions randomly generated, and which are generally very bad, as one would expect a priori from a set of completely random attempts, an iterative process of production of new generations of solutions is started. The generation process consists in a random coupling of the parameters of two parent solutions, randomly selected, and/or in some random variation ("genetic mutation") of the "genome" of one single solution.

At each generation, a check is performed of the residuals generated by each newly born "baby" solution. If it is better than some of those solutions saved until that step, it enters the "top list", whereas the previously worst solution of the parent population is removed from the population.

In this way, after a number of the order of 1 million of "generations", a very good solution is usually found, which produces small residuals and basically solves the inversion problem. Due to the intrinsically random nature of this "genetic" approach, the right solution is not forcedly found always, but if the genetic algorithm is repeatedly applied to the same set of observed data (typically 20 times), then the right solution (giving the minimum residuals in different attempts) is usually found several times.

Based on the above description, it may be now clear why the assumption of a triaxial ellipsoid shape turns out to be very convenient using this approach. The basic advantage is that each of the millions of computations of the brightness corresponding to the observing circumstances of each recorded observation can be made by using analytical formulae, instead of numerical computations of more complex shapes based on an approximation of the shape by means of plane facets, followed

by the computation of the contribution of each facet to the total brightness. Such an approach, in principle feasible and more flexible, would lead in practice to a huge investment in CPU power and in much longer computing times. The apparent illuminated surface of a triaxial ellipsoid seen by any observer in any geometric configuration Sun–asteroid–observer can be instead computed by using analytical formulae, as shown by [96].

The performances of the adopted genetic algorithm for the inversion of sparse photometric data have been tested by means of a large number of numerical simulations and by application to real data. In particular, simulations have been performed considering a large variety of situations for what concerns

- the orbits of the simulated objects, to check whether orbital properties may have important consequences on the number of Gaia detections and on the variety of observing circumstances. In particular, the simulations have also dealt with near-Earth objects, taking into account that many objects of this type will be observed by Gaia, although the majority of detections will refer to main-belt asteroids;
- the rotation periods of the simulated objects;
- the spin axis direction;
- the object shape. Both triaxial shapes characterised by a large variety of axial ratios and more complex shapes, often taken from available observations (optical and radar) of real objects, have been simulated;
- different numbers of available observations, and different photometric uncertainties, in order to test the conditions of applicability of the inversion algorithm when the available data are scarce and/or the error bars of the data are large;
- different light scattering laws, ranging from pure geometric scattering to Lommel–Seeliger and Hapke scattering models.

For what concerns the last item in the above list, Fig. 17 shows the results of simulations of regular triaxial ellipsoid shapes in the simplified case of geometric light scattering (no limb darkening). Simulations produced cases with varying numbers of available observations and different photometric uncertainties for each observation. As can be seen, the results are strongly encouraging in these simplified cases, because they indicate that the inversion method should be applicable even in situations much worse than those expected to hold for Gaia data (numbers of observations of the order of 70, typical photometric uncertainties of the order of 0.01 mag). Even taking into account that the real objects will have surfaces scattering sunlight in a more complicated way, and non-ellipsoidal shapes, it seems nevertheless that the inversion method does a good job and it is not too much constrained by the number of data and their accuracy.

Figure 18 shows what the fit of simulated data can be in cases in which a significant photometric uncertainty of 0.03 mag is added to the computed magnitudes of a simulated triaxial ellipsoid body. As expected, there are not negligible $(O - C)$ residuals, but they are fully corresponding to what should be predicted a priori. The resulting inversion, in fact, turns out to be practically perfect in this case: this

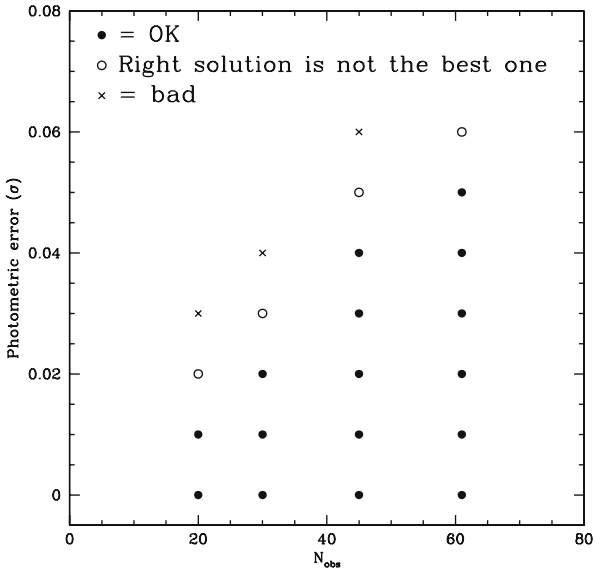

Fig. 17 Conditions of applicability of the photometry inversion method for triaxial ellipsoid shapes and simple geometric light scattering. The horizontal axis gives the number of available observations, while the vertical axis gives the photometric uncertainty of the observations (supposed for simplicity to be the same for each observation). The different symbols in the plot indicate whether the photometry inversion method successfully finds the right solution, if it does not find a unique, right solution or if it fails

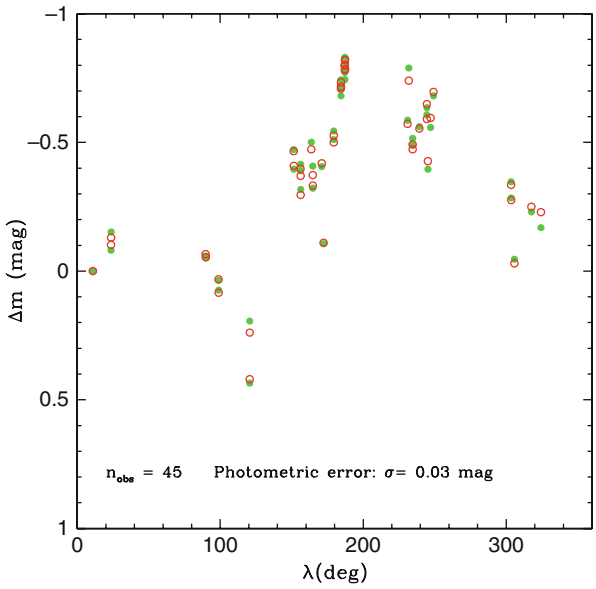

Fig. 18 Results of the inversion of a simulated triaxial ellipsoid asteroid with a fairly large superimposed photometric error of 0.03 mag. *Filled symbols*: simulated observations. *Open symbols*: corresponding computed brightness at each epoch of observation

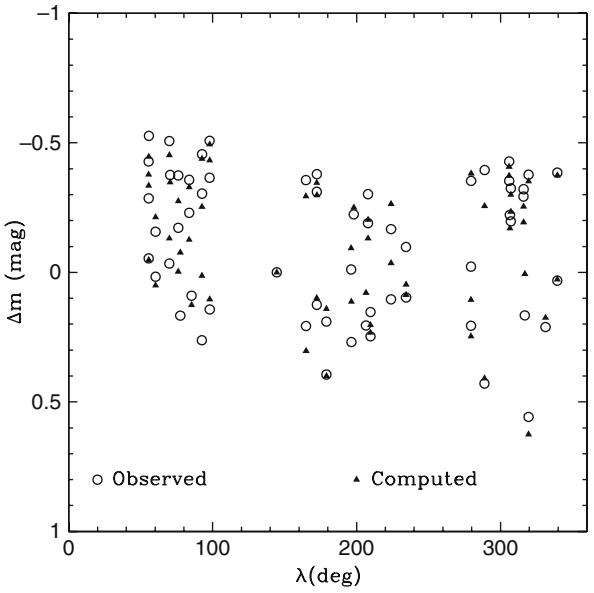

Fig. 19 The same as Fig. 18, but this time the simulated object had the shape of the irregular asteroid (433) Eros as resulting from the images taken by the NEAR-Shoemaker probe. In addition, the simulated surface was assumed to scatter sunlight according to a realistic Hapke scattering law

means a computed period solution within 0.00001 mag from the correct one and pole coordinates within less than 2° from the correct pole solution.

If simulations of triaxial ellipsoids without light scattering effect may seem too poor to assess the effectiveness of the photometry inversion algorithm, Fig. 19 is representative of what is encountered when much more realistic simulated cases are taken into account. The plot shows the simulated photometric data of an object having the same shape of (433) Eros as it has been derived from in situ observations by the NEAR-Shoemaker probe, and simulating a surface scattering of the light according to a Hapke light scattering law with the usual parameters used to represent the photometric behaviour of real asteroids. In the same plot, the magnitudes computed—based on the result of the photometric inversion algorithm—are also shown for a comparison. As expected, there are large values of the $(O-C)$ residuals, like one should expect from a body as irregular as Eros (see Fig. 20) when approximated by a simple triaxial ellipsoid. In spite of the obvious oversimplification of the shape model used in the inversion algorithm, however, the solution turns out to be excellent, with the spin period being exact within a few thousandths of an hour, and the pole solution differing from the simulated one by less than 3°, justifying our previous statement in Sect. 6.3. Even the very elongated, cigar-like shape of Eros is not badly reproduced by the shape solution, with b/a and c/a axial ratios both of the order of 0.4. This is not an isolated success of the inversion method. Equally good results have been obtained also for other complicated simulated shapes and scattering laws, including that of the asteroid (6489) Golevka, an Apollo asteroid

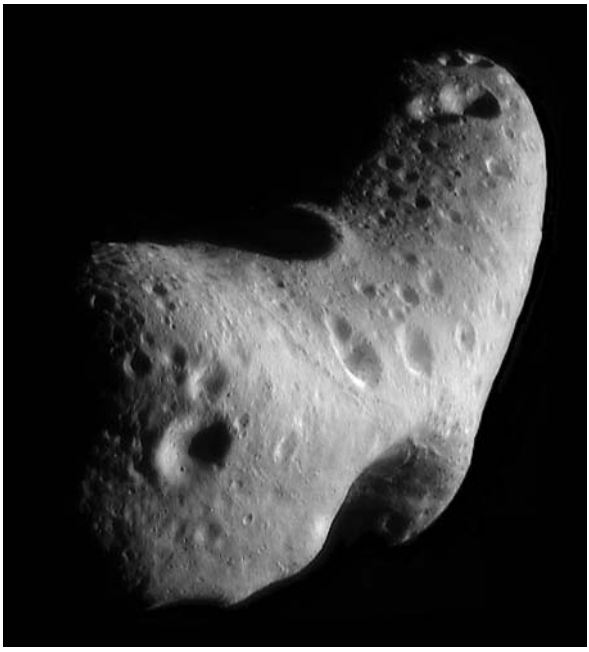

Fig. 20 A NEAR-Shoemaker image of asteroid (433) Eros (Courtesy of NASA)

observed by radar technique for which a detailed and highly non-convex shape model is available.

The most important test so far performed for the effectiveness of the adopted photometry inversion algorithm has been the application to Hipparcos observations. We remind that the Hipparcos satellite is the immediate precursor of Gaia (see Sect. 2.1). What is important in the context of this discussion is that Hipparcos obtained sparse photometric data for a limited number (48) of the brightest asteroids. Among these objects, many were observed only a few times and/or the obtained data had large error bars. If we refer to the results shown in Fig. 17, only 26 objects satisfy the number of observations plus error bar conditions that made inversion possible for simple simulated triaxial ellipsoid shapes without limb darkening. Among these 26 objects, moreover, some were just at the limit of acceptability.

A systematic analysis of these data has demonstrated that a successful inversion of Hipparcos data, in terms of successful determination of the orbital period and pole, has been obtained for 14 objects. In addition, inversion of the two big objects (1) Ceres and (4) Vesta was obtained, but with a resulting spin period equal to twice the correct value. These two results can well be explained by the fact that Ceres and Vesta are peculiar asteroids, whose photometric variation is due to albedo spots on the surface and not to shape effects. Also the results of Hipparcos data inversion seem thus to indicate that the photometry inversion algorithm based on the genetic approach discussed in this section seems quite effective and reliable.

According to the results of these tests, it is reasonable to expect that the inversion of Gaia disc-integrated photometric data will lead to the determination of the spin properties and overall shapes for a number of the order of 10,000 asteroids. Moreover, the triaxial shapes and pole coordinates obtained from photometry inversion will be also used to refine the determination of asteroid sizes as explained in Sect. 5.

7 The Expected Gaia-Based Asteroid Taxonomy

We have considered so far the photometric signal acquired in the broad Gaia photometric G-band and, more specifically, its temporal variation. As seen in Sect. 3.2 there are, after the crossing of the main astrometric field, two additional CCD columns (RP and BP). These will provide low-resolution spectroscopy similar to multiband colour-photometry measurements. One could again analyse variations due to the spin of the asteroids. We focus here on the information that can be obtained on the surface of the body and the taxonomy of asteroids that can be derived.

The relevant spectro-photometric capability of Gaia will be used to obtain spectral reflectance data for a very large number of asteroids. As seen before, there should be a number of the order of 300,000 main-belt asteroids that will exhibit apparent magnitudes brighter than $V \leq 20$ light when detected by Gaia. When passing through the BP and RP detectors of Gaia, the colours of these objects will be recorded. In particular, the whole range of wavelengths covered by the BP and RP detectors is from about 330 up to 1000 nm. This will make it possible to obtain spectro-photometric data covering about 20 bands in the above-mentioned wavelength interval, producing a very valuable data set of spectral reflectance data for asteroids. Of course, many objects will be faint, and, especially at short wavelengths, at some transits across the Gaia field of view the recorded fluxes will be below the detection limit. Moreover, in some cases the objects may move sufficiently fast as to be lost from the observing window before reaching the RP and BP detectors. Although detailed studies have not yet been produced at the moment, however, it seems that taking profit of the fact that each object will be detected several times during the operational lifetime of Gaia (a number of detections of the order of 65 being typical for main-belt asteroids), and some detections will be better than others since the objects will be brighter being at smaller geocentric distances, it is reasonable to expect that for a quite large fraction of the total number of detected bodies a complete coverage of the spectro-photometric behaviour from blue to red wavelengths will be obtained. These data will be very important because they will be used to derive a new taxonomic classification of a very big asteroid sample, much larger than any similar data set currently available.

In particular, Gaia low-resolution spectral data will be obtained for objects over a wide range of sizes, and this will be important to analyse possible size-dependent effects on asteroid colours related to the interplay of collisional and dynamical ages and effects of surface space weathering. For instance, it is known that the surfaces

of young *S*-type asteroids that belong to the near-Earth population have spectral properties that are more similar to those of ordinary chondrite meteorites than those of larger, main-belt object's orbiting belonging to the same taxonomic class.

From the point of view of taxonomy, Gaia will have a couple of excellent properties: first, this very large spectro-photometric database will be obtained using a unique, homogeneous photometric system and not merging together data coming from different instruments. Second, and equally important, the spectral coverage will include the blue region of the reflectance spectrum. This is a very useful property, because the regions of the spectrum corresponding to the classical Johnson *U* and *B* colours, which were included in classical UBV spectro-photometric databases obtained many years ago by means of photoelectric photometers, are currently largely missed by the most recent spectroscopic surveys, like SMASS and SMASS2 [14].

One should consider, in fact, that the blue region of the reflectance spectrum is very important to distinguish between several groups of primitive, low-albedo bodies. Among the many thousands of asteroids that are expected to be classified based on the Gaia spectro-photometric database, a large fraction will consist of primitive, dark objects belonging to the so-called *C* superclass, which dominates the asteroid inventory in the outer regions of the asteroid belt. As opposite to spectroscopic surveys like SMASS and SMASS2, that were limited in practice to an interval of wavelengths between 0.5 and 0.95 μm, Gaia is expected to do a better job in discriminating among different subclasses of the big *C* complex and to determine the relative abundance of these different subclasses.

A typical example of an important class of asteroids that has been lost in recent taxonomies is the *F* class. These objects have low albedos, are generally more abundant in the outer belt, but they exhibit a local overabundance in the inner belt, possibly associated with a dynamical family (Polana [21]). Interestingly, *F* objects have been found to exhibit unusual polarimetric properties [6] and may also have some relations with comets, since comet Wilson–Harrington was classified as an *F* asteroid, before its cometary nature was discovered, and also the near-Earth object (3200) Phaeton, which is known to be associated with the Geminid meteor shower, was also classified as an *F*-type [6].

The Gaia taxonomy will be an important tool for studying the overall compositional gradient of the material which accreted into the planetary bodies in our Solar System [18]. Moreover, the availability of spectral reflectance data and taxonomic classification will also be important for future studies of the identification of asteroid dynamical families. The reason is that classical family searches have been based so far on the distribution of the objects in the space of proper elements. Since many families tend to overlap in this space, spectroscopic data may add a new dimension to the problem and should be very useful to discriminate among members of different families overlapping in the space of proper orbital elements [19].

We note that new taxonomic classifications of the minor bodies of the Solar System based on spectral reflectance data will be obtained in the next years also by the next generation of dedicated ground-based surveys like Pan-STARRS and/or the LSST. The Gaia-based taxonomy, however, in spite of the existence of such

"competitors" will "come for free", being an automatic sub-product of the BP and RP detections. Taking into account the above-mentioned coverage of the *B* region of the spectrum, and the general link with other unique results that will be produced by Gaia asteroid observations, it is certain that Gaia taxonomy will be a very useful addition to the extremely important scientific value of the Gaia mission for asteroid science. For instance, it will be possible to link immediately the obtained average density determinations for about 100 objects with the taxonomic classification of the same objects, a very important result to interpret taxonomy in terms of overall composition of the objects. In this respect, also the determination of Gaia albedos of the same objects will be extremely important to sketch an overall interpretation paradigm of such a wealth of physical information.

8 Dynamical Model Improvement with Gaia

We have seen before in Sect. 4, that the Gaia mission will provide the orbits of a large number of asteroids with high accuracy. Not only will the observations be precise on the CCD (no refraction, no personal equation, well-calibrated measure through one single instrument independently of the northern/southern hemisphere, etc.), but mostly all positions are referred to a very homogeneous reference frame (materialised by QSOs and primary reference stars), avoiding many systematic errors. Such refined orbits (for at least over the period of the Gaia observations) will enable to detect and/or measure small and subtle effect, either dynamical, relativistic or non-gravitational. We will see some aspects for the determination of asteroids mass, test of general relativity and linking of reference frames. Because we have to wait to acquire the real Gaia data all results given below are hence obtained from simulations of the observations, and they mostly concern the precision of parameters estimation. Combining on one side the image simulation on the focal plane and a centroiding precision (Sect. 5.4) to the Gaia scanning law and asteroids ephemerides on the other side, one gets the useful data to derive the quantities presented below.

8.1 Asteroid Mass

As we will see later (Sect. 10), monitoring the orbits of binary and multiple systems will provide their total mass, and in some circumstances, the mass of each component. The other way to derive the mass of an asteroid is to measure the mutual gravitational perturbation during a close encounter [124, 53, 81 and references therein]. This can be done from the perturbations of asteroids on Mars or asteroids–asteroids perturbations. Indeed, this generally involves a large massive asteroid influencing several massless target asteroids, although target asteroids can in some encounters act as perturber too. Given the high precision with which Gaia can derive the orbit of an asteroid (see Fig. 21), one can detect small effects affecting the orbit, in particular small perturbations during close encounters. It is clear that the effect can be detected only if the astrometric observations are obtained before *and*

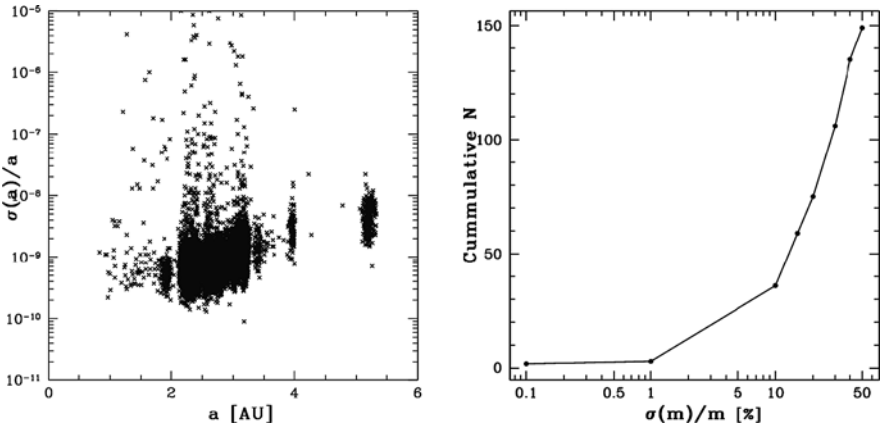

Fig. 21 *Left*: Orbit improvement; relative precision obtained on the semi-major axis from the 5 years of Gaia observations. *Right*: Precision of asteroid mass determination with Gaia from gravitational perturbation during close encounters

after the encounter itself. The more general and straightforward problem would be to integrate the equations of motions (and variational equations) of the dynamical system at once. This means an N-body integration with N of the order of 300,000 plus the eight planets, relativistic effects, the effect of the oblate Sun and a ring of asteroids, etc. The step size would need to be automatic and would be governed by the encounters. However, such high CPU-consuming algorithm is not necessary in our more hierarchical problem, where many asteroid orbits can be considered as perturbed two-body solutions. In fact the gravitational effect of a given asteroid i on a planet or an asteroid j is in most cases negligible and should not be computed, not only to decrease the CPU time but also to avoid unnecessary numerical noise. It is hence better to perform the integration by taking into account only relevant perturbations and thus derive a list of potential perturbers. These perturbing asteroids are to be added to the classical perturbing major planets that are systematically taken into account in the perturbing function [80]. Note also that a perturbing asteroid can still be a target for another one, but the size of the N-body system is decreased by several orders of magnitude. So, the first step is to provide the list of targets for each large asteroid and inversely the list of perturbers to include for a given asteroid. In the second step, one computes the partial derivatives for the additional unknown masses. Note that when the error on the mass is large, the problem is likely to be non-linear. In the other case one starts with a linear least squares treatment of the problem. It has been shown in [79] that Gaia will derive the mass of about 150 asteroids with a relative precision better than 50% (see Fig. 21). Since the number of targets is often large, systematic effects are reduced, so that Gaia will increase the *precision* of the mass determination, but also and mostly its *accuracy*.[5] Some of the targets

[5] Note the semantic. Broadly, the precision corresponds to the dispersion or variance of a quantity around the mean or expectancy, while the accuracy will give the error between the true and the mean. A parameter estimation or determination can be of very high precision, in particular when

can be poorly observed with the Gaia scanning law, or the encounter can take place close to the beginning or the end of the 5 years mission. In the latter case, having only half of the data is totally useless; so it could be interesting to complement the space-based data with additional ground-based astrometry. These should, however, be of good precision and over a limited time span to reduce the effects of systematic errors. Again, conservative simulations have shown that additional masses can be derived from such data combination [83].

The mass of Ceres and Vesta will surely be known with higher accuracy from the future Dawn mission, and they will help to validate the method. Some of the targets are also binary asteroids which will provide additional calibrations. Interestingly other asteroids in the list of Gaia mass determination have currently no mass determination or act as large perturber on Mars [79]. Since the present limitation on the ephemerides of inner planets arises from the poor knowledge on the asteroids mass, Gaia will bring some improvement. Combining the measures of mass and size–shape–volume provides another fundamental physical parameter, the bulk density. With the aid of ground-based high-angular resolution observations this quantity will be obtained with high precision for about 60 targets, covering many taxonomic classes and enabling to test possible links between density and taxonomy. We know for instance that rubble-pile asteroids or highly fractured ones can have substantial porosity making this link non-trivial [12].

8.1.1 Global Solution

Let us stress that—in contrast to previous works on mass determinations—these estimations are obtained while treating the problem globally: all targets for a perturber are treated simultaneously and cross-perturber terms are also taken into account. The system of observational equations takes the form of a sparse diagonal–column block matrix:

$$\mathbf{P}\cdot\begin{pmatrix} \mathbf{A_1} & 0 & \cdots & \cdots & \mathbf{G_1} & \mathbf{M_1} \\ 0 & \mathbf{A_2} & 0 & \cdots & \mathbf{G_2} & \mathbf{M_2} \\ \vdots & & \ddots & & \vdots & \vdots \\ 0 & \cdots & \cdots & \mathbf{A_N} & \mathbf{G_N} & \mathbf{M_N} \end{pmatrix} \cdot \begin{pmatrix} d\mathbf{q_1} \\ d\mathbf{q_2} \\ \vdots \\ d\mathbf{q_N} \\ d\mathbf{g} \\ dm_1 \\ \vdots \\ dm_p \end{pmatrix} = \begin{pmatrix} d\boldsymbol{\lambda}_1 \\ d\boldsymbol{\lambda}_2 \\ \vdots \\ d\boldsymbol{\lambda}_N \end{pmatrix} \tag{19}$$

obtained from a large number of observations, but in severe error if there are uncorrected bias in either the observations or the model. An illustration can be given by a badly built ruler where the graduation starts at 1 not 0, the precision of the measure is about 0.5 mm which is fair, but all measures are systematically wrong by 1 cm, hence inaccurate.

where

- N is the total number of asteroids ($\approx 300,000$);
- $\mathbf{P}$ projection matrix to transform heliocentric (barycentric) position vector $\mathbf{x}$ to observed quantities λ (or RA, Dec);
- $\mathbf{A_i}$ is the Jacobian matrix ($n_i \times 6$) of the partial derivatives with respect to the initial conditions for asteroid i, $\left[\frac{\partial \mathbf{x}}{\partial \mathbf{q_o}}\right]_i$;
- $\mathbf{q_i}$ is the vector of corrections to the six initial conditions for asteroid i;
- $\mathbf{G_i}$ is the Jacobian matrix for asteroid i ($n_i \times n_{par}$) of the partial derivatives with respect to global parameters common to all asteroids $\left[\frac{\partial \mathbf{x}}{\partial \mathbf{g}}\right]_i$;
- $\mathbf{dg}$ is the vector of the n_{par} global parameters;
- $\mathbf{M_i}$ is the Jacobian matrix for asteroid i ($n_i \times p$) of the partial derivatives with respect to the mass corrections relevant to asteroid i. It is again a sparse matrix;
- dm_j is the mass correction for massive asteroid j perturbing asteroid i. As said before the list of perturber has been previously selected from a simulation of close encounters [81, 82].

This system is inverted for all parameters together and combining various target asteroids simultaneously.

8.1.2 Other Small Effects—Estimated and/or Considered Parameters

The photocentre offset corresponds to the difference—projected on the sky—between the centre of gravity of the body (the one to be considered in the equations of motion) and the point derived from some centroiding giving for instance the mode of the light distribution on the focal plane [64, 45]. For a sphere observed at full phase there is no difference as long as the light distribution is radially symmetric. Of course, if the phase is important the difference will be large; it scales with the apparent size of the body and its surface properties, light scattering and albedo variegation. Such effect is generally taken into account for planets and large planetary satellites [77] since it amounts to several mas. Considering only the phase i and the size of the object which is assumed to be spherical (of apparent diameter ϕ) and with no albedo markings on its surface, one can write

$$\delta\lambda_p = C(i)\ \sin(i/2)\ \ \phi/2, \tag{20}$$

where the displacement $\delta\lambda_p$ is given in the direction of the Sun. The function $C(i)$ depends on the scattering properties of the surface; for a perfect and hypothetical Lambertian sphere one has $C(i) \sim \frac{3}{4} + \frac{3}{32}i^2 + o(i^3)$, which generally yields an estimation of the *maximum* offset. This offset is also noticeable in the Hipparcos data for the largest asteroids [46]. With the high precision measure from Gaia, it will be of importance for a larger number of objects. There are three different cases depending on the size of the target, and our knowledge of its size, shape and brightness distribution. If the body is small, the effect can be *negligible*; if the object is bright and large, the effect can be modelled with enough accuracy or some parameters

can even be *estimated* from the inversion of the system of (19) itself [46]. Dynamical (state-vector initial conditions) and physical parameters (size, shape, etc.) can be derived simultaneously for a combination of both astrometric and photometric data [55]; however, as long as they remain largely decoupled one can also proceed iteratively with separate estimations of shape and light scattering properties from the photometry and size estimations from the astrometry. However, there remains the third and intermediate case where the effect (generally systematic, see below) is neither negligible nor can it be fully modelled or estimated for instance from incomplete knowledge of the shape and light scattering at the surface of the body [57]. In this case there will be additional random noise in the modelled photocentre offset which, in turn, will decrease the accuracy of other parameters estimation, this can be handled through consider covariance matrix [112]. The LLS problem to solve can be written as $\mathbf{A_x}.\mathbf{x} + \mathbf{A_c}.\mathbf{c} = \mathbf{b}$, where only $\mathbf{x}$ is the unknown vector to be estimated; the consider vector $\mathbf{c}$ of input parameters is assumed to be known but with some a priori uncertainty, e.g. $\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{c}) = \sigma^{\mathbf{2}}.\mathbf{I}$. Clearly, both the LLS solution for $\mathbf{x}$ and its variance/covariance matrix will depend on the choice of $\mathbf{c}$ and associated a priori variance/covariance matrix. Such procedure also allows to yield the sensitivity of the estimated parameter $\mathbf{x}$ with respect to the considered parameter $\mathbf{c}$.

One should note that the photocentre offset is a *systematic* effect, always directed towards the direction of the Sun, hence typically in the ecliptic plane for an asteroid. Being systematic, such effect should not be compared to the precision of a single observation in the error budget, but to the precision foreseen for global parameters, which—compared to the observation precision—scales as $N^{-1/2}$ and more specifically to their correlations. As analysed in [107, 45] for the Hipparcos data, the photocentre offset effect mimics a global rotation along the ecliptic pole. The situation is very similar to another systematic effect due to thermal inertia of the body, the Yarkovsky effect.

In addition to the perturbations already mentioned, one can consider non-gravitational forces acting on the dynamics of the body (out-gazing comet, asteroid, meteoroid, particle, etc.) such as Poynting–Robertson or solar pressure. Here we consider the Yarkovsky drag which is a thermal effect. An asteroid is not a perfectly reflecting body and is re-emitting part of the solar energy in the thermal infrared. Due to thermal inertia (and the asteroid rotation) this re-emission has some lag in time; to illustrate this, consider for instance the hottest time in the day: it is not at solar noon but a couple of hours later in the day. Thus the photons are re-emitted in a direction different from the (radial) reception direction. Taking the balance of energy one sees that the force created can be decomposed by one part in the radial direction and by another part in the tangent direction and hence create some additional non-gravitational force or acceleration (see Fig. 22). This effect that was previously considered for small particles such as meteoroids has been detected on larger bodies, first on the LAGEOS satellite [100] and later for asteroids [101, 127]. For a fast rotating body or a non-spinning one there is no diurnal effect. The Yarkovsky effect scales with the object size, its thermal inertia (or equivalently thermal conductivity) and lastly its distance to the Sun. The general trend effect is most noticeable in the NEOs population [23]. From its astrometric measurements alone, Gaia will

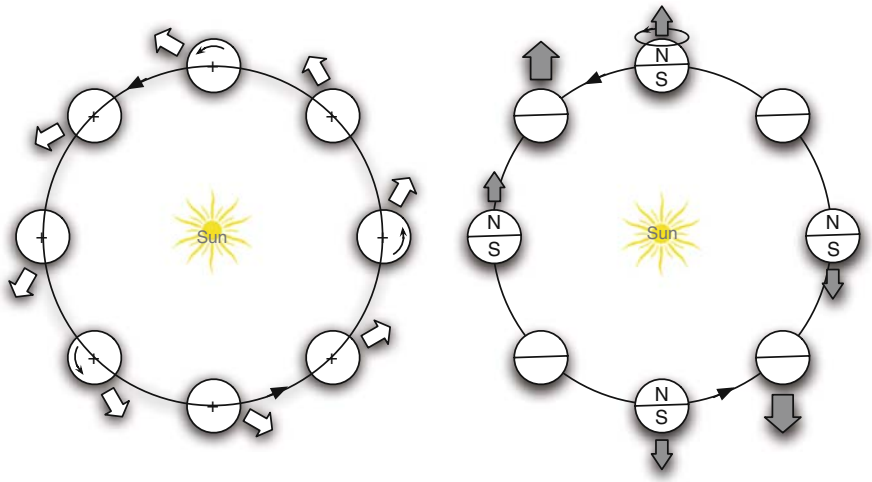

Fig. 22 The diurnal (*left*) and seasonal (*right*) Yarkovsky effects. Depending on the prograde (resp. retrograde) spin axis direction, the force will secularly decrease (resp. increase) the semi-major axis of the asteroid (diurnal). In case of zero obliquity (purely seasonal) one always has $da/dt \leq 0$ (Credits GSFC NASA)

not be able to derive all physical parameters involved in the Yarkovsky effect, but for some NEOs can derive one scaling parameter. This means that if the diameter and spin state are well known, the thermal inertia can be derived for a half dozen of bodies [79]. Again when not completely negligible such systematic effect can affect other parameter estimations, such as those involved in the local test of general relativity.

8.2 Local Tests of General Relativity and Reference Frame Link

There are three historical and fundamental tests that assessed with some success[6] the theory of General Relativity of Einstein (1916): precession of the perihelion of Mercury, light deflection by the Sun and gravitational redshift. A fourth can be added to that list that arrived later, the radar-echo delay, also called Shapiro effect. Last, let us note the importance of the work of Schwarzschild who derived—under given hypotheses—an exact solution to the equations enabling to predict quantitatively these effects.

It was known since Le Verrier in the middle of the nineteeth century that the (Newtonian–Euclidian) planetary theories from celestial mechanics could not match with the precession of the perihelion of Mercury's orbit around the Sun of 43″/cycle.

[6] The two experiments for the light deflection during the Solar eclipse of 1919 were not free of some errors, but nevertheless gave a result close enough to the prediction to convince Eddington and Dyson and further the scientific community. The precession for Mercury's perihelion matches the observations but the part due to the oblateness of the Sun is not well known.

No satisfactory explanation could be given until the theory of general relativity that nicely predicts this very effect. Alternative theories to the GR do exist; a particular class of *metric theories* can be linearised and grouped in a parameterised formalism when gravity is weak and massive bodies move slowly $v \ll c$, the parameterised post-Newtonian formalism [131]. This expansion, in its "first" order[7] corresponds to the classical Newtonian gravity, and in its "second order", to post-Newtonian corrections [76]. Among the parameters of this formalism, γ reflects the light deflection and β the relativistic precession of the perihelion, they are both equal to one in general relativity. Test of general relativity, at least when dealing with phenomena in the Solar System, usually consists in measuring the deviation to the canonical value predicted by GR for these PPN parameters [129, 130].

Excluding the part due to planetary perturbations that can be modelled precisely, the precession of the perihelion of the orbit of a solar system body is governed by the relativistic effect and an additional effect due to the oblateness of the Sun, given by its quadrupole moment J_2. Within the PPN formalism, the secular term for the argument of the perihelion ω is given by

$$\begin{aligned}\Delta\omega &= \Delta\omega_{|PPN} + \Delta\omega_{|J2} \\ &= \left[\frac{6\pi m_\odot}{a^{5/2}\,(1-e^2)}\,\Gamma + \frac{6\pi R_\odot^2}{4}\,\frac{5\cos^2 i - 1}{a^{7/2}\,(1-e^2)^2}\,J_2\right](t-t_0) \\ &= \frac{3m_\odot}{a\,(1-e^2)}\left[\Gamma + \frac{R_\odot^2}{4\,a\,m_\odot}\,\frac{(5\cos^2 i - 1)}{(1-e^2)}\,J_2\right] n\,(t-t_0), \qquad (21)\end{aligned}$$

where $m_\odot = GM_\odot/c^2 \approx 1.48\,\mathrm{km}$, $\Gamma = \frac{2+2\gamma-\beta}{3}$, a is in AU, t in year, n in rad/year and, last, $\Delta\omega$ is in radian. Neglecting the contribution due to the Sun, the relativistic effect from the general relativity is $\Delta\varpi = \Delta\omega = 6\pi m_\odot\,[a\,(1-e^2)]^{-1}\,\lambda_p$ [rad/cycle], which for Mercury yields a precession rate of 43″/cycle. However it clearly appears from the equation above that the effect of the Sun and the relativistic effect are linearly coupled, and one cannot derive both unknown parameters from one single value of the precession of one planet. On the other hand, having different bodies spanning a larger range of eccentricities and semi-major axis improves the separation of the unknowns; moreover, there is an additional effect on the precession of the ascending node $\Delta\Omega = \Delta\Omega_{|J2}$ that arrises from the Sun only, while there is no particular relativistic effect on that argument.

The asteroids act as test particles in the gravity field of the Sun, having hence orbits that are well defined by Gaia enables—in theory—to derive both parameters separately. Performing simulations of the motions and Gaia observations of synthetic populations of NEAs [50] again yields a variance analysis. Since the actual known population of NEAs is not complete to apparent magnitude 20, we should not restrict ourselves to the known population only, but estimate also the NEAs yet to be discovered before Gaia. Heuristic samples have been produced following the (un-correlated) distribution in orbital elements and absolute magnitudes derived by

[7] All small parameters are expanded simultaneously.

[10]. For each possible set of Gaia observations we have determined the variance matrix for the unknown parameters including Γ and J_2 either simultaneously or separately.

Similarly one can add as global parameter a possible variation of the gravitational constant G, a problem similar to Gyldén–Mestchersky problem [71], also present in the analysis of lunar-laser Ranging data and planetary ephemerides [132, 94]. One should note that depending on the definition of the osculating elements, the results for the variation can differ [58, 43]. In addition one can also perform a link of reference frames. All computed positions are given with respect to a frame *defined* by the equations of motions and associated parameters (time scale, masses, etc.): the dynamical reference frame. This frame is materialised by the positions of solar system objects. All observed positions, in contrast, are given with respect to a frame *defined* as kinematically non-rotating. This frame is materialised by the distant QSOs (extra-galactic, quasi-stellar objects) and next by reference stars for which the motion is well modelled. Both reference frames are non-rotating, either dynamically or kinematically; nevertheless, a global rotation might subsist. In any case the reference plane and origin of longitudes in this plane are conventional and completely independent in each frame; in the case of the ICRF (the reference frame defined by the QSO from VLBI astrometry), it is close to the FK5 J2000 mean equator and equinox but the x-axis does not point exactly in the direction of the vernal point. Again, all accurate observations of solar system objects can act as reference point to derive the rotation vector $\mathbf{W} = \mathbf{W_0} + \dot{\mathbf{W}}\,(t - t_0)$, i.e. both the rotation for the link to the dynamical reference frame W_0 at some given reference epoch t_0 and a possible test of rotation rate $\dot{\mathbf{W}}$. One sees from Table 1 that the rotation rate will be derived with a precision similar to the one of the Gaia optical ICRF reference frame itself.

Table 1 Standard deviation ($1\,\sigma$) for various global parameters derived simultaneously by Gaia. The rotation and rotation rate vector components are given in the ecliptic J2000

	β	J_2	$\dot{G}/G$	$\mathbf{W_o}$	$\dot{\mathbf{W}}$	
	–	–	yr^{-1}	μas	μas.yr^{-1}	References
Gaia	5×10^{-4}	2×10^{-8}	2×10^{-12}	2-2-5	1-1-2	[50]
LLR	2×10^{-4}	–	9×10^{-13}	–	100	[132]
EMP	1×10^{-4}	3×10^{-8}	1×10^{-13}	–	–	[94]
INPOP	–	2×10^{-8}	–	–	–	[36]

$1\,\mu$as/yr $\sim 5 \times 10^{-12}$ rad/yr

9 Orbit Determination and Improvement

9.1 Introduction—A Historical Perspective

Determination or more exactly the representation of orbits is an old problem starting, after Babylonian tables, with the epicycles and later with the work of Kepler who derived the correct form of orbits with the conics [106]. Another ancient aspect

connected to this topic is the determination of the orbits of comets which shows some interest, or more recently the orbits and motion of meteoroids and meteor streams.[8] As noted by Gauss in his fundamental "*Theoria Motvs*[9]" preface [39, 38, 40], the problem for the comets (and now, for the asteroids too) is different from that for the planets for which many observations can or have already been gathered. In fact this amount of data collected principally by Tycho Brahe led Kepler to find his eponym laws; now, in the case of the small bodies, the problem to solve is different: we know the orbit has to be an ellipse or an hyperpola and we want to derive its parameters with only few data in hand. As shown by the historical examples of Olbers, and Halley for his famous comet,[10] and also Piazzi, Gauss, von Zachs for Ceres, the practical significance to the determination of an orbit is to be able to compute the ephemeris and hence to be able to track the object in the telescope (consider nowadays typical $\approx 10'$ large/small fields of view) at its next apparition. If not, the body can be lost, necessitating some painful effort to catch it again; one illustrative example is given by the long-lost asteroid 719 Albert, discovered in 1911, it was re-discovered as 2000 JW8 in year 2000, about 9 decades later. With modern archives data mining techniques [117] it becomes also interesting to be able to go back into the past and see whether additional data already exist.

Before the discovery of asteroids and the particular work of Gauss, the history of orbit determination was closely related to comets. Indeed comets were intriguing and furtive objects yet bright and showing large motion in the sky; but even their parallax remained for long unknown. Many attempts to derive their orbits were unsuccessful, mainly because the exact nature of the curve was not known,[11] and often assumed to be a straight line. Newton proposing in his "*Principia*" (1686, book III, prop. XLI) a geometric method based on a description by a parabola with the Sun at the focus paved the way to future successes. While we shall note some contributions from Euler, Lambert and Lagrange in the field, the comet of Halley—or comet of 1759—is the first case for which an orbit and precise prediction was computed. Halley, following the work of Newton, identified in 1705 from the computation of the *parabolic* orbital elements, the apparent periodicity and successive returns of this celestial body (meaning the orbit is an elongated ellipse rather than a parabola). Next, from the calculation of the planetary perturbations on the comet provided by Clairaut (1762), it was possible to give accurate enough predictions for the 1759 apparition with all its spectacular outcome and a beginning of a new era in the cometary science. The oldest apparition attested in observation goes back to 240 BC, in China, possibly 467 BC [29, and reference therein]. We will end

[8] Interestingly, it appears that comets were associated to meteors—in the etymological sense—by ancient philosophers [2].

[9] *Theoria Motvs Corporvm Cœlestivm in Sectionibvs Conicis Solem Ambientivm. Autore Carlo Friderico Gauss.*

[10] Nowadays comets are named after their discoverer, it was not the case at the time of Halley.

[11] Tycho Brahe having observed the comet of 1577 noted it could not be on a circular orbit, hence differing from the planets.

this short historical review by mentioning the later work of Olbers, and by far the most used technique for deriving the orbits of comets. Contemporary to Olbers, Gauss recognised the advances obtained by Newton for the orbit determination of comets: "The great Newton himself, the first geometer of his age [...] he came out of this contest also the victor", but added, however, that having one of the unknown removed (either eccentricity or semi-major axis for the parabola) advantageously reduced the complexity of the problem (there are more equations than unknowns and no transcendental equations). The problem of the determination of an asteroid's orbit generally resides in the determination of the elliptic elements having several topocentric observations of the target spread over a limited interval of time (say a few weeks). In many situations in the Solar System—and in contrast to a full N-body problem—the Keplerian two-body problem studied by Newton gives a satisfactory approach; which can be, if needed, iteratively developed to higher degree of accuracy later on by considering the small perturbations. The two-body problem is of importance also because it has a complete solution in closed form. In the following we shall focus mainly on orbits of asteroids, possibly as test particles in our Solar System, and sometimes massive bodies that show mutual perturbations or gravitational perturbations on Mars and/or other inner planets. We will at some point consider a perturbed two-body problem (the mass and attraction of the Sun is preponderant, the attraction of the planets act as small perturbation) with small additional gravitational forces. We will not develop the case of comets, for which well-known non-gravitational acceleration can be important, neither will we discuss here the problem of the orbit determination of meteoroids or meteors, see [29, Chap. 12]. We will not discuss the case of range and range-rate data although it is of major importance in modern data for NEAs and space vehicles. We will focus mainly on the problem of orbit determination for an asteroid given classical telescopic RA and/or Dec data.

The most impressive work performed on the determination of the orbit of asteroid was performed by a great mathematician and astronomer, Carl Friedrich Gauss.[12] It took some time to Gauss to publish[13] his work in the "*Theoria Motvs*" (1809), where he explains and exposes his method with great details but certainly with additional refinements that were not used in the original work applied to the Ceres case. Indeed the illustrative example of the orbit determination is given not for asteroid (1) Ceres but for (3) Juno [40, Book II, Sect. I]; Ceres is also treated with the additional observations obtained in 1805. One can note a few technical details that maybe are of no more use today: little use of rectangular coordinates, use of manipulations to carry out calculations with the aid of logarithms, use of a fictitious place of the Earth in the plane of the ecliptic to decrease the effect of unknown parallax and reduction of time of observations to the meridian of Paris. After the discovery of the

[12] Karl Friedrich Gauß for the German spelling, this can be helpful for bibliographic research.
[13] It also took some time to have the text translated, 1857, 1859 and 1864 for the English, Russian and French version, respectively. The collected works of Gauss were published starting from 1862.

Fig. 23 G. Piazzi (1746–1826) on the left showing—not Gauss but—his newly discovered minor planet Ceres Fernandinea; and C. F. Gauß (1777–1855) on the *right* in a portrait (extract) painted well after the Ceres story (he was 24 at that time) (Piazzi: (public domain) Gauss: painting from Christian Albrecht Jensen (1792–1870)-copies)

asteroid (or minor planet[14]) Ceres by the astronomer Piazzi at Palermo observatory in January 1801 (see Fig. 23), a large excitation arose to be able to predict its next apparition. Ceres[15] was observed starting from its discovery until the beginning of February down to a solar elongation of 70°, corresponding to a relatively short arc of 3° over 40 days. Being suspected to be a star, then a comet, it appeared clear at the discovery time that its orbit was more likely planetary. With his pioneering work and correspondence with astronomers, Gauss was able to predict Ceres' next apparition in December 1801 with a small 0.1° error as observed by Von Zach and next Olbers 1 year after its discovery. This was better than any other predictions and assessed the success of Gauss' method and collaborator observers[16]. The case of Ceres at discovery is given from his correspondence in the collected works, Band VI [37], with the data from 2 January, 22 January and 11 February 1811. These dates are not too much separated in time and one moreover has $t_2 \approx 1/2\,(t_2 + t_3)$. The presentation of the method of orbit determination is generally separated in two

[14] The discovery of Piazzi is in some sense remarkable, he did find—depending on the nomenclature in use—the first planetoid, the first minor planet the first asteroid, the first dwarf planet, and possibly [69] the first trans-Neptunian that was injected into the inner Solar System!

[15] Originally called Ceres Fernandinea by Piazzi, referring to the King Ferdinand III of Sicily.

[16] All these names are now associated with asteroids: 999 Zachia, 1000 Piazzia, 1001 Gaussia and 1002 Olbersia.

parts, as was done by Gauss; in a first time one develops relation in the orbit and in space from positions or position and velocity, and in a second time one derives the method starting from three or more geocentric observations.

9.2 Orbit Determination

Since the pioneering book of Gauss, several textbooks summarise the problem of the determination of an orbit (elliptic or parabolic) and the evolution with time up to the modern ones allowing the use of fast computing machine [85, 86, 119, 93, 29, 32, 99]. Considering the two-body problem in the rest of this section, there are two classical problems encountered in celestial mechanics and astrodynamics computation of the position and velocity at a given time of a body given the orbital elements, or, inversely, obtain the orbital elements given the initial position–velocity conditions or state vector [99]. The latter form is of particular interest since it says that the (heliocentric or barycentric) orbit is completely defined from the knowledge of such (heliocentric) state vector. Another possibility to fully characterise the orbit is to have two (heliocentric) radius vectors, at any two different epochs (excluding, however, data separated by half periods or colinear vectors). These are the basis of the methods for the determination of orbits developed by Gauss and Laplace. They are in fact very similar as pointed out by H. Poincaré in his preface of the Leçons[17] from Tisserand [120]. The problems of Laplace and Gauss are indeed very similar but technically different and have encountered various success depending on the practical aspects to solve. Both Gauss and Laplace approximations will provide reliable results, possibly with different convergence speed. One can also note here that some practical refinement has been proposed by various authors (in particular [98] for the method of Gauss, or [67] for the method of Olbers). Actually, Gauss method is most widely used (it was already the case before modern computers [95]), although Laplace can be applied with modern CCD observations that allow to estimate the apparent motion of the body, e.g. [15]. It remains dependent on the precision with which one is able to derive the second derivatives of the right ascension and declination [29]. A comparison of the practical aspects with modern computations can be found in [17] (which also introduces a less famous method developed by the Italian astronomer Mossotti). The conclusion of their simulation runs is that Gauss method gives, in general (more than twice times), better results (when compared to the reference orbital parameters a, e, i) than the one of Laplace.

All methods for deriving the orbit of an elliptic motion necessitate to solve non-linear problems, involving transcendental functions, hence with no close form or general algebraic solution. Approximations are thus needed which generally also imply that observations are not too far apart but neither too close to encounter degeneracy of the problem. Gauss method supplants the one of Lagrange, in fact Gauss showed that it is more convenient to reduce the problem to two unknowns x, y

[17] Note that only the methods of Olbers and Gauss are developed in these Leçons.

involved in two equations $X = Y = 0$ as simple as possible. He also gave a method to compute the ratio of sector to triangle (Sect. 9.3.1), one of the fundamental quantities used. These classical methods require three (or more) observations, hence sufficient equations to derive the six unknowns of the problem. One other method of interest has been developed—in the case where only two topocentric observations are available—by Väisälä (in addition to the determination of a circular orbit, e.g. [118, 29]). If the orbital plane is close to the ecliptic the—classical—method fails, Gauss also considered this case and gives the extension to find the orbit from four observations. Before starting the orbit computation itself one can also want to correct the observations (apparent directions given in, say, the true frame of the date, though today such frame will have to be replaced by the CIP and its non-rotating origin) for various classical effects of (stellar) aberration, precession/nutation, parallax. These being in any case small, they can often be ignored on the first step.

We give in the following a sketch of Gauss' method of orbit determination. Let us as before reproduce a remark extracted from the preface of the French translator (E. Dubois [38]) of Gauss' work: "Or il est bien probable que la zone située entre Mars et Jupiter n'est pas encore suffisemment explorée et que le chiffre de 79 auqel on est arrivé, sera encore augmenté. Qui sait ce que réserve l'avenir !!... Bientôt alors les astronomes officiels n'y pourront plus suffire, si des calculateurs dévoués à l'astronomie et à ses progrès ne leurs viennent aussi en aide de ce coté[18]". Well, the future actually gave an increase by several orders of magnitudes of objects in that zone (and a few more in the Trojan and trans-Neptunian region) which number is still increasing exponentially; we shall ask again what will the future hold for us with ongoing surveys like Pans-STARRS, LSST and others?

9.3 Gauss Method—A Sketch

The general scheme of Gauss method is to reduce the problem to a system of two equations $X = Y = 0$ involving two variables x, y; one sets these two variables for two of the observations, derives the orbital parameters and tests whether the equations are satisfied for the third observation. Generally they would not, so the next step is to derive the (small) corrections λ, μ to apply to x, y from the knowledge of the first-order derivatives of $\partial(X, Y)/\partial x$, $\partial(X, Y)/\partial y$ and repeat the process until convergence to a solution that will satisfy all three observations together is reached. Actually Gauss does not give one, but five methods with additional extensions [40, Book II, Sect. I, 124–129] and comparison to the observed geocentric position itself is not always necessary. The first, and more natural, process given by Gauss is to take for the two variables x, y the geocentric distances (or the logarithm of these distances projected on the equator) for the first and last observations. From

[18] It is likely that the zone between Mars and Jupiter is not yet sufficiently explored and that the number of 79 that has been reached will still be increased. Who knows what the future will hold for us!! Soon official astronomers will not suffice, if some computers devoted to astronomy and its progress do not come to their rescue also on that side.

there on derives the heliocentric position vectors of the target, the orientation of the orbital plane in the inertial reference frame and—from the timing and longitudes in this plane—all other elliptical elements. Computing the values for the third–and middle–observation will yield two equations, that, for the computed orbit to be a solution, should satisfy $X = Y = 0$. Another variable that could be tested is the time difference for the third observations, but this is not as precise. The most common process is to take as unknown the heliocentric and geocentric distances of the asteroid at the second observation.

9.3.1 The f and g Series and Sector to Triangle Ratio

A Keplerian orbit is fully characterised from the knowledge of the six elliptic orbital elements or also from the initial conditions of the equations of motion $\mathbf{r}_o$, $\dot{\mathbf{r}}_o$ at some reference epoch t_o. Gauss also showed that it can equivalently be derived from the knowledge of two heliocentric radius vectors $\mathbf{r}_1$, $\mathbf{r}_2$ at time t_1 and t_2.

The so-called f and g series are of fundamental use in such problems of preliminary orbit determination and approximation. Following [27, 32, 99] we shall introduce them broadly here. Starting with the equation of motion

$$\frac{d^2\mathbf{r}}{\mathbf{dt^2}} + \mu\frac{\mathbf{r}}{r^3} = 0 \tag{22}$$

and introducing a transformation of time $\tau = \mu^{1/2}\, t = k\, t$, one can write

$$\frac{d^2\mathbf{r}}{\mathbf{d}\tau^2} + \frac{\mathbf{r}}{r^3} = 0 \tag{23}$$

and

$$\mathbf{r} = f(\tau)\,\mathbf{r}_o + g(\tau)\left(\frac{d\mathbf{r}}{d\tau}\right)_o. \tag{24}$$

Expressing $\mathbf{r}$ in Taylor expansion from the starting position $\mathbf{r}_o$ and expressing also the derivatives $d^n\mathbf{r}/d\tau^n$ from (23) one gets the f, g series:

$$\begin{aligned} f &= 1 - \frac{\tau^2}{2\,r^3} + \frac{\tau^3}{2\,r^4}\frac{dr}{d\tau} + o(\tau^4), \\ g &= \tau\left(1 - \frac{\tau^3}{6\,r^3} + \frac{\tau^3}{4\,r^4}\frac{dr}{d\tau} + o(\tau^4)\right), \end{aligned} \tag{25}$$

note that g is given, without particular reason, to order $O(\tau^4)$. On the another hand, by introducing the eccentric anomalies, one can also derive a closed form for f and g:

$$f = 1 - \frac{a}{r_0}\Big(1 - \cos(E - E_0)\Big),$$
$$g = \tau - \frac{a^{3/2}}{\mu^{1/2}}\Big((E - E_0) - \sin(E - E_0)\Big), \tag{26}$$

where E is the eccentric anomaly. Introducing the true anomalies v in the orbital plane, one can write one of Gauss' fundamental relations giving the ratio of sector to triangle:

$$y = \frac{\sqrt{a\,(1 - e^2)}\,\tau}{|\mathbf{r}_o \times \mathbf{r}|}. \tag{27}$$

Another formulation, not reproduced here, will give a similar—yet more complex—relation with the eccentric anomalies. This ratio can be obtained from continuous fraction of Hansen:

$$y = 1 + \frac{10}{9}\,\frac{h}{1 + \frac{\frac{11}{9}\,h}{1 + \frac{\frac{11}{9}\,h}{1 + \cdots}}} \tag{28}$$

involving the quantity $h = h(r, r_o, \tau)$.

Starting from the knowledge of the two radius vectors $\mathbf{r}_1, \mathbf{r}_2$ at time t_1 and t_2, Gauss provides—from the ratio of sector to triangle relation—a formulation to compute f, g from (26) . We can now write

$$\dot{\mathbf{r}}_2 = (\mathbf{r}_1 - f\,\mathbf{r}_2)/g \tag{29}$$

and consider that the orbit is fully characterised from the knowledge of the state vector $(\mathbf{r}_2, \dot{\mathbf{r}}_2)$ at time t_2.

9.4 Orbit Determination from Three Positions

Since we are looking for a Keplerian orbit, all three position vectors $\mathbf{r}_i$ at time t_i are coplanar, so that one can write

$$\mathbf{r}_2 = \frac{[\mathbf{r}_2, \mathbf{r}_3]}{[\mathbf{r}_1, \mathbf{r}_3]}\,\mathbf{r}_1 + \frac{[\mathbf{r}_1, \mathbf{r}_2]}{[\mathbf{r}_1, \mathbf{r}_3]}\,\mathbf{r}_3, \tag{30}$$

where the coefficients of this liner relation are exactly the ratio of the era of the triangles formed by the respective vectors. Taking into account now Kepler's law of era, Gauss showed that one can get iteratively an approximation of the triangles era $[\mathbf{r}_1, \mathbf{r}_3]$ from the elliptic sectors era $(\mathbf{r}_1, \mathbf{r}_3)$ with high precision. Expressing the areas from the vector cross products, and making use of the f, g series (Sect. 9.3), one will get an approximation for their formulation:

$$c_1 \equiv \frac{[\mathbf{r}_2, \mathbf{r}_3]}{[\mathbf{r}_1, \mathbf{r}_3]} = \frac{g_3}{f_1 g_3 - f_3 g_1},$$
$$c_3 \equiv \frac{[\mathbf{r}_1, \mathbf{r}_2]}{[\mathbf{r}_1, \mathbf{r}_3]} = \frac{-g_1}{f_1 g_3 - f_3 g_1}. \tag{31}$$

Putting

$$\tau_1 = k\,(t_3 - t_2)$$
$$\tau_2 = k\,(t_3 - t_1)$$
$$\tau_3 = k\,(t_2 - t_1) \tag{32}$$

and expressing the f, g coefficient with the modified time, one has

$$c_1 = \frac{\tau_1}{\tau_2}\left(1 + \frac{\tau_2^2 - \tau_1^2}{6\,r_2^3}\right) + o(\tau^3),$$
$$c_3 = \frac{\tau_3}{\tau_2}\left(1 + \frac{\tau_2^2 - \tau_3^2}{6\,r_2^3}\right) + o(\tau^3), \tag{33}$$

which still involves three unknowns. Examination of these relations shows that taking for the time of the second observation, $t_2 = 1/2(t_1 + t_3)$, will derive this approximation with higher precision. Writing the simple triangle vectorial relation Sun, Earth and asteroid position at t_i

$$\mathbf{r}_i = \mathbf{R}_i + \boldsymbol{\rho}_i, \tag{34}$$

where $\mathbf{r}_i$, $\boldsymbol{\rho}_i$ are the heliocentric and geocentric positions of the asteroid and $\mathbf{R}_i$ is the heliocentric position of the Earth, one obtains the system of three equations:

$$\rho_2 - c_1\rho_1 - c_3\rho_3 = c_1\mathbf{R}_1 + c_3\mathbf{R}_3 - \mathbf{R}_2, \tag{35}$$

and after some manipulation, one finally obtains a relation involving only $\mathbf{r}_2$, $\boldsymbol{\rho}_2$ of the form:

$$\boldsymbol{\rho}_2 = A + B/r_2^3, \tag{36}$$

and similar for others ρ_i, r_i.

This provides one equation involving the two unknowns ρ_2, r_2. Simple geometry of (34) at t_2 yields an additional one:

$$r_2^2 = \mathbf{R}_2^2 + 2\,\mathbf{R}_2 \cdot \boldsymbol{\rho}_2 + \rho_2^2. \tag{37}$$

This system of two equations and two unknowns can be solved numerically. Gauss, however, either reduced the system to a single equation of the eighth degree or to a

transcendental equation. Having solved for r_2 and ρ_2 yields the other ρ_i and next all $\mathbf{r_i}$. From the three position vectors at the three dates, or equivalently $\mathbf{r_2}$ and $\dot{\mathbf{r}}_\mathbf{2}$ at t_2, the orbit is fully determined (see Sect. 9.3).

Starting from a first good approximation, it is an iterative method which improves the results at each step. In the next iterations the values for the triangles ratio c_1, c_3 given above to $o(\tau^3)$ will be improved. There are several ways to do so, Gauss developed his use of the sector to triangle ratio Sect. 9.3, and no new equations are to be used. Making use of Kepler's law of era, the ratio of triangles can be expressed as a function of the ratio of sectors:

$$c_1 \equiv \frac{[\mathbf{r}_2, \mathbf{r}_3]}{[\mathbf{r}_1, \mathbf{r}_3]} = \frac{(\mathbf{r}_2, \mathbf{r}_3)}{(\mathbf{r}_1, \mathbf{r}_3)} \frac{y_2}{y_1} = \frac{\tau_1}{\tau_2} \frac{y_2}{y_1}, \tag{38}$$

where now the modified time should include corrections for light-time travel, etc. As noted by Gauss himself, three iterations are generally sufficient. This basic method is very performing when the angular observations are close together (but neither too much, to the point they would correspond to a single observation), the case for which Gauss designed his algorithm(s). Gauss also gives conditions for the method to work [40, Book II, Sect. I, 130]. If the observations are spread over a large interval of time, the method is less well suited and might not converge. One can instead use the "double r-iteration" technique [32] or very similarly the "statistical ranging" technique (Sect. 9.5). In that case one simply takes randomly—with some initial guess on the nature of the object, NEA, MBA, etc.—two values for the geocentric distance at two of the observations. This defines all two radius vectors and one possible orbit. The "double r-iteration" will iteratively correct the initial guess for the distances to converge to a solution, the statistical ranging will more simply proceed to straight trial/error Monte Carlo sampling of possible values.

9.4.1 The Method of Laplace—Briefly

In the method of Laplace one writes the first and second derivatives of the position vector $\mathbf{r}$ considering that the observations can provide the corresponding derivatives for the observed apparent direction. Let us write the heliocentric position vector of (34):

$$\mathbf{r}_i = \mathbf{R}_i + \rho_i \mathbf{u}_i \quad ; \quad |\mathbf{u}_i| = 1. \tag{39}$$

Putting this expression in the equations of motion (22), one has a system of four unknowns $(r, \rho, d\rho/dt, d^2\rho/dt^2)$, the quantities for the Earth being given by the ephemerides.[19] One again will write the additional equation (37) from the triangle. The system can be reduced to an equation of an eighth-degree resultant (as with Gauss' method) involving only r_2 for the observation at time t_2. This is solved

[19] To be found in the "Connaissance des Temps" for the Leçon of F. Tisserand, the "Astronomical Almanach" for Roy, or at URL http://ssd.jpl.nasa.gov and http://www.imcce.fr, here.

numerically by successive approximation. The characteristics of the solution have been studied by Charlier [93, Chaps. 3–21] who gave a geometrical interpretation for the various roots from a strophoid curve. The first and second derivative of $\dot{\mathbf{u}}$ are often derived from an expansion instead of being measured, this is also given by Laplace. The approximation of the geocentric distance will not be obtained as accurately than with Gauss method.

9.5 Other Methods

Before developing the linear case that is well adapted to refinement of asteroids' orbit with large number of observations, of good quality and over large space and time span, we will briefly present non-linear techniques than can be of interest for different purposes, in particular when convergence of the linearisation method is not obtained or when the problem remains highly non-linear and/or non-gaussian. Some of semi-linear or statistical inversion methods are presented in [11]; we also note with particular emphasis a filtering method already discussed some time ago [31]. Some of the methods making use of descent gradient or filters will still need some linearisation parts and/or some relatively good initial guess of the solution. They do not exactly correspond to the case of what we have defined here as orbit determination, because they can fail if the initial conditions taken are too far from the true—and completely unknown—one. Another quite different approach is to tackle the problem of parameter estimation from the other side by some trial/error process: take one set of parameters randomly (either elliptic or equinoctial elements, or state vector, etc.) and test if it is a solution (that is, including some possible error). One illustration can be given, in the one-dimensional case, by the dichotomy method for finding the real root of some monotone function $f(x) = 0$. Here one will sample the $\mathbb{R}$ space randomly with, however, some additional improvement from the fact that $x_i \leq x \leq x_{i+1}$, where $f(x_i).f(x_{i+1}) \leq 0$. Some other methods rely on Monte Carlo technique, MCMC or Bayesian statistical inversion [63], or also genetic algorithms. In this case the full space-phase of the unknowns is sampled *with some adapted strategy* or algorithm to find the solution(s) in terms of χ^2 values. These techniques are fully generic and well adapted to modern computation facilities including possibilities of parallel computations. The genetic algorithm has been introduced before in Sect. 6.4 for deriving unknown parameters from the measured photometric data, they are also used recently in orbit determination of extra-solar planets from measured radial velocities [88, 24]. Statistical ranging [125] is another algorithm of statistical inversion that provides the full set of orbits from Monte Carlo technique. This method is particularly well adapted to the case where observations are scarce and cover short arcs. It is very similar in nature to the method described before for the orbit determination when two positions at two dates are known. Given two observed geocentric directions $\mathbf{u}_i$ at two dates t_1 and t_2 (generally close in time) one will now span the range of plausible geocentric distances ρ_i for both observations to construct two heliocentric positions ($\mathbf{r}_i = \rho_i \mathbf{u}_i + \mathbf{R}_i,\ t_i$). Once the candidate orbit is computed from these two positions, it is tested whether it fits all

available observations. This process is repeated in a Monte Carlo run until sufficient successful trials have been obtained for sampling the actual distribution of the orbital elements. The algorithm is combined to Bayesian approach involving a priori knowledge of the unknown quantities, this yields the full probability density of the solution. Obviously, when the number of observations becomes large, spanning a large range in time for several orbits, and the problem can be linearised locally; any trial/error methods will become not only much less efficient, but useless compared to classical linear least squares.

9.6 Orbit Improvement

The previous section was dealing with the orbit *determination*, i.e. find the orbital elements of the newly discovered asteroid, with little information and few measures, while nothing is known on its actual characteristics (distance to Sun or Earth, eccentricity, etc.). The previous method was based on the assumptions that the observations were not too much separated in time and/or space. When more data are acquired and over a larger time span compared to the orbital period, one shall get a better idea of these elements. Having at our disposal some initial guess of the orbit, we will now look for orbit *improvement* (or differential correction of orbit); this is the aim of the second part of Gauss' book, where he also puts the basis of the least squares method that we are going to present briefly. There are several methods for orbit improvement, the most commonly used being the linear least squares (LLS) from differential correction that will converge rapidly to the least squares solution. If the system is highly non-linear one might use a general least-squares method (e.g. Levenberg–Marquardt). In both cases, the solution is the one that will minimise the L^2 norm of the residuals,[20] and the confidence region around this solution will be an n-axial ellipsoid obtained from a local linear approximation. Last, in more general cases where the problem to invert is non-linear and the input data are not of the same distribution, one can obtain maximum likelihood estimators (MLE) other than the least squares solution.

We will hereafter develop the LLS method for orbit improvement of asteroids and briefly discuss another approach from statistical inversion. We also implicitly assume the object is an asteroid and do not consider here the cases of satellites of the Earth, planetary satellites or comets, for which different adapted methodologies have been developed.

9.7 Linear System—LLS

Starting from the equations of motion and writing the Taylor expansion to first order for the barycentric (resp. heliocentric) position vector denoted by $\mathbf{x}(t)$ (resp. $\mathbf{r}$)

[20] One can also, in case of poor robustness of the least squares solution, consider other norms such as the L^1 one.

$$\mathbf{x}(\mathbf{q_o} + d\mathbf{q_o},\, p_1 \cdots p_m, t) = \mathbf{x}(\mathbf{q_o},\, p_k, t) + d\mathbf{x} \\ = \mathbf{x}(\mathbf{q_o},\, p_k, t) + \frac{\partial \mathbf{x}(t)}{\partial \mathbf{q_o}}\, d\mathbf{q_o} + o(dq_o^2), \qquad (40)$$

where the position/velocity state vector $\mathbf{q_o} = [\mathbf{x}(\mathrm{t_o}), \dot{\mathbf{x}}(\mathrm{t_o})]$ gives the initial conditions at t_o of the system, the one obtained from the orbit determination in Sect. 9.2, and where p_k are other dynamical parameters such as the perturbing planets and asteroids masses, etc. The computed geocentric directions[21] of the object $\mathbf{u}(t_i) = (\alpha_i, \delta_i, t_i)$ are derived from the barycentric (resp. heliocentric) position of (40) $\mathbf{u}(t) = \langle \mathbf{x}(t - \tau) - \mathbf{x}_{\mathrm{E}}(t) \rangle$, where τ accounts for the light-time travel and $\mathbf{x}_{\mathrm{E}}$ is the position of the Earth.

9.8 Partial Derivatives

After correcting these directions for aberration, light deflection, precession/nutation, etc., one can compare them to the observed directions and derive the differences O–C vector $(\Delta\alpha_i \cos\delta_i, \Delta\delta_i, t_i)$. Further, by neglecting all terms of the order of $o(dq_o^2)$ for the small corrections $|d\mathbf{q_o}| \ll 1$, one writes the linear system to solve

$$O - C \equiv \mathbf{b} = \mathbf{P} \,.\, \left[\frac{\partial \mathbf{x}(\mathrm{t})}{\partial \mathbf{q_o}} \right] .\, \mathrm{d}\mathbf{q_o}, \qquad (41)$$

where the expression of $\mathbf{P}$ defining the projection from three-dimensional to the n-dimensional observational space is left to the reader (see, e.g., [13]). The 3×6 Jacobian matrix

$$\mathbf{J} = \left[\frac{\partial \mathbf{x}(\mathrm{t})}{\partial \mathbf{q_o}} \right] \qquad (42)$$

has now to be computed. One can distinguish three different ways to compute such quantity in general, depending whether (a) nothing is known about the function and only tabulated values are available, (b) one can approximate the partial derivative computation to obtain analytical closed-form formulations and (c) we know the function to integrate and compute the variational equations. The first case also corresponds to the case where the variational equations might be too complex to derive and/or integrate.

a. **Finite difference**. The variant or finite difference method is practical for numerical computation either by using the Cartesian form or the elliptical elements for the initial conditions. It is obtained from the limit definition of a derivative:

[21] We restrict the discussion to classical telescopic observations, but these could be other quantities as, e.g., range and range rate with radar, laser ranging, or any other technique, and from other positions in space than the geocentre as well.

$$\frac{\partial f(\mathbf{q})}{\partial h} = \lim_{h \to 0} \frac{f(\mathbf{q}+h)}{h}$$

by

$$\frac{\partial f(\mathbf{q_o}, t)}{\partial h_i} = \frac{f(\mathbf{q_o} + h_i, t) - f(\mathbf{q_o} - h_i, t)}{2h_i} + o(h_i^3), \tag{43}$$

where the small variation h_i is applied to each element of the initial conditions $\mathbf{q}_o$ separately. With this formulation correct to the third order of the small parameter, one has to perform in general three numerical integration or ephemerides computation for $\mathbf{q}$, $\mathbf{q}_o + h_i$ and $\mathbf{q}_o - h_i$. One can reduce this number to two numerical integration with an approximation correct to only $O(h_i)$. The value of the quantity h_i has to be defined by the experience of the user; too small it increases numerical errors, too large it reduces the precision of the approximation. A practical value can be found by successive tests until the results of the partial derivative computation remain robust. It will, in general, be sufficient to use the same step for all objects and all observation dates.

b. **Two-body, analytical.** The analytical formulation with elliptic elements is given in [93, Chap. 7] [13, Chap. 9] and [29, Chap. 11], it is also given with Cartesian elements in [29, 32]. They are obtained from the derivation of the two-body problem. We reproduce here without details the matrix for the more general use as given by [13], it is given as a function of corrections to the elements $d\mathbf{q} = (dl_o + dr, dp, dq, e.dr, da/a, de)$ and the position and velocity $(\mathbf{x}, \dot{\mathbf{x}})$ at time t:

$$d\mathbf{x}(t) = \left[\frac{\dot{\mathbf{x}}}{n} \; ; \; \mathbf{P} \times \mathbf{x} \; ; \; \mathbf{Q} \times \mathbf{x} \; ; \; \frac{1}{e}\left(-\frac{\dot{\mathbf{x}}}{n} + \mathbf{R} \times \mathbf{x} \right) \; ; \; \mathbf{x} - \frac{3}{2} t\, \dot{\mathbf{x}} \; ; \; H\,\mathbf{x} + K\,\dot{\mathbf{x}} \right]. \tag{44}$$

The quantities H, K are given by

$$\begin{aligned} H &= \frac{r - a(1+e^2)}{a\, e(1-e^2)}, \\ K &= \frac{r\, \dot{r}(r + a(1-e^2))}{a^3 n^2\, e(1-e^2)}, \end{aligned} \tag{45}$$

and the vectors $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$ are given by the rotation, transformation from the conventional equatorial frame to the frame associated with the orbital plane and orbit periastron:

$$[\mathbf{P}; \mathbf{Q}; \mathbf{R}] = \mathcal{R}_{\mathbf{z}}(-\Omega).\mathcal{R}_{\mathbf{x}}(-\mathbf{I}).\mathcal{R}_{\mathbf{z}}(-\omega), \tag{46}$$

so that the variation of the angles $(d\Omega, dI, d\omega)$ is represented by new variables (dp, dq, dr) given in Fig. 24 representing an infinitesimal rotation along the

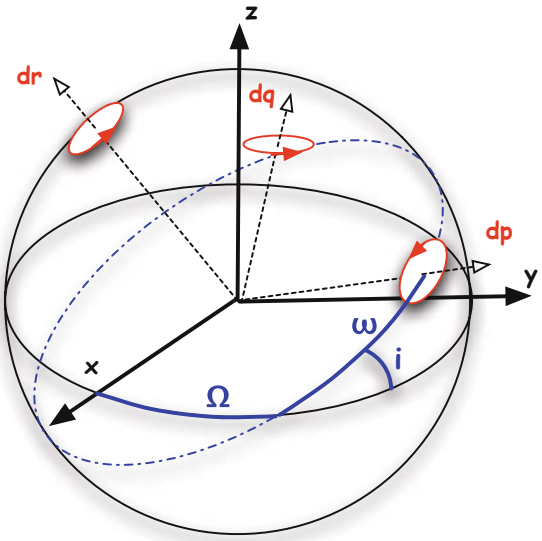

Fig. 24 Elements for the infinitesimal rotation associated with the asteroid orbit

directions of the periastron, the direction directly perpendicular in the orbital plane and the direction of the orbital pole, respectively.

c. **Variational equations**. The variational equations are well adapted to the perturbed two-body problem and are computed during the numerical integration of the equations of motion themselves. The reader may have noticed that we did not yet explicitly mention the system to solve[22]: the equations of motion. In the case of a test particle orbiting a massive central body of mass $M_\odot$ and perturbed by N planetary bodies (planets, dwarf planets, asteroids, etc.) of masses M_i, we can write the equation of motion for the heliocentric position:

$$\ddot{\mathbf{r}} = -GM_\odot \frac{\mathbf{r}}{r^3} + \sum_{i=1}^{N} GM_i \left(\frac{\mathbf{r}_i - \mathbf{r}}{|\mathbf{r}_i - \mathbf{r}|^3} - \frac{\mathbf{r}_i}{r_i^3} \right). \tag{47}$$

Because of its properties one can write the differentiation:

$$\frac{d}{dt^2}\left(\frac{\partial \mathbf{r}}{\partial \mathbf{q}_o}\right) = \frac{\partial \ddot{\mathbf{r}}}{\partial \mathbf{q}_o}.$$

The 3×6 Jacobian matrix $\mathbf{J}$ can hence be obtained by integrating the system

[22] It was only implicitely used in method (*b*), but in its simple algebraic form for the two body approximation.

$$\ddot{\mathbf{J}} = \frac{d}{dt^2}\mathbf{J} = \left[\frac{\partial \ddot{\mathbf{r}}}{\partial \mathbf{r}}\right] \mathbf{J} \tag{48}$$

and where now $[\partial \ddot{\mathbf{r}}/\partial \mathbf{r}]$ is given in closed form and depends on the dynamical system to consider. For our perturbed two-body problem one gets the second-order derivatives [8]:

$$\frac{\partial \ddot{\mathbf{r}}}{\partial \mathbf{r}} = -GM_{\odot}\nabla\left(\frac{\mathbf{r}}{r^3}\right) - \sum_{i=1}^{N} GM_i \nabla\left(\frac{\mathbf{r}_i - \mathbf{r}}{|\mathbf{r}_i - \mathbf{r}|^3}\right) \tag{49}$$

with the operator ∇ on vector $\mathbf{s}$:

$$\nabla\left(\frac{\mathbf{s}}{s^3}\right) = -\frac{3}{s^5}\begin{pmatrix} s_1^2 - s^2/3 & s_1 s_2 & s_1 s_3 \\ s_2 s_1 & s_2^2 - s^2/3 & s_2 s_3 \\ s_3 s_1 & s_3 s_2 & s_3^2 - s^2/3 \end{pmatrix} = \frac{1}{s^3}\left(\mathbf{I} - \frac{3}{s^2}\mathbf{s}\cdot\mathbf{s}'\right), \tag{50}$$

and where both $\mathbf{I}$ the identity matrix and the outer product $\mathbf{s}\cdot\mathbf{s}'$ are 3×3 matrices. One can extend such formulation to other perturbative forces and accelerations and also to the more general N-body problem [8]. The initial conditions associated with the system in (48) are given by $J_{11} = J_{22} = J_{33} = \dot{J}_{14} = \dot{J}_{25} = \dot{J}_{36}$ and $J_{ij} = 0$ elsewhere. The additional equations (49) having been written, one has to solve simultaneously from numerical integration (e.g. some typical methods for ODEs integration: Bülirsch & Stoer, Adams Moulton, Radau, RK-k, etc., cf. [33]) the system of (47) and (48), which will provide simultaneously the ephemerides and the partial derivatives.

Note that the effect of perturbing forces or perturbing bodies is not directly considered in cases (*a*) and (*b*), but through the computation of the position of the target only. This is particularly true for the analytical formulation of Brouwer and Clemence [13] given here, which formulation is somewhat hybrid since the position and velocities can also—and should—be computed not analytically, but from a numerical integration of the perturbed problem. In the case of the perturbed two-body problem, i.e. all forces of the perturbing planets are small and the asteroid does not influence the positions of the planets (no cross terms), the formulations (*b*) and (*c*) remain fully tractable. In the case of a fully N-body problem, more cross terms and equations that have to be integrated appear in formulation (*c*).

9.9 Observational Equations

The partial derivatives $\mathbf{J}$ being computed one can derive the solution and associated errors by classical linear algebra and matrices computations. The observational equations:

$$O - C \equiv \mathbf{b} = \mathbf{P} \,.\, \mathbf{J} \,.\, \mathrm{d}\mathbf{q_o} = \mathbf{A} \,.\, \mathrm{d}\mathbf{q_o} \tag{51}$$

give the observed and measured quantities as a function of the foreseen corrections $d\mathbf{q}_o$. The unweighted[23] least square solution to this linear system is given by:

$$\overline{d\mathbf{q}_o} = (\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\,\mathbf{A}' \,.\, \mathbf{b} \tag{52}$$

and the variance–covariance matrix $\boldsymbol{\sigma}^2(d\mathbf{q}_o)$ for the errors and correlation is the inverse of the normal matrix $(\mathbf{A}'\mathbf{A})^{-1}.\boldsymbol{\sigma}^2(\mathbf{b})$. In the simplest case of one asteroid orbit to improve, this involves a 6×6 matrix inversion. In the more general case where more unknowns (initial conditions to the Cauchy problem and addition dynamical, physical and instrumental parameters) have to be derived different techniques can be adapted such as Cholesky or QR algorithms for dense matrix; in the case of sparse matrices and iterative processes the conjugate gradient, which consist in minimising the quadratic form $f(\mathbf{x}) = (\frac{1}{2}\mathbf{x}'\,\mathbf{A}'\mathbf{A}\,\mathbf{x} - \mathbf{b}'\mathbf{A}\,\mathbf{x})$ will be preferred [97, 41]. Here we will focus on another method, based on the singular value decomposition (SVD), that is not optimal in terms of computation speed, but that is robust and is well adapted to the case of degenerate and/or rank-deficient problems. The SVD also deals with non-square matrix, which is the case here when the number of observations exceeds the number of unknowns. Having a $m \times n$ matrix $\mathbf{A}$, $m \geq n$, of observational equations one can write the SVD decomposition [97]:

$$\mathbf{A} = \mathbf{U} \,.\, \mathbf{W} \,.\, \mathbf{V}', \tag{53}$$

where $\mathbf{U}$, $m \times n$, is orthogonal to the left $\mathbf{U}'\mathbf{U} = \mathbf{1}$; $\mathbf{V}$, $n \times n$, is orthogonal $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{1}$; $\mathbf{W} = [w_i]_\mathrm{D}$, $n \times n$, is diagonal; and the w_i are called the singular values[24] of matrix $\mathbf{A}$. The solution of our LLS problem is given from:

$$\overline{d\mathbf{q}_o} = \mathbf{V} \,.\, [1/w_i]_\mathrm{D} \,.\, \mathbf{U}' \,.\, \mathbf{b} \tag{54}$$

which does not necessitate to explicitly compute the normal matrix, neither to invert any matrix. Obviously if one of the singular value is zero the above expression is not defined, or in other words the determinant of the normal matrix $(\mathbf{A}'\mathbf{A})$ is null too and there is no *unique* solution. Similarly if one singular value is small $(0 < w_i/w_{max} \ll 1)$ the matrix is ill-conditioned. This is encountered, for instance, in case of degeneracy of the problem or of high correlation between two (or more) unknown parameters, showing that given the available data they cannot be determined separately but only one (or more) linear combination can be derived. In such case the matrix can be considered to be rank deficient, and the solution retained—among all possible one—will be the one of minimal norm. The dimension of the

[23] One can weight the equations to better take into account different observations noise by multiplying $\mathbf{A}$ and $(O–C)$ by the diagonal weight matrix $\sqrt{\mathbf{p}} = [1/\sigma_i]_D$ where σ_i is the standard deviation of the observed quantity at time t_i as estimated by the observer.

[24] The singular values of $\mathbf{A}$ are related to the eigenvalues of the positive symmetric normal matrix $\mathbf{A}'\mathbf{A}$, $\lambda_i = w_i^2$.

kernel corresponds to the number of zero singular values, and a vector basis of this sub-space is given by the vectors of $\mathbf{V}$ corresponding to these zero singular values. In practice the SVD has one advantage, since in that case one simply sets to zero all small singular values, without any other modification to the computation software. The solution so obtained is the one of minimal norm, but the solution to the general problem is not unique and is given by $d\mathbf{q} = \overline{d\mathbf{q}_o} + \sum_{k=1,K} \alpha_k \mathbf{v}_k$, where K is the dimension of the (quasi)kernel and $\mathbf{A} \,.\, \mathbf{v_k} \approx \mathbf{0}$. To ensure the validity of the assumption that a singular value is zero one can test whether $\mathbf{A} \,.\, \mathbf{v_k} \approx \mathbf{0}$, and also that the residuals $\mathbf{v} = \mathbf{b} - \mathbf{A} \,.\, \overline{\mathbf{dq_o}}$ do not statistically differ, for instance from its L^2 norm, from the one obtained from the full inversion. Let us also mention the possible introduction of so-called consider covariance matrix which contains additional variance/covariance information on the model. This is useful when some parameters $\mathbf{c}$ were set in the observational or dynamical linearised model, whose parameters cannot be estimated from the data at hand but that, however are known with some uncertainty $\boldsymbol{\sigma}^2(\mathbf{c})$. The uncertainty of these assumed model parameters can increase the formal error of the unknowns [112].

Finally, independent of the method used to compute the partial derivatives of (42), and next to solve the linearised system of Eq. (51), one has a correction to apply to the initial conditions $\mathbf{q_o}$ that improves the asteroid's orbit and fits—to the least squares sense—all available data. However, the linearisation of the equations is valid as long as the corrections are small. Additional iterations will be performed until the corrections to be applied are statistically non-significant.

The least squares estimator (LSE) is only under certain given conditions (often satisfied) equal to the maximum likelihood estimator (MLE). We do not discuss here methods of generalised least squares (GLS) that apply to non-linear systems, these are iterative too, they also need to have a starting point close to the true solution, and last they generally provide a biased solution. In a linear model $\mathbf{Ax} = \mathbf{b} + \epsilon$, where ϵ is the error, the Gauss–Markov theorem states that if the errors are centred $E(\epsilon) = 0$ and also un-correlated and of same variance $\epsilon\epsilon' = \sigma^2\mathbf{I}$ (homoscedasticity) then among all estimators without bias, the LSE is the most *precise*, i.e. of minimum variance.

9.10 Confidence Region

In the previous sections we have derived a solution, an orbit, that fits the data. If one chooses to weight the equation with some particular rule, or equivalently to suppress some data, or to apply another norm than the L^2, the results will be different. If the sensitivity of the solution is high with respect to these changes, the robustness of the solution is poor and should be taken with caution. Moreover the observational data are not free of stochastic (and possibly systematic) errors as recognised by Legendre (1806), their noise can be considered to follow a centred Gaussian distribution $\epsilon \in \mathcal{N}(0, \sigma)$ (as did C. F. Gauss in his pioneering work). So, as for any physical measure, one has to ask how accurate and precise (see Note 5)

is the solution obtained? It is known for instance that in the case of the orbit of planet Neptune [119] or comet Lexell (Le Verrier again!) [121] that the orbital elements are in rather strong error when compared to those that can be derived with large number of observations, but they nevertheless reproduce the available data well and possibly they were the most probable solution. One will hence associate to any solution a confidence region. Other discussion and aspects to the orbit determination and improvement methods connected to observations of various type of asteroids can be found in [11] and [112] for satellites. In the case of a linear system, the observation noise being assumed to be Gaussian, the resulting LSE will follow a Gaussian distribution. The variance–covariance matrix $(\mathbf{A}'\mathbf{A})^{-1}$ defines the n-dimensional probability density function (p.d.f.) of the solution. Let us remind that the multivariate Gaussian distribution is entirely defined by the variance and covariance of the parameter. This is associated to an ellipsoid or to a quadratic form. One can hence decompose this matrix in the space of the proper elements $(\mathbf{A}'\mathbf{A})^{-1} = \mathbf{P}'.\mathbf{D}.\mathbf{P}$ where the matrix $\mathbf{P}$ gives the orientation of the principal axis of the ellipsoid in the frame of the physical parameters and where the diagonal matrix gives the standard deviation 1σ of the parameters in the eigenspace. Also all probabilities or confidence region can be obtained from the 1σ value; for instance $P(|x - \mu| \leq 1\sigma) = 0.68268$; $P(|x - E(x)| \leq 3\sigma) = 0.99730$. In the non-linear case GLS or if the observational noise is non-Gaussian, the LSE will provide one solution but little or no indication on its distribution or error. One can derive an error bar locally from a linear approximation, but if the error is large this estimation of the error or standard deviation can fail, giving rise to the question "What is the error on your error?". The χ^2 for a multivariate distribution can be complex and can be sampled around this solution by Monte Carlo run or by use of the technique given in [3, 61, 128]. Depending hence on the problem to solve, one will have very different topology for the probability distribution that—only in the case of the multivariate Gaussian of the linear case—will be summarised in a simple way from the variance–covariance matrix (see Fig. 25).

10 Binary Stars and Asteroids

We discuss in the following section the case of astrometric observation of binary and multiple systems (stars, asteroids, etc.). The star ζ UMA, Mizar, was the first one found accidentally to appear as a *double star* through a refractor by Italian astronomer G. Riccioli[25] in 1650, well before E. Halley noticed to the attention of the Royal Society that stars do have proper motions (1718). Later, the German-born English astronomer Sir W. Herschel—the one that discovered Uranus and some of its moons with his stupendous telescopes, the one that also coined the denomination "asteroid"—while cataloguing double stars, discovered their motion around one

[25] Riccioli might have been preceded by B. Castelli in 1617, reporting to Galilei about Mizar "una delle belle cose che siano in cielo".

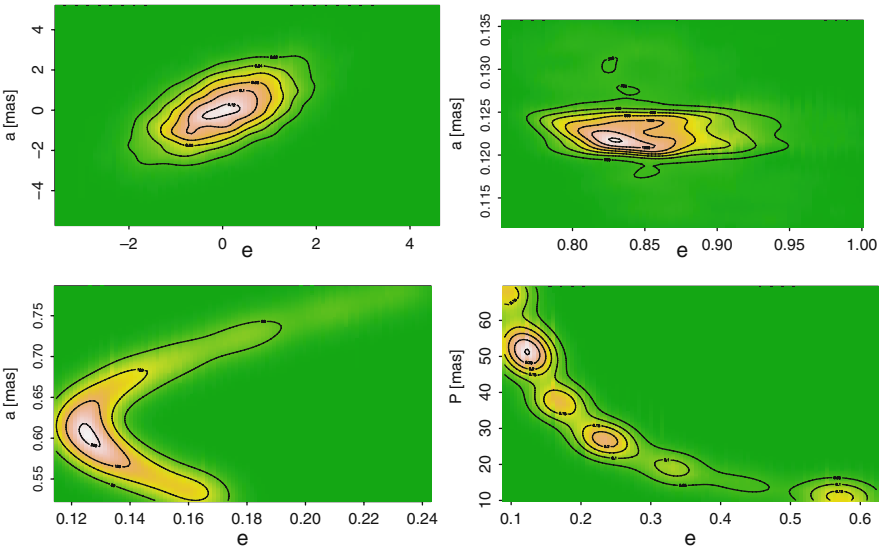

Fig. 25 Examples of two-dimensional projections from a multivariate orbital element distributions for the different cases of linear (*top-left*), semi-linear (*top-right*) and non-linear (*bottom*) problems of orbit determination/improvement

another, defining them as binary stars (1802, 1803) [1]. Since the time of Herschel, such observations and measures of *binary stars'* relative positions, in opposition to *optical double* that happen only by chance to be close on the sky from projective effect, are of high value. Of course they confirm that gravitation is universal and not only present for bodies orbiting around the Sun, for the orbits of satellites around their planets, but also for objects outside of our Solar System. Another very important impact is that it is possible to derive—with good accuracy—a fundamental parameter of the system otherwise inaccessible: the mass. The story for the asteroids in the Solar System is not much different starting with some supposition of their existence with no clear evidence [68], to the accidental discovery of the satellite of the asteroid Ida by the space probe Galileo, and next to detection from ground-based observations [70, and references therein]. Satellites of asteroids are found in the near-Earth asteroid population as well as in the main belt, the Trojans, Centaurs and the trans-Neptunian region. The typical mass ratio, orbital period and separation of these systems are generally different between the various populations, indicating different mechanism of formation (e.g. but neither exclusively nor definitely, YORP spin-up for the NEOs, catastrophic collision in the main belt, chaos assisted capture for the TNOs).

In the zoology of denominations of stellar multiple and binary systems depending on their nature and techniques of observations, we will consider two particular cases: resolved and astrometric binaries.

– *Resolved binaries*, also previously called visual binaries, are systems for which each component is clearly detected or separated in the telescope, either in the visible domain or in other wavelength[26];
– *Astrometric binaries* are, similarly to most of the presently detected extra-solar planets, systems for which only one component is visible (the brightest) but the reflex motion of the photocentre shows a wobble with respect to the barycentre.

There are several techniques of orbit determination, graphical, semi-graphical, analytical, that were adapted to different cases and to different observations techniques. We will develop hereafter the one of Thiele [115] an analytical one that provides the true orbit from three observations. Before that, let us briefly mention the graphical method of Zwiers [93] that derives the true orbit after having drawn the observed and apparent one. In this method one will measure one important parameter characterising the orbit: the areal velocity constant; in order to do so, one shall cut in some paperboard the elliptical sectors and simply weight them on a balance to derive $C' = \rho^2 \dot{\theta}$; this sounds straightforward indeed, but surely not much in use today! In the following we will focus on the orbit determination of a pair once it has been discovered, and/or only few sparse and scarce data is available. In a first approach all orbits can be considered to be Keplerian (two-body problem). We will derive, from statistical inversion, not only one single solution but the (usually broad) bundle of orbits that fit to the data and subsequently their density distribution (Fig. 26).

10.1 Resolved Binary—Thiele–Innes

In the case of a resolved binary, one has at his disposal the relative position of the two components, at the different epochs of observations. This position is often given in polar coordinates (North up and position angle counted positive from North to East) or more recently in Cartesian coordinates. Assume now we have such Cartesian coordinates for n observations (t_i, x_i, y_i); $i \in [1, \ldots, n]$ at time t_i. The objective or problem is to find a solution orbit, i.e. derive the orbital parameters of the orbit that fits the data. In case of under-constrained problem there is no unique solution. In the case of over-constrained problem, there is no exact solution, and we will look for a solution that fits the data in a way statistically acceptable. Based on the "statistical ranging" technique of [126], a method of statistical orbit determination of binary systems has been constructed [51]. This later technique differs from the more recent one used in [42] since it makes use of the well-known Thiele-Innes algorithm [115]. The Thiele–Innes–van den Bos algorithm [1, 44, and references therein] provides, when it exists, the Keplerian solution starting from three observational positions and one assumed orbital period (or conversely the constant of areal velocity). Following [1] and putting:

[26] The reason why the terminology visual was abandoned because of a possible confusion with the "visible domain".

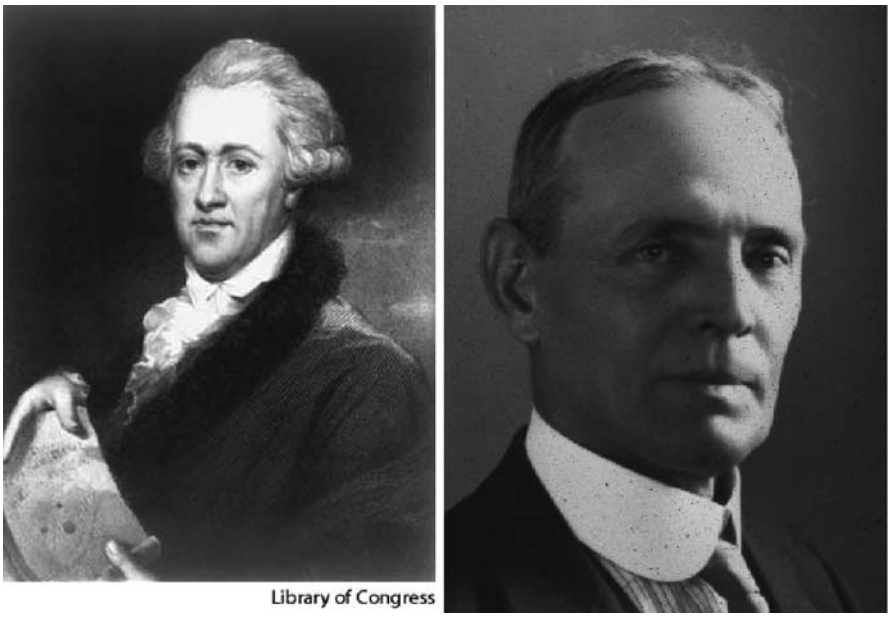

Fig. 26 Two famous binary stars observers from older and modern time. Sir W. Herschel (1738–1822) on the *left* and R. G. Aitken (1864–1951) on the *right* (Aitken: Photo 1923, courtesy Mary Lea Shane Archives, Lick Observatory)

$$\begin{aligned} X &= \cos E - e, \\ Y &= \sqrt{1-e^2}\,\sin E, \end{aligned} \tag{55}$$

one can write the relative position of the secondary:

$$\begin{aligned} x &= A\,X + F\,Y \\ y &= B\,X + G\,Y \\ z &= C\,X + H\,Y, \end{aligned} \tag{56}$$

where the last equation corresponds to the radial (non-observed) quantity which—in contrast to stars—will be of particular use for solar system objects. This last linear system is also convenient to compute the relative position in space at any given epoch and makes use of the Thiele–Innes constants (A, B, F, G, C, H) instead of the usual elliptic elements:

$$\begin{aligned} A &= a\,(\cos\omega\cos\Omega - \sin\omega\sin\Omega\cos i) \\ B &= a\,(\cos\omega\sin\Omega + \sin\omega\cos\Omega\cos i) \\ C &= a\,\sin\Omega\sin i \\ F &= a\,(-\sin\omega\cos\Omega - \cos\omega\sin\Omega\cos i) \\ G &= a\,(-\sin\omega\sin\Omega - \cos\omega\cos\Omega\cos i). \\ H &= a\,\cos\Omega\sin i \end{aligned} \tag{57}$$

Here the angles are referred to the tangent plane and one origin axis in this plane (see [1] for more details). By considering two observations p and q, the double area of the triangle is given by:

$$\Delta_{p,q} = x_p y_q - x_q y_p = (AG - BF)(X_p Y_q - X_q Y_p),$$

it can be related to the eccentric anomalies yielding the fundamental equation of Thiele:

$$t_q - t_p - \mathcal{C}^{-1}\Delta_{p,q} = n^{-1}\left[(E_q - E_p) - \sin(E_q - E_p)\right]. \tag{58}$$

Starting with three observations at time (t_1, t_2, t_3), and a given orbital period, one can then solve a system of three equations:

$$\begin{aligned} t_2 - t_1 - \mathcal{C}^{-1}\Delta_{1,2} &= n^{-1}\left[u - \sin u\right] \\ t_3 - t_2 - \mathcal{C}^{-1}\Delta_{2,3} &= n^{-1}\left[v - \sin v\right] \\ t_3 - t_1 - \mathcal{C}^{-1}\Delta_{1,3} &= n^{-1}\left[u + v - \sin(u + v)\right], \end{aligned} \tag{59}$$

involving the unknown areal constant $\mathcal{C}$ and the two differences in eccentric anomalies $u = (E_2 - E_1)$ and $v = (E_3 - E_2)$, from which one eventually gets:

$$\begin{aligned} e\,\cos E_2 &= \frac{\Delta_{23}\cos(E_2 - E_1) + \Delta_{12}\cos(E_3 - E_2) - \Delta_{13}}{\Delta_{12} + \Delta_{23} - \Delta_{13}}, \\ e\,\sin E_2 &= \frac{\Delta_{23}\sin(E_2 - E_1) - \Delta_{12}\sin(E_3 - E_2)}{\Delta_{12} + \Delta_{23} - \Delta_{13}}, \end{aligned} \tag{60}$$

and the values for e and E_2, next E_1 and E_3, and corresponding mean anomalies from Kepler equation, time of periastron passage, then the values of X and Y from (55) and finally the Thiele–Innes constants and elliptical elements. Note that the solution from this algorithm is almost obtained in closed form. The Keplerian equation is used and remains transcendental, but the numerical solution is easily obtained (at least for reasonable eccentricities of our—supposed—elliptic orbits). The system (59) of three unknowns is non-linear but can be solved with Brown's method with good convergence whatever the starting point in $[0,\ 2\pi]$ and $\mathcal{C} = -1$, whatever the system under consideration, main-belt binary, trans-Neptunian Binary, brown dwarf, etc.

Once the orbital elements are known one can compute subsequent positions for future as well as for past epochs. As said before, (56) will readily provide such positions, but Thiele–Innes constants are related to one particular tangent plane. This plane is invariant for distant stars,[27] but not for a solar system object where, after several months or years, the observer can observe the same system from an opposite

[27] Parallax in this case will introduce, as for the precession, small corrections to add linearly to the nominal solution.

direction. One can take into account this parallax in the Thiele–Innes constants from a transformation from one plane of sky to another one, from a linear relation:

$$\begin{pmatrix} A' & F' \\ B' & G' \\ C' & H' \end{pmatrix} = \mathbf{P} \cdot \mathbf{A} \begin{pmatrix} A & F \\ B & G \\ C & H \end{pmatrix}, \tag{61}$$

where

$$\mathbf{A} = \begin{pmatrix} -\sin\alpha_E & -\cos\alpha_E \sin\delta_E & \cos\alpha_E \cos\delta_E \\ \cos\alpha_E & -\sin\alpha_E \sin\delta_E & \sin\alpha_E \cos\delta_E \\ 0 & \cos\delta_E & \sin\delta_E \end{pmatrix} \tag{62}$$

is a transformation matrix from the POS $(\alpha_E,\ \delta_E)$ to some conventional reference frame (ecliptic, equatorial, etc.) independent of the system. Matrix $\mathbf{P}$ is simply an inverse rotation to some other direction $(\alpha'_E,\ \delta'_E)$. Last, the apparent orbit $(x',\ y')$ in this new POS is given by

$$\begin{aligned} x' &= A'\,X + F'\,Y \\ y' &= B'\,X + G'\,Y \\ z' &= C'\,X + H'\,Y, \end{aligned} \tag{63}$$

where $(X,\ Y)$ are obtained from (55), and the radial coordinate z' is optional. We can note here that (58) involving the classical Thiele–Innes constants (A, B, F, G) does not depend on the sign of the inclination i, so that there are two symmetric true orbits that project to the same apparent one. The coefficient in the new POS, in contrast, depends on the inclination and will not project to the same apparent trajectory, as will be discussed in Sect. 10.3.

10.2 Monte Carlo

The previous algorithm provides one single solution starting from three positions or observations. This is not sufficient or satisfactory for at least two reasons:

- The observations are not free from errors, and one also needs to provide the confidence region around this nominal solution, including the fact that the solution might not be unique. Since the problem to solve is highly non-linear, the error on the orbital parameters should not be Gaussian even from an observational noise that follows a normal distribution. A Monte Carlo (i.e. random) run will provide such information [63, 113].
- There are, in general, more than three observations and one wants to derive the most likely solution (e.g. in the sense of least squares) and also the bundle of orbits that fits the observations with associated probability. Bootstraping or Jackknife without replacement will provide the bundle of orbits.

This goal can be achieved from statistical inversion with a trial/error Monte Carlo technique. At each step of the Monte Carlo computation one chooses a set of three of the observations for the (semi-) analytical computation of the orbit, and additionally one chooses an orbital period generally following a uniform distribution with no particular prior. The computed orbit is then tested for fitting all other available observations. Such statistical inversion will provide the bundle of orbits and their distribution as well as prediction of the position to other epochs and error propagation estimates. All retained solutions are mathematically satisfactory in the sense that they fit the available relative positions data. One can then add additional filter by taking into account either an a priori distribution of the orbital element or prior knowledge on the mass if some other satellite has been better observed, and more efficiently a limit on the total mass (or density) obtained, which in some cases can be physically unrealistic, making it readily adapted to a Bayesian analysis approach.

Note that one can introduce the observational error directly from a normal distribution, and also as uniform distribution with subsequent normalisation filter to be applied to the final set of solutions. Note also that the spurious symmetric solution can generally be removed if the parallax is large enough as in the case of an MBB. Using the Monte Carlo approach also enables one to analyse the propagation of errors to past or future epochs. In the latter case it is possible to derive the best epoch to observe a trans-Neptunian binary (TNB) and remove the spurious solution (see Fig. 27). Indeed, there generally remains four intersections of the two symmetric orbits (i.e. intersection of the apparent trajectory *at the same date*) which observations would not enable to separate the spurious from the true one. This is of particular importance for at least two reasons: prediction of mutual phenomena and test of formation models. Knowing the true inclination of the orbit enables one to predict the mutual phenomena (eclipses/occultation between the primary and secondary) that are of high value to better constrain many of the system's components physical parameters [116, 9]. Let us add that, similarly to equinoxes, such phenomenon occurs twice during a revolution of the binary system around the Sun, approximately every 150 years, hence extremely rare! The second reason why the inclination is an important parameter to know unambiguously is that different models of formation of binary systems exist [84, 133 and references therein] which do not give the same prediction. In particular the angular momentum can show an asymmetric distribution which can be tested from a measure of the relative frequency of prograde/retrogade orbits [62].

10.3 A Summary

We can now resume the complete algorithm in the following sequence:

1. choose randomly three observations among the n available. These should nevertheless belong to a same group or run of observations, not to two very different tangent planes;

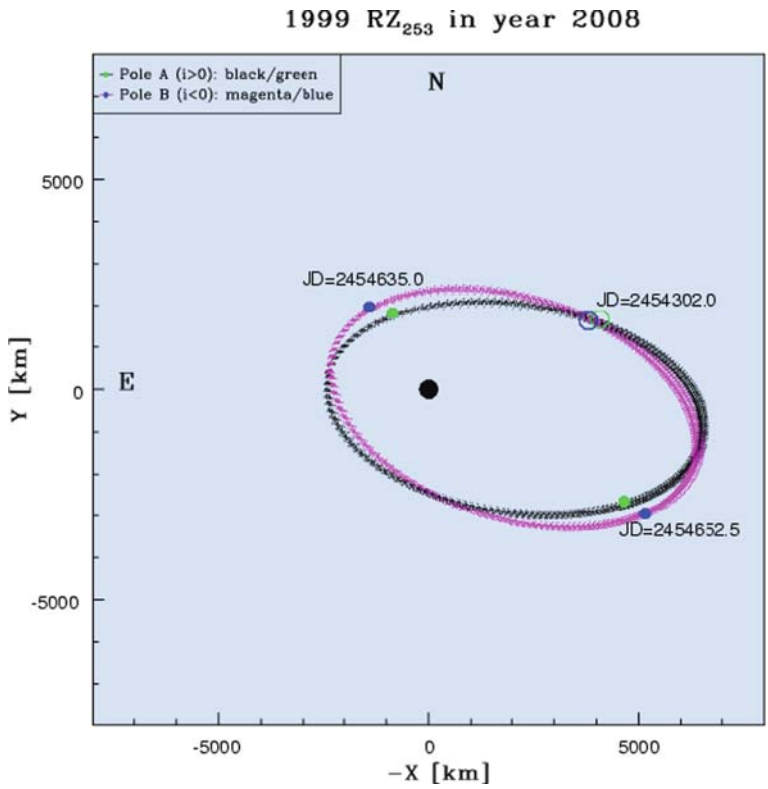

Fig. 27 Two apparent orbits bundle for the TNB Borasisi (ex. 1999 RZ_{253}) corresponding to the two symmetric solutions that fit equally well the data obtained 4 years before. The predicted positions at three different dates are also indicated. It is clear that at the epoch $JD = 2\,454\,302.0$ the orbits (not only the apparent trajectories) intersect, making it again impossible to distinguish the spurious from the true solution

2. add stochastic error from the observational noise model ($\varepsilon \in \mathcal{N}(0, \sigma)$) with Box–Müller method [97, Chap. 7], or prefer a more general uniform distribution that will ease further normalization or introduction of any other noise distribution;
3. choose one orbital period from stochastic distribution (e.g. $P \in \mathcal{U}[P_{min}, P_{max}]$);
4. compute the orbit and Thiele–Innes coefficients from the algorithm described in the previous paragraph. In fact there are two fully symmetric orbits;
5. compute the positions for both symmetric orbits for all other n-3 observations, including parallax and other precession effects;
6. test if the resulting O–C for all these positions is acceptable with respect to the observational noise

   – if no: return to Step 1;
   – if yes: accept the solution with all parameters and return to Step 1.

Because a subset of the observations sample is chosen, one can make use of the jack-knife technique for automatically detecting outlier points. Moreover, since only the orbital period is chosen arbitrarily, this Monte Carlo run is most efficient leaving a one-dimensional space to explore instead of the initial, more general and "brute force", seven-dimension problem to solve. Another possibility is to use the statistical ranging approach [126, 42]; in this case one has to choose two arbitrary relative distances for two different observations epoch. This increases the dimension of the problem, though still to practicable application. Having only *one* parameter of the space-phase to be explored (the orbital period) when all other orbital parameters are easily derived, the algorithm practical and fast in terms of CPU. At the beginning of the process one might have no indication of the orbital period and should span a large interval, e.g. [0, 100] days, depending on the system to be analysed (MBB, TNB, binary star, etc.), or on the other hand, one might reduce the interval to scan if the position angle of the secondary has almost span an entire cycle. As a matter of illustration, a bundle of ≈ 7500 orbits, with 4% efficiency in the trial/error throw (period range $P \in [10; 70]$ days), has been obtained for the data analysis of the trans-Neptunian (136,108) Haumea's second satellite Namaka, with no particular optimisation, in 40 s real time on a personal laptop.

10.4 Astrometric Binary

Let us briefly mention in this last paragraph a subset of problem closely related to resolved or visual binaries: the astrometric binaries. The first detection of such systems was made by Bessel [7] from his analysis of the abnormal proper motions of Sirius and Procyon. Luminosity, as he said, is not a straight propriety of stellar mass.[28] Astrometric binaries are binary system for which both components are not observed separately (either the secondary is too faint to be detected in the instrument, or similarly it is too close to the bright primary), but instead one observes the photocentre of the system. This photocentre differs periodically from the centre of mass with an amplitude that depends on the mass ratio for the position of the barycentre, the brightness distribution and ratio for the position of the photocentre, and the inclination of the orbit [122, 1]. What is observed is not the Keplerian orbit a but the photometric orbit α which are related by

$$f = \alpha/a + \beta \quad ; \quad \beta = (1 + 10^{0.4\Delta V})^{-1}, \tag{64}$$

where f is the fractional mass, and β is the fractional light. The photometric orbit is thus scaled from the Keplerian one by $(f - \beta)$. In the case of a stellar system one will consider particular mass–luminosity relation [44], a key parameter that links mass and luminosity. In the case of solar system objects, considering that both components are spherical with same albedo and there is no strong phase effect, the

[28] Let us add that it seems to be the same for the matter in the Universe.

relative position of the barycentre and photocentre is given as a function of mass ratio $0 < q \leq 1$:

$$\alpha = \left(\frac{1}{1+q^{-2/3}} - \frac{1}{1+q^{-1}} \right) a. \tag{65}$$

It can be seen from the equations above that there is no astrometric signal in the two extreme cases $q = 0$ and $q = 1$, the peak being at $q \sim 0.15$. One can put this additional parameter in the Monte Carlo run for resolved binaries, having now two dimensions (P, q) to explore which remains tractable. In the case of asteroids and if one solution orbit exists, one will end up with a possible separation that will have to be compared to prior knowledge. If the separation is relatively large it can plausibly be a binary system, if the "separation" is small, it will likely be a non-symmetric single object.

In contrast to the situation in the main belt (and assumed as such by us) where multiple systems involve mostly small moonlet, or in the near-Earth region where separations are small making them difficult to distinguish from a single body, the trans-Neptunian and Centaur binaries can have much larger separation and mass ratio, well in the detection range for modern astrometric observations. The other parameter of importance in this study is the orbital period which can be large [92]. Given the ratio of known (resolved) binaries in TNO population, one could follow the statement of Bessel for stars [66] and wondering how many targets should be suspicious of being binaries in our outer Solar System?

11 Conclusion

Starting with the Hipparcos catalogue birthday, we have reviewed in this lecture the different aspects of the Gaia mission, its payload, its instruments and observations, and the results to be expected from the direct observations of asteroids. From the highly accurate astrometry and photometry gathered over the 5 year mission duration, Gaia will provide a breakthrough in our knowledge of these bodies and subsequently on the formation of the Solar System and its dynamical evolution. Gaia will provide a wealth of data and results from the direct observations (photometric in many bands from the low-resolution spectra, astrometric and imaging; the high-resolution spectroscopy is very marginal) of asteroids and small bodies of the Solar System. We will have a clearer view of their dynamical and physical characteristics, for an incredibly and yet unprecedented number of objects of different kinds. One of the high impacts is obtained from the large number of targets observed, from which one will get simultaneously sharp information and large statistical analysis, all programme that could not be achieved by a single team from ground-based observations, or a space probe rendezvous, and surely not over such short time span of observations.

Moreover Gaia will also provide a new area in asteroids and small bodies science from astrometric catalogue of stars. The current—and severe—limitation to the use

of the Hipparcos or Tycho2 astrometric catalogues in the reduction of photographic or CCD plates is their poor numbers of star or low density: there are tiny chances to have one Hipparcos star in a typical $12' \times 12'$ field of view, but this is useless because at least three are needed. The situation will be considerably improved with Gaia, so that next generation post-Gaia era astrometry will have a gain of one order of magnitude in the classical astrometric reductions. Surveys that go to deeper magnitudes (Pan-STARRS, LSST, etc.) will dramatically benefit the Gaia astrometric catalogue. Similarly the computations of both asteroids and stars ephemeris will be increased, yielding much more accurate stellar occultation predictions for different kinds of bodies, MBAs to TNOs, with or without atmospheres, and again a huge step in our understanding of the physical characteristic of small and faint bodies that were not even observable with Gaia.

Acknowledgments The authors wish to thank Marco Delbò (OCA, Nice) and Serge Mouret (IMCCE, Paris) for their various contributions and help, undergraduate students who also took part in the project, and the members of the Gaia-CU4/SSO and Gaia-REMAT at large. We also acknowledge Marc Fouchard (IMCCE, Paris) for his careful reading of the manuscript and improving the quality of the text. Let us express our gratitude to the editors and also organisers of this school in Bad Hofgastein (Austria).

Acronyms

AF	astrometric field
BP/RP	blue/red photometry
CCD	charge-coupled device
CTE	charge transfer efficiency
PSF	point spread function
RVS	radial velocity spectrometer
SM	sky mapper
TDI	time-delayed integration
GLS	general least squares
GR	general relativity
LLS	linear least squares
mas	milli-arcsecond
MCMC	Monte Carlo Markov chain
MLE	maximum likelihood estimator
O-C	observed-minus-calculated
ODE	ordinary differential equation
p.d.f.	probability distribution function
PPN	parameterised post-newtonian
SNR	signal-to-noise ratio
SVD	singular value decomposition
TI	thermal inertia
FK5	Fundamental Katalog, 5th version

ICRF International Celestial Reference Frame
IAU/UAI International Astronomical Union – Union Astronomique Internationale
LSST large-aperture synoptic survey telescope
MBA main-belt object
MBB main-belt binary
NEO near-earth object
QSO quasi-stellar object
SSO solar system object
SSSB small solar system bodies
TNB trans-Neptunian binary
TNO trans-Neptunian object

References

1. Aitken, R.G.: The binary stars. Dover Publication, New York (1964) 324, 325, 327, 331
2. Arago, F.: Les comètes. Astronomie Populaire. A. Blanchard, Paris;, livre xvii, nouveau tirage edition, 1986. (1st ed. 1858) 306
3. Avni, Y.: Energy spectra of X-ray clusters of galaxies. Astrophys. J. **210**, 642–646, December (1976) 323
4. Bange, J.: An estimation of the mass of asteroid 20-Massalia derived from the HIPPARCOS minor planets data. Astron. Astrophys. **340**, L1–L4 (1998) 254
5. Batrakov, Y.V., Chernetenko, Y.A., Gorel, G.K., Gudkova, L.A.: Hipparcos catalogue orientation as obtained from observations of minor planets. Astron. Astrophys. **352**, 703–711, December (1999) 254
6. Belskaya, N., Shkuratov, Y.G., Efimov, Y.S., Shakhovskoy, N.M., Gil-Hutton, R., Cellino, A., Zubko, E.S., Ovcharenko, A.A., Bondarenko, S.Y., Shevchenko, V.G., Fornasier, S., Barbieri, C.: The F-type asteroids with small inversion angles of polarization. Icarus. **178**, 213–221, November (2005) 297
7. Bessel, F.W.: Extract of a letter from on the proper motions of Procyon and Sirius. MNRAS. **6**, 136–141, December (1844) 331
8. Beutler, G.: Methods of celestial mechanics. Astronomy and Astrophysics Library. Springer, Berlin, In cooperation with Leos Mervart and Andreas Verdun (2005) 320
9. Binzel, R.P.: Hemispherical color differences on Pluto and Charon. Science. **241**, 1070–1072, August (1988) 329
10. Bottke, W.F., Morbidelli, A., Jedicke, R., Petit, J.M., Levison, H.F., Michel, P., Metcalfe, T.S.: Debiased orbital and absolute magnitude distribution of the near-Earth objects. Icarus. **156**, 399–433, August (2002) 305
11. Bowell, E., Virtanen, J., Muinonen, K., Boattini, A.: Asteroid Orbit Computation. In Asteroids III, Britt, D.T., Yeomans, D., Housen, K., Consolmagno, G. (eds.), pp. 27–43 (2002) 315, 323
12. Britt, D.T., Yeomans, D., Housen, K., Consolmagno, G.: Asteroid density, porosity, and structure. In Asteroids III, Bottke W.F. Jr., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) University of Arizona Press, Tucson, pp. 485–500 (2002) 300
13. Brouwer, D., Clemence, G.M.: Methods of celestial mechanics. Academic Press, New York (1961) 317, 318, 320
14. Bus, S.J., Binzel, R.P.: Phase II of the small main-belt asteroid spectroscopic survey the observations. Icarus. **158**, 106–145, July (2002) 297
15. Bykov, O.: Initial orbit determinations of neo-2005 with pulkovo amp-method. In MEOTUS Conf., Paris, France (2006) 309
16. Capaccioni, F., Cerroni, P., Coradini, M., Farinella, P., Flamini, E., Martelli, G., Paolicchi, P., Smith, P.N., Zappalà, V.: Shapes of asteroids compared with fragments from hypervelocity impact experiments. Nature. **309**, 832–834, June (1984) 289

17. Celletti, A., Pinzari, G.: Four classical methods for determining planetary elliptic elements: A comparison. Celestial Mech. Dynam. Astron. **93**, 1–52, September (2005) 309
18. Cellino, A.: Minor bodies: Spectral gradients and relationships with meteorites. Space Sci. Rev. **92**, 397–412, April (2000) 297
19. Cellino, A., Bus, S.J., Doressoundiram, A., Lazzaro, D.: Spectroscopic properties of asteroid families. In Asteroids III, Britt, D.T., Yeomans, D., Housen, K., Consolmagno, G. (eds.), pp. 633–643 (2002) 297
20. Cellino, A., Tanga, P., Dell'Oro, A., Hestroffer, D.: Asteroid science with Gaia: Sizes, spin properties, overall shapes and taxonomy. Adv. Space Res. **40**, 202–208 (2007) 263
21. Cellino, A., Zappalà, V., Doressoundiram, A., di Martino, M., Bendjoya, P., Dotto, E., Migliorini, F.: The puzzling case of the Nysa-Polana family. Icarus. **152**, 225–237, August (2001) 297
22. Chernetenko, Y.A.: Orientation of the Hipparcos frame with respect to the reference frames of the DE403/LE403 and DE405/LE405 ephemerides based on asteroid observations. Astron. Lett. **34**, 266–270, April (2008) 254
23. Chesley, S.R., Vokrouhlický, D., Ostro, S.J., Benner, L.A.M., Margot, J.-L., Matson, R.L., Nolan, M.C., Shepard, M.K.: Direct estimation of Yarkovsky accelerations on Near-Earth asteroids. LPI Contributions, **1405**, 8330 (2008) 302
24. Cochran, W.D., Endl, M., Wittenmyer, R.A., Bean, J.L.: A planetary system around HD 155358: The lowest Metallicity planet host star. Astrophys. J., **665**, 1407–1412, August (2007) 315
25. Dell'Oro, A., Cellino, A.: Asteroid sizes from Gaia observations. In Turon, C., O'Flaherty, K.S., Perryman, M.A.C. (eds.) The three-dimensional universe with Gaia, volume 576 of ESA Special Publication, pp. 289–+, January (2005) 263
26. Dell'Oro, A., Cellino, A.: Numerical simulations of asteroid signals on the GAIA focal plane. Adv. Space Res. **38**, 1961–1967 (2006) 267, 268
27. Deutsch, R.: Orbital dynamics of space vehicles. Prentice-Hall International Series in Space technologies. Prentice Hall Inc., Englewood Cliffs, NJ (1963) 311
28. Di Sisto, R.P., Orellana, R.B.: Determinación de posiciones de asteroides utilizando el catálogo Hipparcos. Boletin de la Asociacion Argentina de Astronomia La Plata Argentina. **43**, 8–13 (1999) 254
29. Dubyago, D.: The determination of orbits. The Macmillan Company, New York (1961) 306, 307, 309, 310, 318
30. Dunham, D.W., Goffin, E., Manek, J., Federspiel, M., Stone, R., Owen, W.: Asteroidal occultation results multiply helped by Hipparcos. Memorie della Societa Astronomica Italiana. **73**, 662–+, September (2002) 254
31. Dvorak, R., Edelman, C.: Statistical method for the determination of asteroid and satellite orbits. Astron. Astrophys. **77**, 320–326, August (1979) 315
32. Escobal, P.R.: Methods of orbit determination. Wiley, New York (1965) 309, 311, 314, 318
33. Exertier, P., Coulot, D.: L'Intégration Numérique en Calcul d'Orbites. Technical report, Cours de l'Ecole GRGS 2002, http://igsac-cnes.cls.fr/documents/gins/GINS_Doc_Algo_V4.html (2002) 320
34. Farinella, P., Paolicchi, P., Tedesco, E.F., Zappalà, V.: Triaxial equilibrium ellipsoids among the asteroids. Icarus. **46**, 114–123, April (1981) 289
35. Farinella, P., Paolicchi, P., Zappalà, V.: The asteroids as outcomes of catastrophic collisions. Icarus. **52**, 409–433, December (1982) 289
36. Fienga, A., Manche, H., Laskar, J., Gastineau, M.: INPOP06: A new numerical planetary ephemeris. Astron. Astrophys. **477**, 315–327, January (2008) 305
37. Gauss, F.: Carl Friedrich Gauss' Werke Hereausgeber von der Königl. Gesellschaft der Wissenschaften zu Göttingen. Astronomische Nachrichten, **57**, 53–54 (1862) 308
38. Gauss, F.: Théorie du mouvement des corps célestes parcourant des sections coniques autour du soleil/trad. du "Theoria motus" de Gauss; suivie de notes, par Edmond Dubois. Paris: A. Bertrand, 1855; trad. E.-P. Dubois; in 8, April (1864) 306, 310

39. Gauss, K.F.: Theoria motvs corporvm coelestivm in sectionibvs conicis solem ambientivm. Hambvrgi, Svmtibvs F. Perthes et I.H. Besser, December (1809) 306
40. Gauss, K.F.: Theory of the motion of the heavenly bodies moving about the sun in conic section/translated and with Appendix by C.H. Davis. Dover, New York (1963) 306, 307, 310, 314
41. Golub, G.H., van Loan, C.F.: Matrix computations. Johns Hopkins University Press, Baltimore (Johns Hopkins studies in the mathematical sciences) (1996) 321
42. Grundy, W.M., Noll, K.S., Virtanen, J., Muinonen, K., Kern, S.D., Stephens, D.C., Stansberry, J.A., Levison, H.F., Spencer, J.R.: (42355) Typhon–Echidna: Scheduling observations for binary orbit determination. Icarus. **197**, 260–268, September (2008) 325, 331
43. Gurfil, P., Belyanin, S.: The gauge-generalized Gyldén–Meshcherskii Problem. Adv. Space Res. **42**, 1313–1317, October (2008) 305
44. Heintz, W.D.: Double stars/Revised edition/volume 15 of Geophysics and Astrophysics Monographs. Reidel Publishing Company, Dordrecht (1978) 325, 331
45. Hestroffer, D.: Astrométrie et photométrie des astéroïdes observés par le satellite Hipparcos. Apportà l'élaboration d'un système de référence dynamique. PhD thesis, Paris Observatory, France (1994) 253, 301, 302
46. Hestroffer, D.: Photocentre displacement of minor planets: Analysis of HIPPARCOS astrometry. Astron. Astrophys. **336**, 776–781, August (1998) 254, 301, 302
47. Hestroffer, D., Berthier, J.: Determination of the PPN beta and Solar quadrupole from Asteroid Astrometry. In ESA SP-576: The Three-Dimensional Universe with Gaia, pp. 297–300, January (2005) 263
48. Hestroffer, D., Mignard, F.: Photometry with a periodic grid. I. A new method to derive angular diameters and brightness distribution. Astron. Astrophys. **325**, 1253–1258, September (1997) 254
49. Hestroffer, D., Morando, B., Høg, E., Kovalevsky, J., Lindegren, L., Mignard, F.: The HIPPARCOS solar system objects catalogues. Astron. Astrophys. **334**, 325–336, June (1998) 253
50. Hestroffer, D., Mouret, S., Berthier, J., Mignard, F.: Relativistic tests from the motion of the asteroids. In Kleinert, H., Jantzen, R.T., Ruffini, R. (eds.) The Eleventh Marcel Grossmann Meeting. On recent developments in theoretical and experimental general relativity, gravitation and relativistic field theories, November (2008) 304, 305
51. Hestroffer, D., Vachier, F., Balat, B.: Orbit determination of Binary Asteroids. Earth Moon and Planets. **97**, 245–260, doi: 10.1007/s11038-006-9097-3, December (2005) 325
52. Hestroffer, D., Viateau, B., Rapaport, M.: Minor planets ephemerides improvement. From joint analysis of Hipparcos and ground-based observations. Astron. Astrophys. **331**, 1113–1118, March (1998) 254
53. Hilton, J.L.: Asteroid masses and densities. In Asteroids III, Bottke Jr. W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) University of Arizona Press, Tucson, pp. 103–112 (2002)
54. Hoeg, E., Fabricius, C.: Sampling in the new MBP. Technical report, Gaia Technical Report, GAIA-CUO-139 (2004) 269
55. Kaasalainen, M., Hestroffer, D., Tanga, P.: Physical Models and Refined Orbits for Asteroids from Gaia Photometry and Astrometry. In ESA SP-576: The Three-Dimensional Universe with Gaia, pp. 301–+, January (2005) 302
56. Kaasalainen, M., Mottola, S., Fulchignoni, M.: Asteroid models from disk-integrated data. In Asteroids III, Bottke Jr. W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) University of Arizona Press, Tucson, pp. 139–150 (2002) 281, 288, 289
57. Kaasalainen, M., Tanga, P.: Photocentre offset in ultraprecise astrometry: Implications for barycentre determination and asteroid modelling. Astron. Astrophys. **416**, 367–373, March (2004) 302
58. Kholchevnikov, C., Fracassini, M.: Le problème des deux corps avec G variable selon l'hypothèse de Dirac. Conf. Oss. Astron. Milano-Merate, Ser. I, No. 9, 50 p., 9 (1968) 305
59. Kovalevsky, J.: *Astrométrie Moderne*. Lect. Notes Phys. **358**, Springer Verlag, Berlin (1990) 253
60. Kovalevsky, J.: Modern astrometry. Astronomy and astrophysics library, 2nd edn. Springer, Berlin, New York (2002) 253

61. Lampton, M., Margon, B., Bowyer, S.: Parameter estimation in X-ray astronomy. Astrophys. J. **208**, 177–190, August (1976) 323
62. Lee, E.A., Astakhov, S.A., Farrelly, D.: Production of trans-Neptunian binaries through chaos-assisted capture. MNRAS. **379**, 229–246, July (2007) 329
63. Leonard, T., John, S.J.: Bayesian Methods, volume 4 of Cambridge series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1999) 315, 328
64. Lindegren, L.: Meridian observations of planets with a photoelectric multislit micrometer. Astron. Astrophys. **57**, 55–72 (1977) 301
65. Lindegren, L.: The astrometric instrument of Gaia: Principles. In Turon, C., O'Flaherty, K.S., Perryman, M.A.C. (eds.) The three-dimensional universe with Gaia, volume 576 of ESA Special Publication, pp. 29–34, January (2005) 258
66. Lippincott, S.L.: Astrometric search for unseen stellar and sub-stellar companions to nearby stars and the possibility of their detection. Space Sci. Rev. **22**, 153–189, July (1978) 332
67. Lœwy, M.: Détermination des orbites des comètes. Gauthier-Villars, Paris (1872) 309
68. Magnusson, P., Barucci, M.A., Drummond, J.D., Lumme, K., Ostro, S.J.: Determination of pole orientations and shapes of asteroids. In Asteroids II, pp. 67–97 (1989) 281, 289, 324
69. McKinnon, W.B.: Could Ceres be a Refugee from the Kuiper Belt? Asteroids, Comets, Meteors 2008 held July 14–18, 2008 in Baltimore, Maryland. LPI Contributions, **1405**, 8389 (2008) 308
70. Merline, W.J., Weidenschilling, S.J., Durda, D., Margot, J.-L., Pravec, P., Storrs, A.D.: Asteroids do have satellites. In Asteroids III, Bottke Jr. W.F., Cellino, A., Paolicchi, P., Binzel, R.P. (eds.) University of Arizona Press, Tucson, pp. 289–312 (2002) 324
71. Mestschersky, J.: Über die Integration der Bewegungsgleichungen im Probleme zweier Körper von veränderlicher Masse. Astronomische Nachrichten, **159**, 229–242, September (1902) 305
72. Mignard, F.: Observations of solar system objects with Gaia.I. Detection of NEOS. Astron. Astrophys. **393**, 727–731, October (2002) 255, 274
73. Mignard, F.: Overall science goals of the Gaia mission. In ESA SP-576: The three-dimensional universe with Gaia, pp. 5–+, January (2005) 255
74. Mignard, F.: The Gaia mission: Science highlights. In Seidelmann, P.K., Monet, A.K.B. (eds.) Astrometry in the age of the next generation of large telescopes, volume 338 of Astronomical Society of the Pacific Conference Series, pp. 15–+, October (2005) 255
75. Mignard, F., Cellino, A., Muinonen, K., Tanga, P., Delbò, M., Dell'Oro, A., Granvik, M., Hestroffer, D., Mouret, S., Thuillot, W., Virtanen, J.: The Gaia mission: Expected applications to asteroid science. Earth Moon and Planets, **101**, 97–125, December (2007) 255, 263
76. Misner, W., Thorne, K.S., Wheeler, J.A.: Gravitation. W.H. Freeman and Co., San Francisco (1973) 304
77. Morrison, D., Matthews, M.S.: Satellites of Jupiter. Space Science Series, University of Arizona Press, Tucson, edited by Morrison, David; Matthews, Mildred Shapley (ass.) (1982) 301
78. Morrison, L.V., Hestroffer, D., Taylor, D.B., van Leeuwen, F. Check on JPL DExxx Using HIPPARCOS and TYCHO Observations. Highlights in Astronomy, **11**, 554+ (1998) 254
79. Mouret, S.: Investigations on the dynamics of minor planets with Gaia. PhD thesis, Observatoire de Paris (2007) 299, 300, 303
80. Mouret, S., Hestroffer, D., Mignard, F.: Asteroid mass determination with the Gaia mission. In Valsecchi, G.B., Vokrouhlický, D. (eds.) IAU Symposium, volume 236 of IAU Symposium, pp. 435–438 (2007) 299
81. Mouret, S., Hestroffer, D., Mignard, F.: Asteroid masses and improvement with Gaia. Astron. Astrophys. **472**, 1017–1027, September (2007) 301
82. Mouret, S., Hestroffer, D., Mignard, F.: Asteroid mass determination with the Gaia mission. IAU Symposium, **248**, 363–366 (2008) 301
83. Mouret, S., Hestroffer, D., Mignard, F.: Asteroid mass determination with the Gaia mission. A simulation of the expected precisions. Planet. Sp. Sci. **56**, 1819–1822, November (2008) 300

84. Noll, K.S., Grundy, W.M., Chiang, E.I., Margot, J.-L., Kern, S.D.: Binaries in the Kuiper Belt. In The Solar System Beyond Neptune, pp. 345–363 (2008) 329
85. Oppolzer, T.: Lehrbuch zur Bahnbestimmung der Kometen und Planeten. W. Engelmann, Leipzig (1882) 309
86. Oppolzer, T.: Traité de la détermination des orbites des Comètes et des Planètes. Gauthier-Villars, Paris; E. Pasquier trad. (1886) 309
87. Pace, O.: Gaia: The Satellite and Payload. In Turon, C., O'Flaherty, K.S., Perryman, M.A.C. (eds.) The three-dimensional universe with Gaia, volume 576 of ESA Special Publication, pp. 23–28, January (2005) 257
88. Pepe, F., Correia, A.C.M., Mayor, M., Tamuz, O., Couetdic, J., Benz, W., Bertaux, J.-L., Bouchy, F., Laskar, J., Lovis, C., Naef, D., Queloz, D., Santos, N.C., J.-Sivan, P., Sosnowska, D., Udry, S.: The HARPS search for southern extra-solar planets. VIII. μ Arae, a system with four planets. Astron. Astrophys. **462**, 769–776, February (2007) 315
89. Perryman, M.A.C., de Boer, K.S., Gilmore, G., Høg, E., Lattanzi, M.G., Lindegren, L., Luri, X., Mignard, F., Pace, O., de Zeeuw, P.T.: Gaia: Composition, formation and evolution of the Galaxy. Astron. Astrophys. **369**, 339–363 (2001) 268, 272, 273
90. Perryman, M.A.C., ESA (eds.): The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission, volume 1200 of ESA Special Publication. ESA (1997) 253
91. Perryman, M.A.C.: Astronomical applications of astrometry: A review based on ten years of exploitation of the Hipparcos satellite data. Cambridge University Press, Cambridge (2008) 253
92. Petit, J.-M., Kavelaars, J.J., Gladman, B.J., Margot, J.L., Nicholson, P.D., Jones, R.L., Parker, J.W., Ashby, M.L.N., Campo Bagatin, A., Benavidez, P., Coffey, J., Rousselot, P., Mousis, O., Taylor, P.A.: The extreme Kuiper belt binary 2001 QW_{322}. Science, **322**, 432–434, October (2008) 332
93. Picart, L.: Calcul des orbites et des éphémérides. Octave Doin et fils, Paris; Encyclopédie scientifique (1913) 309, 315, 318, 325
94. Pitjeva, V.: Relativistic effects and solar Oblateness from radar observations of planets and spacecraft. Astron. Lett. **31**, 340–349 (2005) 305
95. Poincaré, H.: Mémoires et observations. Sur la détermination des orbites par la méthode de Laplace. Bulletin Astronomique, Serie I, **23**, 161–187 (1906) 309
96. Pospieszalska-Surdej, A., Surdej, J.: Determination of the pole orientation of an asteroid – The amplitude-aspect relation revisited. Astron. Astrophys. **149**, 186–194, August (1985) 292
97. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical recipes in FORTRAN. The art of scientific computing, 2nd edn. University Press, Cambridge (1992) 321, 330
98. Radau, R.: Mémoires et observations. Sur la détermination des orbites. Bulletin Astronomique, Serie I, **2**, 5–15 (1885) 309
99. Roy, E.: Orbital motion, 4th edn. IoP, Bristol and Philadelphia (2005) 309, 311
100. Rubincam, D.P.: Yarkovsky thermal drag on LAGEOS. JGR. **93**, 13805–13810, November (1988) 302
101. Rubincam, D.P.: Yarkovsky thermal drag on small asteroids and Mars-Earth delivery. JGR. **103**, 1725–+, January (1998) 302
102. Safa, F.: A summary of the Gaia spacecraft design. Technical report, GAIA Technical Report, GAIA-CH-TN-EADS-FS-001-1 (2006) 272
103. Seidelmann, P.K., Archinal, B.A., A'Hearn, M.F., Conrad, A., Consolmagno, G.J., Hestroffer, D., Hilton, J.L., Krasinsky, G.A., Neumann, G., Oberst, J., Stooke, P., Tedesco, E.F., Tholen, D.J., Thomas, P.C., Williams, I.P.: Report of the IAU/IAG Working Group on cartographic coordinates and rotational elements: 2006. Celestial Mechanics and Dynamical Astronomy, **98**, 155–180, July (2007) 288
104. Shao, M.: SIM: the space interferometry mission. In Reasenberg, R.D. (ed.) Proc. SPIE vol. 3350, pp. 536–540, Astronomical Interferometry, Robert D. Reasenberg; Ed., volume 3350 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, pp. 536–540, July (1998). 255

105. Shao, M.: Science overview and status of the SIM project. In Traub, W.A. (ed.) New Frontiers in Stellar Interferometry, Proceedings of SPIE Volume 5491. Edited by Wesley Traub, A.: Bellingham, WA: The International Society for Optical Engineering, 2004., p. 328, volume 5491 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, pp. 328–+, October (2004) 255
106. Simon, B.: How Kepler determined the orbit of the Earth around the Sun (French Title: Comment Kepler a determiné l'orbite de la Terre autour du Soleil). Observations et Travaux, **62**, 23–25, June (2006) 305
107. Söderhjelm, S., Lindegren, L.: Inertial frame determination using minor planets. A Simulation of Hipparcos Observations. Astron. Astrophys. **110**, 156–162 (1982) 254, 302
108. Stone, R.C., Monet, D.G., Monet, A.K.B., Harris, F.H., Ables, H.D., Dahn, C.C., Canzian, B., Guetter, H.H., Harris, H.C., Henden, A.A., Levine, S.E., Luginbuhl, C.B., Munn, J.A., Pier, J.R., Vrba, F.J., Walker, R.L.: Upgrades to the flagstaff astrometric scanning transit telescope: A fully automated telescope for astrometry. Astron. J. **126**, 2060–2080, October (2003) 254
109. Tanga, P., Delbò, M.: Asteroid occultations today and tomorrow: Toward the GAIA era. Astron. Astrophys. **474**, 1015–1022, November (2007) 263
110. Tanga, P., Delbò, M., Hestroffer, D., Cellino, A., Mignard, F.: Gaia observations of Solar System objects: Impact on dynamics and ground-based observations. Adv. Space Res. **40**, 209–214 (2007) 262
111. Tanga, P., Hestroffer, D., Delbò, M., Frouard, J., Mouret, S., Thuillot, W.: Gaia, an unprecedented observatory for Solar System dynamics. Planss **56**, 1812–1818, November (2008) 263
112. Tapley, D., Schutz, B.E., Born, G.H.: Statistical orbit determination. Elsevier Academic Press, Amsterdam (2004) 302, 322, 323
113. Tarantola, A.: Inverse problem theory. Methods for model parameters estimation. SIAM, Philadelphia (2005) 328
114. Tedesco, F., Noah, P.V., Noah, M., Price, S.D.: The Supplemental IRAS Minor Planet Survey. Astron. J. **123**, 1056–1085 (2002) 275
115. Thiele, T.N.: Neue Methode zur Berechnung von Doppelsternbahnen. Astronomische Nachrichten. **104**, 245–254 (1883) 325
116. Tholen, J., Buie, M.W., Binzel, R.P., Frueh, M.L.: Improved orbital and physical parameters for the Pluto-Charon system. Science, **237**, 512–514, July (1987) 329
117. Thuillot, W., Berthier, J., Vaubaillon, J., Vachier, F., Iglesias, J., Lainey, V.: Data mining for the improvement of orbits of the NEOs. In IAU Symposium, volume 236 of IAU Symposium, August (2006) 306
118. Tisserand, H.: Mémoires et observations. Sur la détermination des orbites circulaires. Bulletin Astronomique, Serie I, **12**, 53–59 (1895) 310
119. Tisserand, H.: Traité de mécanique céleste. Perturbations des planètes d'après la méthode de la variation des constantes arbitraires, volume 1. 2nd edn. Gauthier-Villars, Paris (1960). (1st ed. 1889) 309, 323
120. Tisserand, F., Perchot, J.: Leçons sur la détermination des orbites. Gauthier-Villars, Paris (1899) 309
121. Valsecchi, B.: 236 years ago... In Valsecchi, G.B., Vokrouhlický, D. (eds.) IAU Symposium, volume 236 of IAU Symposium, pp. D17+ (2007) 323
122. van de Kamp, P.: Unseen astrometric companions of stars. Ann. Rev. Astron. Astrophys. **13**, 295–333 (1975) 331
123. van Leeuwen, F.: Hipparcos, the new reduction of the raw data. Astrophys. Space Sci. Library. vol. 350, 20 Springer, Dordrecht (2007) 253
124. Viateau, B.: Apport des observations faites à Bordeaux à l'amélioration des orbites des astéroïdes. Utilisation de ces orbites. PhD thesis, PhD Thesis, Observatoire de Paris (in French), January (1995)
125. Virtanen, J.: Asteroid orbital inversion using statistical methods. PhD thesis, University of Helsinki, Finland (2005) 315

126. Virtanen, J., Muinonen, K., Bowell, E.: Statistical ranging of asteroid orbits. Icarus. **154**, 412–431 (2001) 325, 331
127. Vokrouhlicky, D.: Diurnal Yarkovsky effect as a source of mobility of meter-sized asteroidal fragments. I. Linear theory. Astron. Astrophys. **335**, 1093–1100, July (1998) 302
128. Wall, V., Jenkins, C.R.: Practical statistics for astronomers. Cambridge University Press, Cambridge, November (2003) 323
129. Will, M.: The confrontation between general relativity and experiment. Living Rev. Relativity. 9, March (2006) 304
130. Will, M.: The confrontation between general relativity and experiment. In Oscoz, A., Mediavilla, E., Serra-Ricart, M. (eds.) EAS Publications Series, volume 30 of EAS Publications Series, pp. 3–13 (2008) 304
131. Will, M., Nordtvedt, K.J.: Conservation laws and preferred frames in relativistic gravity. I. Preferred-frame theories and an extended PPN formalism. Astrophys. J. **177**, 757–774, November (1972) 304
132. Williams, J.G., Turyshev, S.G., Boggs, D.H.: Progress in Lunar Laser Ranging tests of relativistic gravity. Phys. Rev. Lett. **93**(26), 261101–1–261101–4, December (2004) 305
133. Young, J.S., Baldwin, J.E., Boysen, R.C., Haniff, C.A., Pearson, D., Rogers, J., St-Jacques, D., Warner, P.J., Wilson, D.M.A.: Measurements of the changes in angular diameter of Mira variables with pulsation phase. In IAU Symp. 191: Asymptotic Giant Branch Stars, volume 191, pp. 145–+ (1999) 329

Cometary Dynamics

H. Rickman

Abstract We present a review of cometary dynamics focusing on the long-term evolutions of cometary orbits that are responsible for the transfer of comets between different parts of the Solar System. The underlying mechanisms are described with particular emphasis on planetary perturbations. In order to place the dynamical theory into its proper context, we also discuss the distribution of observed cometary orbits and how this is affected by discovery biases. We end the review by a preliminary discussion of two current problems dealing with cometary dynamics: the formation and evolution of the scattered disk and the Oort Cloud and the capture of comets into short-period orbits.

1 Introduction

A comet is defined to be a small body of the Solar System, which develops an outgassing activity. The force that governs its orbital motion will hence include both the gravity of the Sun and other objects and the jet force caused by gases leaving the nucleus. This is one distinctive feature of cometary dynamics as compared with, e.g., planetary dynamics. However, the nongravitational force is relatively small and is of interest mainly when discussing models for the thermophysical behaviour of the cometary nucleus or the linkage of all apparitions of a periodic comet during a long time interval (see, e.g., [88]). It is rarely of importance when dealing with the large-scale evolution of cometary orbits, which this review focuses on, so we will not pay much attention to it.

Another issue is what we shall mean by cometary orbits. A straightforward definition would be the orbits of observed comets, but these evolve into or have evolved from very different orbits, quite unlike those of the observed comets, with much larger perihelion distances such that no cometary activity can be expected. In fact, it is natural to also include these other kinds of orbits. We thus treat some aspects of the dynamics of Centaurs and transneptunian objects as well as Oort Cloud objects way

H. Rickman (✉)
PAN Space Research Center, Warsaw, Poland; Uppsala Astronomical Observatory, Uppsala, Sweden

Rickman, H.: *Cometary Dynamics*. Lect. Notes Phys. **790**, 341–399 (2010)
DOI 10.1007/978-3-642-04458-8_7

beyond the limits of cometary activity, regarding these as true features of cometary dynamics.

The mechanisms of orbital transfer include both gravitational scattering during encounters with planets or passing stars and the secular effects of planetary gravity or the gravity of the entire Galactic disk. We will also pay some attention to how the transfer of comets may have differed in the early Solar System from what it is now, but we shall not indulge into the complicated dynamics that characterized the newly formed comets, when they were immersed into the solar nebula with its effects of gravity and aerodynamic drag—see, e.g., [8]).

2 General Transfer Scenarios

It is important to understand the basic role of perihelion and aphelion distances in cometary dynamics—these are the minimum and maximum distances from the Sun in an elliptic heliocentric orbit. In particular, the perihelion distance (q) governs the amount of outgassing activity of a comet such that, in general, comets are only observable if $q \lesssim 2.5$–3 AU, well inside the orbit of Jupiter. On the other hand, the aphelion distances (Q) are practically always larger than Jupiter's mean distance from the Sun, so that the observed comets are practically always Jupiter-crossing. This cometary property is quite distinctive when compared to other small Solar System bodies, but is of course shared by the particles of cometary meteor streams. Apart from the Trojans, which may indeed have an origin similar to that of comets [71], only a small minority of asteroids share the Jupiter-crossing type of orbits that characterize the periodic comets, and these are generally believed to be defunct cometary nuclei based on their dynamical and spectral properties [62, 79].

Jupiter-crossing short-period comets sometimes avoid encountering the planet due to mean motion resonance (in case of small inclination) or libration of the argument of perihelion (in case the inclination is substantial), but long-term integrations show that these protections do not last for very long. Thus one may think of these comets as dominated by *close encounters with Jupiter* and as part of a larger population of objects—not always observable—sharing the same property and linked to the observed comets by special dynamical routes. As we shall see, this is a fruitful idea when studying cometary origins. In fact, the idea of the giant planets as powerful transformers of cometary orbits [47] through successive encounters is at the heart of the currently preferred scenario for linking the observed Jupiter Family (see Sect. 3.1) to the transneptunian scattered disk [19].

To a large part, cometary dynamics can be understood as the chaotic migration of comets on a complicated network of routes connecting different orbits, of which we see only those that come closest to the Sun. But the picture would remain mysterious without considering certain integrable, secular effects that do play essential roles. Consider Fig. 1, which is copied from Rickman and Froeschlé [84]. These two plots, showing $\lg q$ versus $\lg Q$, were devised to illustrate the main transfer routes of comets in the Solar System, as far as they were known at that time. Note the absence

of the transneptunian population, which remained speculative and is indicated by a large question mark. The main idea was that external influences (mainly the so-called Galactic tide) perturb the angular momenta of Oort Cloud comets (upper right) so that some have their perihelion distances reduced practically to zero, thus falling into the inner Solar System and getting perturbed by the giant planets. Then the comets enter into a random walk in orbital energy (mainly along the abscissa), caused by close and distant encounters with the planets. This may finally lead to ejection into interstellar space ("unstable orbits") or into "rapid destruction" by falling into the Sun or disintegrating in the solar heat, but some comets will find their ways into the short-period population like the Jupiter Family (lower left).

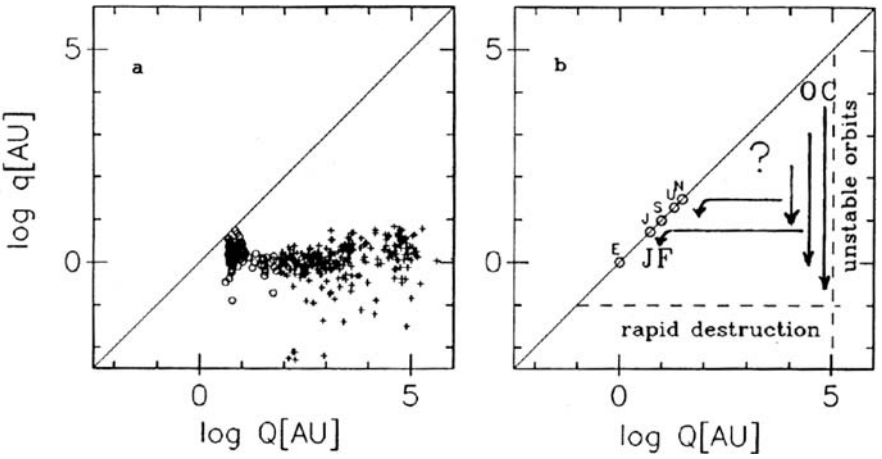

Fig. 1 Summary of cometary dynamics according to [84]. The *left* diagram shows the comets discovered at that time, and the *right* one illustrates the main dynamical transfer routes believed to be important for delivering observable comets at that time

In this picture, an essential ingredient of secular dynamics was the angular momentum transfer by the Galactic tide, but the rest could in fact be understood as resulting from planetary encounters. However, the progress made since Fig. 1 was first published has been enormous, and we now need to replace the plots by new ones shown in Fig. 2. First, of course, the transneptunians have been discovered, and these have been included into Fig. 2a even though they are not counted as comets. Moreover, there has been a drastic increase of the number of Centaurs, which we have also included into the new plot. The emerging theoretical scenario, shown in Fig. 2b, is now different from before. A central role is played by *the scattered disk*, which was probably much more massive, when the Solar System was young. Objects have leaked away from this reservoir, both inwards into the Centaur and Jupiter Family populations and outwards into the Oort Cloud. Some leakage also occurred into orbits inside the Oort Cloud like that of (90377) Sedna—especially during the early days of the Solar System, when the scattered disk was massive and the influences due to neighbouring stars may have been very large [7]. The Oort Cloud is still

generally considered to be the main source of long-period and Halley-type comets (see Sect. 3.1), but it now appears to be to a large extent a secondary structure that has arisen from the scattered disk [32].

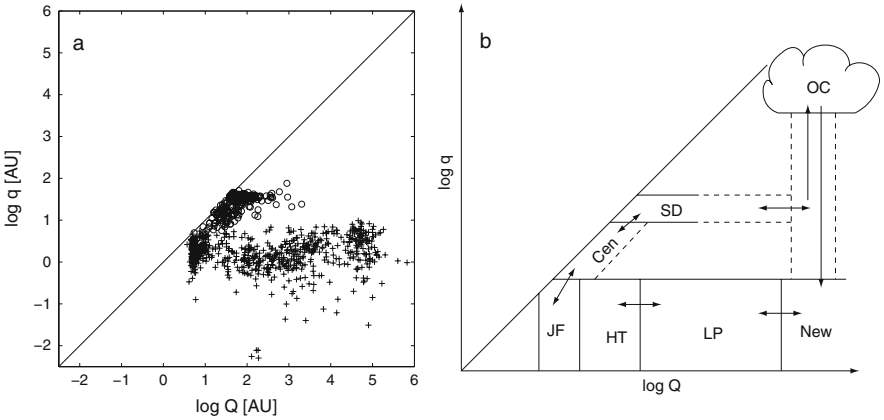

Fig. 2 (**a**) Comets discovered until the end of 2005 (*plus signs*), and Centaurs and transneptunians discovered until the end of 2007 (*open circles*), in a diagram of the log of the perihelion distance versus the log of the aphelion distance, both counted in AU. (**b**) A rough picture of the limits of cometary and related populations in the same kind of diagram, along with arrows aimed to indicate the current picture of the transfer between the different populations. See the text for details

The role of secular cometary dynamics appears more prominent today than 20 year ago. It is almost exclusively a question of so-called *Kozai cycles* (see Sect. 5.2), which lead to coupled variations of eccentricity and inclination so that the perihelion distance may undergo large changes. The Galactic tide can basically be seen as an example of a Kozai cycle, but there are other examples too. For instance, secular effects are important in the dynamics of the scattered disk, even though a decisive role is played by close encounters with Neptune. There are resonant routes linking the disk with the exterior Edgeworth–Kuiper Belt, so that a certain exchange of objects is unavoidable over long periods of time. In this regard, Kozai cycles may be important in bringing perihelia away from the vicinity of Neptune's orbit [36]. Another example is the case of *sungrazers*, i.e. comets falling into the Sun or into orbits with perihelia so close to the Sun that the objects disintegrate rapidly. This has been shown to be a fairly common end state of comets [2], caused by Kozai cycles. As a final, somewhat related example, the famous comet D/Shoemaker-Levy 9, which fell into Jupiter in 1994, was experiencing a temporary satellite capture (see Sect. 4.3), where the orbit changed due to a Kozai cycle caused by the solar perturbations.

Moreover, the role of indirect perturbations in contributing to the scatter in orbital energy for new comets from the Oort Cloud or long-period comets with small perihelion distances is now better understood (see Sect. 5.1), thereby adding a component to this part of cometary dynamics that is not induced by planetary encounters. A mapping for cometary dynamics, *i.e.*, an algebraic transformation

of orbital elements between successive revolutions, based on indirect perturbations was devised by Chambers [12], in particular to study the dynamics of Halley-type comets.

Finally, we have to realize that transient features may play an important role in cometary dynamics. The reason is that there are phase space domains where the "usual" perturbations are totally negligible, so normally these are isolated from the regions where active dynamical transfer proceeds. However, on rare occasions comets can be stored there or return into the active regions. For instance, Hills [42] pioneered the study of temporary peaks in the influx of Oort Cloud comets due to rare, close stellar passages ("comet showers"), assuming that the cloud has a massive inner core that is inert to the action of usual stars passing at large distances. Recent investigations of Oort Cloud formation [7, 8] show that the dense stellar environment of the early Solar System may indeed have created such inner populations, like that of Sedna-type objects, so the picture of "dynamically dormant" cometary populations that may be "awakened" by close enough stellar encounters remains highly relevant.

After this introductory description of how comets are believed to be transferred between different orbits, we next describe in some detail how the distributions of orbital elements of observed comets are interpreted in terms of possible source populations, selection effects, and observable lifetimes. Then we will concentrate on the effects of close planetary encounters, stellar passages, Galactic tides, and Kozai cycles, and finally we will discuss some current problems plaguing the current understanding of the main transfer mechanisms.

3 Orbital Elements of Observed Comets

3.1 Short-Period Comets and the Tisserand Parameter

One of the most striking features of cometary orbits is that they span an enormous range of periods (P), from just over 3 yr (comet 2P/Encke) to an upper limit that is hard to define for reasons to be demonstrated below, but in any case extends to millions of years. Separating comets into classes with different orbital periods is natural and may actually have some dynamical relevance, but the classification used is rather made for practical reasons. Long-period comets are those with $P > 200$ yr, and short-period comets have $P < 200$ yr. Sometimes the short-period comets are referred to as "periodic", because they have usually been seen during more than one perihelion passage ("multi-apparition comets"), or they will likely be seen again soon. On the other hand, for long-period comets the previous perihelion passage happened before there were any serious attempts at discovering comets, and the next one will occur far into the future. There is only one case of a comet that has been observed during two consecutive perihelion passages separated by more than 200 yr, namely, comet 153P/Ikeya-Zhang with perihelia in 1661 and 2002, and this is listed with the periodic comets. Another peculiar case is comet D/1770 L1 (Lexell), which was observed during one perihelion passage with an orbital period of 5.6 yr but

encountered Jupiter shortly afterwards and was expelled into a long-period orbit with $P \sim 300$ yr, where it is still performing its first revolution [47].

The latter example illustrates an inconvenience of the period-based classification, namely, that comets can easily change categories following close encounters with Jupiter, as the orbital period can then be drastically perturbed. The same remark holds for attempts in old literature to classify the short-period comets into families associated with the different giant planets based on where the aphelion is located. Such families have no dynamical meaning, since Jupiter is almost always the planet that dominates the orbital perturbations on observed comets, independent of their aphelion distance. A much more useful quantity can be formulated using the fact that Jupiter is the dominant planet. In the approximation where Jupiter's orbit is circular and no other planets have any effects (the so-called *circular restricted three-body problem*) one can write down the Jacobi integral, which is the energy integral in the corotating frame, and Tisserand [90] found an expression for this in terms of orbital elements (the Tisserand criterion) that holds approximately as long as the comet is not in the vicinity of either Jupiter or the Sun. The quantity in question is called the *Tisserand parameter*:

$$T = \frac{a_J}{a} + 2\sqrt{\frac{a}{a_J}(1 - e^2)}\cos i. \tag{1}$$

Here a and a_J are the orbital semi-major axes of the comet and Jupiter, e is the eccentricity, and i is the inclination of the cometary orbit. Since the circular restricted three-body problem is a fairly good approximation for observed comets, T remains nearly constant even in the presence of close encounters between the comets and Jupiter. Thus, using T, one may classify comets in a way that reflects their orbital history as well as their present orbits.

The prime example of the use of T concerns the separation of short-period comets into two types, the *Jupiter Family* and the *Halley-type* comets (usually abbreviated JF and HT, respectively). Historically, the JF was recognized as a peak around a_J in the histogram of aphelion distances of short-period comets, and thus the JF was the prototype of the above-mentioned comet families. Figure 3a illustrates this, showing the Q distribution for short-period comets discovered before 1980. Figure 3b shows the bivariate distribution of i versus Q for the same sample of comets, and here we see something quite peculiar. The concentration of comets with $Q \simeq a_J$ shows a very strong preference for small inclinations, but going to larger values of Q, there is a sharp transition to a different regime, where all inclinations exist in relatively equal proportions. Thus, with the historical definition of the JF, there follows the peculiar property of a very flattened orbital distribution that is not shared by other comets.

In fact, low-inclination comets generally tend to remain at low inclination even after close encounters with Jupiter that may change the orbits considerably. This means that their orbital evolution can be approximately traced in a parametric plane such as (a, e) or (Q, q), putting $\cos i \equiv 1$. In Fig. 4 we show the (Q, q) plane with evolutionary curves derived from (1) for different values of T. We also show the

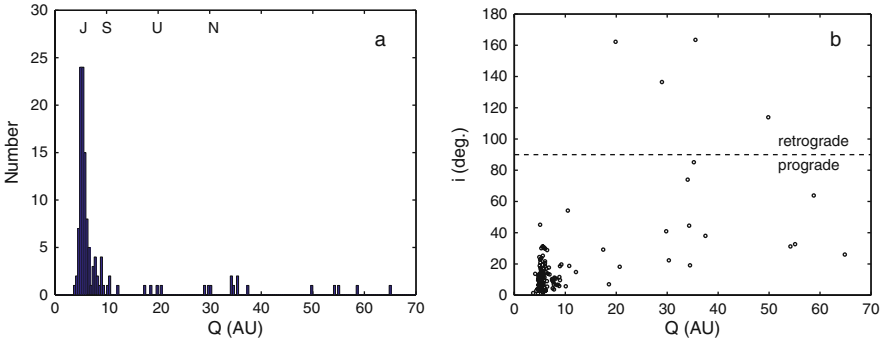

Fig. 3 (**a**) Histogram of aphelion distances for short-period comets discovered before 1980. The mean distances from the Sun of the giant planets are marked at the *top*. (**b**) Inclinations of the same comets plotted versus their aphelion distances

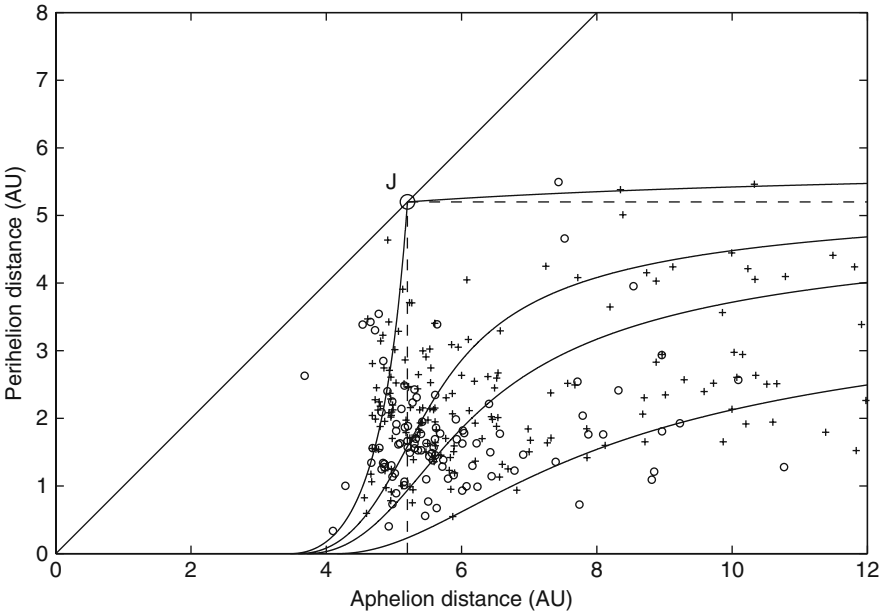

Fig. 4 Orbital evolutionary curves in a diagram of perihelion distance versus aphelion distance, as given by the Tisserand criterion with zero inclination. The *four curves* represent, from top to bottom, $T = 3.0$, 2.9, 2.8, and 2.5, respectively. Symbols have been plotted for comets with inclinations $i < 30^\circ$ discovered before 1980 (*open circles*) and after 1980 (*plus signs*). The leftmost symbol corresponds to the peculiar comet 133P/Elst-Pizarro

observed comets with $i < 30^\circ$, using different symbols for discovery dates before and after 1980.

Since $\cos i$ is close to unity, the actual T values are well approximated by interpolation between the curves, and we see that the comets forming the concentration near $Q = a_J$ mostly fall between $T = 2.5$ and $T = 3$. A general feature of the

evolutionary curves for $2.5 < T < 3$ is that with increasing orbital periods, the perihelion distances increase towards $q \simeq a_J$. The concentration of observed comets seen at low q values and $Q \simeq a_J$ is hence expected to continue towards the upper right in the diagram, where the objects are much less easily observable, and Jupiter may have "captured" comets into their observed orbits along the evolutionary curves in question. The start of a verification of this scenario can be seen in the diagram, since the more recent discoveries tend to dominate in the region to the upper right.

This allows us to introduce a dynamical definition of the Jupiter Family as short-period comets with $T > 2$ [11, 54], which is indeed the accepted definition nowadays. The rest of the short-period comets (with $T < 2$) are referred to as Halley-type comets. The HT comets generally avoid the range of orbital periods covered by Fig. 4, but a few exceptions exist. Moreover, the recent improvements in detection techniques that have allowed the discovery of JF comets with larger perihelion distances (Fig. 4) tend to wash out the concentration to $Q \simeq a_J$. Thus, the latter is only an effect of observational selection, and the JF actually extends to orbital periods typical of the HT comets although these members remain difficult to observe. Figure 5 illustrates the mixture that exists in orbital period between JF and HT comets—mainly due to recent discoveries. Less than 30 yr ago the $P = 20$ yr line was an excellent proxy for the $T = 2$ line, since comets were practically absent in the 1st and 3rd quadrants.

Due to the stability of T, very few comets are able to transit across the $T = 2$ border [56]. Thus, even if the JF and HT groups may or may not have belonged to the same orbital population a long time ago, they are currently distinct and have arrived via different pathways. Let us note an interesting physical interpretation of this classification. We use (1) with the customary units of the restricted three-body problem (see [15]), so that $a_J = 1$. If a comet approaches Jupiter, its velocity V in a fixed frame satisfies the relation

$$V^2 = 2 - \frac{1}{a} \tag{2}$$

expressed in the same units. We also recognize that the component of cometary angular momentum perpendicular to Jupiter's orbit is

$$L_z = \sqrt{a(1-e^2)}\cos i = V_T, \tag{3}$$

where V_T is the cometary velocity component parallel to Jupiter's velocity upon encounter, so that the Tisserand parameter can be expressed as

$$T = 2 - V^2 + 2V_T. \tag{4}$$

The relative encounter velocity U between the comet and Jupiter can be written as

$$U^2 = (\mathbf{V} - \mathbf{C})^2 = V^2 + 1 - 2\mathbf{V}\cdot\mathbf{C} = V^2 + 1 - 2V_T, \tag{5}$$

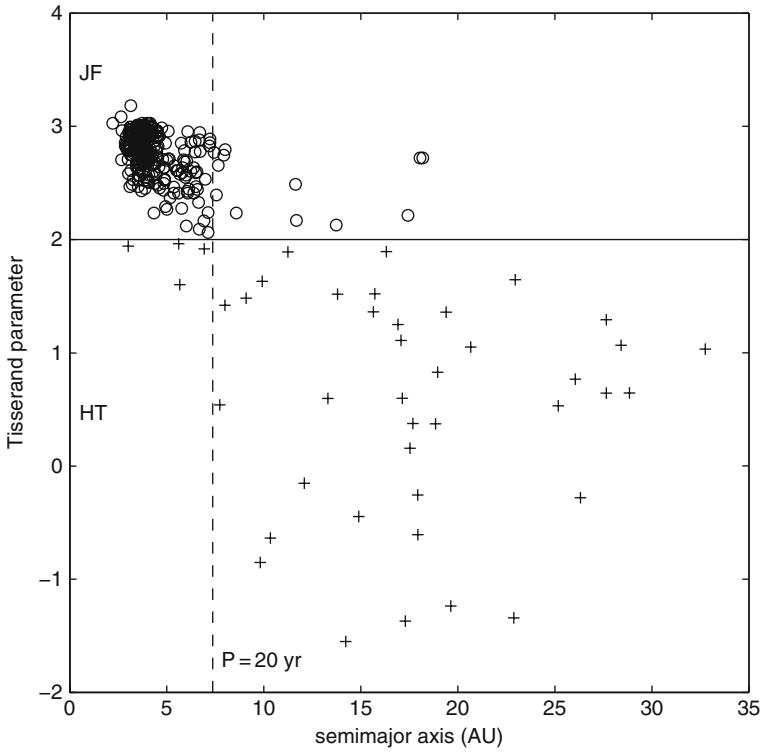

Fig. 5 Tisserand parameters versus semi-major axes of short-period comets. The *horizontal line* marks the difference between Jupiter Family and Halley-type comets, and the *vertical dashed line* marks the orbital period of 20 yr, which has been used as a proxy for the distinction of the two types. *Open circles* show Jupiter Family comets, and *plus signs* show Halley-type comets

where **C** is Jupiter's velocity vector ($C = 1$), and from (4) and (5) we get

$$U^2 = 3 - T. \tag{6}$$

This relation [74] shows that T can be seen as a way to express the relative encounter velocity as a comet approaches Jupiter in units of Jupiter's orbital speed. If Jupiter's orbit were circular, $T = 3$ would correspond to zero-velocity encounters, and Jupiter crossers would have to have $T < 3$. The fact that some comets appear to have $T > 3$ in Fig. 4 is partly an illusion, since the comets do not have zero inclinations. A small number actually do have $T > 3$, but they are still able to encounter Jupiter due to the planet's orbital eccentricity. However, these encounters occur at very low velocities and often involve temporary satellite captures. The Jupiter Family is—quite generally—composed of those comets that encounter Jupiter at low velocities (smaller than Jupiter's orbital speed), while Halley-type comets have high encounter velocities.

3.2 Inclinations and Perihelion Distances

In connection with Fig. 3b we noted that comets with aphelia near Jupiter's mean distance from the Sun tend to have low inclinations, while short-period comets with aphelia further out do not share this property. The former correspond closely to the Jupiter Family, and the latter to the Halley-type comets. If we use $P = 20$ yr as a proxy for the separation of JF from HT comets (Fig. 5), we can illustrate the different inclination distributions of all observed comets by means of a plot of i versus $1/a$. Figure 6 shows such a diagram. Retrograde comets ($\cos i < 0$) are in the lower half of the plot, and prograde comets ($\cos i > 0$) are in the upper half.

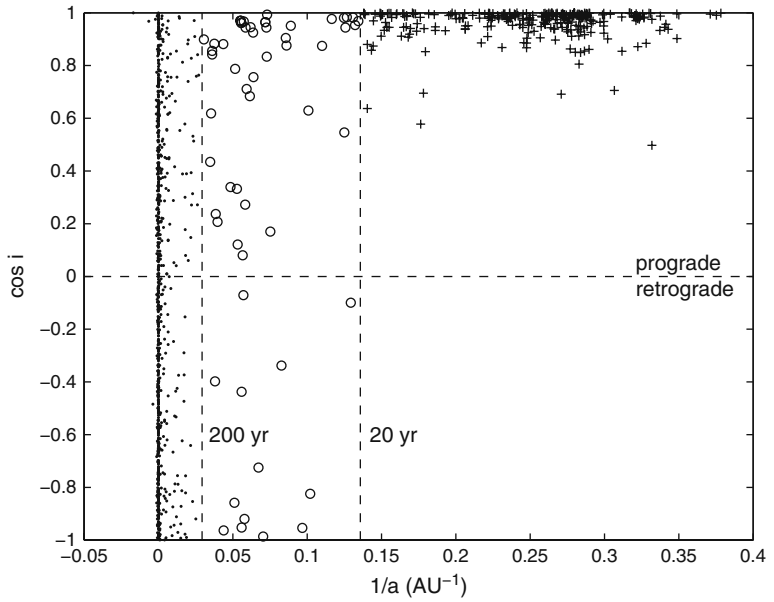

Fig. 6 The cosine of the inclination versus the inverse semi-major axis for all the comets. Different symbols are used for long-period comets ($P > 200$ yr), short-period comets with $P > 20$ yr, and short-period comets with $P < 20$ yr. The last two groups closely correspond to the Halley-type and Jupiter Family comets, respectively

We see that, indeed, JF comets are extreme in their preference for low inclinations. The HT comets appear intermediate in the sense that they keep some such preference although a much smaller one, while the long-period comets ($P > 200$ yr) appear to have a more or less flat distribution of $\cos i$. This would be characteristic of a uniform distribution of the orbital planes, but a uniform distribution is not necessarily implied. The distribution of $\cos i$ may be flat, even if the orbital planes are not uniformly distributed (see Sect. 5.2). Near $1/a = 0$ we see a characteristic feature of long-period comets, to which we shall return in Sect. 3.3, namely, a sharp pile-up in a narrow range of orbital energies. The comets contributing to this pile-up have a very flat distribution of $\cos i$.

The difference of inclination distributions between JF and HT comets is in line with our above statement that the two populations are basically distinct from each other and that they have followed different dynamical routes into their current orbits. We shall return to a discussion of these routes in Sect. 6.2.

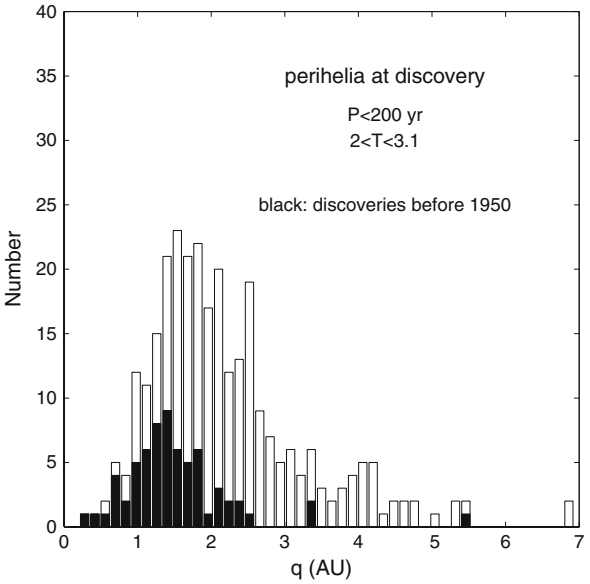

Fig. 7 The distribution of perihelion distances of Jupiter Family comets is shown by the *white histogram*, while the *black histogram* shows the same distribution limited to comets discovered before 1950. Note that there are relatively few recent discoveries of comets with small perihelion distances—instead, a large fraction concern comets with $q \gtrsim 2.5$ AU

Let us now concentrate on the perihelion distances and recent trends in extending observations of cometary activity to larger heliocentric distances. Figure 7 shows how q is distributed for JF comets, and since these are often subject to large perturbations of q at close encounters with Jupiter, we have plotted the value of q of each comet at the time when it was discovered. Thus the diagram is useful as an indicator of the most important discovery bias for comets—that of selecting small enough (but not too small) perihelion distances. This bias obviously depends on the observational facilities used for comet discoveries, and as the diagram shows, recent developments in this regard have tended to push the limit of detection of a JF comet outwards. Before 1950 there were practically no JF comets that had been discovered with $q > 2.5$ AU (although a few more were being observed with $q > 2.5$ AU after having been discovered closer in). But nowadays a significant fraction are being discovered with $q > 4$ AU, and the main peak in the histogram extends to $q \simeq 3$ AU.

It has to be emphasized, however, that a comet at 4 AU from the Sun is not the same thing as one at 1 AU. The advantage of recent techniques is not just

that they reach to fainter magnitudes, but they are able to detect and characterize a low-level cometary activity around objects that might otherwise be taken for distant asteroids. In fact, many of the recent discoveries at large heliocentric distance have happened in a peculiar way. Some asteroid search programme (e.g. LINEAR) finds an apparently asteroidal object, but orbit determinations indicate a "cometary" orbit, and observers are alerted to this fact. They quickly find that the object actually has cometary activity, so a new comet has been found, but it often continues to carry its preliminary, asteroidal designation.

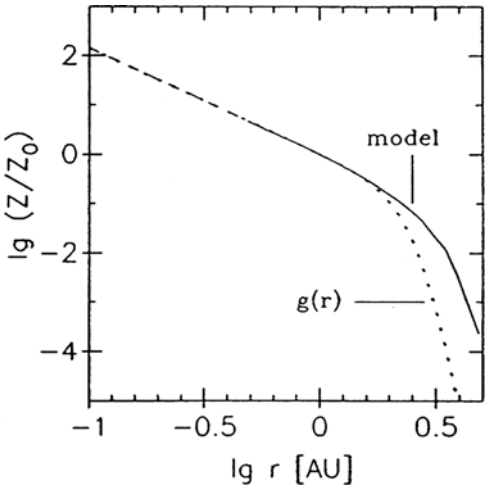

Fig. 8 A thermal model of H_2O sublimation flux (Z) versus heliocentric distance (r) yields the *solid curve*, and the *dashed extension* closest to the Sun is an extrapolation following a r^{-2} dependence. The *dotted curve* shows the $g(r)$ function commonly used in non-gravitational orbit determinations for comets. Both *curves* have been scaled to $\log(Z/Z_0) = 0$ at $r = 1$ AU. From [78]

In Fig. 8 we show the results of theoretical model calculations of the H_2O sublimation rate from an icy object as a function of its heliocentric distance. The "knee" in the curve marks a shift of the surface energy balance such that at smaller distances from the Sun, where the insolation rate is higher, most of the energy absorbed goes into sublimation of ice, while at larger distances, the energy goes mostly into thermal radiation and the sublimation rate drops rapidly. Therefore, the position of the knee at $r \sim 2.5$–3 AU is often interpreted as the effective limiting distance of cometary activity—at least that driven by H_2O sublimation. One can easily understand why the activity sometimes seen at larger distances is at a lower level and has been hard to observe until recently. There are also exceptional cases of comets that show a strong activity very far away, and these are believed to be comets with a large amount of volatiles—in particular, CO.

Comet Hale-Bopp (C/1995 O1) was an important example among the long-period comets, but there are short-period comets too. Historically the most famous example is comet 29P/Schwassmann-Wachmann 1, which has a low-eccentricity orbit between those of Jupiter and Saturn and is well known since its discovery in the late 1920s for irregularly occurring outbursts of activity. During the intervening periods of quiescence the comet is active too [46], and the gas that causes this activity was identified in 1993. Radio observations at millimetre wavelengths [87]

then showed that the comet is outgassing CO at a significant rate at all times, but the reason for the outbursts is still unknown.

For a very long time comet 29P was a unique short-period comet with its large perihelion distance, but the situation changed with the discovery of 95P/Chiron in the late 1970s. This object was first listed with the asteroids with the number 2060, but its orbit with perihelion near Saturn's orbit and aphelion near that of Uranus was clearly not asteroidal at all. It rather appeared to be a possible precursor of the JF comets, in case Saturn would capture it along an evolutionary curve with T between 2.5 and 3 with respect to Saturn, so that the perihelion would come closer to Jupiter's orbit and Jupiter could take control of the evolution. Discovery of cometary activity in the form of a grain coma had to wait until 1990 [63], but whether CO vapour is the main driver remains to be convincingly demonstrated. It is true that Chiron with its estimated radius of ∼100 km is much larger than any of the JF comet nuclei, and the nucleus of comet 29P is extremely large as well, but it was evident already in the 1980s that the distant source populations of the Jupiter Family in the regions between the giant planet orbits should be much more numerous than the JF, and the largest objects should then be much larger too [77].

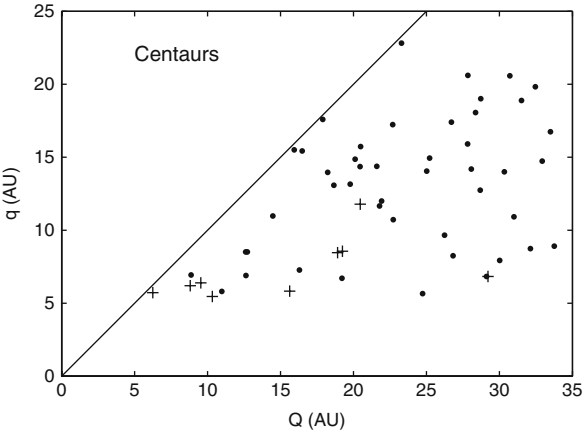

Fig. 9 The orbital distribution of Centaurs with aphelion distances $Q < 35$ AU, showing the ones that have exhibited cometary activity by *plus signs* (the leftmost one is comet 29P) and the so far inactive ones by *black dots*. An upper limit of 25 AU has been adopted for the perihelion distances

More recently there have been many discoveries of objects moving among the orbits of the giant planets, and they have come to be called *Centaurs*. With the usual definition of this term, all objects with perihelia beyond Jupiter's orbit but inside that of Neptune are Centaurs, so both comets 29P and 95P are included. But what about the other objects? If this is really a protocometary population, it is obviously of interest to find out if they show some signs of cometary activity like the members that we already discussed. Indeed such activity has been looked for and found in several cases, and Fig. 9 shows a plot of the Centaur orbits in a (Q, q) diagram (all the inclinations are low), identifying those that have so far shown cometary activity.

Thus we see that the observations of comets are being pushed to larger distances than before, and we are beginning to explore the source populations from where the comets have been captured. This of course includes the transneptunian population

too, although cometary activity has not been observed there except—somewhat marginally—in the case of the giant object (134 340) Pluto with its tenuous and partially escaping atmosphere. The latter, however, has not been imaged like one usually does with cometary comae. The Centaur comets on the other hand have in fact received official cometary designations, so they must be regarded as short-period comets, but their T values are usually very large since they are moving far beyond Jupiter's orbit (see Fig. 4). It does not appear appropriate to include them into the JF, so one tends to include an upper limit of T into the JF definition, and this may be taken to be 3 or slightly larger.

3.3 The Oort Cloud and Cometary Fading

Let us now turn to the most distant cometary reservoir in the Solar System. The evidence for its existence is of a very special kind. One starts from orbit determinations for long-period comets, which lead to osculating elements valid at a time near perihelion passage, when the observations were made. This orbit can then be integrated backwards or forwards in time, allowing for the perturbations due to all the planets (mainly the giant planets) in a heliocentric frame. After this is done one can transform the heliocentric orbits before entry into and after exit from the planetary system into the barycentric frame of the Solar System. Those orbits are called the *original* and *future* orbits, respectively. They are of great interest, since they show whether the comet approached the Sun along an elliptic or a hyperbolic orbit and, thus, whether it came as a member of the Solar System or as an intruder from interstellar space. They also show if the comet will remain in the Solar System or escape into interstellar space.

Such calculations—initially using simple numerical integrators but no electronic computers—were first done about a century ago by Elis Strömgren and his colleagues at Copenhagen Observatory and Gaston Fayet at Paris Observatory. Of course it took a long time before the number of resulting orbits started to grow considerably, but today we have access to more than 400 of them. Figure 10 shows combined histograms of original and future inverse semi-major axes ($1/a_{\rm ori}$ and $1/a_{\rm fut}$, respectively), and it reveals a very peculiar feature of the original orbits. There is a strong tendency for the values of $1/a_{\rm ori}$ to pile up in the interval from 0 to 10^{-4} AU^{-1}. Those orbits are very weakly bound to the Solar System, and they extend to very large distances, the semi-major axes being larger than 10,000 AU.

There is no counterpart of this peak among the future orbits. These scatter over both positive and negative values of $1/a_{\rm fut}$ with nearly equal probabilities as a result of the planetary perturbations experienced during their visits in the planetary system. The average planetary perturbation (mainly due to Jupiter) is much larger than the width of the original peak, so the latter is completely wiped out. Nearly half the comets contributing to these statistics are ejected from the Solar System by the planets, demonstrating one essential characteristic of cometary dynamics—the risk of ejection into interstellar space. In fact, this is a fate that has been shared by most

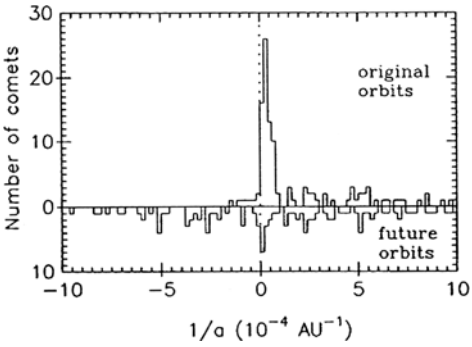

Fig. 10 Double histogram showing the distributions of the original inverse semi-major axes of long-period comets on top and the future ones (after exit from the planetary system) inverted below, using data listed in [65]. From [78]

comets, which entered in a similar way during the age of the Solar System, and the enrichment of interstellar space by comets may be considerable, in case other stars tend to have systems of comets and planets like our own.

Evidently the rest of the comets will return to perihelia in the future, and apparently their distribution of original orbits will then look like the tail seen to the right of the peak in the $1/a_{\rm ori}$ distribution. Thus it is natural to think that the comets contributing to this tail are returning comets, but there clearly needs to be a special source for the comets of the peak, which must be newcomers injected from this source. From the size of the orbits, and the fact that the comets spend most of the time in relative proximity of their aphelia more than 20,000 AU away, we conclude that the source must be a vast region of space in the outskirts of the Solar System. From the flat distribution of $\cos i$ in Fig. 6 we see that it extends in all directions. Jan Oort [72] was the first to plot the distribution of original orbits (he actually had only 19 orbits in his sample!) and to realize the significance of the pile-up. After him, the peak is called the *Oort peak*, and the distant source from which the new comets arrive is called the *Oort Cloud*.

The mechanism envisaged by Oort for inserting comets from the Oort Cloud into observable orbits is "stellar perturbations", i.e. the gravitational impulses received by comets in the cloud from passing stars. Oort was able to estimate that the frequency of close passages is large enough to cause a significant scatter of the angular momentum vectors of comets at distances of $\sim 10^4$ AU, and thus it was clear both that the cloud should be essentially isotropic and that there is a way to remove angular momentum from some members so that they finally penetrate into the observable region with small perihelion distances. We look closer at these perturbations in Sect. 5.1, and in Sect. 6.1 we consider the question about the interaction of individual passing stars and the tidal influence of the entire Galactic disk, as well as the importance of nongravitational effects when determining the original orbits. All this is very important when investigating the size and mass of the Oort Cloud.

Let us now use the Oort peak to illustrate the phenomenon of cometary fading or observable lifetimes, which is of essence for linking cometary dynamics to the observed distributions of orbital elements. As a test of the above idea of new and returning comets, consider the following simulation. From the IAU/MPC

Catalogue of Cometary Orbits [67] we extract two data sets. One is the original inverse semi-major axes for a sample of 279 orbits of quality class 1,[1] and the other is the differences $1/a_{\rm fut} - 1/a_{\rm ori}$ for the corresponding sample of 418 comets representing all the quality classes. During each time step we introduce a constant number of new comets with random values of $1/a_{\rm ori}$ uniformly chosen between 0 and $0.8 \cdot 10^{-4}$ AU^{-1}. These get perturbed by quantities $\Delta(1/a)$ that we pick at random from the second data set. If the new orbit is hyperbolic, or if the new semi-major axis is larger than 200, 000 AU $\simeq$ 1 pc, the comet is considered lost from the Solar System. Otherwise, it comes back after one orbital period in an elliptic orbit and gets perturbed again.

If this is allowed to run for a very long time, a steady state is reached so that no matter which time interval we pick for extracting the statistics of inverse semi-major axes of new and returning comets, the resulting distribution will approximate a constant parent distribution. In Fig. 11 we show both the observed $1/a_{\rm ori}$ distribution for the class 1 sample (first data set above) and the simulated steady-state distribution scaled to the same peak height. It is obvious that the two do not agree. In particular, the background flux of returning comets is far too large in the simulated distribution. So where is the error?

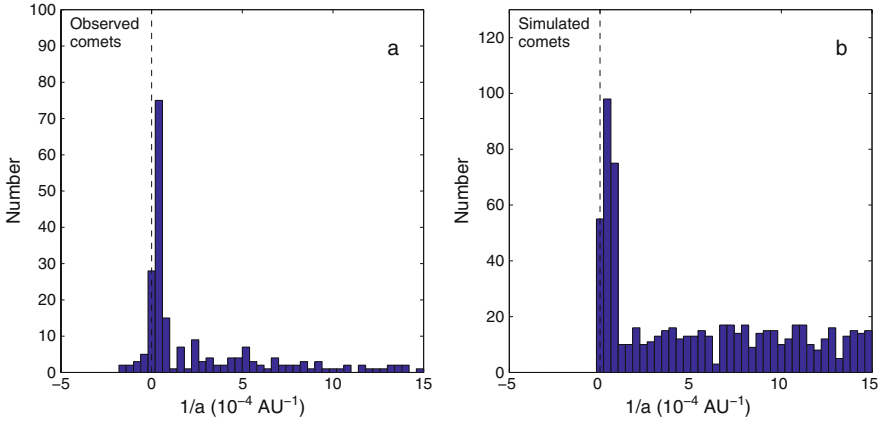

Fig. 11 (**a**) Histogram of original inverse semi-major axes for the class 1 comets of [67]. This corresponds to the upper histogram in Fig. 10, but the sample is considerably larger. (**b**) The corresponding histogram for a simulated sample of fictitious comets injected from the Oort Cloud at a constant rate and subject to the planetary perturbations shown in Fig. 12 at each perihelion passage. See the text for details

There is nothing wrong about the dynamics. The sample distribution of $\Delta(1/a)$ (shown in Fig. 12) is just as symmetric about zero as the one to be expected during

[1] The catalogue divides the comets into quality classes that are meant to indicate the reliability or accuracy of the derived $1/a_{\rm ori}$ values. These are assigned mainly in dependence of the mean residual of the astrometric observations with respect to the ephemeris provided by the orbit and the length of the underlying observational arc. Class 1 contains the best determined orbits.

a very long time interval, so we are not biasing against ejections from the Solar System and thereby favouring the return of comets. It is true that the limited size of our sample makes the distribution miss the far tails of the parent distribution entirely (see Sect. 5.1), but this does not lead to too many returning comets. Another concern would be warranted if the successive perturbations experienced at the different returns of a comet are not uncorrelated, as we have assumed. But we shall return to this point in Sect. 5.1, and it will be seen that real sequences of perturbations are indeed stochastic for long-period comets. Thus we are correct in modelling the evolution as a random walk.

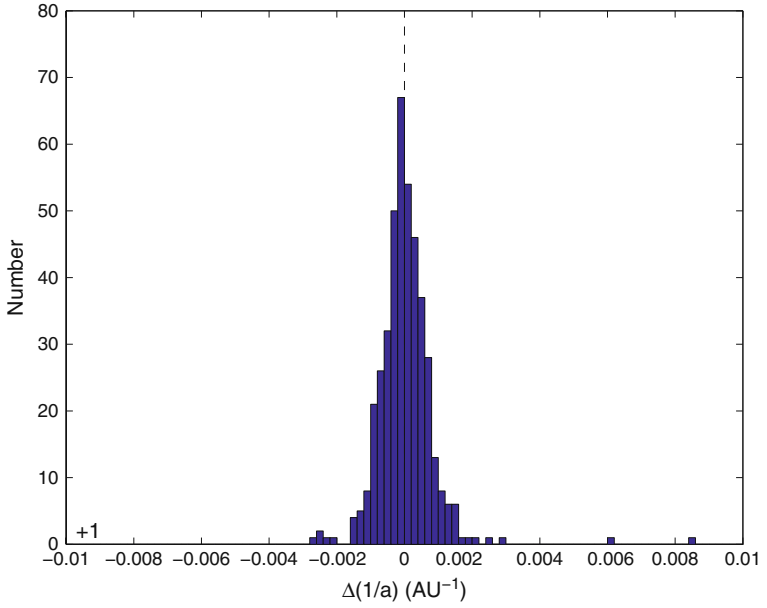

Fig. 12 Histogram of planetary perturbations of $1/a$ taken from [67] and used for the dynamical simulation of Fig. 11b. The number at the lower left indicates one extra comet that experienced $\Delta z \simeq -0.016$ AU^{-1}

Only two possibilities remain. Either we are not in a steady state or comets disappear (i.e. become unobservable) before they return. The first scenario is unattractive, because the comets come from such large distances that the reservoir should indeed be thermalized by stellar passages. Thus either it would recently have been captured by the Solar System, which is an extremely unlikely event to assume, or it would recently have experienced a very strong perturbation caused by a very massive or slow-moving object that passed. This again is unlikely, and we see no evidence for such a recent passage in the observations of stars and gas clouds in the solar neighbourhood of the Galaxy.

The importance of the loss of comets by some physical mechanism for explaining the large contrast between the Oort peak and the returning background has been realized since a long time. It was discussed already by Oort and Schmidt [73] in

terms of a rapid fading. The intrinsic brightness of a comet may decrease from one apparition to the next, and as a result the comet may be missed, if it does not reach the limiting magnitude of the search telescopes used. Modern theories about the origin of cometary activity often see fading as a result of the burial of subsurface gas production beneath a thickening dust layer, and for the case in point it may be that the new comets have nuclei with special chemical or physical properties.

Observational statistics of short-period comets also show that these often enter into a "dormant" state—sometimes for many orbits around the Sun, before they may eventually be woken up again [49]. Thus, at first sight, the idea of cometary fading is attractive, but it may be a problem that asteroid search programmes have not detected the amount of long-period asteroids that one would expect on the basis of fading and dormancy of long-period comets [59]. An alternative physical scenario that would avoid this problem is the sudden splitting or disintegration of cometary nuclei, which is known to occur due to numerous observations and would have the same effect of limiting the lifetimes of returning comets.

Whatever is the true mechanism, or combination of mechanisms, that shapes the distribution of orbital energies of long-period comets, it is clear that both observational selection and limited lifetimes must generally be taken into account, when interpreting observed orbital distributions of comets in terms of dynamical pathways. When it comes to short-period comets, rapid fading based on comparison of observed absolute magnitudes at different apparitions has been both claimed [96] and disputed [50]. It is very difficult to reach any safe conclusions based on these magnitudes, and the picture appears more complicated than a simple time dependence. In Fig. 13 we show evidence for an interplay of physical and dynamical evolution, based on the well-known orbital histories of some Jupiter Family comets during about a century backwards and forwards [4].

We plot the perihelion distances of each comet relative to that during the discovery apparition versus the number of revolutions before or after discovery. While the perihelia are seen to have diffused both inwards and outwards after discovery, there is a strong tendency for the pre-discovery perihelia to have been much further out. That is not enough to demonstrate anything interesting, because if the population of JF comets is already known to a good degree of completeness, newly discovered members will likely be found directly after they are captured by means of a close encounter with Jupiter that reduces the perihelion distance. One may think of a greedy child who has already eaten all the goodies in a candy box, and each time a new candy is put there, it is eaten at once. Thus all the candies that are eaten would be fresh. What we need is instead a situation, where the child is presented with a box containing both new and old candies, where nothing has been eaten yet. In this case the child would not have any reason to pick a fresh candy rather than one that has already come of age, unless the candies really deteriorate.

In Fig. 13 we have attempted to simulate this situation by plotting comets discovered before 1950 with perihelion distances in excess of 1.5 AU, because from the black part of the histogram in Fig. 7 we see that this sample is far from completeness. The fact that we nonetheless see a clear preference for "fresh captures" indicates that there is indeed a physical evolution such that the comets "deteriorate"

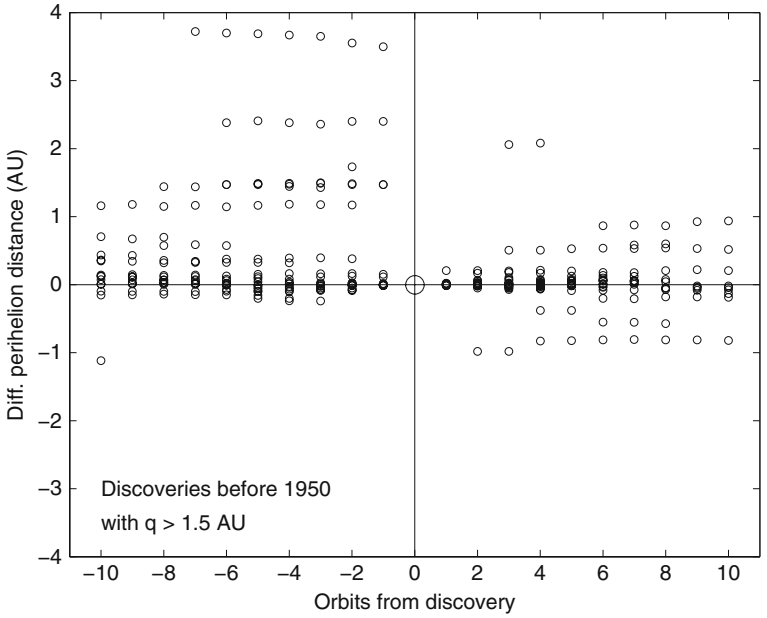

Fig. 13 Perihelion distances of Jupiter Family comets discovered before 1950 with $q_{\rm disc} > 1.5$ AU relative to $q_{\rm disc}$, plotted versus the number of orbital revolutions counted from the discovery apparition. The data used are from [4]

with time elapsed since the capture occurred. This indicates that one must take care about the definition of any sample of JF comets used for estimating the necessary flux of captures and the size of the source population.

4 Close Encounter Dynamics

Although most of the concepts to be described in this section are valid for encounters with any planet, we will focus on encounters with Jupiter. These provide the principal agent for changing the semi-major axes of observable comets. As an introduction we consider the equation of motion of a comet in the heliocentric frame and in the presence of one perturbing planet:

$$\ddot{\mathbf{r}} = -\nabla U_{\odot} - \nabla R_p, \tag{7}$$

where $U_{\odot} = -GM_{\odot}/r$ is the solar gravitational potential and

$$R_p = -GM_p \left\{ \frac{1}{|\mathbf{r} - \mathbf{r}_p|} - \frac{\mathbf{r}_p \cdot \mathbf{r}}{r_p^3} \right\} \tag{8}$$

is the planetary perturbing function. Taking the scalar product of the heliocentric velocity vector $\dot{\mathbf{r}}$ with (7), we obtain

$$\dot{E} = -\dot{\mathbf{r}} \cdot \nabla R_p, \tag{9}$$

with $E = T + U_\odot = -GM_\odot/2a$ denoting the total orbital energy. Let a_p denote the semi-major axis of the planetary orbit. We can then write

$$\frac{d}{dt}\left(\frac{1}{a}\right) = \frac{2m_p}{a_p}\,\dot{\mathbf{s}} \cdot \left\{ \frac{(\mathbf{s}_p - \mathbf{s})}{|\mathbf{s}_p - \mathbf{s}|^3} + \frac{\mathbf{s}_p}{s_p^3} \right\}, \tag{10}$$

where m_p is the mass of the planet in solar masses, and $\mathbf{s}$ and $\mathbf{s}_p$ are dimensionless position vectors of the comet and the planet defined by $\mathbf{s} = \mathbf{r}/a_p$ and $\mathbf{s}_p = \mathbf{r}_p/a_p$.

We see that the time derivative of the inverse semi-major axis is expressed as a scaling factor proportional to m_p/a_p times the time derivative of a dimensionless quantity that depends on the geometrical configuration Sun–comet–planet. The integration of this quantity with respect to time is a complicated matter especially when the comet and the planet experience a close approach, but we may note that the result scales as m_p/a_p, if different planets are considered with similar geometrical configurations.[2] This brings out the fact that Jupiter, with a ratio m_p/a_p that by far surpasses those of the other planets, is the dominant perturber of cometary orbits. In fact, if distributions of $\Delta(1/a)$ like that shown in Fig. 12 were plotted separately for the contributions from the different planets, we would be able to verify the proportionality just derived.

Before embarking on a description of close encounter dynamics, let us note the role that such encounters play in promoting chaos in the orbital evolutions of comets. There are different sources of chaos or unpredictability affecting different kinds of cometary orbits. For long-period comets successive perturbations $\Delta(1/a)$ are like random choices from the parent distribution, essentially because due to the long periods involved, the position of Jupiter at the next perihelion passage is very sensitive to minor changes in the perturbation applied (see Sect. 5.1). For Oort Cloud comets the regular dynamics normally imposed by the Galactic tides is interrupted by stellar passages in a way that has to do with the distribution of stars in the Galaxy but is completely independent on how the comets move. Thus the sequence of passages is random and represents an externally imposed chaos.

Finally, coming to short-period comets, close encounters with Jupiter happen frequently, and even a slight change in the outcome of one such encounter will lead to a major change of the circumstances around the next encounter, implying completely different evolutions. This is the reason why it is in fact impossible to trace the evolutions of individual JF comets over more than a couple of centuries (the typical interval between close encounters) from the observed motion even with very

[2] Note that this result is not exact concerning the closest encounters because of the finite extent of the planets, as will be discussed in Sect. 4.2.

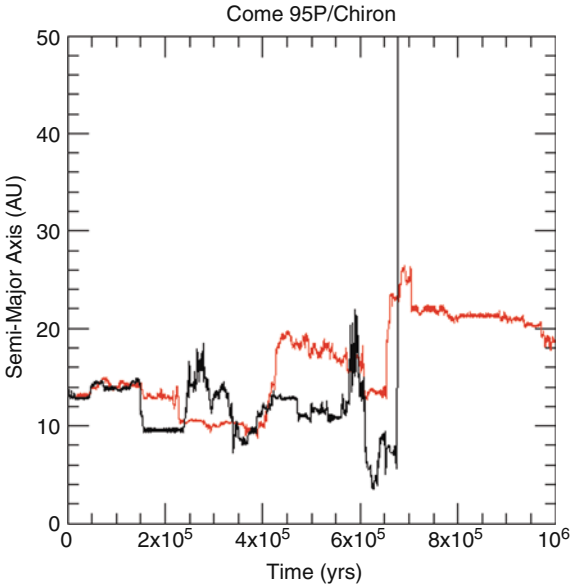

Fig. 14 Long-term evolution of the semi-major axis of Chiron's orbit. Two variants, plotted by *red* and *black curves*, were integrated with a minute difference in the initial conditions, using the same planetary system model. Due to close encounters with the planets, the orbits diverge considerably after $\sim 10^5$ yr. The two objects have lost all memory of their initial vicinity, and the dynamics is clearly chaotic. Courtesy H.F. Levison

accurate starting orbits and the best numerical integrators and dynamical models. An illustration is provided in Fig. 14. This shows the orbital evolution of 95P/Chiron, which is a Centaur (Sect. 3.2) affected only by Saturn and Uranus, but the phenomenon of extreme sensitivity to initial conditions is evident even in the absence of close encounters with Jupiter.

4.1 Sphere of Influence

Consider the acceleration of a comet situated close to a planet. The equations of motion in an inertial frame for the three-body system Sun–planet–comet are

$$\ddot{\mathbf{r}}_\odot = \frac{GM_p}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}\left(\mathbf{r}_p - \mathbf{r}_\odot\right), \tag{11}$$

$$\ddot{\mathbf{r}}_p = -\frac{GM_\odot}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}\left(\mathbf{r}_p - \mathbf{r}_\odot\right), \tag{12}$$

$$\ddot{\mathbf{r}}_c = -\frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}\left(\mathbf{r}_c - \mathbf{r}_\odot\right) - \frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}\left(\mathbf{r}_c - \mathbf{r}_p\right). \tag{13}$$

The acceleration of the comet in the heliocentric frame is thus

$$\ddot{\mathbf{r}}_c - \ddot{\mathbf{r}}_\odot = -\frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}(\mathbf{r}_c - \mathbf{r}_\odot) - \frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}(\mathbf{r}_c - \mathbf{r}_p) - \frac{GM_p}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot). \tag{14}$$

$$CENTRAL \Longrightarrow \qquad PERTURBING \Longrightarrow$$

Since $|\mathbf{r}_c - \mathbf{r}_p| << |\mathbf{r}_p - \mathbf{r}_\odot|$, the last term can be neglected, and we have approximate values of the magnitudes of the other terms:

$$central \quad \tfrac{GM_\odot}{a_p^2}; \qquad perturbing \quad \tfrac{GM_p}{\Delta^2},$$

where Δ is the planet–comet, distance and a_p is the Sun–planet distance . The value of Δ, for which the two accelerations are equal, is

$$\Delta_h = a_p \cdot \left(\frac{M_p}{M_\odot}\right)^{1/2}. \tag{15}$$

The acceleration of the comet in the planetocentric frame is

$$\ddot{\mathbf{r}}_c - \ddot{\mathbf{r}}_p = -\frac{GM_p}{|\mathbf{r}_c - \mathbf{r}_p|^3}(\mathbf{r}_c - \mathbf{r}_p) - \frac{GM_\odot}{|\mathbf{r}_c - \mathbf{r}_\odot|^3}(\mathbf{r}_c - \mathbf{r}_\odot) + \frac{GM_\odot}{|\mathbf{r}_p - \mathbf{r}_\odot|^3}(\mathbf{r}_p - \mathbf{r}_\odot). \tag{16}$$

$$CENTRAL \Longrightarrow \qquad PERTURBING \Longrightarrow$$

Both $|\mathbf{r}_c - \mathbf{r}_\odot|$ and $|\mathbf{r}_p - \mathbf{r}_\odot|$ are $\simeq a_p$, so we get the approximate magnitudes of the accelerations:

$$central \quad \tfrac{GM_p}{\Delta^2}; \qquad perturbing \quad \tfrac{GM_\odot \Delta}{a_p^3}.$$

The value of Δ, for which the two accelerations are equal, is

$$\Delta_p = a_p \cdot \left(\frac{M_p}{M_\odot}\right)^{1/3}. \tag{17}$$

Finally, the ratio of central to perturbing acceleration is for the heliocentric frame:

$$\frac{M_\odot}{M_p} \cdot \left(\frac{\Delta}{a_p}\right)^2,$$

and for the planetocentric frame:

$$\frac{M_p}{M_\odot} \cdot \left(\frac{\Delta}{a_p}\right)^{-3}.$$

Equality of the two ratios occurs for

$$\Delta_{h/p} = a_p \cdot \left(\frac{M_p}{M_\odot} \right)^{2/5} . \tag{18}$$

For instance, for Jupiter the three values are $\Delta_h = 0.16$ AU; $\Delta_p = 0.52$ AU; and $\Delta_{h/p} = 0.33$ AU.

We see that there is a small region around Jupiter ($\Delta < \Delta_h$), where the heliocentric orbit gets unstable. There is a larger region ($\Delta < \Delta_p$), where the jovicentric orbit is reasonably stable. For $\Delta_h < \Delta < \Delta_p$, both orbits may be called stable, though only marginally so. The three zones are shown schematically in Fig. 15. Similar *spheres of influence* may be drawn around any planet, and the radii then scale as shown above.

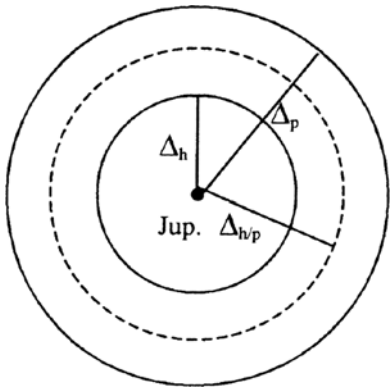

Fig. 15 Rough sketch of the spheres of influence around Jupiter. Δ_h marks the region of instability of the heliocentric Keplerian orbit, Δ_p is the maximum distance where the planetocentric Keplerian orbit is reasonably stable, and $\Delta_{h/p}$ is the distance where both orbits are equally stable

Δ_p represents a stability criterion for planetocentric motion that differs somewhat from that of the Hill sphere (the largest closed zero-velocity surface around the planet in the circular restricted three-body problem). The limiting distance is a factor $3^{1/3} \simeq 1.44$ smaller for the Hill sphere, which is the more stringent of the two criteria, when discussing the stability of satellite motion.

4.2 Hyperbolic Deflections

An approximate way to treat close encounters is to follow the unperturbed heliocentric orbit until $\Delta = \Delta_{h/p}$, and then shift to a hyperbolic planetocentric orbit that is followed until $\Delta = \Delta_{h/p}$ again, and a new heliocentric orbit is computed. This means that the close encounters can be treated using "hyperbolic deflections" in a scattering problem analogous to, e.g., nuclear particle scattering. Now recall the result we derived in Sect. 3.1, namely, $U^2 = 3 - T$ for the square of the encounter velocity. The conservation of T is equivalent to that of U, so that it can be seen to follow from a hyperbolic deflection approach to the close encounters as well as from the Jacobi integral.

When following the hyperbolic deflection, or "matched conic sections", approach, we may assume the interaction to be instantaneous so that Jupiter remains in the same place during the event. We thus see that it is simply a matter of keeping U constant while turning the relative velocity vector by an angle that follows from the impact parameter and geometry of the encounter. Analytic applications of this are common in cometary research and are based on a fundamental theory developed by Ernst Öpik [74, 75].

Before illustrating some results of the Öpik theory, let us note a condition which a cometary orbit must fulfil, if Jupiter shall be able to eject the comet from the Solar System by close encounters, based on the assumption that Jupiter's orbit is circular. In the units usually adopted for the circular restricted three-body problem, the velocity of escape from the Solar System at Jupiter's orbit is $V_E = \sqrt{2}$. The encounter velocity must hence obey $U > \sqrt{2} - 1$, and we get $3 - T > (\sqrt{2} - 1)^2$, which implies $T < 2\sqrt{2} \approx 2.82$.

Let us for simplicity consider only zero-inclination encounters, i.e. encounters where the planet's velocity vector is situated in the plane of the comet's hyperbolic planetocentric orbit. Figure 16 illustrates the deflection of the relative velocity vector by an angle θ, which we may—as an approximation—take to be the full angle between the asymptotes of the hyperbola, even though of course the sphere of influence does not extend to infinity.

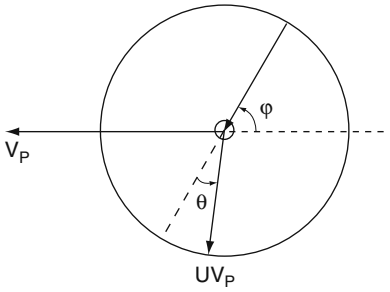

Fig. 16 Schematic diagram showing the incoming and outgoing velocity vectors of a comet that encounters a planet. These velocity vectors are planetocentric, while the vector V_p is the heliocentric orbital velocity of the planet. The incoming velocity of the comet makes the angle φ with V_p, while the outgoing one has been deflected by the angle θ in the direction of φ. Note that θ may be positive or negative, depending on which side of the planet the comet passes on

Introducing V_p as the orbital velocity of the planet, the figure shows how the planetocentric velocity of the comet is a vector of length UV_p, whose initial direction makes an angle φ with that of the planet's orbital motion. After the encounter this angle is changed to $\varphi + \theta$. If the encounter is characterized by an impact parameter b, the absolute value of the deflection angle is given by

$$\tan\frac{|\theta|}{2} = \frac{GM_p}{b(UV_p)^2} \tag{19}$$

if M_p is the mass of the planet. The notations used here for the angles in Fig. 16 differ from those used by Öpik [75] and in subsequent papers, e.g., [94].

Let us investigate the maximum deflection angle achievable with a certain encounter velocity U for a planet of radius R_p and mass M_p. This occurs for a grazing encounter, where the planetocentric hyperbola has a pericentre distance $= R_p$. If the impact parameter for such an orbit is b_g, and the comet's velocity at pericentre is V_g, we get from angular momentum conservation

$$R_p V_g = U V_p b_g \tag{20}$$

and from energy conservation

$$V_g^2 = V_e^2 + U^2 V_p^2, \tag{21}$$

where V_e is the escape velocity from the planetary surface. We introduce $E = V_e/V_p$ as a parameter describing the planet's capability of deflecting orbits, i.e. the ratio of the escape velocity and the orbital velocity of the planet. For the Earth we have $V_e \simeq 11$ km/s and $V_p \simeq 30$ km/s, while for Jupiter $V_e \simeq 60$ km/s and $V_p \simeq 13$ km/s, so we see that $E_J \gg E_E$, Combining (20) and (21), we find

$$b_g = R_p \cdot \sqrt{1 + \frac{E^2}{U^2}} \tag{22}$$

and inserting this into (19), we get

$$\tan \frac{|\theta_g|}{2} = \frac{E^2}{2U\sqrt{U^2 + E^2}}. \tag{23}$$

Figure 17 shows how the maximum deflection angle $|\theta_g|$ varies with the encounter velocity U for the two cases of the Earth and Jupiter. This illustrates a general difference between terrestrial and giant planets. Even for grazing encounters, large deflection angles can only be achieved with very small encounter velocities for terrestrial planets, while such angles are the rule for nearly grazing encounters with giant planets.

Let us now use the deflection angle approach to estimate perturbations of orbital energy during close encounters. When a comet encounters a planet moving in a circular orbit with velocity V_p, its heliocentric velocity is given by

$$V^2 = V_p^2\Big(2 - \frac{a_p}{a}\Big), \tag{24}$$

from which we derive, with the aid of Fig. 16,

$$\Delta\Big(\frac{a_p}{a}\Big) = -\frac{\Delta V^2}{V_p^2} = 2U\Big\{\cos\varphi - \cos(\varphi + \theta)\Big\} = 4U \sin\frac{\theta}{2}\sin\Big(\varphi + \frac{\theta}{2}\Big), \tag{25}$$

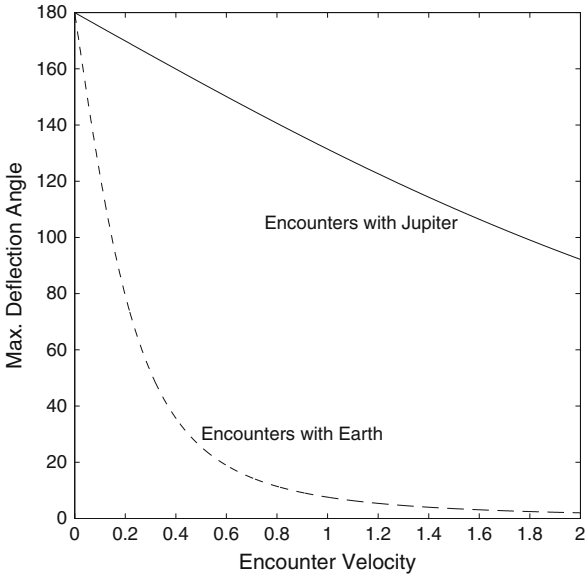

Fig. 17 The maximum attainable deflection angle (in degrees) for grazing encounters with Jupiter (*full-drawn curve*) and the Earth (*dashed curve*), as a function of the approach velocity at infinite distance (in units of the planet's orbital velocity)

and we see that encounters with very small values of U yield small energy perturbations, even though $|\theta|$ may be large. On the other hand, (23) shows that, as $U \to \infty$, $\sin\theta/2 \to 0$ proportional to U^{-2}. Hence $\Delta(a_p/a)$ also approaches zero according to (25). Figure 18 illustrates the behaviour for close encounters with Jupiter, using a constant impact parameter $b = 100\ R_J$.

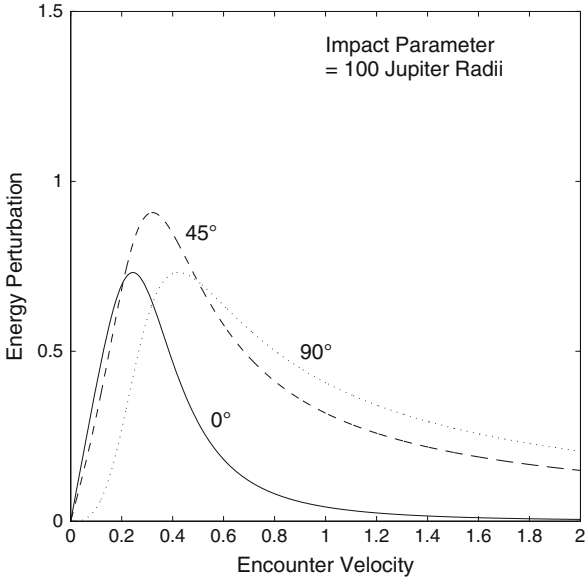

Fig. 18 Perturbations of a_J/a, computed from (19) and (25) for encounters with Jupiter, as functions of U for different approach directions given by the angle φ, as marked at each of the *three curves*. A constant value of $b = 100\ R_J$, typical of close encounters, has been used

Note that for this case the largest energy perturbations are obtained with encounter velocities $U \simeq 0.3 - 0.4$. In fact, this result holds for close encounters in general, as long as they are not extremely close, i.e. nearly grazing. It is interesting to translate this into a range of T using (6), because we thus get $T \simeq 2.8$–2.9 in good agreement with many of the Jupiter Family comets.

4.3 Slow Encounters

One may build a probability distribution of $\Delta(1/a)$ due to close encounters for any particular range of U (and, thus, of T) in the three-dimensional case by purely analytic means [93] using a generalization of (25), where the angles φ and θ are supplemented by an angle ψ that defines the orientation of the planet's orbital velocity vector with respect to the plane of the hyperbolic deflection. One may also choose many random values of U, b, φ, and ψ, and calculate one value of $\Delta(1/a)$ for each combination. These values would define the probability distribution of energy perturbations for random encounters.

Methods of this kind have been widely used in cometary dynamics or related problems (e.g. [33, 34]) and are remarkably accurate in view of the approximations made. In fact, the Öpik formulae give a good representation of the perturbation distribution even for distant encounters, where b is comparable to Δ_p (Fig. 15). However, there still are situations where the outcome of the encounter cannot be accurately estimated in terms of hyperbolic deflection [38], and this is generally the case for very slow encounters by comets with $T \simeq 3$.

Such encounters can be considered in the framework of the zero-velocity surfaces of the restricted three-body problem—in particular those that penetrate into the vicinity of the planet. As explained in celestial mechanics textbooks (e.g. [15]), T is an approximation of the Jacobi constant C of the circular restricted three-body problem, and for $C \gg 3$ the motion of the massless object can occur either far outside the planet, within a large ovoid around the Sun, or within a small ovoid engulfing the planet. Let us now consider the case of Jupiter. When C decreases towards 3, it first reaches the value $C_1 = 3.0388$, when the two ovoids meet at the inner Lagrangian point L_1 situated 0.35 AU from Jupiter, and then the value $C_2 = 3.0375$, when the jovicentric ovoid opens up on the outer side at L_2 (0.36 AU outside Jupiter). This is illustrated in Fig. 19, which shows zero-velocity curves in Jupiter's orbital plane for the two critical values of C. The smaller of the jovicentric ovoids corresponds to what is usually called the *Hill sphere*.

Now consider comets that approach Jupiter from the outside or inside with T values in the vicinity of C_1 and C_2. If Jupiter's orbit were circular and the influences of the other planets were nil, the zero-velocity surface would remain unchanged. Thus, if there was a large gap at the Lagrangian point in question, this gap would remain constant, and if the comet would enter into the ovoid, it could escape as easily. If there was a gap at the opposite Lagrangian point, exit could also occur there, and the comet could transit between orbits inside and outside Jupiter's orbit. These

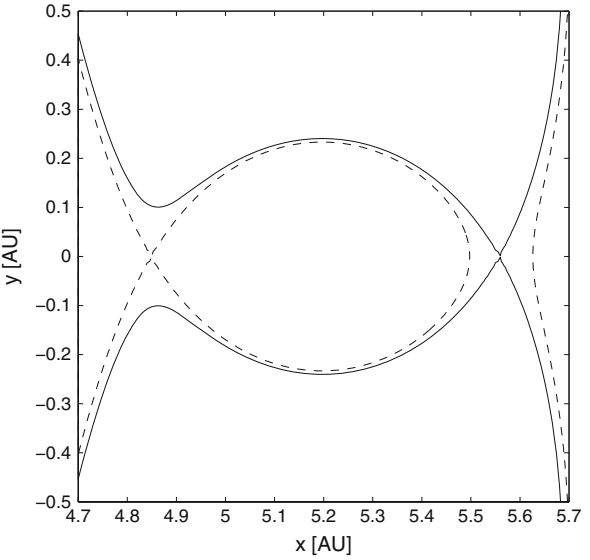

Fig. 19 Zero-velocity curves in the region around Jupiter have been plotted in Jupiter's orbital plane for two values of the Jacobi constant, corresponding to the Lagrangian points L_1 (*dashed*) and L_2 (*full-drawn*). Distances are expressed in AU, the Sun is at the origin, and Jupiter has been placed at $x = 5.2$

orbits would in fact be nearly tangent to Jupiter's orbit at aphelion and perihelion, respectively.

With a non-circular jovian orbit and non-negligible perturbations by other planets (primarily Saturn), the ovoid that applies at any moment is no longer constant but may close and open up in phase with Jupiter's changing heliocentric distance and the changing relative position of Saturn. Thus comets in the relevant range of T may easily experience *temporary satellite captures* (TSCs) around Jupiter, when they stay for extended periods of time in the ovoid region, orbiting around Jupiter with some degree of gravitational binding. Such captures in the near past or future are known for several Jupiter Family comets, and an example is shown in Fig. 20. The relationship between TSCs and nearly tangent encounters with Jupiter was first explored and explained by Carusi and Valsecchi [10].

Another example worth noting is comet D/Shoemaker-Levy 9, which was experiencing a long-lasting TSC [14] during which the solar perturbations imposed a fast oscillation between orbits of low eccentricity and high inclination on the one hand and high eccentricity and low inclination on the other. It was the decrease of the perijove distance caused by this Kozai cycle that eventually brought the comet into tidal break-up and collision with Jupiter.

When comets exit from Jupiter's vicinity after TSCs or other slow encounters, they enter into a special range of heliocentric orbits, as explained by Tancredi et al. [89]. They tend to get new aphelion distances close to that of Jupiter's L_1 point ($Q \simeq 4.7$ AU) and orbital periods close to 2/3 that of Jupiter. Since this resonance is also shared by the Hilda group of asteroids, the comets in question are often called *quasi-Hildas*. They are profoundly distinct from their asteroidal counterparts in that no protection from close encounters with Jupiter exists. Thus they are just temporary visitors, but the slight concentration seen in Fig. 4 at $q \simeq 3.5$ AU and $Q \simeq 4.7$ AU

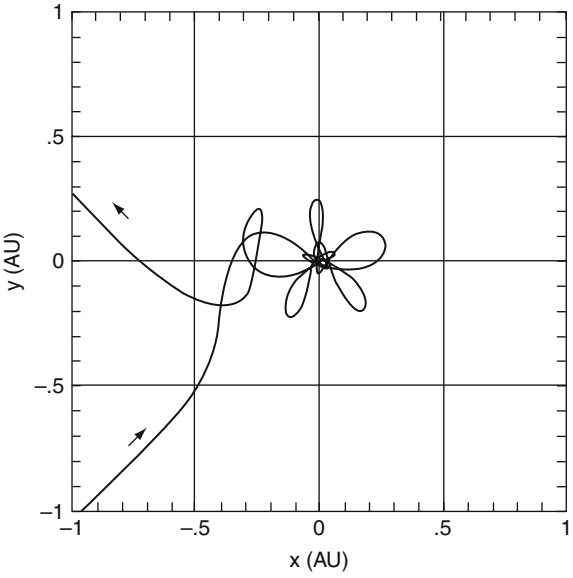

Fig. 20 Trajectory of comet 111P/Helin-Roman-Crockett with respect to Jupiter, plotted at the origin, during a temporary satellite capture predicted to occur around 2075. The frame is rotating so that the Sun is always on the negative x-axis. Both entry and exit occur in the general vicinity of the L_1 point. From [89]

is likely a real feature caused by the particular dynamics of slow encounters with Jupiter.

This behaviour of the quasi-Hilda comets illustrates a more general phenomenon that applies to all comets in aphelion- or perihelion-tangent orbits with respect to Jupiter, i.e. a strong asymmetry of the probability distribution of energy perturbations upon close encounters. From Fig. 16 it is seen that the heliocentric velocity of the comet just before the encounter is $V_c = V_p(1 + U\cos\varphi)$. Suppose that the comet is perihelion-tangent. Then we have $\varphi \simeq 0$ and $\cos\varphi \simeq 1$, and no matter how the jovicentric velocity is deflected, the orbital energy (as expressed by V_c^2) cannot increase substantially. On the other hand, it may decrease by a large amount in case U and $|\theta|$ are large, and the post-encounter orbit may even be aphelion-tangent with $\cos(\varphi + \theta) \simeq -1$. The opposite holds if the initial orbit is aphelion-tangent. Then the orbital energy cannot decrease substantially but may increase all the way to a perihelion-tangent orbit.

Another way to describe the same phenomenon is to look at Fig. 4 and realize that close encounters are restricted to crossing orbits. Therefore, if a comet encountering Jupiter is situated close to one of the dashed lines (perihelion- or aphelion-tangent orbits), it is near an extreme of the allowed portion of its evolutionary curve (with constant T or U), and thus it can only evolve in one direction.

Let us end this section by mentioning two examples of comets that have transited between perihelion- and aphelion-tangent orbits and illustrate what has just been said. Comet 39P/Oterma illustrates the quasi-Hilda behaviour with $T \simeq 3$. It was captured as a result of a slow, long-lasting encounter with Jupiter with closest approach in 1937 from an outer orbit with $q \simeq 5.8$ AU and $Q \simeq 8$ AU into an inner orbit with $q \simeq 3.4$ AU and $Q \simeq 4.5$ AU, and it was discovered in 1943. After three

orbits a new slow encounter with closest approach in 1963 transferred the comet back into an outer orbit with $q \simeq 5.5$ AU and $Q \simeq 9$ AU. It was rediscovered in this distant orbit in 2001.

The second example involves encounters at much higher velocity and is of great historical interest [92]. Comet D/1770 L1 (Lexell) was captured in 1767 at a close encounter with Jupiter that changed the orbit from $q \simeq 2.9$ AU and $Q \simeq 5.9$ AU into $q = 0.67$ AU and $Q \simeq 5.6$ AU. The Tisserand parameter is $T \simeq 2.6$. Since the post-capture orbit was in 2/1 resonance with Jupiter, a new close encounter took place in 1779 and led to ejection into an orbit with $q \simeq 5.2$ AU and $Q \sim 80$ AU [4].

5 Long-Term Orbital Evolution

The long-term evolution of cometary orbits can usually be described in terms of Kozai cycles or random walks in energy or angular momentum. Sometimes the combination of these regular and stochastic phenomena—when they take place simultaneously—has an important influence on the results. In this section we give an introduction to all of these topics, explaining the background of the dynamical processes and how they work, in particular for long-period comets. We also discuss the limitations of the simple concepts (Kozai cycles and random walks) and describe some circumstances where they do not apply in a strict sense.

5.1 Random Walks

5.1.1 Orbital Energy

Let us first consider the near-parabolic cometary orbits discussed in Sect. 3.3. Their distribution of $\Delta(1/a)$ due to planetary perturbations was presented in Fig. 12. This corresponds to a sample of comets with a nearly isotropic distribution of orbital planes, where the giant planets had random orientations with respect to the comets, as the latter passed perihelion. In the simulation represented in Fig. 11b it was used as an approximation of the parent distribution, from which the sequences of $\Delta(1/a)$ values experienced by the test comets were drawn at random. But is such a random walk model relevant? Does it approximate the sequences experienced by the real comets?

The answer is that it does, if an insignificant change in $\Delta(1/a)$ at one apparition may lead to a completely different sample from the parent distribution at the next apparition. Using this criterion, we can take the half-width of the bins in Fig. 12 (0.0001 AU^{-1}) as a finite but rather insignificant difference of $\Delta(1/a)$, since we are sampling a minor fraction of the distribution. Now, if the perturbation leads to an orbit with period $P \simeq 10{,}000$ yr ($a \simeq 500$ AU), the difference of P will be $\Delta P \simeq$ 500–1000 yr, and the positions of the giant planets at the next perihelion passage are completely undetermined. The criterion for a random walk is thus fulfilled. If the new period is $P \simeq 1000$ yr ($a \simeq 100$ AU), we instead get $\Delta P \simeq 15$ yr,

and we may still consider the criterion as marginally fulfilled, since Jupiter is the dominating planet with a period $P_J \lesssim \Delta P$.

We conclude that the random walk scenario is relevant as long as the comets stay in the interval $0 < 1/a \lesssim 0.01$ AU^{-1}. Brasser et al. [8] derived an even larger value for $1/a_c$ as the limit of energy diffusion in a somewhat different way (for Jupiter it was ~ 0.04 AU^{-1}), but considering the arbitrary choice of the bin width in Fig. 12 that we applied in the above estimate, we should consider the two analyses to be in reasonable agreement. There are several ways to model the random walk of comets numerically, and the Monte Carlo simulation scheme used in Sect. 3.3 is one. Another method is a Markov chain [82], where one divides the range of orbital energy into bins of $1/a$. From a numerical representation of the parent distribution $f(\Delta 1/a)$ obtained by integrating a large number of fictitious comets, one derives a matrix of jump probabilities per unit time between the different bins. The orbital distribution can be represented by a state vector containing the number of comets in each bin, and this vector is evolved by multiplication with the jump matrix and addition of a source vector, representing injection of new comets from the Oort Cloud.

More detailed information on the perturbation distributions are given in Fig. 21, which uses samples of 100,000 accurate integrations of comets passing through a planetary system made up of the four giant planets on realistic orbits. The initial semi-major axis is always $a_o = 25{,}000$ AU, and the distributions of $\cos i$, ω, and Ω are uniform. The time of perihelion passage with respect to the planetary phases is also random. One sample has q uniformly distributed between 5 and 6 AU, and and the other sample has $0 < q < 1$ AU. The four panels in Fig. 21 show distributions of $\Delta(1/a)$ for two subsamples of each q range: one with low-inclination prograde orbits, where T is close to its maximum possible value, and one for strongly retrograde orbits, where T is close to its minimum. All cases of close approach to Saturn, Uranus, or Neptune have been excluded, while all close encounters with Jupiter are kept. We can thus consider the distributions to approximate those of jovian perturbations.

By comparing the panels we find the following main features. The far tails are most prominent in panel (a) for low-inclination, nearly Jupiter-tangent orbits, and they are nearly non-existent in panel (b), where the comets are strongly retrograde. The orbits with small perihelion distances are intermediate, and we can see that one needs very large samples of perturbations in order to get a good statistical representation of such tails. The central peaks have quite different widths as well, and interestingly, these are largest for small perihelion distances. Thus, the part of the random walk that is caused by perturbations of normal size is quicker for such orbits than for orbits with larger q. In particular, the large-q, retrograde orbits have an extremely narrow central peak in addition to the almost complete lack of far tails.

The differences of the tails are easy to understand using the above results on close encounters. The comets of panel (a) have very small encounter velocities, which makes such encounters efficient in changing the orbital energy. Those of panels (c) and (d) are intermediate, and those of panel (b) have by far the highest velocities, making them very difficult to perturb. Let us now analyse the central

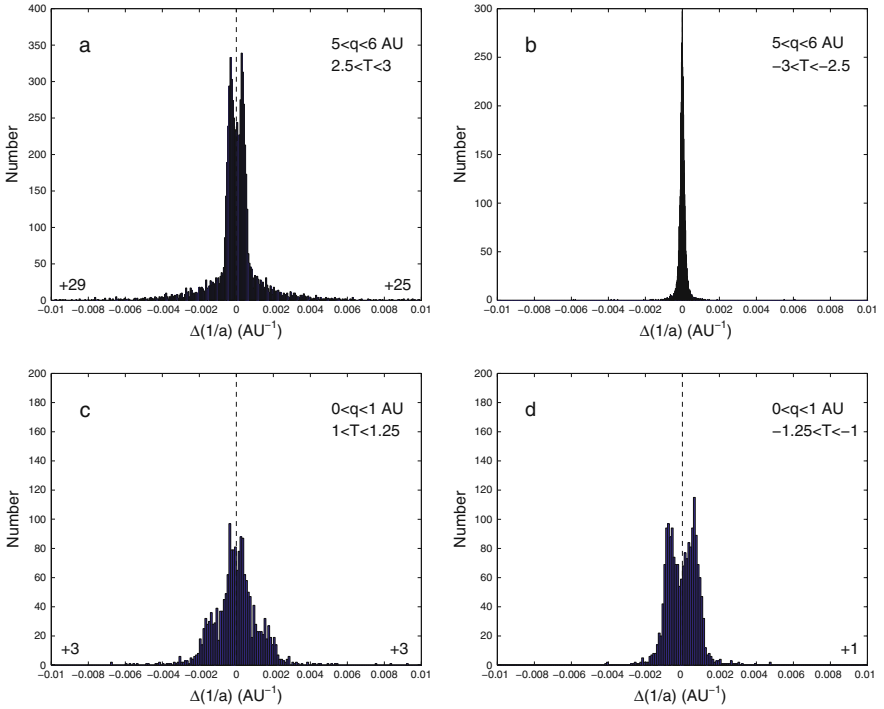

Fig. 21 Histograms of the perturbations of inverse semi-major axis for large samples of randomly distributed orbits typical of new Oort Cloud comets. Panels (**a**)–(**d**) are for subsamples with different perihelion distances and Tisserand parameters as indicated in each plot. Whenever relevant, the numbers of outliers (perturbations larger than the plotted range) are printed at the *bottom left* and *right*

peaks in some more detail. A double-peaked structure is evident in panels (a) and (d) and might possibly exist in panel (c). In general, the shapes of the central peaks are not quasi-Gaussian but much more intricate. They can be understood by looking into the geometrical circumstances of the cometary passages with respect to Jupiter.

Figure 22 illustrates a simplified picture, where Jupiter is made to move on a circular orbit and the comet follows an unperturbed parabola. The latter assumption would break down if a close encounter would occur, but we will now exclude such cases. For simplicity we will only consider the co-planar case, so that the cometary inclination is either 0° or 180°. It is evident that the perturbations experienced by the comet in this case only depend on one parameter, i.e. the angle ϕ describing Jupiter's position at the time of perihelion passage of the comet. Taking a particular value of ϕ, one can integrate (10) over the time period when the comet passes the orbital arc inside a given distance (to be taken as 11 AU). Both terms within the curly brackets can be integrated separately, thus giving the direct and indirect parts of the energy perturbation as well as the total value. This calculation is done in the same spirit as the Keplerian estimator of perturbations by Rickman and Froeschlé [83].

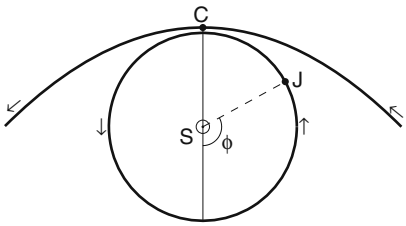

Fig. 22 Simplified sketch of the geometry of a cometary passage through the planetary system along a parabolic orbit. A circular planetary orbit—supposed to represent Jupiter—is shown, and the positions of Jupiter and the comet at the latter's perihelion are marked. Jupiter's position can be specified by the angle ϕ

In Fig. 23 we plot the results of such calculations for four cometary orbits typical of the samples used in the four panels of Fig. 21. Consider first the indirect perturbations shown by dotted curves. These always follow a sinusoidal variation with ϕ, and the amplitude is much higher for prograde orbits than for retrograde ones (note the different scales of the plots). The largest negative perturbations for prograde orbits occur around $\phi \simeq 90^\circ$, when both the x and y components of the cometary velocity tend to be directed opposite to the radius vector of Jupiter. For retrograde orbits, however, the x and y components of the scalar product are of opposite signs, and depending on the perihelion distance the variation can be of either equal or opposite phase, as shown by panels (b) and (d).

The direct perturbations are always small in case the comet stays far from close approach for the entire orbit, but they grow considerably in the proximity of encounters. The intervals that have been excluded due to close encounters correspond to cases when the unperturbed comet–Jupiter distance decreases below 1 AU for prograde orbits and 0.25 AU for retrograde orbits. As a general pattern, the perturbation is positive—corresponding to a net deceleration of the comet—for small values of ϕ and negative for the largest values of ϕ. For an external orbit such as in panel (a) the close encounters occur around $\phi = 180^\circ$ with large perturbations showing the start of the far positive tail to the left and opposite ones corresponding to the negative far tail to the right. For a crossing orbit such as in panels (c) and (d) there are two intervals of close encounters, one on either side of $\phi = 180^\circ$, and between them there is a region where the perturbation grows from large negative to large positive values.

Note that the frequency function of $\Delta(1/a)$ would be obtained by adding up the contributions from all relevant values of ϕ, and thus it increases for values where the derivative $d\Delta(1/a)/d\phi$ is small. The double peaks in Fig. 21 are thus the result of the fact that in certain intervals of ϕ the combination of direct and indirect perturbations leads to a slow variation of $\Delta(1/a)$. The lack of a double peak in Fig. 21b can be explained by noting that almost all the orbits have even smaller perturbations than the one shown in Fig. 23b, and thus any central dip must be very narrow and probably washed out by effects of the eccentricity of Jupiter's orbit. The complicated appearance of the central peak for prograde orbits

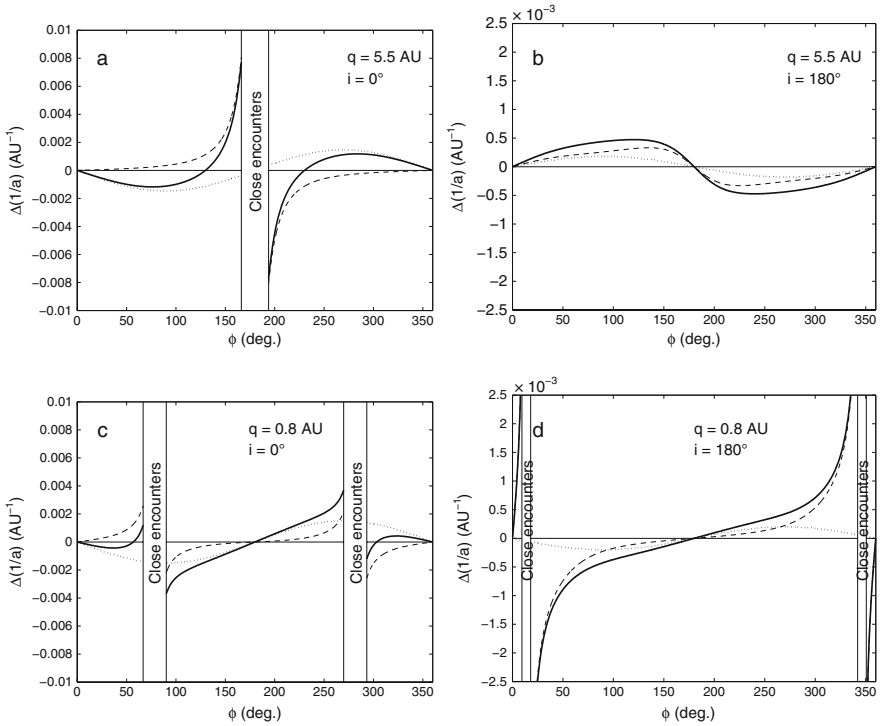

Fig. 23 Plots of direct and indirect perturbations of $1/a$ due to Jupiter on a circular orbit for comets passing perihelion on a parabolic orbit, as functions of the angle ϕ defined in Fig. 21. Panels (**a**)–(**d**) are for different orbital elements of the comet, as indicated in each plot. The direct perturbations are shown by *dashed curves* and the indirect perturbations by *dotted curves*. The *thick solid curves* show the total perturbations. Intervals of ϕ that have been excluded due to close encounters are indicated

with small q is very well explained by Fig. 23c, where it is seen that the interval with $|\Delta(1/a)| \lesssim 0.0005$–$0.001$ AU^{-1} must be highly populated, followed by the one directly on the outside extending to $|\Delta(1/a)| \simeq 0.002$ AU^{-1}. For even larger values we are in the close encounter dominated regime, where $\Delta(1/a)$ varies rapidly with ϕ.

We conclude this discussion by noting that Rickman et al. [85], after finding that the orbits with the smallest perihelion distances stand a relatively high chance of being "captured" into $a < 1000$ AU on their first perihelion passage, suggested that this had to do with the relatively large indirect perturbations experienced thanks to the large velocity by which the comets pass their perihelia. We now find the same phenomenon but a slightly different explanation. The effect comes not from the size of the indirect perturbations but from the way they cooperate with direct perturbations in the ϕ range between the close encounters, making the total perturbation grow at relatively small slope with respect to ϕ all the way to ~ 0.002 AU^{-1}.

5.1.2 Angular Momentum

In Sect. 3.1 we demonstrated that the Tisserand parameter can be expressed as

$$T = 2L_z - E, \tag{26}$$

where $E = -1/a$ is the orbital energy and L_z is the angular momentum component perpendicular to the planetary orbit (4). Hence, since T is a quasi-conserved quantity, the random walk in E discussed in the previous section is directly coupled to a random walk in L_z. Independent of this, a random sequence of planetary perturbations must cause a random walk of the entire angular momentum vector **L**. However, if we consider long-period comets, the relation

$$L = \sqrt{GM_\odot a(1-e^2)} = \sqrt{GM_\odot q(1+e)} \tag{27}$$

shows that the planetary perturbations—requiring $q \lesssim a_p$—effectively limit L to its smallest possible values, since $q \ll a$, and the allowed range of L extends to $L_{\max} = \sqrt{GM_\odot a}$.

The fundamental reason why planets are efficient perturbers of the orbital energy of long-period comets but inefficient in changing their angular momenta is that they act when the comets are near their perihelia. Perturbations applied near the aphelia would behave in the opposite way—the orbital energies would be relatively insensitive, while the angular momenta might be strongly affected. The latter is often the case for the stellar perturbations mentioned in Sect. 3.3, to which we will now turn our attention.

Figure 24 shows the inverse frequency of stellar encounters within a distance R of the Sun as a function of R. This is an approximate relation based on a given number density n_* in the solar neighbourhood and a given mean encounter velocity $\langle v_* \rangle$ for these stars. The statistics is dominated by red dwarfs, which are low-mass stars

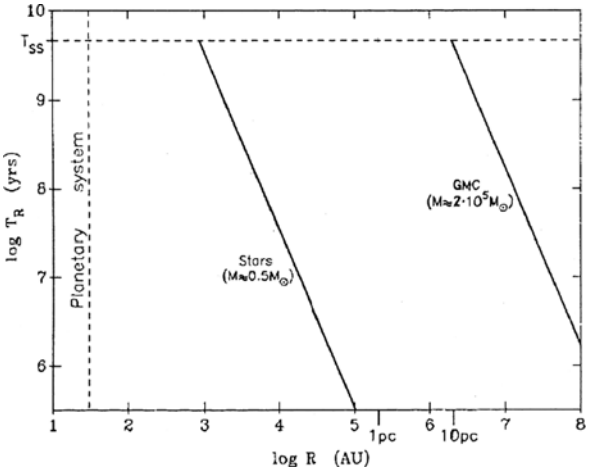

Fig. 24 The typical time interval between consecutive encounters between the Sun and Galactic objects as a function of the miss distance. Both quantities are plotted on log scales. The two lines show the relations for the field stars of the current solar neighbourhood and for Giant Molecular Cloud complexes, respectively

passing, typically, at high velocities. This means that they are relatively inefficient perturbers of cometary orbits, and early-type main sequence stars with high masses and low velocities contribute significantly to the overall effect in spite of their low number density [80]. A full discussion of the stellar perturbations must therefore account for the different contributions of different stellar types rather than lumping all the stars together as in Fig. 24.

Another possible shortcoming concerns the very close encounters with intervals of $\gtrsim 1$ Gyr. Obviously, the frequency of such encounters should correspond to the time average of n_* and $\langle v_* \rangle$ over the relevant intervals, but these may be quite different from the current values used in Fig. 24. For instance, the Sun has probably spent most of the time at larger distances from the Galactic midplane with lower values of n_* and $\langle v_* \rangle$. Moreover the infant Sun may have had an orbit closer to the Galactic centre than it currently has, and it may even have been born in a stellar cluster with very large n_* and small $\langle v_* \rangle$ [8].

When a star passes at relatively close range, each comet in the Oort Cloud suffers an externally induced chaotic perturbation. To evaluate these perturbations, in most cases the method used has been the *Classical Impulse Approximation* (CIA), which considers the comet at rest with respect to the Sun, while the star passes at a high velocity (the near-aphelion velocities of Oort Cloud comets are indeed much smaller than typical stellar encounter velocities). In addition, the heliocentric orbit of the star, which really is a high-eccentricity hyperbola ($e_* \gg 1$), is approximated as a straight line ($e_* \to \infty$) with constant speed. Impulses are thus imparted to the Sun and the comet (Fig. 25), and the heliocentric impulse of the comet is

$$\Delta \mathbf{v} = \frac{2GM_*}{v_*} \left\{ \frac{\hat{\mathbf{d}}_c}{d_c} - \frac{\hat{\mathbf{d}}_\odot}{d_\odot} \right\} \tag{28}$$

[76], showing the importance of the stellar mass to velocity ratio for the strength of the perturbation. In principle it is easy to understand the limitations of the CIA, and

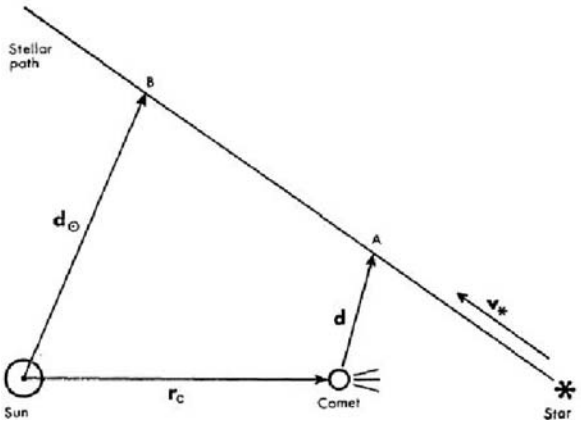

Fig. 25 The geometry of a stellar passage, as used to develop the Classical Impulse Approximation. $\hat{\mathbf{d}}_c$ and $\hat{\mathbf{d}}_\odot$ are unit vectors pointing from the comet and the Sun, respectively, towards the points of closest approach of the star. From [76]

the problems mainly concern stars with slow and/or close encounters. The straight-line motion of the star may then become a bad approximation, and neglecting the motion of the comet during the stellar encounter may not be warranted. It is true that the cases in point are rare, but they represent the largest effects, and thus it is important to try to model them with a good accuracy.

It has been common in recent years to resort to numerical integration of the equations of motion, but of course there is a limitation to this approach too in terms of CPU time, when very extensive simulations are performed. However, other analytical approximations with better performance than the CIA have been developed and found useful. Dybczyński [21] introduced an improved version of the CIA, where the impulses received by the comet and the Sun are calculated from the hyperbolic deflections of their asterocentric motions. Eggers and Woolfson [23] were the first to introduce a sequential treatment of the stellar passages, where separate impulses were computed using the CIA for finite steps along the stellar path. A combination of both these improvements, called the Sequential Impulse Approximation, was developed by Rickman et al. [81] and was found to give results of high accuracy in almost all cases while saving a large majority of the CPU time expended in accurate numerical integrations.

When the stellar encounter is relatively distant, as illustrated in Fig. 26, (28) can be recast in approximate form as

$$\Delta\mathbf{v} \approx \frac{2GM_* r_c \sin\alpha}{v_* d_\odot^2} \cdot \hat{\mathbf{d}}_\odot. \tag{29}$$

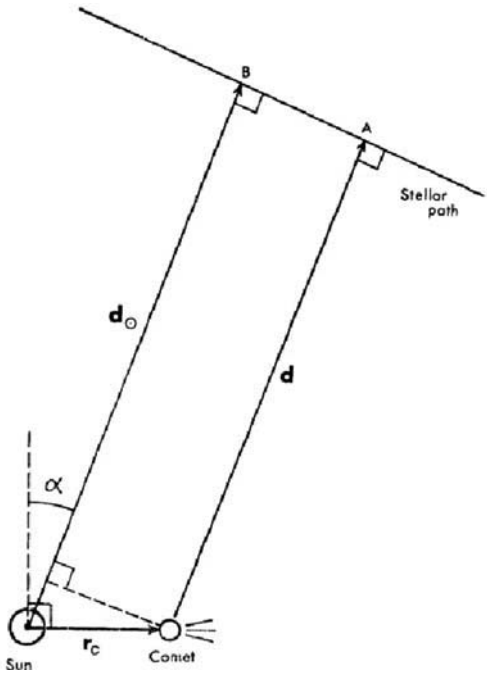

Fig. 26 The geometry of a distant stellar passage, as used to develop the approximate equation (29). In this case the differential impulses on the comet and the Sun correspond to a tidal perturbing effect. From [76]

Within this approximation, using the fact that the angular momentum perturbation is $\Delta\mathbf{L} = \mathbf{r}_c \cdot \Delta\mathbf{v}$, we see that its absolute value $|\Delta L|$ is proportional to r_c^2. Thus, statistically, over a given period of time, when a large number of stars pass with different geometries and at different distances, we get the relationship $|\Delta L| \propto a^2$ in terms of the semi-major axis a of the cometary orbit. Since the perturbation of the perihelion distance is $\Delta q \propto \sqrt{q}\,\Delta L$, we get the statistical relationship

$$\mathcal{E}(|\Delta q|) \propto \sqrt{q} \cdot a^2 \tag{30}$$

for the expectation of the variation of perihelion distance during the interval in question. Figure 27 shows that such a relationship is a good approximation at least for the small to normal size range of star-induced Δq per passage, while the approximation is a bit worse for the largest perturbations, where the encounter distances are too small for the "tidal" formula of (29) to work. In such cases, as mentioned by Rickman et al. [81], one should rather expect $|\Delta q| \propto a$ according to (28).

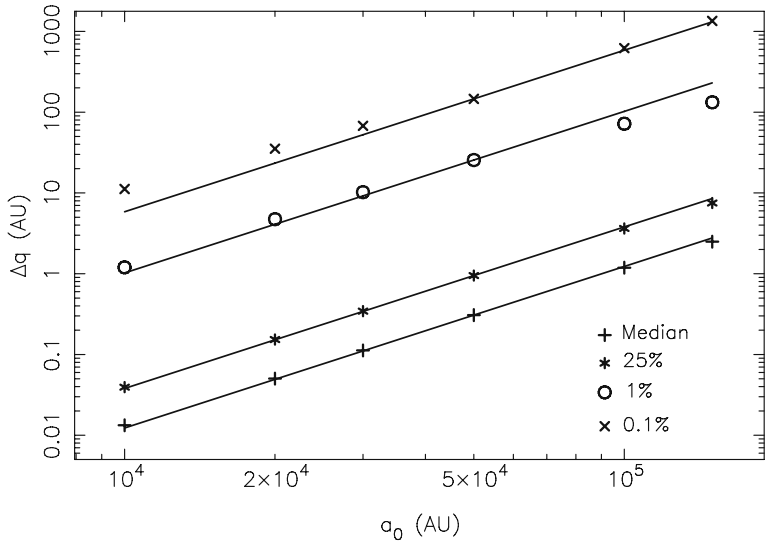

Fig. 27 Median, upper quartile, 99th percentile, and 99.9th percentile of the distribution of $|\Delta q|$ for samples of nearly 16 million Sun–comet–star interactions, computed with the Sequential Impulse Approximation. All the comet orbits have $q = 100$ AU, and the results for five different values of a are shown by symbols. The *straight lines* are fits with a slope of 2. From [81]

Note that about one perturbation in 1000 has a value 1000 times larger than the median. Consider a random walk picture, where the total $|\Delta q|$ experienced is the result of a large number N of random perturbations, and consider such results based on median-size perturbations ($|\Delta q|_M$) and on the very large perturbations ($|\Delta q|_L$) separately. Those results will be approximately the respective individual perturbation size times the square root of the number of perturbations (N_M and N_L, respectively). We thus find that the largest long-term effect will be given by the

largest perturbations, once again emphasizing the need to treat these perturbations accurately.

If one considers just one orbital revolution by the comet, it appears justified to use the a^2 dependence and multiply the result by the orbital period $P \propto a^{3/2}$, so one gets the expected change in q per orbital period

$$|\Delta q|_1 \propto \sqrt{q} \cdot a^{7/2}. \tag{31}$$

Now, return to the phenomenon found in Sect. 3.3 (Fig. 10), i.e. that the newcomers from the Oort Cloud that form the spike in the distribution of original inverse semi-major axes ($1/a_{\rm ori}$) are removed from this range by planetary perturbations. In fact, the same thing would happen even if the perihelion distances were as large as 10–15 AU [26]. The phenomenon was known already to Oort in 1950, and he introduced the concept of *the loss cylinder*, which means a narrow cylinder along the radial direction in velocity space, corresponding to the smallest perihelion distances and angular momenta. This is expected to be efficiently cleared by planetary perturbations, and the existence of the observed newcomers with very small perihelion distances indicates that there is a mechanism to inject comets deep inside the loss cylinder on a time-scale of one orbital revolution.

Oort realized that stellar perturbations can do this, but the very steep relation with a in (31) means that only the largest values of a can be considered realistic. It is therefore no surprise that the spike is situated in close vicinity of the parabolic limit. However, imagine that the Oort Cloud also contains many comets with smaller semi-major axes, say, with $a \lesssim 10{,}000$ AU where usually there is no way to inject them into observable orbits. This hypothetical population has been termed the *inner core* of the Oort Cloud. It is easy to see that the inner core may be activated from time to time as a consequence of very close stellar encounters, which temporarily invalidate (31) by encountering many comets in the inner core and bringing them into the depths of the loss cylinder. This phenomenon was first investigated by Hills [42] and is referred to as *comet showers*. From Fig. 24 we see that one may expect such temporary enhancements of the flux of newcomers on time-scales of $\sim 10^8$ yr.

5.2 Kozai Cycles

Let us now consider the secular perturbations in cometary dynamics. They are not very important in the case of low-inclination, Jupiter-crossing orbits like those of the Jupiter Family, where the dynamics is dominated by close encounters. We rather have to concentrate on orbits of large inclinations and moderate to high eccentricities like those of the Halley-type and long-period comets. The secular dynamics in such cases differs from that of main belt asteroids, where the slow precession of the apsidal and nodal lines induces oscillations of eccentricity and inclination, respectively, and resonances with the orbital precession rates of Jupiter and Saturn may bring the objects into Earth-crossing orbits. In the cometary case the orbit is

Jupiter-crossing to begin with, and the chance of close encounters—like the general perturbations—depends on the argument of perihelion (ω).

Jupiter's orbital eccentricity and the orientation of its apsidal line do not matter very much in this case, and when short-period terms are averaged out, the cometary orbit essentially feels the attraction of a circular ring of mass around Jupiter's orbit. This perturbing force does not change the orbital energy, and since it is confined by circular symmetry to the meridional plane spanned by the normal vector to the planet's orbital plane and the comet's heliocentric radius vector, the associated torque does not change the comet's angular momentum component perpendicular to the planetary orbit. Thus the quantity $L_z = \sqrt{a(1-e^2)}\cos i$ in (3) is conserved, as is a, and the secular perturbations are limited to a coupled periodic oscillation of e and i, resulting from the variation of ω. Due to the conservation of L_z, the maximum eccentricity occurs at the time of minimum inclination and vice versa.

Two kinds of motion are possible. Either ω circulates or it librates around $\pm\pi/2$. If the latter occurs, it means that there is a 1:1 resonance between the precessional periods of the nodal and apsidal lines. This is commonly referred to as the *Kozai resonance*, since it was first investigated by Yoshihide Kozai [48],[3] and the associated—often substantial—variations of e and i are called *Kozai cycles*.

5.2.1 Galactic Tides

A special case of Kozai cycles—though an extremely important one—occurs in the Oort Cloud due to the fact that the Solar System is immersed in the Galactic disk, and the disk potential—varying smoothly with distance from the midplane—causes a tidal force on Oort Cloud comets in the direction perpendicular to this plane. Figure 28 illustrates both this disk tide and the radial tide (with respect to the Galactic centre) that arises from the central force field in the plane.

The strength of the tide can be expressed in terms of the local density of the disk (ρ_o) and the kinematic parameters of Galactic differential rotation—the so-called Oort constants A and B. For example, using cartesian coordinates (x', y', z) centred on the Sun such that the unit vectors $\hat{\mathbf{x}}'$ and $\hat{\mathbf{y}}'$ point towards the Galactic anticentre and transversely along the local circular velocity in the Galactic plane, and $\hat{\mathbf{z}}$ is perpendicular to this plane, the equation of motion can be written as

$$\ddot{\mathbf{r}} = -\nabla U_\odot + (A-B)(3A+B)x'\hat{\mathbf{x}}' - (A-B)^2 y'\hat{\mathbf{y}}' - [4\pi G\rho_o - 2(B^2-A^2)]z\hat{\mathbf{z}}. \quad (32)$$

With modern estimates of $A = +13$ km/s/kpc and $B = -13$ km/s/kpc [39] and $\rho_o = 0.1\, M_\odot\ \mathrm{pc}^{-3}$ [44] we realize that (1) only the term involving ρ_o is nonvanishing in the z component; (2) this term is almost ten times larger than the coefficients of the x' and y' components. Thus the disk tide is much stronger than the radial tide.

[3] It has recently been recognized that this resonance was first studied in Russia by M.L. Lidov, whose first paper in English [61] appeared almost simultaneously with that of Kozai. It thus seems more correct to refer to the *Lidov–Kozai resonance* or *Lidov–Kozai cycles*.

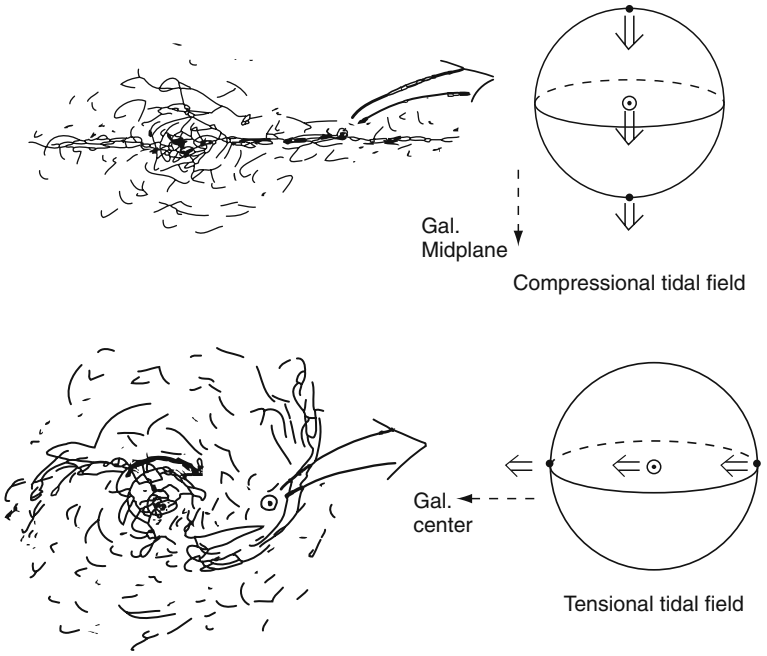

Fig. 28 Illustration of the tidal action of the Galaxy on comets in the Oort Cloud. The *upper panel* shows the disk tide due to the attraction of the Galactic disk, and the *lower panel* shows the radial tide due to the gravity of the Galactic bulge and inner parts of the disk

By neglecting the latter, one may thus get a first approximation to the long-term dynamical behaviour of Oort Cloud comets by using the equation

$$\ddot{\mathbf{r}} = -\nabla\left\{-\frac{GM_{\odot}}{r} + 2\pi G\rho_o z^2\right\} \tag{33}$$

[41]. In this paper Heisler and Tremaine developed a theory for the motion by using orbital averaging of the corresponding Hamiltonian, thus eliminating its dependence on the mean anomaly and securing an energy integral. Furthermore, due to the conservation of the angular momentum component L_z perpendicular to the Galactic plane that follows from the absence of any in-plane perturbing force component, they found the problem to be integrable with only one degree of freedom, where the angular variable can be taken as the Galactic argument of perihelion ω_G. The orbital evolution was found to follow patterns like the ones illustrated in Fig. 29, so that the eccentricity e and the Galactic inclination i_G vary in phase with ω_G.

The curves in each panel correspond to different values of the energy constant

$$C = 1 - e^2 + 5e^2 \sin^2 i_G \sin^2 \omega_G \tag{34}$$

and a common value of the perpendicular angular momentum constant

$$L_z = \sqrt{1 - e^2}\cos i_G. \tag{35}$$

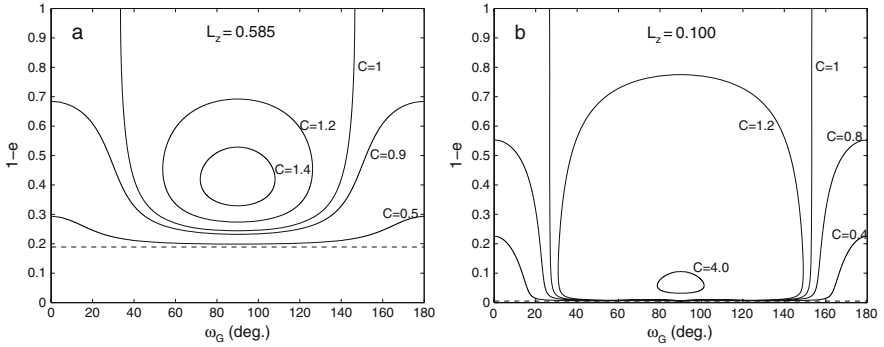

Fig. 29 Variations of eccentricity due to libration or circulation of the Galactic argument of perihelion due to the Galactic disk tide. Panels (**a**) and (**b**) show two examples for different values of the perpendicular component of the angular momentum, and the curves show projections of the phase space trajectories for different values of the Hamiltonian. The dashed lines correspond to $L \equiv L_{\min} = L_z$

The separatrix between ω_G libration and circulation corresponds to $C = 1$. An interesting relation can be derived for the time rate of change of the perihelion distance as a function of the semi-major axis a and the Galactic latitude of perihelion β_G. Analogous to the discussion of stellar perturbations in Sect. 5.1, we have

$$\frac{dq}{dt} \propto \sqrt{q}\frac{dL}{dt} \tag{36}$$

and from the expression for the averaged Hamiltonian [41] we obtain a maximum rate of change of the angular momentum:

$$\left|\frac{dL}{dt}\right| \simeq 5\pi G\rho_o a^2 |\sin 2\beta_G|. \tag{37}$$

We thus find that the perturbation of the perihelion distance during one orbital revolution is generally expressible as

$$|\Delta q|_1 \propto \sqrt{q} \cdot a^{7/2} \cdot |\sin 2\beta_G|, \tag{38}$$

analogous to the corresponding equation for stellar perturbations (31). In order to inject comets directly into observable orbits from outside the loss cylinder (Sect. 5.1), i.e. $|\Delta q|_1 \gtrsim 10$ AU, one generally needs to have $a \gtrsim 3 \cdot 10^4$ AU.

Comparing the relative magnitudes of the stellar and Galactic perturbations of q, Duncan et al. [20] found the latter to be larger, although not by a very large margin. Hence one would expect there to be a Galactic signature in the latitudes of perihelia (β_G) of new Oort Cloud comets due to the presence of the factor $|\sin 2\beta_G|$. Comets with $\beta_G \simeq \pm\pi/4$ should most easily be perturbed into the depths of the loss cylinder and therefore be dominating the statistics of β_G for new comets. Delsemme [16]

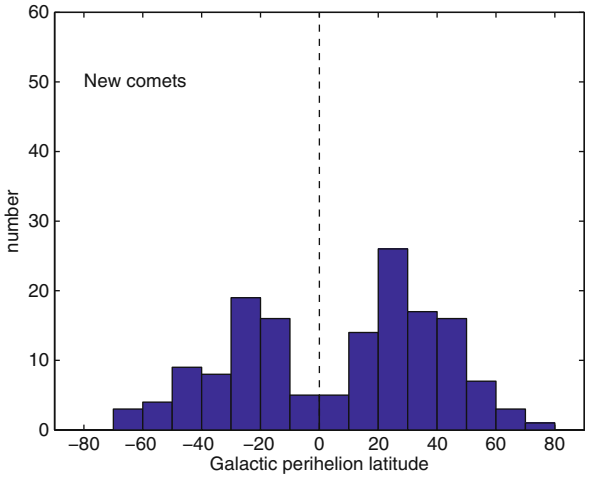

Fig. 30 Histogram of Galactic latitudes of perihelion for new comets from the Oort Cloud, using original orbital elements from [67]. The double peak is consistent with the prediction based on injection of comets via the Galactic disk tide

was the first to show evidence for this effect, and a more recent distribution of β_G is shown in Fig. 30. It is important to consider possible effects of discovery biases, but it is generally concluded that the effect is real and thus the Galactic disk tide plays an important role in providing new comets from the Oort Cloud into observable orbits. The fact that the histogram in Fig. 30 peaks at $\pm 30^\circ$ instead of $\pm 45^\circ$ is explained by the fact that the perturbation efficiency variation of (38) has to be convolved with $\cos \beta_G$ for an isotropic distribution of orbits in the Oort Cloud.

Finally, from the expression given by Heisler and Tremaine [41] for the period of ω_G libration we obtain an approximate estimate of

$$P_{\text{lib}} \sim 6 \cdot 10^8 \left(\frac{a}{20,000} \right)^{-3/2}, \tag{39}$$

with P_{lib} in yr and a in AU. For the active, outer part of the Oort Cloud with $a \gtrsim 30{,}000$ AU, the period is a few hundred million years or less, so the current age of the Solar System spans many cycles. But for the inner core with $a < 10{,}000$ AU, the period starts to approach the age of the Solar System. We see that the long-term evolution of the Oort Cloud should involve such slow oscillations of e and i_G, as long as the integrable approximation holds, i.e. the orbital averaging of the Hamiltonian is a relevant procedure. This will, however, break down when the orbital period becomes non-negligible compared to the period of the oscillation. For the outermost parts of the Oort Cloud this is the case—the orbital periods are counted in tens of million years, while the oscillation period is $\sim$ 100 Myr or less. The effect is that the oscillations are still seen but the maximum and minimum values are no longer constant, since the C and L_z integrals break down. In addition, the radial tide plays

an important role and may even cause the loss of comets from the Solar System through migration out of the Sun's Hill sphere in the Galaxy [1].

5.2.2 Sungrazing Comets

As already mentioned, the secular effect of planetary perturbations on typical cometary orbits is that of a circular annulus with uniform density, and this will lead to orbital variations like those caused by the Galactic tide—i.e. a coupled oscillation of eccentricity and inclination, this time with respect to the ecliptic. As seen in Fig. 29, the largest eccentricity variations occur near the separatrix, especially if the perpendicular angular momentum component is small. In the case of the Kozai cycles caused by planetary perturbations, this means high-inclination orbits with i in the general vicinity of 90°. Such comets do not exist in the Jupiter Family, but the orbits are common among Halley-type and long-period comets.

Evidently, objects with $L_z \simeq 0$ and $C \lesssim 1$ will reach eccentricities very near unity, when ω passes 90° or 270°, and the perihelion distance may then fall to practically zero. Actually, comets with such orbits have been known since centuries, and they are usually referred to as *sungrazing comets*. As noted by Marsden [66] in a recent review of sungrazers, the first cometary orbit computed using Newton's law of gravity (comet C/1680 V1) belonged to a sungrazer. Other comets of this type were discovered until recently at a rate of several per century from the ground or from airplanes, but coronagraphic observations from satellites and space probes (especially SOHO) have caused an enormous surge of discoveries during the last decades, so that the number of sungrazing comets is now well over 1000.

Almost all the sungrazing comets seen before the 1980s had similar orbits—in particular with regard to the perihelion direction (latitude $+35^\circ$ and longitude 283°). For the best observed cases, orbital periods had also been established and found to be in the range of $500-1000$ yr. In recognition of the seminal work by Kreutz [51, 52] the sungrazers were often collectively called the Kreutz group. It was natural to think that they had originated from the tidally induced splitting of a common parent comet, and this idea was explored by, e.g., Marsden [64]. But until much more recently, there was no real progress in understanding how this parent comet had come into such a peculiar orbit.

The turning point came when long-term numerical integrations for short-period comets started to be performed on a routine basis. Comet 96P/Machholz 1 was discovered in 1986 in an uncommon orbit with a high inclination ($i \simeq 60^\circ$) and a small perihelion distance ($q \simeq 0.12$ AU) in spite of a period $P < 6$ yr. With the above definitions it is a Halley-type comet, although the period is typical of the Jupiter Family. Rickman and Froeschlé [84] noted with surprise a secular evolution unlike that of Jupiter Family comets and dominated by a large-scale oscillation of q and i, which would take the comet to $q_{\min} \simeq 0.04$ AU around the year 2400. The explanation came when Bailey et al. [2] showed that this evolution is due to the Kozai resonance. Figure 31 illustrates their results for comet 96P, showing that the comet may actually fall into the Sun more than 10,000 yr from now. The oscillation does not have a constant period or amplitude, perhaps due to a breakdown of the

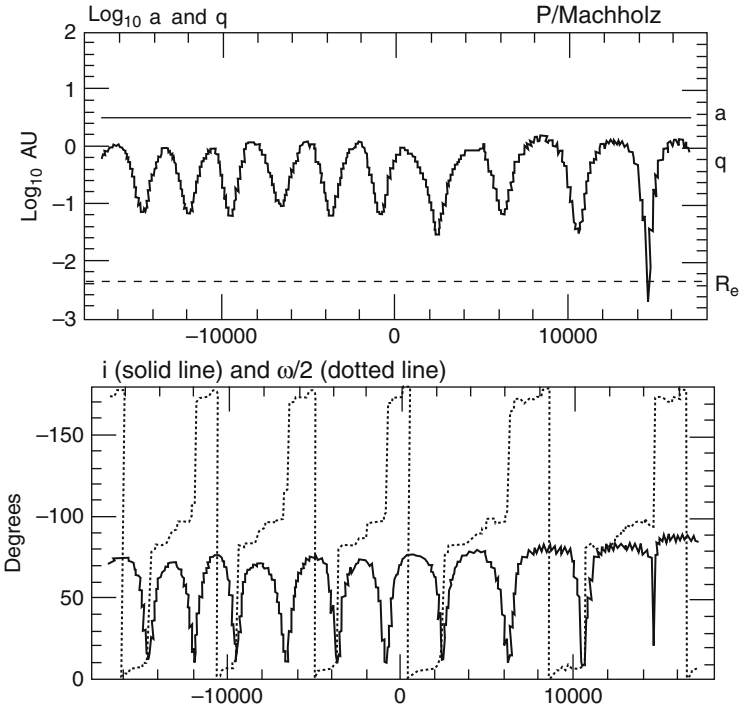

Fig. 31 Orbital elements of comet 96P/Machholz 1 resulting from numerical integrations by Bailey et al. [2], plotted versus time in yr A.D. The semi-major axis and perihelion distance, in AU, are plotted on a log scale in the *upper panel*, and the inclination and half the argument of perihelion are shown in the *lower panel*. Courtesy M.E. Bailey

integrable approximation caused by the closeness of the 9/4 mean motion resonance with Jupiter. The variation of ω is a circulation in proximity of the resonance.

In the same paper, Bailey et al. [2] also showed that similar phenomena occur for a few other short-period comets and, importantly, for the sungrazers. Their conclusion was that evolution into a sungrazing state is a common feature for comets that start with inclinations not very far from 90° and $q \lesssim 2$ AU, so that this should be an important end state for comets, and the occurrence of the Kreutz group—if due to the splitting of just one parent that was perturbed into a sungrazing orbit—is not a very peculiar feature.

During the last decade the space-based discoveries of sungrazing comets have led to the identification of a few more groups with somewhat larger perihelion distances than the Kreutz group. These other groups (the Meyer, Marsden, and Kracht groups) are also believed to have arisen from tidal splittings of parent comets. The Marsden and Kracht groups have similar perihelion directions, and both are similar to the one of comet 96P/Machholz 1 [66], suggesting a relationship of all these comets (implying that the two sungrazing groups are also of short period) with the Quadrantid meteor stream [68, 37].

6 Current Problems

6.1 Source Populations: Formation and Evolution

Already Oort [72] commented on possible mechanisms of formation of his proposed reservoir of comets at extremely large distances. He favoured the expulsion of icy material from the nascent planetary system but placed the region of origin (near the asteroid belt) likely too close to the Sun. Since then there have been many models and suggestions for the creation of the Oort Cloud, and while some of them have even considered comet formation outside the Solar System and capture of the cloud from the star cluster where the Sun was born (e.g. [99]), the main line of thought has remained the one favoured by Oort. However, the formation of comets is now believed to have occurred either in the same region where the giant planets grew or somewhat outside so that they could be gravitationally scattered into larger orbits by these planets.

The mechanism of emplacement thus involves two parts. First, there has to be a sequence of close encounters with a giant planet so that the orbital energy random walks towards the parabolic limit. When the semi-major axis is large enough, and before ejection into a hyperbolic escape orbit occurs, the external influence of stellar encounters or the Galactic tide raises the perihelion distance away from the vicinity of the planetary orbit. Hence, eventually, the comet may end up in the Oort Cloud.

In earlier works [86], where only stellar passages typical of the present Solar System were considered for raising the perihelion distances, a severe problem was identified. If comets had been formed near the orbits of Jupiter or Saturn, where there was certainly enough material to form them in huge numbers, the gravitational scattering would have been so efficient that practically all the comets would have been ejected into interstellar space. The chance of halting in the narrow range of $1/a$ where the external perturbers were efficient would have been very small. Thus the concept arose that Uranus and Neptune are more likely providers of the Oort Cloud comets, so that the most likely place of origin is near their orbits. However, there remains the problem of efficiency of emplacement, because due to the smaller masses of these planets, the amount of material available to form the comets was quite limited.

Some worry also arose from the possibility that encounters with Giant Molecular Clouds [6] might have disrupted the outer parts of the Oort Cloud, so that the currently active cloud producing observable comets might be only a bleak shadow of what it was, when it had just been formed. If thus a very large initial mass would be required, the problem of emplacing it from the Uranus–Neptune region would perhaps be unsurmountable. This worry was only partially relieved by the study of Hut and Tremaine [45], who found that the parameters of the GMCs are too uncertain to say, if the disruption problem is severe or not.

After the importance of the Galactic disk tide was realized, these problems were somewhat reduced. In their study of Oort Cloud formation, Duncan et al. [20] found that the cloud would start at $a \sim 3000$ AU, because this is where the timescales

of orbital diffusion in $1/a$ and tidal torquing of q are equal. Thus the cloud would necessarily form with an inner core. Such a core would serve as a reservoir to bring new comets into the outer parts, if the latter were disrupted by GMC encounters. However, if the Oort Cloud still contains a massive inner core, its total mass may be much higher than one estimates using the flux of new comets from the outer parts (see below). The picture of the current Oort Cloud that emerges from such a scenario is illustrated in Fig. 32 from [29]. While the relaxation of the angular momentum distribution in the outer parts of the cloud is complete, and thus the cloud is spherically symmetric, the inner core may not yet have been fully thermalized and might still show some concentration towards the ecliptic plane, thus forming a transition towards the flattened shape of the transneptunian scattered disk.

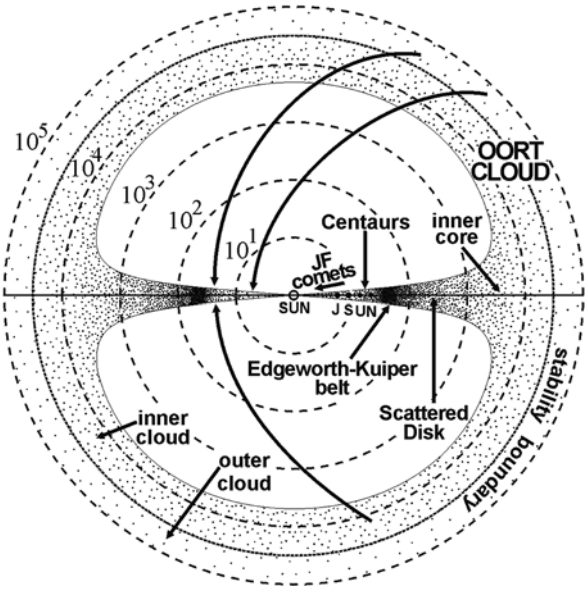

Fig. 32 Sketch of the semi-major axis and inclination distribution of the Oort Cloud, the scattered disk, and populations (Centaurs and JF comets) interior to the latter being captured from it. Note the log scale of distances from the Sun. The ecliptic plane is marked by the horizontal line passing through the Sun. From [29] with the author's permission

In view of the remaining problems to understand the formation of the Oort Cloud, Fernández [27] introduced a new idea, i.e. that the association of the new-born Solar System with its parent molecular cloud and a surrounding star cluster would have made the perihelion extraction process much more efficient. Thus, even if the scattering of icy planetesimals from the region of the giant planets would actually make Jupiter and Saturn responsible for most of the outwards diffusion, the range of $1/a$ where extraction occurs would be wide enough to allow for a rather efficient emplacement process. Fernández and Brunini [30] found that a few Earth masses of

comets might have been emplaced into an initial Oort Cloud extending from a few hundred AU with most of the comets residing in the inner core.

The discovery in 2003 of the very large object (90377) Sedna in an orbit with $q \simeq 76$ AU and $a \sim 490$ AU [9] made a further case for such an Oort Cloud formation scenario, since very close and very slow stellar encounters could have happened when the Solar System was young, and the extraction of a Sedna-type population from the early scattered disk thus appeared as a viable idea [70]. This population might thus be identified with the innermost part of the Oort Cloud's inner core. However, when Brasser et al. [7, 8] attempted to model the extraction of both Sedna-type objects and the Oort Cloud with the assumption that the new-born Solar System was immersed in an embedded star cluster of the type invoked by Fernández [27], they ran into new problems.

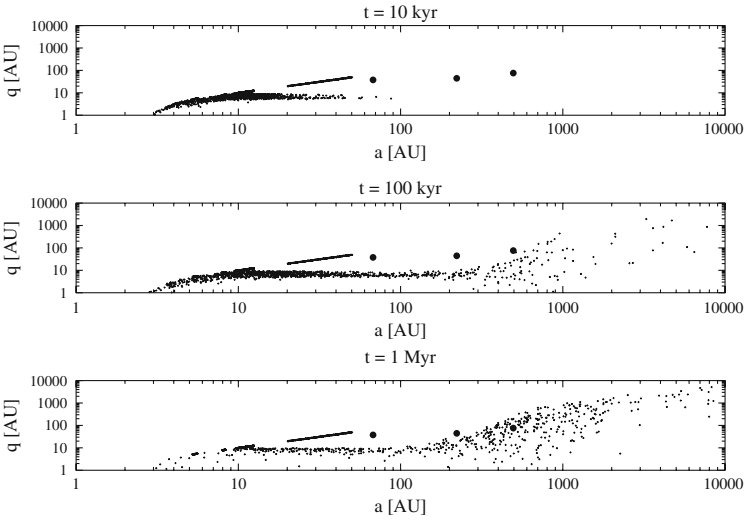

Fig. 33 Three snapshots of the distribution of q and a from the simulation of Oort Cloud formation by Brasser et al. [7]. Note the growing scatter of q for semi-major axes of several hundred AU or more. The three points indicate extended scattered disk objects, the rightmost one being Sedna. Courtesy R. Brasser

In the first paper [7] they found that the extraction mechanism works nearly perfectly, considering a timescale of $\sim$1 Myr during which the comets are perturbed by the planets, the encounters with cluster stars, and the tidal action of the entire cluster (see Fig. 33). But when, in the second paper [8], they considered also the gravity and gas drag effect of the Solar System gas disk (the "solar nebula"), they found that emplacement into the outer regions is practically impossible for objects the size of typical comet nuclei ($\sim$ a few kilometres), because the gas drag circularizes the orbits much faster than the gravitational scattering can proceed. Thus, during this early phase, only very large objects would have been able to reach into

the Oort Cloud, and essentially the whole cloud would have to be explained in a different way.

Currently it seems unclear if the embedded cluster environment could have outlived the solar nebula, since both have similar estimated lifetimes of a few million years. A more likely scenario for Oort Cloud formation appears to involve the emplacement of much more material into the early scattered disk than just the residual planetesimals from Neptune's own accretion zone. This might have occurred, e.g., as a consequence of Neptune's migration through the outer planetesimal disk in the "Nice Model" [91] or due to the gas drag effects in the model of Brasser et al. [8]. Thus, via the intermediary of an early, massive scattered disk, one may still consider an Oort Cloud with a mass of $\sim$1 Earth mass to be a viable idea. It would have arisen as a byproduct of the "erosion" of the scattered disk, when Neptune scattered objects both inwards and outwards and thus caused many objects to diffuse into orbits with large semi-major axes. Fernández et al. [32] showed that this process is efficient in populating the Oort Cloud when applied to the current scattered disk. Hence, the creation of the Oort Cloud may be an ongoing process, even though the rate should have been much higher long ago, when the Solar System was young and the scattered disk was much more massive than today.

Charnoz and Morbidelli [13] discussed the consequences of such a picture of Oort Cloud formation for the collisional evolution of the transneptunian population. They concluded that, if the mass deficit of the classical Kuiper Belt had been caused by collisional grinding (as would be the case for a steep size distribution), the scattered disk would have been severely depleted too, and the estimated mass of the Oort Cloud would be hard to explain. Thus, either there is still a problem to explain the Oort Cloud or the Kuiper Belt has rather been depleted by dynamical mechanisms. It is obviously very important to evaluate the number of comets in the Oort Cloud and its total mass as realistically as possible.

However, this is easier said than done. The basic equation from which the number of Oort Cloud comets (N_{OC}) can be derived using the observed rate of perihelion passages (R_p) of new Oort Cloud comets with perihelion distances less than q_o is

$$R_p = N_{OC} \times \int_{a_{\min}}^{a_{\max}} \varphi(a) \cdot f_{lc}(a) \cdot \frac{2q_o}{a} \cdot a^{-3/2}\, da, \tag{40}$$

if the cloud extends from $a_{\min}$ to $a_{\max}$ in semi-major axis with a frequency function $\varphi(a)$ and if f_{lc} denotes the filling factor of the observable part of the loss cone with $q < q_o$ [3]. This factor expresses the ratio between the actual population of this phase space domain and the one that would apply if the cloud was completely thermalized.

One first has to estimate R_p and then apply a dynamical model of the cloud with assumed values for $a_{\min}$, $a_{\max}$, and $\varphi(a)$ in order to find $f_{lc}(a)$ and finally derive N_{OC}. Note that the value used for R_p will refer to some limit in the brightness of the comets, which may be roughly associated with a minimum mass of the nuclei. Using this, one may estimate an average mass $\langle M \rangle$ and compute the mass of the Oort Cloud as $M_{OC} = N_{OC} \cdot \langle M \rangle$.

Francis [35] analysed recent comet discovery statistics with particular emphasis on the results of the LINEAR survey programme. His conclusion was that $R_p \simeq 0.8$ new comets per AU of perihelion distance per year with absolute magnitudes $H_{10} \lesssim 11$, which is lower than most previous estimates. Concerning dynamical models, we may compare several of those found in the literature with regard to the ratio between the value used for R_p and the number N_{OC} found for $a \gtrsim 20{,}000$ AU. If we refer to this as the injection efficiency E_p of the outer Oort Cloud, we can see a trend towards increasing values during the past decades. Weissman [97] found $E_p \sim 2 \cdot 10^{-12}$, Bailey and Stagg [3] found $E_p \sim 3 \cdot 10^{-12}$, Heisler [40] and Wiegert and Tremaine [99] found $E_p \sim 4 \cdot 10^{-12}$, Dones et al. [17] found $E_p \sim 6 \cdot 10^{-12}$, and most recently Emel'yanenko et al. [25] found $E_p \sim 15 \cdot 10^{-12}$ (the unit is comets/AU/yr).

The latter simulation is the only one that has included a realistic treatment of planetary perturbations. Earlier models considered the whole loss cylinder to be dynamically opaque, i.e. all the comets that enter there are immediately removed by the planets. But in reality there is a chance of survival, so that some new comets may have made previous passages in the loss cylinder [22]. It appears that this extra contribution of "new" comets is quite important in Emel'yanenko et al.'s study.

Comets with $H_{10} < 11$ may be estimated to have nuclei with radii $R \gtrsim 1$ km [3, 98], and with an average density of ~ 500 kg/m^3 we expect masses $M \gtrsim 2 \cdot 10^{12}$ kg. Combining the Francis [35] and Emel'yanenko et al. [25] results, we get $N_{OC} \sim 6 \cdot 10^{10}$ comets in the outer Oort Cloud ($a > 2 \cdot 10^4$ AU) corresponding to a total mass $M_{OC} \sim 0.1$ Earth masses. Even with an important inner core the total mass would not be much larger than 1 Earth mass. This does not appear to be excessive, but a full study of the problem—yet to be made in a realistic way—should include the evolutionary histories of the Oort Cloud and the scattered disk.

6.2 The Capture of Comets

Let us return to some of the issues mentioned in Sect. 2 and in particular the origin of the short-period comets. This is traditionally called the *capture problem*. We have already seen in Sect. 3 that the Jupiter Family and Halley-type populations appear to have little in common, so we will discuss them separately, starting with the Halley-types.

This capture process begins with the injection of Oort Cloud comets into planet-crossing orbits—in particular those that have small enough perihelion distances to be observable. We have seen that as such comets enter, they are perturbed by the giant planets so that the ensuing random walk in orbital energy may eventually lead into the HT population. Recent simulations of the dynamical history of the Oort Cloud using 10^6 comets during 5 Gyr perturbed either by the Galactic tides, by passing stars, or by both in common by Rickman et al. [80] reveal that the injections are largely due to synergy effects between the two perturbing mechanisms. This causes an increase of the injection efficiency and a shift of the median semi-major

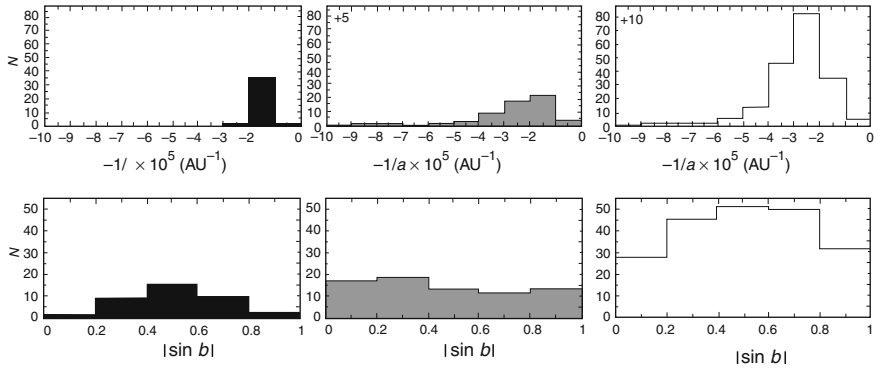

Fig. 34 Distributions of $-1/a$, where a is the cometary semi-major axis (*top panels*) and $|\sin b|$, where b is the Galactic latitude of perihelion (*bottom panels*), for simulated comets entering into observable orbits during a 170 Myr time interval. When present, numbers in the top-left corners of $-1/a$ distribution panels correspond to comets with $-1/a < -1 \cdot 10^{-4}$ AU^{-1}. The left column corresponds to a model with Galactic tide alone, the middle column to passing stars alone, and the right column to a model with both effects combined. From [80]

axis of new comets to lower values compared to earlier results. Figure 34 illustrates these findings for a time interval that is probably representative of the current Solar System containing no significant comet shower.

These computations were made without including planetary perturbations, and all the injected comets have jumped directly from $q > 15$ AU to $q < 5$ AU. However, it would be very interesting to use a full dynamical model to get an indication of the ratio of such "jumpers" to the "creepers", i.e. comets that have passed one or more perihelia in the outer parts of the loss cylinder before reaching the observable orbits, as a function of semi-major axis. One obvious reason is that it might then be possible to check the following interesting suggestion on the origin of HT comets.

Levison et al. [55] found that capture of HT comets from an isotropic flux of new Oort Cloud comets—even allowing for the possibility of finite active lifetimes [31]—leads to the wrong inclination distribution for the Halley-types. Figure 35 illustrates this discrepancy, showing that the observed HT comets have a much stronger preference for prograde orbits than the simulated sample. Levison et al. [58] found a possible solution of the dilemma by invoking the same outwards diffusion from the scattered disk as studied by Fernández et al. [32]—only that they considered the objects for which the Galactic tide decreases the perihelion distance rather than increasing it. This process should yield an additional source of new comets that come almost directly from the scattered disk and thus keep a preference for low-inclination orbits.

One may ask if this suggestion is not in contradiction with the observed flat distribution of $\cos i$ for new comets (see Fig. 6), and this question is still open. Fernández [28] analysed the inclination distribution of long-period comets by separating them into subsamples with different original semi-major axes. He argued that the tidal injection limit from $q \gtrsim 15$ AU is at $a_{\rm ori} \simeq 30{,}000$ AU and thus the comets with

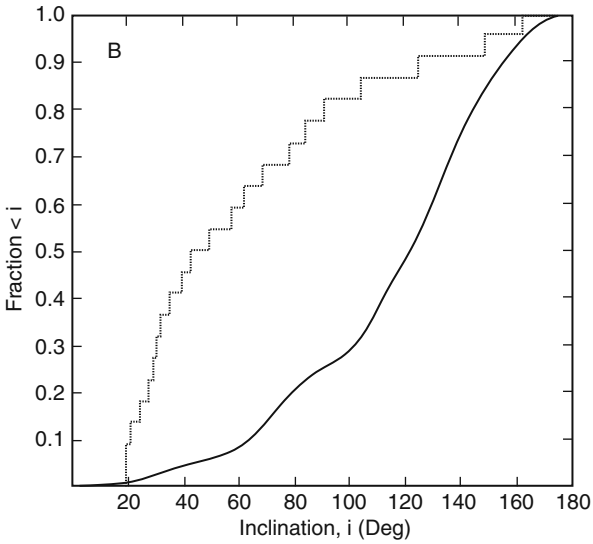

Fig. 35 Cumulative distributions of inclination of Halley-type comets. The observed sample is shown by a *dotted line*, and the sample of captured HT comets from orbital integrations of new Oort Cloud comets is shown by the *solid line*. From [55] with the authors' permission

$a_{\rm ori} \gtrsim 30{,}000$ AU should be jumpers, while those with $a_{\rm ori} < 30{,}000$ AU should be creepers. The latter must have a preference for retrograde orbits, since the chance of survival in the outer parts of the loss cylinder is larger due to the smaller planetary perturbations (cf. Fig. 21). This was indeed seen, but Fernández also found an unexpected preference for prograde orbits among the jumpers, which Levison et al. [58] argued to be evidence for a new comet flux directly from the scattered disk.

However, there are several reasons to cast doubt on this interpretation. One is the evidence from Fig. 34 that jumpers have a much broader distribution of $a_{\rm ori}$ than previously believed, and another is the result that non-gravitational effects have a strong influence on the $a_{\rm ori}$ values of comets with $q \sim 1$ AU as soon as they are included into the orbital solutions [53]—the general trend being that including such effects leads to smaller $a_{\rm ori}$. Thus great care is needed in order to reach any conclusion on whether there is evidence for a flattened source of new comets in addition to the Oort Cloud.

But the implications may be far-reaching. As noted above, the scattered disk may be providing an important source of replenishment of the Oort Cloud, and it is quite likely that it also contributes to the flux of new comets, thus reducing even further the number of comets in the Oort Cloud. At present the constraints are too few and vague to reach a proper understanding of these relationships, but progress in dynamical modelling and observed statistics of the orbits of new comets may soon lead to better insight.

Finally, the origin of Jupiter Family comets is likewise open to debate. The only thing that is clear is that a population of low-inclination Neptune-crossers may form a relevant source, if it is rich enough to supply the required capture rate in order for the JF to be in a steady state. Figure 36 shows a rough sketch of how comets can then be handed over by one giant planet to the next inner one along evolutionary curves

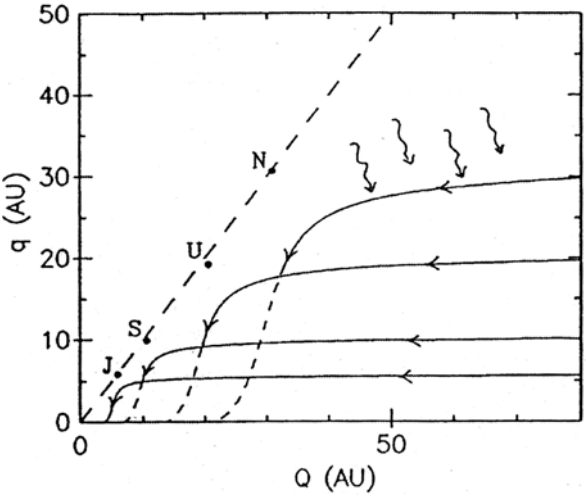

Fig. 36 Sketch of possible evolutionary routes leading from low-inclination Neptune-crossers into the Jupiter Family via a multi-planet capture process, as suggested by Duncan et al. [19]. Each planet dominates the evolution as long as the perihelion is near or inside the orbit of that planet but has not reached the orbit of the next inner planet. Thus the evolution proceeds along curves of constant Tisserand parameter with respect to the governing planet. From [78]

in the (Q, q) plane with constant T referring to the first planet ($T \simeq 2.8$–2.9) in a chain that leads from Neptune to Jupiter and inwards.

An important though difficult question is, what is the origin of those Neptune-crossers, and how many comet-sized objects need there be in the source population in order to maintain the Jupiter Family? There are several different ideas. The classical Kuiper Belt is able to provide a flux of objects into Neptune-crossing orbits via resonant eccentricity pumping (the region at $a \simeq 40$–42 AU is a possible source due to overlapping secular resonances), and Holman and Wisdom [43] and Levison and Duncan [57] estimated the necessary number of comet-sized objects as 5–$7 \cdot 10^9$. Duncan and Levison [18] instead considered the scattered disk with emphasis on the part of it that may encounter Neptune ($q \simeq 30$ AU), and in this case only $6 \cdot 10^8$ objects would be enough. Since the populations of large, $\sim$ 100-km objects in the Kuiper Belt and the scattered disk appear to be of similar sizes, the conclusion was that the scattered disk is the prime source of JF comets [19].

Emel'yanenko et al. [24] found that the part of the scattered disk corresponding to near-Neptune high-eccentricity objects is unlikely to have evolved solely from the Kuiper Belt source (note the paucity of non-resonant Kuiper Belt objects with $q < 40$ AU in Fig. 37), but that it may provide a relevant source of JF comets with an estimated population of 10^{10} objects of cometary size with perihelion distances $28 < q < 35.5$ AU. In addition, Morbidelli [69] found that Neptune-crossing Plutinos leak out of the 2:3 resonance on Gyr timescales due to slow, chaotic diffusion, thus possibly providing an important source of JF comets, and the number of comet-sized Plutinos should then be $\sim 10^8$–10^9.

However, it is currently not clear—perhaps not even likely—that the above numbers fit with reality. Volk and Malhotra [95] emphasize that the break in the transneptunian size-distribution power law with a shallower slope at smaller sizes, suggested by the HST observations by Bernstein et al. [5], leads to a number of kilometre-sized scattered disk objects that is orders of magnitude too small compared to the above requirements. It is possible that an alternative source of Jupiter Family comets is called for, such as the above-mentioned Plutinos or the fragmentation of large scattered disk objects as they approach the Sun during the capture process [95].

Another problem was mentioned by Fernández et al. [32]. Figure 37 shows an update and extension of their Fig. 1 based on the data files of the MPC web site of 15 July 2008. Most scattered disk objects have perihelion distances too large to be Neptune-encountering in the sense of Duncan and Levison's [18] investigation, and Fernández et al. [32] found that such objects tend to diffuse outwards, eventually feeding the Oort Cloud as discussed in Sect. 6.1, rather than being captured into the Centaur and JF comet populations. This would mean that the requirement on the total number of comet-sized scattered disk objects in order to feed the Jupiter Family would be even harder.

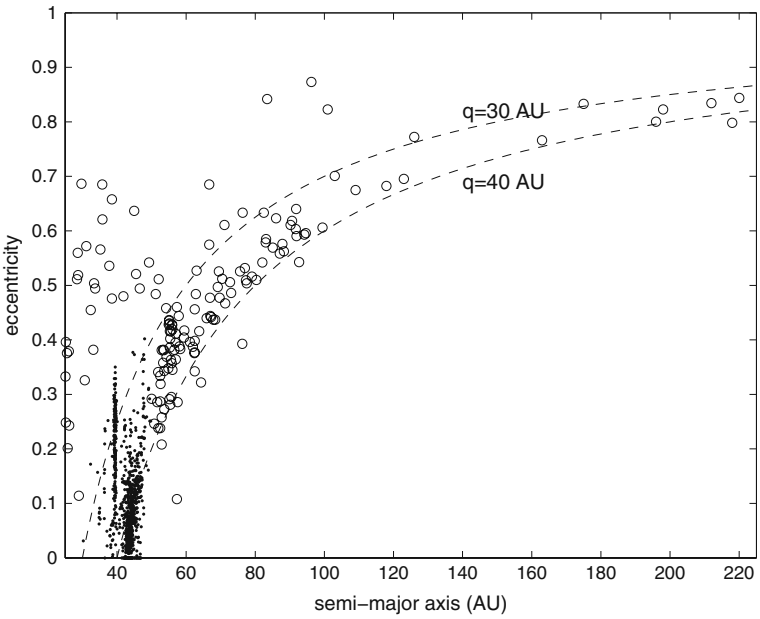

Fig. 37 Orbital distribution of transneptunians and Centaurs as plotted on the (a, e) plane. The *dots* show Kuiper Belt objects and the *open circles* show scattered disk objects and Centaurs, all with at least 5 days' observational arc. The *dashed curves* indicate crude perihelion distance limits of the scattered disk. *Open circles* above and to the left indicate Centaurs, while those below and to the right indicate the "extended scattered disk". There are five objects to the right of the diagram, whereof one scattered disk object, two Centaurs, and two extended scattered disk objects: (90377) Sedna and 2004 VN112. Data from the IAU Minor Planet Center lists

It appears that there are still several important, unanswered questions regarding the capture of short-period comets, and the picture shown in Fig. 2b, which is intended to represent current thinking, is not likely to be the final word. Further studies of the evolution of the scattered disk and its relations with the Oort Cloud and the fate of Oort Cloud comets upon entry into the planetary system are warranted on the theoretical side. Concerning observations, the access to larger and better telescopes aiding in the search for distant Solar System objects in the near future is bound to yield significantly improved statistics both for the orbital distribution of transneptunians and for the detailed structure of the Oort peak.

There is thus reason to hope for continued, rapid progress in understanding both the real workings of cometary dynamics and the origin and shaping of the Solar System's cometary populations. This should aid importantly in the interpretation of recent and future cometary space mission results, which provide a lot of detailed insights about the chemical and physical build-up of cometary nuclei—thus playing an essential part in the use of comets as cosmogonic probes.

Acknowledgments I am greatly indebted to Björn Davidsson for providing me with the plot for Fig. 19 and to Marc Fouchard and Giovanni Valsecchi for letting me use their orbit integration results to prepare Fig. 21. I also thank Giovanni Valsecchi for helping me find and correct several small errors in my first manuscript and for drawing my attention to Lidov's work on ω librations.

References

1. Antonov, V.A., Latyshev, I.N.: Determination of the Form of the Oort Cometary Cloud as the Hill surface in the Galactic Field. In Chebotarev, G.A., Kazimirchak-Polonskaya, E.I., Marsden, B.G. (eds.) The motion, evolution of orbits, and origin of comets, Proc. of IAU Symp. 45, Reidel, Dordrecht, pp. 341–345 (1972) 384
2. Bailey, M.E., Chambers, J.E., Hahn, G.: Origin of sungrazers – A frequent cometary end-state. Astron. Astrophys. **257**, 315–322 (1992) 344, 384, 385
3. Bailey, M.E., Stagg, C.R.: Cratering constraints on the inner Oort cloud – Steady-state models. Mon. Not. R. Astron. Soc. **235**, 1–32 (1988) 389, 390
4. Bel'yaev, N.A., Kresák, L'., Pittich, E.M., Pushkarev, A.N.: Catalogue of short-period comets. Slovak Academy of Sciences, Astronomical Institute, Bratislava (1986) 358, 359, 370
5. Bernstein, G.M., Trilling, D.E., Allen, R.L., Brown, M.E., Holman, M., Malhotra, R.: The size distribution of trans-Neptunian bodies. Astron. J. **128**, 1364–1390 (2004) 394
6. Biermann, L.: Dense interstellar clouds and comets. In Reiz, A., Anderson, T. (eds.) Astronomical papers dedicated to Bengt Strömgren, pp. 327–335. Copenhagen Observatory, Copenhagen (1978) 386
7. Brasser, R., Duncan, M.J., Levison, H.F.: Embedded star clusters and the formation of the Oort cloud. Icarus **184**, 59–82 (2006) 343, 345, 388
8. Brasser, R., Duncan, M.J., Levison, H.F.: Embedded star clusters and the formation of the Oort cloud. II. The effect of the primordial solar nebula. Icarus **191**, 413–433 (2007) 342, 345, 371, 376, 388, 389
9. Brown, M.E., Trujillo, C., Rabinowitz, D.: Discovery of a candidate inner Oort cloud planetoid. Astrophys. J. **617**, 645–649 (2004) 388
10. Carusi, A., Valsecchi, G.B.: Numerical simulations of close encounters between Jupiter and minor bodies. In Gehrels, T. (ed.) Asteroids, pp. 391–416. Univ. Arizona Press, Tucson (1979) 368

11. Carusi, A., Valsecchi, G.B.: Dynamical evolution of short-period comets. In Ceplecha, Z., Pecina, P. (eds.) Interplanetary matter, pp. 21–28. Astron. ústav CSAV, Praha (1987) 348
12. Chambers, J.E.: A simple mapping for comets in resonance. Celest. Mech. Dynam. Astron. **57**, 131–136 (1993) 345
13. Charnoz, S., Morbidelli, A.: Coupling dynamical and collisional evolution of small bodies 2: Forming the Kuiper Belt, the Scattered Disk and the Oort Cloud. Icarus **188**, 468–480 (2007) 389
14. Chodas, P.W., Yeomans, D.K.: The orbital motion and impact circumstances of Comet Shoemaker-Levy 9. In Noll, K.S., Weaver, H.A., Feldman, P.D. (eds.) The collision of Comet Shoemaker-Levy 9 and Jupiter, Proc. of IAU Coll. 156, pp. 1–30. Cambridge University Press, Cambridge (1996) 368
15. Danby, J.M.A.: Fundamentals of celestial mechanics, 2nd ed. Willmann-Bell, Richmond, VA (1988) 348, 367
16. Delsemme, A.H.: Galactic tides affect the Oort cloud – An observational confirmation. Astron. Astrophys. **187**, 913–918 (1987) 382
17. Dones, L., Weissman, P.R., Levison, H.F., Duncan, M.J.: Oort cloud formation and dynamics. In Johnstone, D., Adams, F.C., Lin, D.N.C., Neufeld, D.A., Ostriker, E.C. (eds.) Star formation in the interstellar medium, pp. 371–379. Astronomical Society of the Pacific, San Francisco (2004) 390
18. Duncan, M.J., Levison, H.F.: A scattered comet disk and the origin of Jupiter family comets. Science **276**, 1670–1672 (1997) 393, 394
19. Duncan, M.J., Levison, H.F., Dones, L.: Dynamical evolution of ecliptic comets. In Festou, M.C., Keller, H.U., Weaver, H.A. (eds.) Comets II, pp. 193–204. Univ. of Arizona Press, Tucson (2004) 342, 393
20. Duncan, M., Quinn, T., Tremaine, S.: The formation and extent of the solar system comet cloud. Astron. J. **94**, 1330–1338 (1987) 382, 386
21. Dybczyński, P.A.: Impulse approximation improved. Celest. Mech. Dynam. Astron. **58**, 139–150 (1994) 377
22. Dybczyński, P.A.: Dynamical history of the observed long-period comets. Astron. Astrophys. **375**, 643–650 (2001) 390
23. Eggers, S., Woolfson, M.M.: Stellar perturbations of inner core comets and the impulse approximation. Mon. Not. R. Astron. Soc. **282**, 13–18 (1996) 377
24. Emel'yanenko, V.V., Asher, D.J., Bailey, M.E.: High-eccentricity trans-Neptunian objects as a source of Jupiter-family comets. Mon. Not. R. Astron. Soc. **350**, 161–166 (2004) 393
25. Emel'yanenko, V.V., Asher, D.J., Bailey, M.E.: The fundamental role of the Oort cloud in determining the flux of comets through the planetary system. Mon. Not. R. Astron. Soc. **381**, 779–789 (2007) 390
26. Fernández, J.A.: New and evolved comets in the solar system. Astron. Astrophys. **96**, 26–35 (1981) 379
27. Fernández, J.A.: The formation of the Oort Cloud and the primitive galactic environment. Icarus **129**, 106–119 (1997) 387, 388
28. Fernández, J.A.: Changes in the inclination-distribution of long-period comets with the orbital energy. In Warmbein, B. (ed.) Proceedings of Asteroids, Comets, Meteors – ACM (2002) ESA SP-500, pp. 303–304. ESA Publications Division, Noordwijk (2002) 391
29. Fernández, J.A.: Comets – nature, dynamics, origin and their cosmological relevance, ASSL Vol. 328, Springer, Dordrecht (2005) 387
30. Fernández, J.A., Brunini, A.: The buildup of a tightly bound comet cloud around an early Sun immersed in a dense Galactic environment: numerical experiments. Icarus **145**, 580–590 (2000) 387
31. Fernández, J.A., Gallardo, T.: The transfer of comets from parabolic orbits to short-period orbits: Numerical studies. Astron. Astrophys. **281**, 911–922 (1994) 391
32. Fernández, J.A., Gallardo, T., Brunini, A.: The scattered disk population as a source of Oort cloud comets: evaluation of its current and past role in populating the Oort cloud. Icarus **172**, 372–381 (2004) 344, 389, 391, 394

33. Fernández, J.A., Ip, W.-H.: Dynamical evolution of a cometary swarm in the outer planetary region. Icarus **47**, 470–479 (1981) 367
34. Fernández, J.A., Ip, W.-H.: Some dynamical aspects of the accretion of Uranus and Neptune – the exchange of orbital angular momentum with planetesimals. Icarus **58**, 109–120 (1984) 367
35. Francis, P.J.: The demographics of long-period comets. Astrophys. J. **635**, 1348–1361 (2005) 390
36. Gomes, R.S.: The origin of the Kuiper Belt high-inclination population. Icarus **161**, 404–418 (2003) 344
37. Gonczi, R., Rickman, H., Froeschlé, C: The connection between Comet P/Machholz and the Quadrantid meteor stream. Mon. Not. R. Astron. Soc. **254**, 627–634 (1992) 385
38. Greenberg, R., Carusi, A., Valsecchi, G.B.: Outcomes of planetary close encounters – A systematic comparison of methodologies. Icarus **75**, 1–29 (1988) 367
39. Gunn, J.E., Knapp, G.R., Tremaine, S.D.: The global properties of the Galaxy. II – the Galactic rotation parameters from 21-cm H I observations. Astron. J. **84**, 1181–1188 (1979) 380
40. Heisler, J.: Monte Carlo simulations of the Oort comet cloud. Icarus **88**, 104–121 (1990) 390
41. Heisler, J., Tremaine, S.: The influence of the galactic tidal field on the Oort comet cloud. Icarus **65**, 13–26 (1986) 381, 382, 383
42. Hills, J.G.: Comet showers and the steady-state infall of comets from the Oort cloud. Astron. J. **86**, 1730–1740 (1981) 345, 379
43. Holman, M.J., Wisdom, J.: Dynamical stability in the outer solar system and the delivery of short period comets. Astron. J. **105**, 1987–1999 (1993) 393
44. Holmberg, J., Flynn, Ch.: The local density of matter mapped by Hipparcos. Mon. Not. R. Astron. Soc. **313**, 209–216 (2000) 380
45. Hut, P., Tremaine, S.: Have interstellar clouds disrupted the Oort comet cloud?. Astron. J. **90**, 1548–1557 (1985) 386
46. Jewitt, D.: The persistent coma of Comet P/Schwassmann-Wachmann 1. Astrophys. J. **351**, 277–286 (1990) 352
47. Kazimirchak-Polonskaya, E.I.: The major planets as powerful transformers of cometary orbits. In Chebotarev, G.A., Kazimirchak-Polonskaya, E.I., Marsden, B.G. (eds.) The motion, evolution of orbits, and origin of comets, Proc. of IAU Symp. 45, Reidel, Dordrecht, pp. 373–397 (1972) 342, 346
48. Kozai, Y.: Secular perturbations of asteroids with high inclination and eccentricity. Astron. J. **67**, 591–598 (1962) 380
49. Kresák, L.: Dormant phases in the aging of periodic comets. Astron. Astrophys. **187**, 906–908 (1987) 358
50. Kresák, L., Kresáková, M.: The absolute total magnitudes of periodic comets and their variations. In Rolfe, E.J., Battrick, B. (eds.) Symposium on the Diversity and Similarity of Comets, ESA–SP 278, pp. 37–42 (1987) 358
51. Kreutz, H.: Untersuchungen uber das Cometensystem 1843 I, 1880 I und 1882 II I. Theil. Publ. Sternw. Kiel, No. 3 (1888) 384
52. Kreutz, H.: Untersuchungen uber das Cometensystem 1843 I, 1880 I und 1882 II. Theil. Publ. Sternw. Kiel, No. 6 (1891) 384
53. Królikowska, M.: Non-gravitational effects in long-period comets and the size of the Oort Cloud. Acta Astron. **56**, 385–412 (2006) 392
54. Levison, H.F.: Comet taxonomy. In Rettig, T.W., Hahn, J.M. (eds.) Completing the inventory of the solar system, Astron. Soc. of the Pacific, pp. 173–191 (1996) 348
55. Levison, H.F., Dones, L., Duncan, M.J.: The origin of Halley-type comets: Probing the inner Oort cloud. Astron. J. **121**, 2253–2267 (2001) 391, 392
56. Levison, H.F., Duncan, M.J.: The long-term dynamical behavior of short-period comets. Icarus **108**, 18–36 (1994) 348
57. Levison, H.F., Duncan, M.J.: From the Kuiper Belt to Jupiter-Family comets: The spatial distribution of ecliptic comets. Icarus **127**, 13–32 (1997) 393
58. Levison, H.F., Duncan, M.J., Dones, L., Gladman, B.J.: The scattered disk as a source of Halley-type comets. Icarus **184**, 619–633 (2006) 391, 392

59. Levison, H.F., Morbidelli, A., Dones, L., Jedicke, R., Wiegert, P.A., Bottke, W.F.: The mass disruption of Oort cloud comets. Science **296**, 2212–2215 (2002) 358
60. Lidov, M.L.: The evolution of orbits of artificial satellites of planets under the action of gravitational perturbations of external bodies. Planet. Space Sci. **9**, 719–759 (1962)
61. Lupishko, D.F., Di Martino, M., Binzel, R.P.: Near-Earth objects as principal impactors of the Earth: Physical properties and sources of origin. In Milani, A., Valsecchi, G.B., Vokrouhlický, D. (eds.) Near Earth objects, our celestial neighbors: opportunity and risk, Proc. of IAU Symp. 236, pp. 251–260. Cambridge University Press, Cambridge (2007) 380
62. Luu, J.X., Jewitt, D.C.: Cometary activity in 2060 Chiron. Astron. J. **100**, 913–932 (1990) 342
63. Marsden, B.G.: The sungrazing comet group. Astron. J. **72**, 1170–1183 (1967) 353
64. Marsden, B.G.: Catalogue of cometary Orbits, 6th ed., IAU Central Bureau for Astronomical Telegrams, Cambridge, MA (1989) 384
65. Marsden, B.G.: Sungrazing comets. Annu. Rev. Astron. Astrophys. **43**, 75–102 (2005) 355
66. Marsden, B.G., Williams, G.V.: Catalogue of cometary orbits, 16th edn. IAU Minor Planet Center, Cambridge, MA (2005) 384, 385
67. McIntosh, B.A.: Comet P/Machholtz and the Quadrantid meteor stream. Icarus **86**, 299–304. 356, 357, 383
68. Morbidelli, A.: Chaotic diffusion and the origin of comets from the 2/3 resonance in the Kuiper Belt. Icarus **127**, 1–12 (1997) 385
69. Morbidelli, A., Levison, H.F.: Scenarios for the origin of the orbits of the trans-Neptunian objects 2000 CR105 and 2003 VB12 (Sedna). Astron. J. **128**, 2564–2576 (2004) 393
70. Morbidelli, A., Levison, H.F., Tsiganis, K., Gomes, R.: Chaotic capture of Jupiter's Trojan asteroids in the early solar system. Nature **435**, 462–465 (2005) 388
71. Oort, J.H.: The structure of the cloud of comets surrounding the Solar System and a hypothesis concerning its origin. Bull. Astron. Inst. Neth. **11**, 91–110 (1950) 342
72. Oort, J.H., Schmidt, M.: Differences between new and old comets. Bull. Astron. Inst. Neth. **11**, 259–269 (1951) 355, 386
73. Öpik, E.J.: Collision probability with the planets and the distribution of planetary matter. Proc. R. Irish Acad., Sect. A, **54**, 165–199 (1951) 357
74. Öpik, E.J.: Interplanetary encounters: close-range gravitational interactions. Elsevier Scientific Pub. Co., Amsterdam, New York (1976) 349, 364
75. Rickman, H.: Stellar perturbations of orbits of long-period comets and their significance for cometary capture. Bull. Astron. Inst. Czech. **27**, 92–105 (1976) 364, 365
76. Rickman, H.: Interrelations between comets and asteroids. In Carusi, A., Valsecchi, G.B. (eds.) Dynamics of comets: Their origin and evolution, Proc. of IAU Coll. 83, pp. 149–172, D. Reidel, Dordrecht (1985) 376, 377
77. Rickman, H.: Physico-Dynamical Evolution of aging Comets. In Benest, D., Froeschlé, C. (eds.) Interrelations between physics and dynamics for minor bodies in the solar system, pp. 197–263. Société Française des Spécialistes d'Astronomie, Paris (1992) 353
78. Rickman, H., Fernández, J.A., Tancredi, G., Licandro, J.: The cometary contribution to planetary impact rates. In Marov, M.Ya., Rickman, H. (eds.) Collisional processes in the solar system, Astrophysics and Space Science Library, vol. 261, pp. 131–142. Kluwer, Dordrecht (2001a) 352, 355, 393
79. Rickman, H., Fouchard, M., Froeschlé, Ch., Valsecchi, G.B.: Injection of Oort cloud comets: the fundamental role of stellar perturbations. Celest. Mech. Dynam. Astron. **102**, 111–132 (2008) 342
80. Rickman, H., Fouchard, M., Valsecchi, G.B., Froeschlé, Ch.: Algorithms for Stellar Perturbation Computations on Oort Cloud Comets. Earth, Moon, Planets. **97**, 411–434 (2005) 376, 390, 391
81. Rickman, H., Froeschlé, C.: Orbital evolution of short-period comets treated as a Markov process. Astron. J. **84**, 1910–1917 (1979) 377, 378
82. Rickman, H., Froeschlé, C.: A Keplerian method to estimate perturbations in the restricted three-body problem. Moon Planets. **28**, 69–86 (1983) 371

83. Rickman, H., Froeschlé, C.: Cometary dynamics. Celest. Mech. **43**, 243–263 (1988) 372
84. Rickman, H., Valsecchi, G.B., Froeschlé, C.: From the Oort cloud to observable short-period comets – I. The initial stage of cometary capture. Mon. Not. R. Astron. Soc. **325**, 1303–1311 (2001b). 342, 343, 384
85. Safronov, V.S.: Oort's cometary cloud in the light of modern cosmogony. In Delsemme, A.H. (ed.) Comets, asteroids, meteorites: Interrelations, evolution and origins, Proc. of IAU Coll. 39, pp. 483–484, University of Toledo, Toledo, OH (1977) 374
86. Senay, M.C., Jewitt, D.: Coma formation driven by carbon-monoxide release from comet Schwassmann-Wachmann 1. Nature **371**, 229–231 (1994) 386
87. Szutowicz, S., Rickman, H.: Orbital linkages of Comet 6P/d'Arrest based on its asymmetric light curve. Icarus **185**, 223–243 (2006) 352
88. Tancredi, G., Lindgren, M., Rickman, H.: Temporary satellite capture and orbital evolution of Comet P/Helin-Roman-Crockett. Astron. Astrophys. **239**, 375–380 (1990) 341
89. Tisserand, F.F.: Mémoires et observations. Sur la théorie de la capture des cométes périodiques. [Suite et fin.]. Bull. Astron. Sér. I, **6**, 289–292 (1889) 368, 369
90. Tsiganis, K., Gomes, R., Morbidelli, A., Levison, H.F.: Origin of the orbital architecture of the giant planets of the solar system. Nature **435**, 459–461 (2005) 346
91. Valsecchi, G.B.: 236 years ago.... In Milani, A., Valsecchi, G.B., Vokrouhlický, D. (eds.) Near Earth objects, our celestial neighbors: opportunity and risk, Proc. of IAU Symp. 236, pp. xvii–xx. Cambridge University Press, Cambridge (2007) 389
92. Valsecchi, G.B., Milani, A., Gronchi, G.F., Chesley, S.R.: The distribution of energy perturbations at planetary close encounters. Celest. Mech. Dynam. Astron. **78**, 83–91 (2000) 370
93. Valsecchi, G.B., Milani, A., Gronchi, G.F., Chesley, S.R.: Resonant returns to close approaches: analytical theory. Astron. Astrophys. **408**, 1179–1196 (2003) 367
94. Volk, K., Malhotra, R.: The scattered disk as the source of the Jupiter family comets. Astrophys. J. **687**, 714–725 (2008) 365
95. Vsekhsvyatskij, S.K.: Physical characteristics of comets. Israel Program for Scientific Translations (1958) 394
96. Weissman, P.R.: Dynamical evolution of the Oort cloud. In Carusi, A., Valsecchi, G.B. (eds.) Dynamics of comets: their origin and evolution, Proc. of IAU Coll. 83, pp. 87–96, D. Reidel, Dordrecht (1985) 358
97. Weissman, P.R.: The Oort cloud. In Rettig, T.W., Hahn, J.M. (eds.) Completing the inventory of the solar system, pp. 265–288, Astron. Soc. of the Pacific (1996) 390
98. Wiegert, P., Tremains, S.: The Evolution of Long-Period Comets. Icarus **137**, 84–121 (1999) 390
99. Zheng, J.-Q., Valtonen, M.J., Valtaoja, L.: Capture of comets during the evolution of a star cluster and the origin of the Oort Cloud. Celest. Mech. Dynam. Astron. **49**, 265–272 (1990) 390

Dynamical Features of the Oort Cloud Comets

M. Fouchard, C. Froeschlé, H. Rickman, and G. B. Valsecchi

Abstract The Oort cloud which corresponds to the outer boundary of our Solar system, is considered to be the main reservoir of long period comets. At such distance from the Sun (several times 10 000 AU), the comet trajectories are affected by the galactical environment of the Solar System. Two main effects contribute to inject comets from the Oort cloud to the inner Solar system where comets may become observable: the Galactic tide which is due to the difference of the gravitational attraction of the entire Galaxy on the Sun and on the comets, and the gravitational effects of stars passing close to the Sun. In this lecture the characteristics and the long term effects of these two mechanisms, taken independently and simultaneously, will be illustrated.

1 Introduction

In 1950 from the distribution of semi-major axes of 22 well determined of observed comets, Oort [32] showed that a clear peak, named later the "Oort peak" around 100,000 AU, was present. The orbital energy of such comets is such that the planetary perturbations by Jupiter or Saturn can easily remove the comets from this region, i.e., the comets go either in the interstellar medium or are sent on a much tighter orbit to the Sun. Consequently the comets in the peak are "new," they were

M. Fouchard (✉)
LAL-IMCCE/Université de Lille, 1 impasse de l'Observatoire F-59000 Lille, France,
fouchard@imcce.fr

C. Froeschlé
Observatoire de la Côte d'Azur, BP 4229, FR-06304 Nice, France, froesch@obs-nice.fr

H. Rickman
Uppsala Astronomical Observatory, Box 515, SE-75120 Uppsala, Sweden,
hans@astro.uu.se

G.B. Valsecchi
INAF-IASF, Via Fosso dell Cavaliere 100, 00133 Roma, Italy,
giovanni@iasf-roma.inaf.it

Fouchard, M. et al.: *Dynamical Features of the Oort Cloud Comets*. Lect. Notes Phys. **790**, 401–430 (2010)
DOI 10.1007/978-3-642-04458-8_8

entering the planetary region of the Solar system for the first time and should form a reservoir surrounding the Sun between 10^4 and 10^5 AU : the *Oort cloud*.

Once in this reservoir, since the comets are so far from the Sun, Oort showed that only perturbations from random passing stars, can change significantly the angular momenta of comets and send some of them into the planetary region. Thus from this time stellar perturbations were the only mechanism considered to inject comets in the planetary region (e.g., [35, 40, 16, 26, 34]).

However, since 1983 the importance of galactic tides was pointed out first by Byl [9]. Then several papers [10, 23, 39] have confirmed the main influence of the tides. Moreover an observational confirmation of the action of the vertical galactic tide was pointed out by Delsemme [11], who studied the distribution of the galactic latitudes of perihelia of 152 known original orbits of comets and found that these new Oort Cloud comets present a double-peaked distribution that is a characteristic of the disk tide.

Duncan et al. [12] have shown that the characteristic timescale for changing the perihelion distance, whatever the semi-major axis, is shorter for the galactic tide than for the stellar perturbations. Further numerical integrations [23] have confirmed the dominant role of the galactic tide to inject comets in the planetary region. Moreover the result obtained by [12] has been verified by analytical work [18].

Consequently from that time, stellar perturbations have been neglected when cometary injection is concerned; however, as shown by Hill [26] very close stellar passages may produce a short but strong increase (many orders of magnitude) of the inward flux of "new comets." During these so-called comet showers, comets coming from the inner part of the Oort cloud, i.e., below 10–20,000 AU are observable.

In fact three main perturbers may disturb the Oort cloud comets:

- The Giant Molecular clouds, but it has been shown that an encounter of the Solar System with such a cloud will generate huge perturbations of the Oort cloud. But such encounters are so rare [29] that they are almost never taken into account when dynamics of the Oort cloud comets is concerned.
- The Galactic tide, which is due to the difference between the gravitational attraction of the entire Galaxy on the Sun and of course on the Oort cloud surrounding the Sun.
- The passing stars, though as said previously they are often neglected, we will see that they may have an important role when they are considered together with the galactic tide.

Thus the present lecture is devoted to the presentation of the dynamics generated by the Galactic tide and the passing stars. In Sect. 2, we will first consider the effects of the Galactic tide, starting with the development of the equations of motion (Sect. 2.1). We will see that accordingly some assumptions which are considered lead to the integrability of the long-term dynamics (Sect. 2.2). However, we will see (Sect. 2.3) that, when the full tide is at work and none assumptions are done, the integrability does not hold any more.

The stellar perturbations are introduced in Sect. 3. We will first discuss some general characteristics of the stellar encounters, leading to a very simple model of its effects on cometary orbits (Sect. 3.1). Then, in Sect. 3.2 we will focus on the cumulative effects of stellar encounters on long timescales.

In Sect. 4, we will show that when both perturbers are at work, i.e., the full galactic tides and stellar perturbations are considered, a synergy takes place. The main conclusions are summarized in Sect. 5.

2 The Galactic Tide

2.1 Equations of Motion

The Galactic tide which is due to the difference between the gravitational attraction of the Galaxy applied on the Sun and that applied on the comets was not considered until 1983 [9]. Such difference may be obviously neglected when one considers an object close to the Sun such as planets, asteroids, or Kuiper belt objects, however, for the Oort cloud which may extent as far as 1 pc from the Sun, the tide may turn out to be one of the main disturbers of the cometary orbits. From [23, 12] the Galactic tide has had a growing role (see [3] for a review). It is now considered as the main perturber of the Oort cloud [42] and many studies were devoted to its effects [4, 5, 20, 29, 30].

When the tide was first considered, only the central bulk of the Galaxy was taken into account [9], but it appeared [10, 23] that the Galactic mid-plane in the solar neighborhood generates a major contribution to the Galactic tide. We will now derive the equations of motion of a comet orbiting the Sun under the influence of the Galactic tide, taking into account both the central bulk and the mid-plane of the Galaxy. First of all, in order to simplify the equations of motion, and in the limit of our knowledge of the Galactic neighborhood of the Sun on long timescale, the following assumptions are made:

- the potential does not depend on time;
- the Galaxy is axisymmetric; and
- the Sun moves around the Galactic center with an uniform speed on a circular orbit in the Galactic mid-plane.

Let R_0 and Ω_0 be the radius and the angular speed of the Sun trajectory around the Galactic center. One then defines two different frames. The first one called the rotating frame is defined as a heliocentric frame with the $\hat{x}'$ axis in the radial direction pointing toward the galaxy center, $\hat{z}$ axis normal to the galactic plane pointing toward the north galactic pole, and the $\hat{y}'$ axis completing a right-handed system. The second one, referred as the fixed frame $(\hat{x}, \hat{y}, \hat{z})$, heliocentric as well, is such that it coincides with the rotating frame $(\hat{x}', \hat{y}', \hat{z})$ at $t = 0$ and then keeps its directions fixed (see Fig. 1). If α_r is an angle in the galactic plane measured in the rotating frame from $\hat{x}'$, and α the corresponding angle measured in the fixed frame from $\hat{x}$

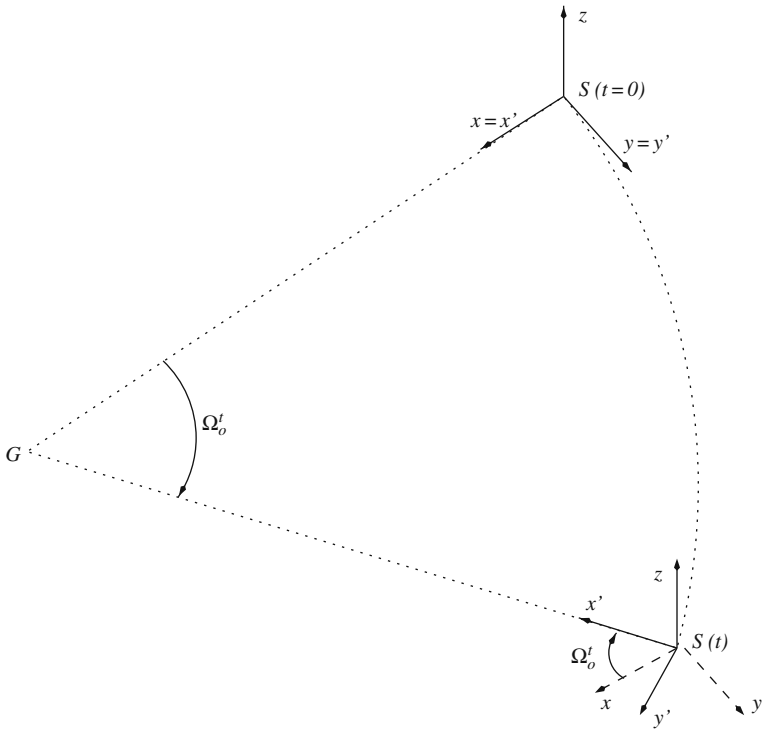

Fig. 1 The fixed frame $\hat{x}, \hat{y}, \hat{z}$ and the rotating frame $\hat{x}', \hat{y}', \hat{z}$ used in this lecture. S denotes the Sun angular, G the galactic center. Ω_0 is the Sun angular speed around the galactic center. The relation of an angle α measured in the fixed frame from the $\hat{x}$ vector and the corresponding angle α_r measured in the rotating frame from the $\hat{x}'$ vector is $\alpha = \alpha_r + \Omega_0 t$ at time t

at time t then one has the following relation $\alpha = \alpha_r + \Omega_0 t$. One should note that in both frames, since the motion of the Sun around the galaxy is retrograde, Ω_0 is negative. The subscript r will denote an angle measured in the rotating frame.

Because the Galaxy is axisymmetric, the Galactic potential may be written as $U_g(R, z)$ where R is the distance to the Galactic axe of rotation and $z = 0$ is the Galactic mid-plane. Following [23] the Galactic potential is developed at order 2 in the neighborhood of the Sun, i.e., a development of order 1 of the force deriving from the potential U_g.

The gravitational attraction of the Galaxy on a comet and on the Sun are, respectively,

$$\mathbf{f}_c = -\nabla U_g(R, z),$$
$$\mathbf{f}_\odot = -\Omega_0^2 \mathbf{R}_0,$$

where R is the distance between the comet and the Galactic axe of rotation, z is the third coordinate of the comet in any frame, and $\mathbf{R}_0$ is the Galactic center—Sun vector.

The force per unit of mass on a test particle orbiting the Sun is then given by

$$\mathbf{F} = -\frac{\mu M_\odot}{r^3}\mathbf{r} + \mathbf{f}_f - \mathbf{f}_\odot,$$

where $\mathbf{r}$ is the Sun–comet vector of length r, $M_\odot$ the mass of the Sun, and μ the gravitational constant.

Hence, one has

$$\mathbf{F} = -\frac{\mu M_\odot}{r^3}\mathbf{r} - \nabla U_g + \Omega_0^2 \mathbf{R}_0. \tag{1}$$

Taking into account that the angular speed of the Sun is given by

$$\Omega_\odot = \left[\frac{1}{R}\frac{\partial U_g(R,0)}{\partial R}\right]_{R_\odot}^{1/2},$$

Equation (1) writes

$$\mathbf{F} = -\frac{\mu M_\odot}{r^3}\mathbf{r} - \left[\frac{\partial^2 U_g}{\partial R^2}\right]_{R_0} x'\hat{x}' - \left[\frac{1}{R}\frac{\partial U_g}{\partial R}\right]_{R_0} y'\hat{y}' - \left[\frac{\partial^2 U_g}{\partial z^2}\right]_{R_0} z\hat{z}, \tag{2}$$
$$+0(x^2, y^2, z^2),$$

where $(x, y, z)^{\mathrm{T}}$ and $(x', y', z)^{\mathrm{T}}$ are the coordinates of the particle in the non-rotating and rotating frame, respectively (thus $x' = x\cos(\Omega_0 t) + y\sin(\Omega_0 t)$ and $y' = -x\sin(\Omega_0 t) + y\cos(\Omega_0 t)$).

Let us consider the usual Oort cloud constants defined by

$$A = -\left[\frac{R}{2}\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right]_{R_0},$$
$$B = -\left[\Omega\frac{R}{2}\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right]_{R_0},$$

and the total density in the Solar neighborhood ρ_0 obtained through the Poisson's equation:

$$4\pi\mu\rho_0 = \left[\frac{\partial}{\partial R}\left(\frac{\partial U_g}{\partial R}\right) + \frac{\partial^2 U_g}{\partial z^2}\right]_{R_0} = 2(B^2 - A^2) + \left[\frac{\partial^2 U_g}{\partial z^2}\right]_{R_0}.$$

Then, (2) becomes

$$\mathbf{F} = -\frac{\mu M_\odot}{r^3}\mathbf{r} + (A-B)(3A+B)x'\hat{x}' - (A-B)^2 y'\hat{y}' - [4\pi\mu\rho_0 - 2(B^2 - A^2)]z\hat{z},$$

where the terms of order 2 in x, y, and z are neglected.

The force $\mathbf{F}$ may be further simplified using the constants $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ given by

$$\begin{aligned}\mathcal{G}_1 &= -(A-B)(3A+B)\\ \mathcal{G}_2 &= (A-B)^2\\ \mathcal{G}_3 &= 4\pi\mu\rho_0 - 2(B^2-A^2).\end{aligned}$$

Hence

$$\mathbf{F} = -\frac{\mu M_\odot}{r^3}\mathbf{r} - \mathcal{G}_1 x'\hat{x}' - \mathcal{G}_2 y'\hat{y}' - \mathcal{G}_3 z\hat{z}, \tag{3}$$

from which the Cartesian equations of motion may be easily obtained. It is common to call the component of the tide in the Galactic mid-plane depending on $\mathcal{G}_1$ and $\mathcal{G}_2$ the *radial component*, and the component normal to the Galactic mid-plane depending on $\mathcal{G}_3$ the *normal component*.

As regards the values of the Oort constants A and B, and of the Galactic density ρ_0, the following values are commonly used [27]: $\rho_0 = 0.1\ M_\odot\ \mathrm{pc}^{-3}$, an angular velocity of the Sun around the Galaxy center $\Omega_0 = B - A = -26\ \mathrm{km\ s}^{-1}\ \mathrm{kpc}^{-1}$ and with the approximation $A = -B$ (thus $\mathcal{G}_1 = -\Omega_0^2$). Thus, one gets

$$\begin{aligned}\mathcal{G}_1 &= -7.0706 \times 10^{-16}\mathrm{years}^{-2}\\ \mathcal{G}_2 &= -\mathcal{G}_1\\ \mathcal{G}_3 &= 5.6530 \times 10^{-15}\mathrm{years}^{-2}.\end{aligned}$$

The values of $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ tell us that the radial component of the tide is almost ten times smaller than the normal one. This is the reason why many authors neglected the radial component, i.e., $\mathcal{G}_1 = \mathcal{G}_2 = 0$. The special case where $\mathcal{G}_1 = \mathcal{G}_2 = 0$ will be investigated in Sect. 2.2.

As an alternative to the Cartesian coordinates, one can also use an Hamiltonian formalism to write down the equations. The complete Hamiltonian writes

$$\begin{aligned}H &= H_{\mathrm{kep}} + H_{\mathrm{tide}},\\ H_{\mathrm{kep}} &= = -\frac{\mu M_\odot}{2a},\\ H_{\mathrm{tide}} &= \mathcal{G}_1\frac{x'^2}{2} + \mathcal{G}_2\frac{y'^2}{2} + \mathcal{G}_3\frac{z^2}{2}.\end{aligned}$$

Some useful Hamiltonian variables are the so-called Delaunay's elements L, G, Θ, M, ω, Ω defined by

$$\begin{aligned}M,\quad & L = \sqrt{\mu M_\odot a}\\ \omega,\quad & G = L\sqrt{1-e^2}\\ \Omega,\quad & \Theta = G\cos i,\end{aligned}$$

where M is the mean anomaly of the comet, a the semi-major axis, e the eccentricity, ω the argument of perihelion, and Ω the longitude of the ascending node. The Hamiltonian equations are then obtained from

$$\begin{aligned} \frac{\mathrm{d}M}{\mathrm{d}t} &= \frac{\partial H}{\partial L}, & \frac{\mathrm{d}L}{\mathrm{d}t} &= -\frac{\partial H}{\partial M}, \\ \frac{\mathrm{d}\omega}{\mathrm{d}t} &= \frac{\partial H}{\partial G}, & \frac{\mathrm{d}G}{\mathrm{d}t} &= -\frac{\partial H}{\partial \omega}, \\ \frac{\mathrm{d}\Omega}{\mathrm{d}t} &= \frac{\partial H}{\partial \Theta}, & \frac{\mathrm{d}\Theta}{\mathrm{d}t} &= -\frac{\partial H}{\partial \Omega}. \end{aligned} \tag{4}$$

For a typical Oort cloud comet, one has $H_{\mathrm{tide}}/H_{\mathrm{kep}} \sim 10^{-3}(a/20{,}000)^3$, where a is the semi-major axis of the comet, thus one may neglect short-period perturbations and average the Hamiltonian H over one orbital period with respect to the mean anomaly. One obtains the averaged Hamiltonian $\langle H \rangle$ given by

$$\begin{aligned} \langle H \rangle = -\frac{\mu^2}{2L^2} + \frac{L^4}{4\mu^2}\Bigg\{ & \mathcal{G}_3\Big(1-\frac{\Theta^2}{G^2}\Big)\Big[\frac{G^2}{L^2}+5\Big(1-\frac{G^2}{L^2}\Big)\sin^2\omega\Big] \\ & +\Big(\mathcal{G}_1\cos^2\Omega_r+\mathcal{G}_2\sin^2\Omega_r\Big)\Big[\frac{G^2}{L^2}+5\Big(1-\frac{G^2}{L^2}\Big)\cos^2\omega\Big] \\ & +\Big(\mathcal{G}_1\sin^2\Omega_r+\mathcal{G}_2\cos^2\Omega_r\Big)\Big[\frac{G^2}{L^2}+5\Big(1-\frac{G^2}{L^2}\Big)\sin^2\omega\Big]\frac{\Theta^2}{G^2} \\ & -10\Big(\mathcal{G}_1-\mathcal{G}_2\Big)\Big(1-\frac{G^2}{L^2}\Big)\cos\omega\sin\omega\cos\Omega_r\sin\Omega_r\frac{\Theta}{G}\Bigg\}. \end{aligned} \tag{5}$$

The averaged Hamiltonian equations of motion may be deduced from (4) using the averaged Hamiltonian $\langle H \rangle$. The solutions will correspond to averaged elements of the cometary orbit. But, as far as Oort cloud comets dynamics is concerned, we are more interested in long-term behavior and statistical properties rather than in high accuracy. Hence, in this frame, averaged elements are a good approximation of the osculating elements of the cometary orbit at any time (except for the mean anomaly obviously) as long as they do not diverge from osculating elements and conserve the same statistical properties on long timescale. For the complete equations of the Hamiltonian system one is referred to [20], where different sets of Hamiltonian equations are also used.

From a computational point of view it turns out that a Lie–Poisson formalism, which is more general than the Hamiltonian one, allows to build a very efficient integrator. For a general discussion on the numerical integration of the galactic tide effects on Oort cloud comets one should read [7, 19].

The dynamics generated by the averaged equations of motion is also discussed in [6] where the existence and stability of stationary orbits are investigated.

2.2 The Integrable Case

From the averaged Hamiltonian $\langle H \rangle$ given in (5), one easily deduces that L is conserved with time (see (4)). In addition, when the radial component of the tide is neglected (which corresponds to $\mathcal{G}_1 = \mathcal{G}_2 = 0$), Ω cancels in the averaged Hamiltonian, thus the third component of the angular momentum Θ is also conserved.

Consequently, one gets three integrals of motion which means that under the hypothesis that (i) the average of the Hamiltonian is justified and (ii) the radial component is negligible, the dynamics is completely integrable.

In this case, one notes that L and Θ being constant, the dynamics is completely described in the G, ω variables. The Hamiltonian equations of motion for these variables can be writen as

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\mathcal{G}_3 \frac{5\,L^2}{4\,\mu^2} \left(L^2 - G^2\right) \left(1 - \frac{H^2}{G^2}\right) \sin 2\omega, \tag{6}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \mathcal{G}_3 \frac{L^2\,G}{2\,\mu^2} \left[1 - 5\,\sin^2\omega \left(1 - \frac{L^2\,H^2}{G^4}\right)\right]. \tag{7}$$

The dynamics generated by (6) and (7) was extensively studied by analytical and numerical methods [4, 5, 8, 23, 29, 30]. We will summarize the main implications of the integrability when long-term dynamics is concerned.

Figure 2 shows the phase portrait of the dynamics in the $(\omega, G/L)$ plane with $\Theta/L = \sqrt{1-e^2}\cos i = 1.7321 \times 10^{-2}$ and $a = 30{,}000$ AU. The value of Θ/L tells us that for these orbits [8, 30], the lowest bound of the minimum of the perihelion distance may be equal to 4.5 AU [8, 30].

One notes that different families of orbits may be observed (stationary, circulating, librating). A detailed discussion on the different families of orbit may be found in [5, 8]. Each orbit has a perihelion distance librating with a period given by [8, 29, 30]:

$$P_{\mathrm{peri}} = \frac{\mathbf{K}(k)}{\gamma}, \tag{8}$$

where $\mathbf{K}$ is the complete elliptic function of the first kind and:

$$k^2 = \frac{\min(G_0^2, G_2^2) - G_1^2}{\max(G_0^2, G_2^2) - G_1^2},$$

$$\gamma = \frac{2\pi\rho_0 L^2 \sqrt{\max(G_0^2, G_2^2) - G_1^2}}{\mu},$$

$$G_0^2 = \Theta^2 + \left(1 - \frac{\Theta}{G}\right)\left(G^2 + 5\,(L^2 - G^2)\sin^2\omega\right),$$

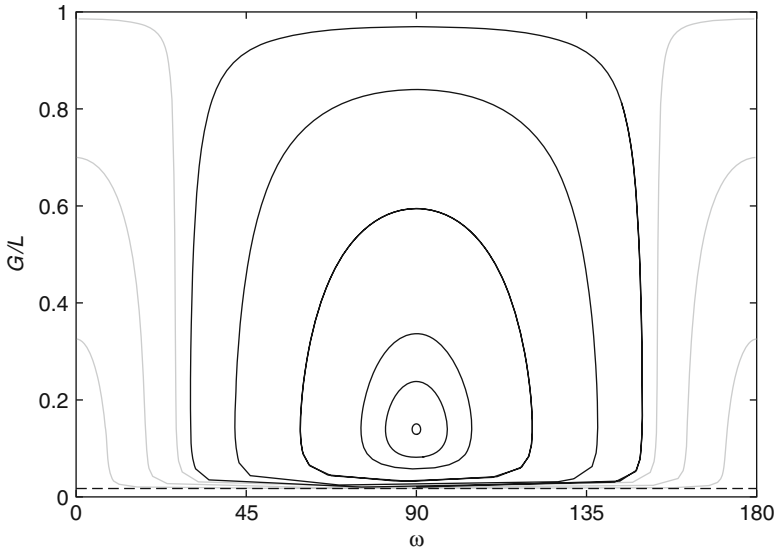

Fig. 2 Phase portrait of the dynamics in the $G/L, \omega$ plane with $\Theta/L = 1.7321 \times 10^{-2}$ and $a = 30{,}000$ AU. The two main families of orbits are shown: *gray curves* correspond to circulating argument, whereas *black curves* correspond to librating argument, The *horizontal dotted line* in the bottom of the frame corresponds to $G/L = 1.7321 \times 10^{-2}$

$$G_1^2 = \frac{1}{8}\left(5\,L^2 + 5\,\Theta^2 - G_0^2 - \sqrt{(5L^2 + 5\Theta^2 - G_0^2)^2 - 80L^2\,\Theta^2}\right),$$
$$G_2^2 = \frac{1}{8}\left(5\,L^2 + 5\,\Theta^2 - G_0^2 + \sqrt{(5L^2 + 5\Theta^2 - G_0^2)^2 - 80L^2\,\Theta^2}\right).$$

Figure 3 shows the period of the perihelion cycle P_{peri} given by (8) versus the minimum value reached by the perihelion over one cycle $q_{\min}$ which is directly obtained from G_1 which is the minimum of G over one cycle [30]. The gray part of the curve corresponds to circulating argument of perihelion, whereas the black part to librating argument of perihelion. One notes that the period P_{peri} goes to infinity for orbits getting closer to the homoclinic motion between orbits with circulating argument of perihelion and orbits with librating argument [8]. The black horizontal dotted line on Fig. 3 corresponds to the minimal value $P_{\text{peri}}^{\min}$ of P_{peri} when $a = 30,000$ AU. This lower bound is given by [30]:

$$P_{\text{peri}}^{\min} = \frac{\pi}{2\sqrt{5}\mu\rho_0} \cdot \frac{1}{P_{\text{orbi}}}, \tag{9}$$

where P_{orbi} is the orbital period of the comet. Thus $P_{\text{peri}}^{\min}$ depends only on the semi-major axis a and is proportional to P_{orbi}^{-1}.

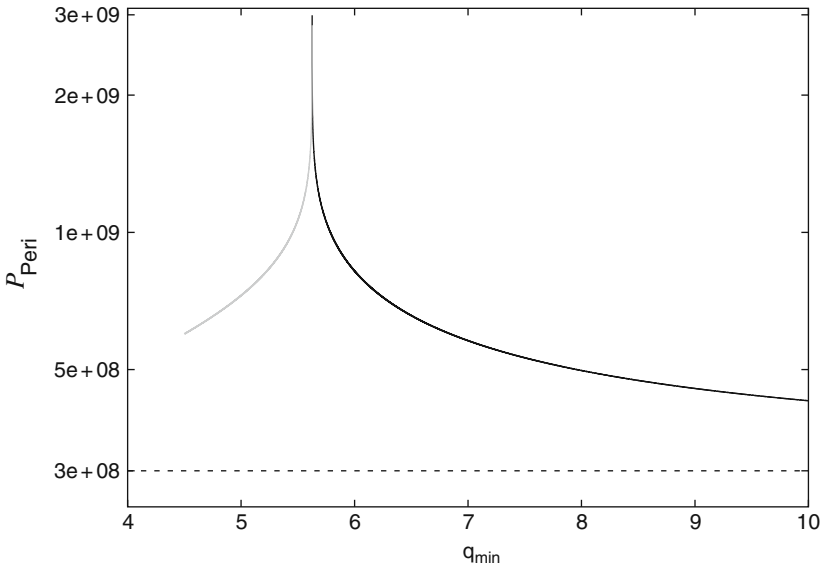

Fig. 3 Period of the perihelion cycle versus the minimum value reached by the perihelion on one cycle. The *gray part* of the curve correspond to circulating argument of perihelion, whereas the *black part* to librating argument of perihelion. The *dotted curve* is the minimal value $P_{\mathrm{peri}}^{\mathrm{min}}$ of P_{peri} when $a = 30,000$ AU

Consequently, under the assumption that the dynamics is integrable, when long-term effects of the Galactic tide are concerned, there are two main points that one should keep in mind [29, 30]:

(i) the minimum heliocentric distance reached by the perihelion of a comet may be computed analytically from the initial conditions,
(ii) the lower bound $P_{\mathrm{peri}}^{\mathrm{min}}$ of the perihelion cycles is proportional to P_{orbi}^{-1}, where P_{orbi} is the orbital period of the comets.

An illustration of these properties may be observed on Fig. 4 where the evolution of the semi-major axis and the perihelion distance versus time of two comets are shown. All the initial elements are the same except the semi-major axes which are, respectively, equal to 30,000 AU (top panel) and 20,000 AU (bottom panel). For both comets the semi-major axis remains constant for 5 Gyr. The perihelion distance librates perfectly with a longer period for the comet having the lower semi-major axis and will never be smaller than the critical value $q_c = 15$ AU (see later for the choice of this value).

Let us now consider the effects of an integrable tide(*the radial component of the tide is neglected*) on a hypothetical Oort cloud of comets.

We consider as [36] a thermalized Oort cloud of 10^6 comets with the following initial osculating elements:

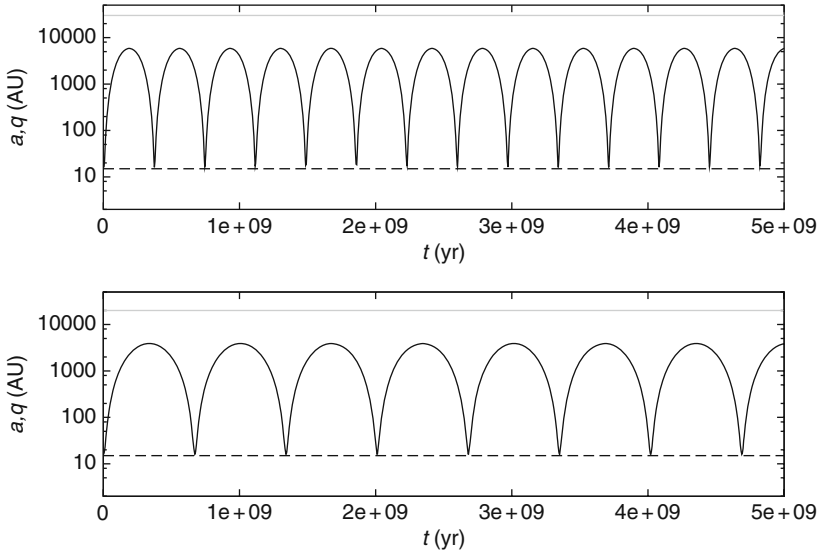

Fig. 4 Semi-major axis a (*gray curve*) and perihelion distance q (*black curve*) versus time for the same initial condition except for the semi-major axis a_0: *top panel:* $a_0 = 30{,}000$ AU, *bottom panel:* $a_0 = 20{,}000$ AU. The *dotted curve* corresponds to $q_c = 15$ AU

- the initial semi-major axes a_0 are between 3000 and 100,000 AU with a distribution $\propto a_0^{-1.5}$;
- the starting eccentricity e_0 is chosen with a density probability $\propto e_0$;
- the initial inclination i_0 is such that the distribution of $\cos i_0$ is uniform between -1 and 1;
- the initial argument of perihelion ω_0, longitude of ascending node Ω_0, and the mean anomaly M_0 are uniformly distributed in the range $0 - 2\pi$.

The comets are integrated during 5 Gyr unless the heliocentric distance of a comet becomes $r < q_c = 15$ AU (the comet is lost due to planetary perturbations) or the comet reaches $r = 4 \times 10^5$ AU (it escapes directly into the interstellar space).

Since only the vertical tide is considered and the averaging over one orbital period is valuable, i.e., the dynamics is integrable. Figure 5 shows the number of comets entering the target region (heliocentric distance smaller than 15 AU) and the observable region (heliocentric distance smaller than 5 AU) per period of 50 Myr versus time. For each period of 50 Myr, the black area is proportional to the number of comets with semi-major axis $a < 20{,}000$ AU—the *inner* Oort cloud—, the gray area is proportional to the number of comets with semi-major axis $20{,}000 < a < 50{,}000$ AU—*central* cloud—and the white area to the comets with $a > 50{,}000$ AU or on hyperbolic orbit—the *outer Oort* cloud (in the present case the conservation of a prohibits the existence of such orbits).

If one first considers the flux toward the target zone, one clearly observes a decreasing of the flux with time. In addition the decrease is faster for the outer

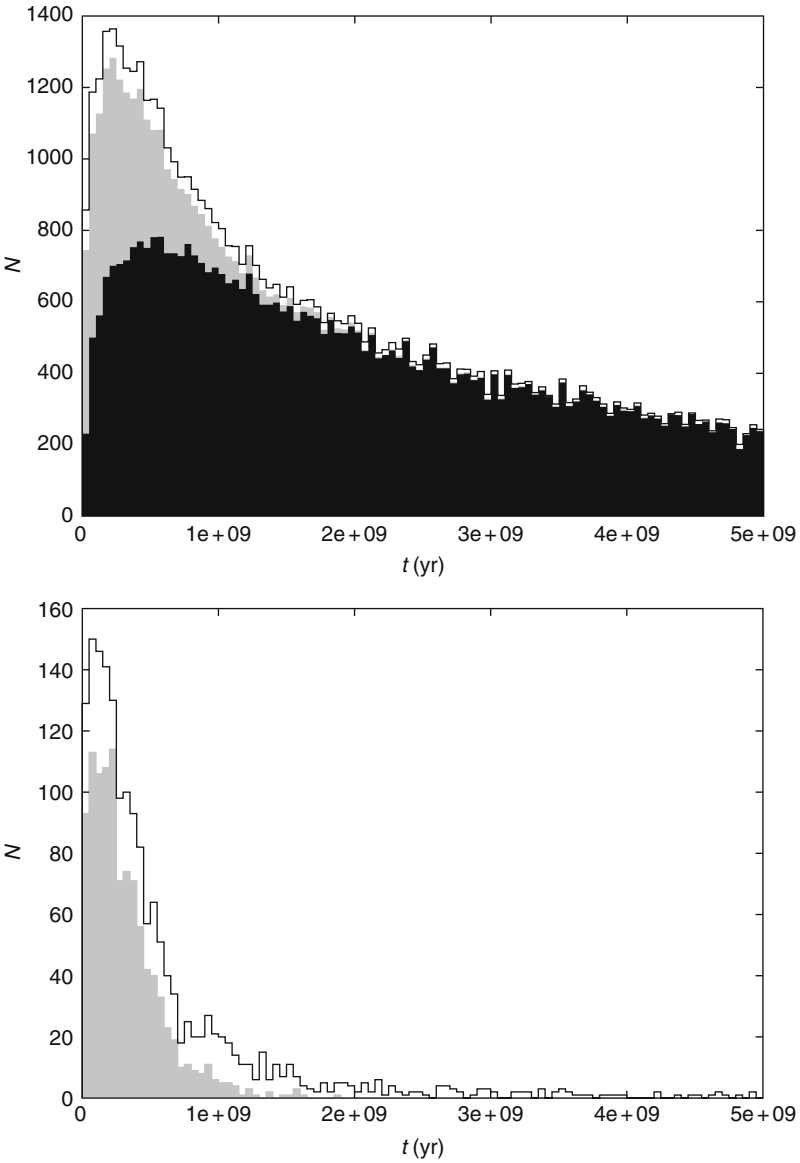

Fig. 5 *Top panel*: number of comets entering the target region ($q < 15$ AU) per period of 50 Myr versus time. *Bottom panel*: number of comets entering the observable zone ($q < 5$ AU) per period of 50 Myr versus time. For each period of 50 Myr, the *black area* is proportional to the number of comets with semi-major axis $a < 20{,}000$ AU (the *inner* Oort cloud), the gray area is proportional to the number of comets with semi-major axis $20{,}000 < 50{,}000$ AU (*central* cloud), and the *white area* to the comets with $a > 50{,}000$ AU (*outer* cloud)

parts of the cloud rather than for the inner parts. This is easily explained from the two points (i) and (ii) stated previously. Indeed, only the infeed trajectories, i.e., the trajectories for which the minimum of the perihelion distance $q_{\min}$ over one cycle is smaller than 15 AU, may enter the target zone. Consequently, when time goes on, these comets enter the target zone and are removed from the Oort cloud. Thus the number of infeed trajectories in the Oort cloud decreases with time, which induces the decrease of the flux toward the target zone. In addition, we have seen that the lower bound of the periods of the perihelion cycle is proportional to P_{orbi}^{-1}, thus the majority of the infeed trajectories with large semi-major axis will enter the target zone more rapidly than infeed trajectories with moderate semi-major axis. This last point explains why the decrease is steeper for the outer cloud rather than for the inner cloud.

As regards the flux toward the observable region, the same arguments explain the decrease of the flux. In addition one should note that, because the entrance in the target zone corresponds to an end state, the comets in the observable zone have had their perihelion distance decreased from outside the target zone to inside the observable zone in less than one orbital period. However, the variation of the perihelion distance over one orbital period is proportional to $a^{7/2}$ (see (6)), thus only the comets with a sufficiently large semi-major axis may enter the target region. This is why, no comet from the inner cloud is found in the observable region.

One notes a marginal flux from the outer cloud until the end of the integration. This contradicts that the flux from the outer cloud should decrease faster than the flux from the central cloud. However, this is easily explained by the fact that the heliocentric distance of the comets itself is considered, rather than the perihelion distance, for the flux toward the observable and the target zone. Let $q_{\min}$ be the minimal value of the perihelion distance over one cycle, if $q_{\min} < q_c$ the lapse of time Δt_{peri} during which the perihelion distance $q < q_c$ may be considered as a fraction of P_{peri}, that is, $\Delta t_{\text{peri}} = c \cdot P_{\text{peri}}$, where c may be considered as independent of a. Similarly, when $q < q_c$, the lapse of time Δt_{orbi} per orbital period during which the comet is at heliocentric distance smaller than q_c may be considered as independent of a, since all the orbits are quasi-parabolic. Thus, on an average, during a lapse time T the comet is at heliocentric distance smaller than q_c for $c\, P_{\text{peri}}\, \Delta t_{\text{orbi}}\, T / P_{\text{orbi}}$.

In conclusion, using $P_{\text{peri}}^{\min}$ instead of P_{peri}, and using (9), when $q_{\min} < q_c$, the lapse of time during which the comets is at heliocentric distance smaller than q_c over one orbital period is proportional to P_{orbi}^{-1}.

Because the number of orbital periods during a fixed lapse time T is also proportional to P_{orbi}^{-1} one can say that, statistically, the lapse of time spent by a comet at heliocentric distance smaller than q_c when $q_{\min} < q_c$ during T is proportional to P_{orbi}^{-2}. In other words, it will need more time for the comet with a large semi-major axis to get in the target zone rather than for comets with low semi-major axis.

2.3 Non-Integrable Case

The assumptions over which the integrability relies are usually not fulfilled. First of all, the radial component is not negligible; and furthermore, for a comet with a

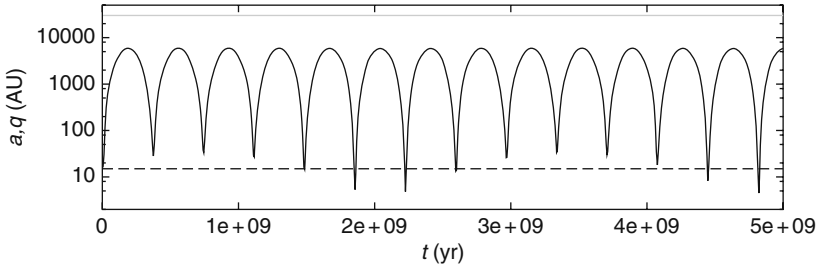

Fig. 6 Semi-major axis a (*gray curve*) and perihelion distance q (*black curve*) versus time for the same initial condition as for Fig. 4 (top) but when the effects of the tide is due to its two component

semi-major axis $a \sim 50{,}000$ AU, at aphelion one may have $H_{\rm tide}/H_{\rm kep} \sim 0.125$; which means that the average is not justified anymore.

One example for which the integrability is broken is shown on Fig. 6. The comet with the same initial conditions as that of Fig. 4 (top) is integrated over 5 Gyr, but both components of the tide are taken into account and without averaging the equations of motion. Contrarily to Fig. 4, the minima of the perihelion distance q are not equal, since the radial component has broken the integrability of the system. In the case of Fig. 6 the perihelion distance becomes smaller than 15 AU after 1.5 Gyr and may become as small as 4.5 AU after 4.8 Gyr.

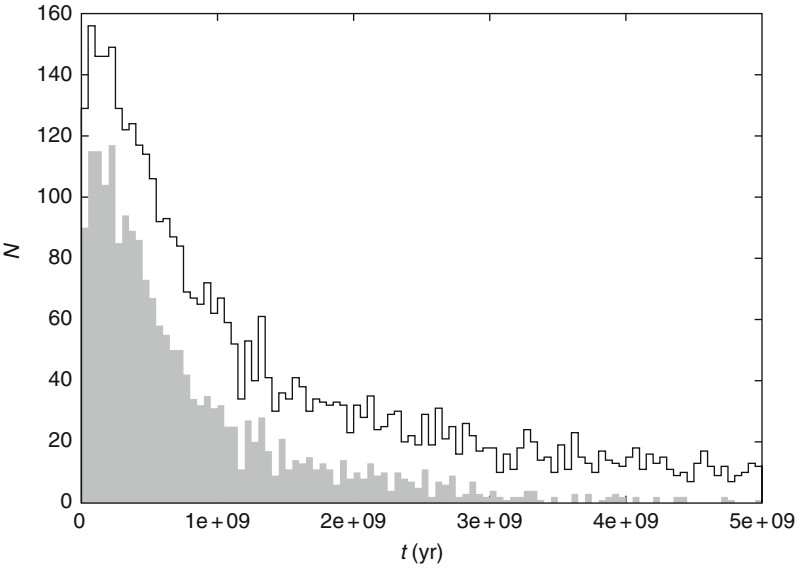

Fig. 7 Number of comets entering the observable zone ($q < 5$ AU) per period of 50 Myr versus time. For each period of 50 Myr, the gray area is proportional to the number of comets with semi-major axis $20{,}000 < a < 50{,}000$ AU (central cloud) and the *white area* to the comets with $a > 50{,}000$ AU. No comet with $a < 20{,}000$ AU is injected in the observable region

Now the same simulation as in Sect. 2.2 is performed but considering a full tide (the vertical and radial components are taken into account) rather than an integrable tide. Figure 7 shows the flux of comets toward the observable region per period of 50 Myr versus time, i.e., Fig. 7 is the same as Fig. 5 (bottom) but considering the two components of the tide.

There still be no comets coming from the inner cloud. Indeed, the fact that the perturbation of the perihelion distance over one orbital period is proportional to $a^{7/2}$ is still valid for the full tide, consequently the tide is still unable to send *directly* comets from the inner cloud into the observable region. The fast decrease of the flux in less than 2 Gyr for the *central* cloud shows that the discussion presented in Sect. 2.2 for the integrable tide is still valid. However, this time there is a marginal flux until quite the end of the integration which was not found in Fig. 4 (bottom). This marginal flux is due to the fact that for the *central* Oort cloud the radial component of the tide breaks the integrability [20].

For the outer cloud, the decrease of the flux is gradual until 2.5 Gyr and almost constant until the end of the integration. This behavior shows that the previous discussion made in Sect. 2.2 is not valid any more. Indeed, for the outer Oort cloud the integrability is broken by both the presence of the radial component and the fact that the procedure of averaging is not valid anymore. As a consequence the infeed trajectories toward the observable region are efficiently refilled, which induces that the flux is quite constant during the last 2.5 Gyr.

3 Stellar Perturbations

3.1 The Stellar Impulse on a Cometary Orbit

As underlined in the introduction random passing stars are external perturbers which may affect the dynamical evolution of Oort cloud comets. Oort [32] considered that the stars are the only perturbers able to inject a comet from the Oort cloud to the planetary region, decreasing drastically its perihelion distance. Close or penetrating passages through the Oort cloud can deflect large numbers of comets on orbits that enter the planetary region and consequently producing a cometary shower [26].

The heliocentric velocity of a star is about several thousands larger than the velocity of a comet. In effect a star has a typical velocity of about 40 km s^{-1}, whereas the heliocentric velocity of Oort cloud comets is of the order of 10 m s^{-1}. Thus the effects of a stellar passage in the Oort cloud was modeled has an impulsion given to the heliocentric velocity of the comets. To estimate stellar perturbations Oort [32] used the so-called impulse approximation, introduced in [33], to investigate the influence of stellar encounters on a cloud of meteoroids or comets. This approximation allows to obtain analytical solutions using simplifying assumptions, namely:

- the star velocity is constant and the motion follows a straight line;
- the star velocity is large enough that the comet and the Sun can be considered to be at rest during the stellar passage.

This approximation was used in a large number of papers (e.g., [2, 26, 28, 35, 41]) and have been found to be useful as a quick estimator in numerical Monte Carlo simulations of cometary orbital evolutions (e.g., [12, 16, 17, 22, 24, 27, 31, 34, 38, 40]).

Under this approximation, the comet is held fixed with respect to the Sun, while the star passes with constant velocity along the straight line defined by the impact parameter b^*, a unit vector $\hat{\mathbf{b}}^*$ that defines the direction of closest approach, and the velocity vector $\mathbf{V}^*$ of the star with respect to the Sun. The impulse of the comet relative to that of the Sun caused by the time-integrated stellar attraction is computed from

$$\Delta\mathbf{v} = \frac{2GM_*}{V^*}\left\{\frac{\hat{\mathbf{b}}_c}{b_c} - \frac{\hat{\mathbf{b}}^*}{b^*}\right\}; \tag{10}$$

adding this value of $\Delta\mathbf{v}$ to the heliocentric velocity of the comet at the orbital position in question, one obtains a new orbital velocity and thus new values of the orbital elements. Figure 8 illustrates the geometry considered. As for the Galactic tide, accurate and fast methods have been developed (see [37] for a detailed discussion).

Figure 9 gives some examples of the effects of two stellar passages on typical Oort cloud comets. The stellar passages were modeled with a more accurate method than the classical impulse for which the stellar heliocentric trajectory is considered as hyperbolic, and the comet is allowed to move on its trajectory [37].

For each panel, the motion of the comet and the star are shown on the plane containing the stellar path and the Sun and such that the star comes from the right with a velocity at infinity anti-parallel to the x axis. The stellar effects are taken into account when the distance of the star to the plane $x = 0$ (the impact plane) is less than 1×10^6 AU. During this period the comet moves on its trajectory. This motion corresponds to the black portion of the comets trajectories shown on Fig. 9. The unperturbed orbits of the comets prior and after the stellar passages are also shown. The two top panels correspond to a star S_{weak} with an impact parameter $b_* =$ 51,000 AU, the velocity at infinity $V^* = 96$ km s^{-1}, and mass 0.9 $M_\odot$, whereas the two bottom panels to a star S_{strong} with impact parameter $b_* = 6600$ AU, $V^* =$

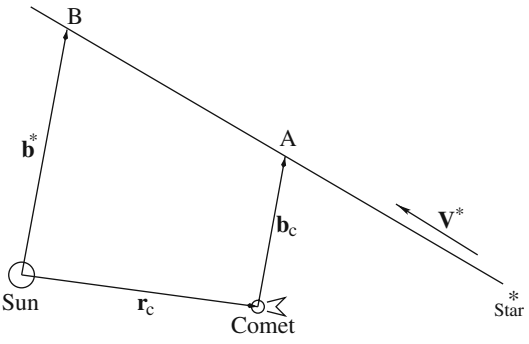

Fig. 8 Geometry of a stellar passage considered for the Classical Impulse Approximation

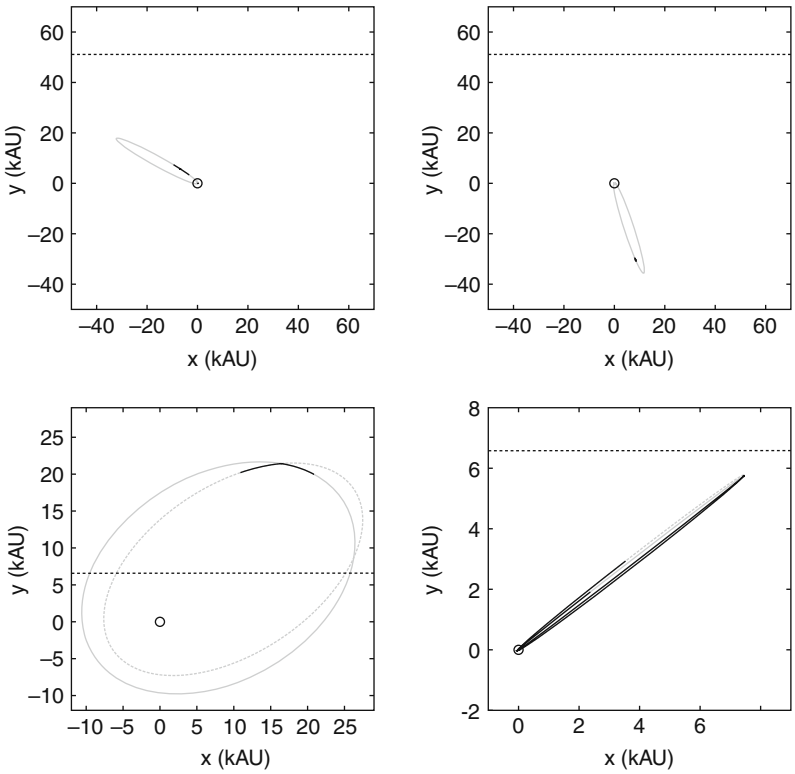

Fig. 9 Stellar perturbations on typical comets of the Oort cloud. The star comes from infinity on the x axis from the right. For each passage the plane contains the star trajectory and the Sun. The *filled gray ellipse* is the unperturbed cometary orbit prior to the stellar perturbations, whereas the *dotted gray ellipse* is the unperturbed trajectory after the stellar passage. The *top panels* correspond to a star of mass 0.9 $M_{\odot}$, an impact parameter $b_* = 51{,}000$ AU and the velocity at infinity $V^* = 96$ km s^{-1}. The two comets have an initial semi-major axis equal to 20,000 AU. The two *bottom panels* correspond to a star of mass 2.1 $M_{\odot}$, an impact parameter $b_* = 6600$ AU, and the velocity at infinity $V^* = 24$ km s^{-1}. The initial semi-major axis and perihelion distance are $a = 20{,}000$ AU, $q = 10{,}000$ AU for the comet of the *bottom left panel* and $a = 5000$ AU, $q = 70$ AU for the *right bottom panel*

24 km s^{-1}, and mass 2.1 $M_{\odot}$. For the star S_{weak} the two comets have the same initial semi-major axis $a = 20{,}000$ AU and the same perihelion distance $q = 100$ AU. For both comets the perihelion perturbation is smaller than 1 AU. For the star S_{strong} the initial semi-major axis and perihelion distance are $a = 20{,}000$ AU, $q = 10{,}000$ AU for the comet of the bottom left panel, and $a = 5000$ AU, $q = 70$ AU for the right bottom panel. The perturbations on the perihelion distance are, respectively, $\Delta q = -2600$ AU and $\Delta q = -66$ AU. One remarks the wide amplitude of these perturbations. The star is even able to decrease the perihelion distance of a comet with initial semi-major axis $a = 5000$ AU from 70 AU to ~ 4 AU. Such passages

are rather rare and are responsible of the so called comets showers [26] during which comets from the inner Oort cloud may become observable.

From (10) one easily sees that the stars are mainly able to change the angular momentum $\mathbf{h} = \mathbf{r} \times \mathbf{v}$ of a comet, and in a minor importance, the orbital energy. When the encounter is distant with respect to the sun ($b^* > b_c$), this change is proportional to $M_*/(V^* b_c)$. Consequently, if one assumes a uniform spatial distribution of the comets in the Oort cloud, the number of comets for which the stellar impulse will be greater that an arbitrary threshold will be proportional to $(M_*/V^*)^2$. It means that apart of the impact distance of the star with the Sun, the stellar mass and the stellar velocity are key parameters to measure or predict the strength of a comet shower.

3.2 Cumulative Stellar Impulsions

Obviously during its life time, the Solar System and its Oort cloud suffer many close encounters with stars. When recent history or close future are investigated one may relay on observations to deduce the sequence of encounters of the Solar System with the neighboring stars [21]. However, on long timescales (of the order of 1 Gyr), it is obviously impossible to know the encounters of the Solar System with stars. Consequently, if one wants to investigate the evolution of the Oort cloud during a timescale of 5 Gyr, one should make some hypothesis, namely that (i) the Solar neighborhood is statistically constant over its lifetime and that (ii) the neighborhood at present time may be used to build a statistical sequence of stellar encounters of the Solar System during its lifetime.

In this way, we have built a set of 197,906 stellar encounters. The encounters occur at random times during a lapse of $t_{max} = 5 \times 10^9$ yr, with random solar impact parameters up to $d_{max} = 4 \times 10^5$ AU, and with random stellar masses and velocities. We use 13 categories of stellar type as in [38] with parameters listed in Table 1. To each category we associate one value of the stellar mass. These masses are generally taken as those of the archetypal spectral classes along the main sequence (see Table 1) according to [1]. The relative encounter frequencies f_i of Table 1 are taken from [21], where they were derived from the respective products of number density and mean velocity, $n_i \langle v_i \rangle$. A random number ξ_i is used to pick a stellar category i with the probability $f_i / \sum f_i$.

Considering the above sequence of stellar encounters on the comet already used in Sect. 2 with semi-major axis $a = 30{,}000$ AU, one obtains the evolution of semi-major axis and perihelion distance versus time shown on Fig. 10. One notes that the perihelion distance and the semi-major axis show a behavior similar to a random motion, which highlights the stochastic aspect of the stellar perturbations. In particular, the perihelion of the comet becomes as small as 5 AU in less than 200 Myr and then takes value greater than 1000 AU in a 100 Myr lapse time. As regards the semi-major axis, the comet starts in the central cloud, but after 3 Gyr it has shifted to the inner cloud with semi-major axis $\sim$ 18,000 AU. This example shows that the stellar perturbations are able to induce wide drifts of the perihelion distance on

Table 1 Stellar parameters. The types are mostly MK types for main sequence stars; "wd" indicates white dwarfs and "gi" indicates giant stars. The encounter frequencies are given in number per Myr within 1 pc. The following two columns list the solar apex velocity with respect to the corresponding type and the spherical Maxwellian velocity dispersion. The last two columns give the mean heliocentric encounter velocity and its standard deviation according to our results

Type	Mass ($M_\odot$)	Enc. freq	$v_\odot$ (km/s)	σ_* (km/s)	$\langle V \rangle$ (km/s)	σ_V (km/s)
B0	9	0.005	18.6	14.7	24.6	6.7
A0	3.2	0.03	17.1	19.7	27.5	9.3
A5	2.1	0.04	13.7	23.7	29.3	10.4
F0	1.7	0.15	17.1	29.1	36.5	12.6
F5	1.3	0.08	17.1	36.2	43.6	15.6
G0	1.1	0.22	26.4	37.4	49.8	17.1
G5	0.93	0.35	23.9	39.2	49.6	17.9
K0	0.78	0.34	19.8	34.1	42.6	15.0
K5	0.69	0.85	25.0	43.4	54.3	19.2
M0	0.47	1.29	17.3	42.7	50.0	18.0
M5	0.21	6.39	23.3	41.8	51.8	18.3
wd	0.9	0.72	38.3	63.4	80.2	28.2
gi	4	0.06	21.0	41.0	49.7	17.5

timescales comparable to that of the Galactic tide and also to be responsible of a stochastic diffusion of the cometary semi-major axis.

Let us now model the same sequence of stellar pertubations on our hypothetical initial Oort cloud. Figure 11 shows the comets entering the observable zone per period of 50 Myr (top panel) and 10 Myr (bottom panel) versus time when only the stellar perturbations are included. One notes that the flux is very sporadic and is characterized by a sequence of comets showers occurring at random. The strongest showers are evidenced on the bottom panel. The characteristics of the stars which are responsible of the showers are shown on Table 2. The shower S_2 is due to the star S_{strong} already considered in Sect. 3.1.

However, if one looks to the semi-major axis of the comets injected, one remarks that for the central and outer cloud, the stellar perturbations induce a flux which is less affected by the showers. Indeed, the flux from these regions exhibits rather a

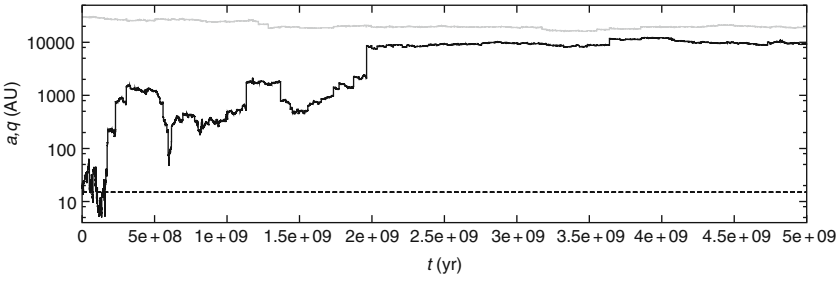

Fig. 10 Semi-major axis a (*gray curve*) and perihelion distance q (*black curve*) versus time for the same initial condition as for Fig. 4 (top)

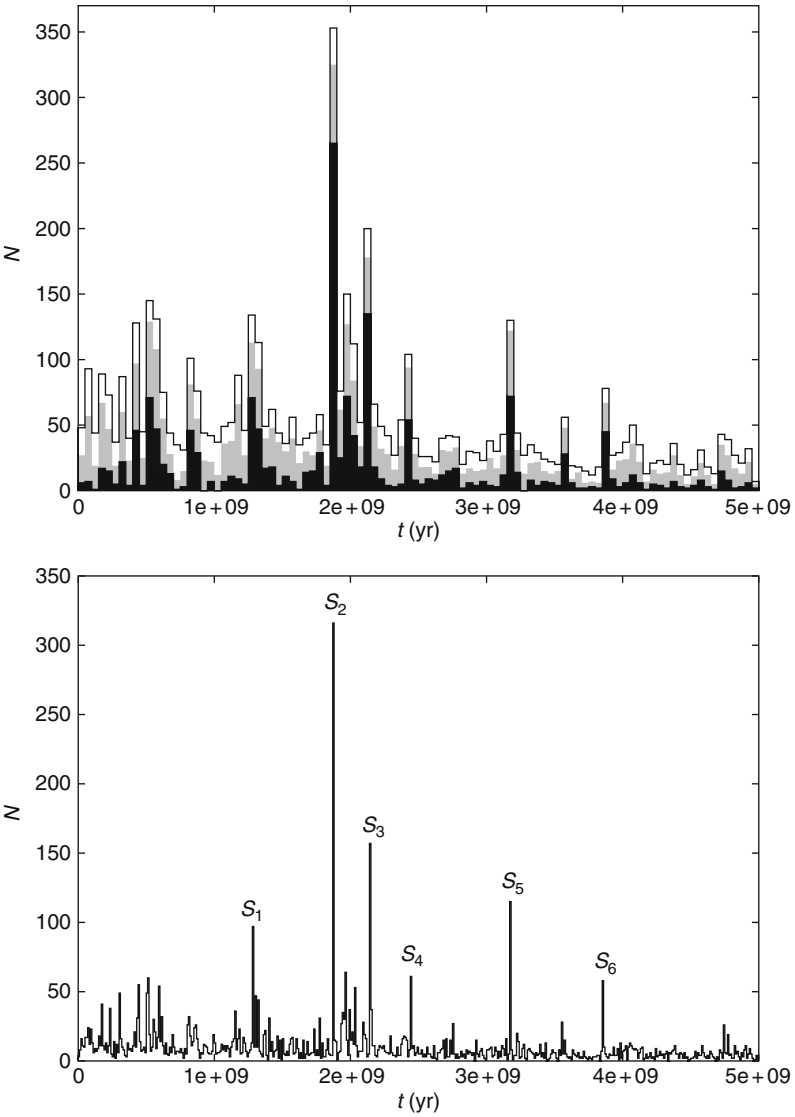

Fig. 11 *Top panel*: same as Fig. 7 but only the stellar pertubations are taken into account. *Bottom panel*: the total flux toward the observable region per period of 10 Myr versus time. The strongest showers are evidenced by letters. See Table 2 for the characteristics of the stars causing the showers

small and regular decrease on long timescale. This decrease is due to the depletion of the Oort cloud with time under stellar perturbations.

On the contrary the inner part of the cloud ($a < 20{,}000$ AU) is very sensitive to close stellar passages. The increase of the flux from this region during a close passage of a star to the Sun lasts for a very short time (less than 10 Myr) and may be several orders of magnitude higher than the flux out of the showers. However, a flux from this region of the cloud is nevertheless observed even when no showers

Table 2 Characteristics of the stars causing the showers shown on the bottom panel of Fig. 11

Star	Type	d_* (AU)	V^* (km s^{-1})
S_1	M5	1400	25
S_2	A5	6600	24
S_3	G5	3150	35
S_4	M5	1700	17
S_5	gi-M5	13,400–2300	37/20
S_6	M5	2000	18

are observable. This distinction between a background flux coming mainly from the central and outer cloud under distant passage of star in the Oort cloud and a sporadic flux coming from the inner cloud during a close passage to the Sun was already identified by [22, 26].

4 The Combined Effects of Galactic and Stellar Perturbations

Let us now consider the dynamics of the Oort cloud comets when both the galactic tide and the stellar perturbations are at work. Figure 12 shows again the evolution of the semi-major axis and the perihelion distance versus time for the same comet as for Fig. 6. The two effects (full tide and stellar pertubations) are now at work. During the first 1.5 Gyr one clearly sees the cycle of the perihelion due to the tide. Obviously the dynamics is not integrable and the minimum that the perihelion distance may reach is now almost a random quantity. In the present case the comet could reach heliocentric distances as small as 0.5 AU because of the combined effects of both the galactic tide and stellar perturbations. After the first 1.5 Gyr the semi-major axis is increasing due to stellar perturbations which affects deeply the regime of the perihelion cycle which becomes much shorter. However, the minimum of the perihelion distance of one cycle is above 100 AU, i.e., very far from our dynamical barrier at 15 AU. This comet is still in the outer Oort cloud after 5 Gyr.

Let us now consider the dynamical evolution of our fictive Oort cloud of 10^6 comets under the combined effects of Galactic tide and stellar perturbations. The upper part of Fig. 13 shows a histogram plot of the number of comets injected

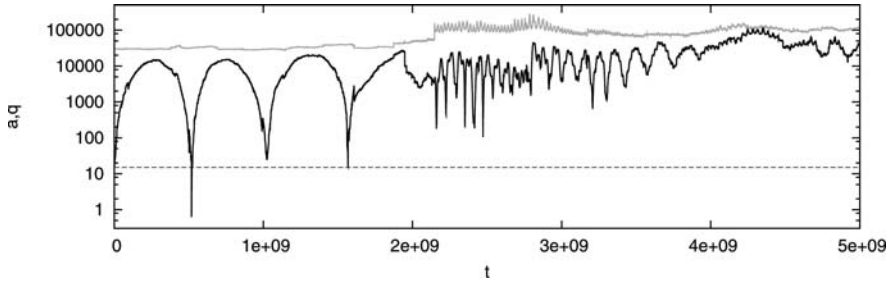

Fig. 12 Semi-major axis a (*gray curve*) and perihelion distance q (*black curve*) versus time for the same initial condition as for Fig. 4 (top)

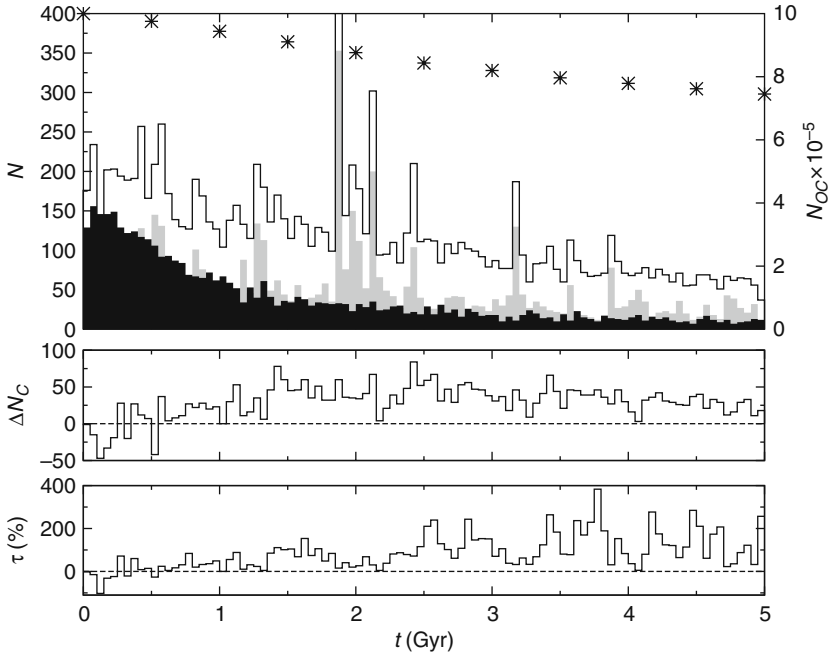

Fig. 13 The *upper diagram* shows the number of comets entering the observable zone per 50 Myr versus time. The *white histogram* corresponds to the combined model, the *black histogram* to the Galactic tide alone, and the *gray histogram* to the passing stars alone. The *asterisks* indicate the number of comets remaining in our simulation for the combined model at every 500 Myr with scale bars to the right. The *middle diagram* shows the excess number of injections into the observable region per 50 Myr in the combined model with respect to the sum of the stars-only and tides-only models. The *lower diagram* shows this excess expressed in percent of the mentioned sum

into the observable region per period of 50 Myr as a function of time from the beginning till the end of the simulation. Three histograms are shown together: the one in black corresponds to the model with only Galactic tides, and the gray one to a model including only stellar perturbations—they correspond to the white histogram of Figs. 7 and 11, respectively. Finally, the top, white histogram corresponds to the combined model that includes both tides and stars.

Because the stars are the same in the two simulations where stellar perturbations are at work, we see the same comet showers appearing and the same quasi-quiescent periods in between. The white area at the top of each bin corresponds to the extra contribution of the combined model as compared with that of the stars only. If the numbers plotted in the white, gray, and black histograms are called N_C, N_S, and N_G, respectively, we can define $\Delta N_C = N_C - N_S - N_G$ as an absolute measure of this extra contribution [1]. Already at first glance, looking at the later part of the

[1] Towards the end of our simulation the number of Oort Cloud comets has decreased in all three models but most in the combined one. We then have about 930, 000, 840, 000, and 760, 000 comets

simulation, we see that this is very significant. In the two lower panels of the figure, we plot histograms of ΔN_C and $\tau = \Delta N_C/(N_S + N_G)$,s i.e, the extra contribution expressed as a fraction of $N_S + N_G$.

The basic observations are as follows. While during the first Gyr the level of N_G is generally higher than that of N_S, this situation gets reversed after more than two Gyr. Even outside the main showers, N_S is then at a somewhat higher level than N_G. The white histogram, showing N_C, shares the spikes of the strongest showers, but the contrast between the spikes and the background is less than in the gray histogram. Indeed, the ΔN_C histogram shows no spikes at all. Therefore, during the later part of the simulation, the τ parameter shows fluctuations anticorrelated with those of N_S. It reaches a few hundred percent, when N_S drops to its lowest levels, but sometimes decreases to nearly zero during the peaks of N_S.

In order to smooth out those fluctuations we present in Table 3 time averages of τ over 1 Gyr periods along with the corresponding integrals of N_C, N_S, and N_G. During the first Gyr the flux of the combined model is not much larger than the sum of the two fluxes with separate effects, and the difference is just a small fraction of the total flux. But toward the end the synergy effect of the combined model, as measured by ΔN_C, has grown—on the average—to nearly the same level as $N_S + N_G$. During the last Gyr we find that $\langle N_C \rangle$ is about 2.5 times larger than $\langle N_S \rangle$ in fair agreement with earlier estimates by [24, 22]. After an initial, relatively fast decrease due to the emptying of the tidal infeed trajectories, $\langle N_C \rangle$ continues to decrease approximately in proportion to the total number of Oort Cloud comets (N_{OC}), and $\langle N_S \rangle$ and $\langle N_G \rangle$ show similar behaviors.

Looking in detail at the ΔN_C and τ histograms in Fig. 13 for the beginning of our simulation, we see that they start from negative values and turn into positive ones after ~ 0.5 Gyr. Thus, in the very beginning, the sum of the separate fluxes is larger than the combined flux. This phenomenon was found by [28], whose calculations were limited to only 5 Myr, and as they explained, it is typical of a situation where both tides and stars individually are able to inject comets into the observable region to a high degree.

Table 3 Number of comets entering the observable region during periods of 1 Gyr. Model G corresponds to the Galactic tide alone, S to passing stars alone, and C to Galactic tide and passing stars together. $\langle \tau \rangle$ is the increment from the sum of the two first rows (Galactic tide plus passing stars separately) to the third row (Galactic tide and passing stars together)

Model	[0 − 1] Gyr	[1 − 2] Gyr	[2 − 3] Gyr	[3 − 4] Gyr	[4 − 5] Gyr
G	2128	797	481	307	248
S	1425	1555	1030	717	511
C	3618	3141	2412	1733	1274
$\langle \tau \rangle$	1.8%	33.6%	59.6%	69.2%	67.9%

in the tides-only, stars-only and combined models, respectively. This means that ΔN_C actually underestimates the extra contribution of the combined model.

The large amount of synergy ($\tau \sim 70\%$) seen in the later part of our simulation is remarkable and indicates that both the tides and the stars on their own are seriously inefficient in injecting comets in the observable region. It is only by means of the synergy of both effects that we are able to explain an important degree of the flux of comets toward the observable region at the current epoch.

The most important synergy mechanism of the Galactic tide and stellar perturbations is that the latter are able to repopulate the critical phase space trajectories that in the quasi-regular dynamics imposed by the tide lead into the observable region [13, 18]. Our results appear to verify and quantify this picture.

Figure 14 shows the distributions of the opposite of the inverse semi-major axis ($-1/a$) and the sine of the Galactic latitude of perihelion (for clarity we use the absolute value $|\sin b|$) of the comets entering the observable region, i.e., heliocentric distance smaller than 5 AU, during a typical 170 Myr interval near the end of our simulation, where no strong comet showers are registered. We show an average of three such periods, i.e., 4.38−4.55 Gyr, 4.55−4.72 Gyr, and 4.80−4.97 Gyr. In fact, comparing the three data sets, we find a rather good agreement, so that tentatively, the expected error of the mean is not very large. Three histograms are shown for each quantity: the one in black is for the model with Galactic tides only, the gray one is for the model with only stellar perturbations, and the white one shows the combined model.

After more than 4 Gyr the Galactic tides alone are practically only able to inject comets into the observable region if $a > 50{,}000$ AU, so that the non-integrable part of the tides may provide new comets into the emptied infeed trajectories of the vertical component. Thus the feeble flux of new observable comets is strictly confined to the outermost parts of the Oort cloud. If only the stellar perturbations are at work, the injected comets are almost as few as in the case of the Galactic tides. However, the distribution of $-1/a$ shows that the stellar perturbations are relatively efficient injectors of comets with semi-major axes in the whole range from 25,000 to more than 100,000 AU, and there is some marginal infeed all the way into the inner core. Note that this concerns a time interval without any strong comet showers.

When both the processes are at work, the number of comets entering the observable zone is 206, about 86% more than the sum of the two separate contributions ($39 + 72$). This estimate of τ is a bit higher than for the entire 1 Gyr interval, listed in Table 3, because the three intervals have been selected as particularly calm, avoiding even the smaller peaks of N_S that can be seen in Fig. 13. We have shown above that larger values of N_S lead to smaller values of τ. The distribution of $-1/a$ is as wide as for the stellar perturbations alone. However, the picture has changed, since the additional 86% of the comets are strongly concentrated to the interval from

Fig. 14 (continued) during 170 Myr near the end of the simulation. When present, numbers in the top-left corners of $-1/a$ distribution panels correspond to comets with $-1/a < -1 \times 10^{-4}$ AU^{-1}. The *left column* corresponds to the model with Galactic tide alone, the *middle column* to passing stars alone, and the *right column* to the model with both effects

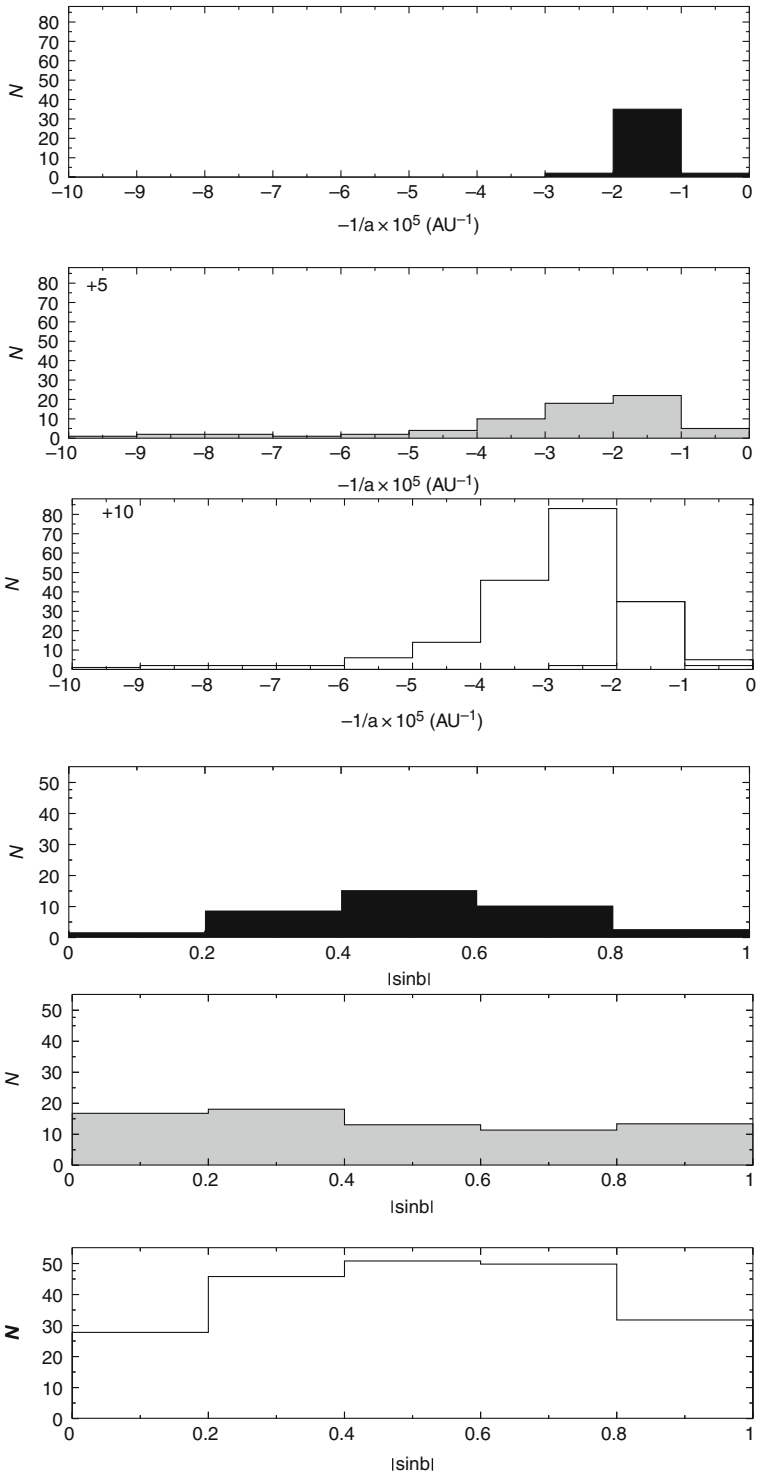

Fig. 14 Distributions of $-1/a$, where a is the semi-major axis (*top panels*), and $|\sin b|$, where b is the Galactic latitude of perihelion (*bottom panels*), for the comets entering the observable region

-4×10^{-5} to -2×10^{-5} AU^{-1} ($25,000 < a < 50,000$ AU). The local values of ΔN_C for the five $1/a$ intervals (0−1), (1−2), (2−3), (3−4), and (4−5)$\times 10^{-5}$ AU^{-1} are −2, −22, +63, +36, and +10, respectively. The negative values of N_C for the two outermost bins of semi-major axis may be explained in the same way as for the negative values of N_C at the begining of the integration, i.e., for the outer Oort cloud both the tides and the stellar perturbations are independently able to inject efficiently comets into the observable region.

We see that the mechanism of synergy that increases the flux of injections in the combined model prefers the range of semi-major axis ($a > 30{,}000$ AU) where the vertical Galactic tide is able to provide the injections, once the relevant trajectories are populated. But there is an important extension of the synergy to smaller semi-major axes as well, extending at least to $a \simeq 20{,}000$ AU. We conclude that another synergy mechanisms must be at work. The repopulation mechanism is obviously important, but the shift to smaller semi-major axes can only be explained by a "constructive interference" mechanism which decrease the threshold from which the tide is able to eject comet toward the observable region due to the help of stellar perturbation (see [36] for more details).

Looking at the distributions of $|\sin b|$, indeed the signature of the Galactic tide is clearly present in the left diagram and absent in the middle one. However, it appears again to some extent in the right-hand diagram, where the combined model is presented. Thus we have evidence that the synergetic injection of comets in the combined model carries an imprint in the latitudes of perihelia similar to that of the Galactic tide, though the feature is strongly subdued.

5 Conclusion

The effects of the Galactic tide and the stellar perturbations on the long-term behavior of Oort cloud comet trajectories have been investigated. Our study has two main limitations. We do not treat encounters with very massive Galactic perturbers, such as star clusters or Giant Molecular Cloud complexes, the justification being that they occur so rarely that the current Solar System is unlikely to feel the direct reverberations of any such encounter and that even if they modify the structure of the Oort cloud, our interest is not primarily in its dynamical history but rather in the way stars and Galactic tides currently interact when injecting observable comets. Moreover, we do not treat planetary perturbations in any direct manner. Like most previous investigators (e.g., [22]) we use a dynamical barrier defined by a limiting perihelion distance (in our case, 15 AU) outside which no planetary effects are included and inside which all comets are considered lost from the cloud through perturbations by Jupiter and Saturn. In terms of "transparency" of the planetary system [14, 15], our model is completely opaque ($P = 1$). This means that we are limiting our attention to a subset of the observed population of "new" Oort cloud comets, i.e., those that have jumped directly from $q > 15$ AU into their observed orbits with $q < 5$ AU.

We have first investigated the effects of the Galactic tide and the stellar perturbations separately in order to highlight their own characteristics. For each perturbator,

after having observed how it acts on individual case, we have simulated the evolution of the Oort cloud over 5 Gyr, using for initial conditions a relaxed model with a distribution of semi-major axis $f(a) \propto a^{-1.5}$ within the interval 3000–100,000 AU. This model is based on the results of simulation of Oort Cloud formation and evolution by [12].

As regards the Galactic tide, the following conclusions may be drawn:

- it generates a quasi-integrable dynamics,
- it induces large oscillation of the perihelion distance, whereas the semi-major axis is almost constant,
- the period of the oscillation is shorter for large semi-major axis,
- the larger the semi-major axis is, the greater is the variation of the perihelion distance over one orbital period.

From these observations, one deduces that

- the efficiency of the tide to inject comet in the observable region increases with the semi-major axis of the comets (no comet with $a < 20{,}000$ AU—the inner Oort cloud—were found in the observable region with the tide only model),
- the feeding zone from which the tide is able to inject comet in the observable region, get depleted as time increases, the outer feeding zone (in the outer Oort cloud: $a > 50{,}000$ AU) being depleted more rapidly than the central ones (in the central Oort cloud: $20{,}000 < a < 50{,}000$ AU).

As regards the stellar perturbations, we have seen that they generate a stochastic process characterized by comet showers due to close encounters with stars. Such stars were able to inject comet into the observable region even from the inner part of the Oort cloud. The total flux of comets to the observable region during the shower being several orders of magnitude higher than when no showers are observed, which correspond to a background flux, mainly coming from the outer and central part of the cloud. The stellar perturbations, on the contrary to the Galactic tides, are able to change the semi-major axis of the comets, inducing an exchange of populations between the different regions of the Oort cloud.

The model with the combined effects of the Galactic tides and the stellar perturbations, has highlight that an efficient synergy takes place. We have indeed shown that, during the later parts of our simulation, there is a very important synergy effect of the Galactic tide and stellar perturbations such that the combined injection rate is on the average $\sim 70\%$ larger than that of the stars alone plus that of the tide alone. This synergy is strongest for semi-major axes between $\sim 20{,}000$ *and* 50,000 AU but continues all the way into the inner core. During comet showers the synergy effect in the outer parts of the cloud practically disappears, but the one affecting the inner parts becomes very important.

We have identified two mechanisms for the synergy during quiescent periods in the outer parts of the Oort Cloud. One is that the stellar perturbations provide a supply of new comets that replenishes the depleted tidal infeed trajectories, and the other is that the gain of comet injections, when stellar perturbations decrease the perihelion distance, dominates over the loss caused by opposing perturbations. For

the synergy of the inner cloud we hypothesize that the Galactic tides provide the material for stellar injections by slowly feeding the region of phase space in the vicinity of the loss cone. Thus, the general picture spawned by our results is that injection of comets from the Oort Cloud is essentially to be seen as a team work involving both tides and stars. It appears meaningless to rank the two effects in terms of strength or efficiency.

Indeed, for the smaller semi-major axes the Galactic tide does not dominate the injection of comets, contrary to the conclusions of [24] and [22][2]. It only contributes to a synergy with stellar perturbations, and without the stars one would not have any injections of comets with $a \lesssim 20,000$ AU.

The distribution of Galactic latitudes of perihelia of the observable comets exhibits a maximum for $|\sin b| \simeq 0.5$ as expected in the tides-only model, but in the combined model this feature can hardly be seen at all. The tides form part of the synergetic injection, but their imprint is largely washed out by the stellar contribution.

References

1. Allen, C.W.: Astrophysical Quantities, 3rd edn. Athlone Press, London (1985) 418
2. Bailey, M.E.: The mean energy transfer rate to comets in the Oort cloud and implications for cometary origins. Mon. Not. R. Astron. Soc. **218,** 1–30 (January 1986) 416
3. Bailey, M.E., Clube, S.V.M., Napier, W.M.: The Origin of Comets, 599 p. Pergamon Press, Oxford, England and Elmsford, NY (1990) 403
4. Brasser, R.: Some properties of a two-body system under the influence of the Galactic tidal field. MNRAS **324,** 1109–1116 (July 2001) 403, 408
5. Breiter, S., Dybczyński, P., Elipe, A.: The action of the galactic disk on the Oort cloud comets. A&A **315**, 618–624 (1996) 403, 408
6. Breiter, S., Fouchard, M., Ratajczak, R.: Stationary orbits of comets perturbed by Galactic tides. MNRAS **383**, 200–208 (2008) 407
7. Breiter, S., Fouchard, M., Ratajczak, R., Borczyk, W.: Two fast integrators for the Galactic tide effects in the Oort cloud. MNRAS **377**, 1151–1162 (2007) 407
8. Breiter, S., Ratajczak, R.: Vectorial elements for the galactic disc tide effects in cometary motion. MNRAS **364**, 1222–1228 (2005) 408, 409
9. Byl, J.: Galactic perturbations on nearly-parabolic cometary orbits. Moon Planets **29**, 121–137 (October 1983) 402, 403
10. Byl, J.: The effect of the Galaxy on cometary orbits. Earth Moon Planets **36**, 263–273 (November 1986) 402, 403
11. Delsemme, A.H.: Galactic tides affect the Oort cloud—an observational confirmation. Astron. Astrophys. **187**, 913–918 (November 1987) 402
12. Duncan, M., Quinn, T., Tremaine, S.: The formation and extent of the solar system comet cloud. Astron. J. **94**, 1330–1338 (November 1987) 402, 403, 416, 427

[2] The main reason for this discrepancy is that the Heisler papers considered injections into the loss cone–mainly by slight perturbations of q across the limiting value $q_c = 10$ AU—while we consider large jumps from $q > 15$ AU into the observable region with $q < 5$ AU. Interestingly, [25] commented that the injection into orbits with $a \lesssim 20,000$ AU and $q < 2$ AU is indeed dominated by stellar perturbations.

13. Dybczyński, P.A.: Simulating observable comets. I. The effects of a single stellar passage through or near the Oort cometary cloud. Astron. Astrophys. **396**, 283–292 (December 2002) 424
14. Dybczyński, P.A.: Simulating observable comets. II. Simultaneous stellar and galactic action. Astron. Astrophys. **441**, 783–790 (October 2005) 426
15. Dybczyński, P.A., Prętka, H.: The galactic disk tidal force: Simulating the observed Oort cloud comets. In: Wytrzyszczak, I.M., Lieske, J.H., Feldman, R.A. (eds.) IAU Colloq. 165: Dynamics and Astrometry of Natural and Artificial Celestial Bodies, p. 149 Kluwer Academic Publishers, Netherlands (1997) 426
16. Fernández, J.A.: Evolution of comet orbits under the perturbing influence of the giant planets and nearby stars. Icarus **42**, 406–421 (June 1980) 402, 416
17. Fernández, J.A.: Dynamical aspects of the origin of comets. Astron. J. **87**, 1318–1332, (1982) 416
18. Fernández, J.A. (ed.) Comets—Nature, Dynamics, Origin and their Cosmological Relevance, volume 328 of Astrophysics and Space Science Library (2005) 402, 424
19. Fouchard, M., Froeschlé, Ch., Breiter, S., Ratajczak, R., Valsecchi, H., Rickman, G.: Methods to study the dynamics of the Oort cloud comets II: Modelling the galactic tide. In: Benest, D., Froeschlé, Cl., Lega E. (eds.) Topics in Gravitational Dynamics, Vol. 729, Lect. Notes Phys., 271–293. Berlin Springer Verlag, in press (2007) 407
20. Fouchard, M., Froeschlé, Ch., Valsecchi, G., Rickman, H.: Long-term effects of the Galactic tide on cometary dynamics. Celest. Mech. Dyn. Astron. **95**, 299–326 (September 2006) 403, 407, 415
21. García-Sánchez, J., Weissman, P.R., Preston, R.A., Jones, D.L., Lestrade, J.-F., Latham, D.W., Stefanik, R.P., Paredes, J.M.: Stellar encounters with the solar system. Astron. Astrophys. **379**, 634–659 (November 2001) 418
22. Heisler, J.: Monte Carlo simulations of the Oort comet cloud. Icarus **88**, 104–121 (November 1990) 416, 421, 423, 426, 428
23. Heisler, J., Tremaine, S.: The influence of the galactic tidal field on the Oort comet cloud. Icarus **65**, 13–26 (January 1986) 402, 403, 404, 408
24. Heisler, J., Tremaine, S., Alcock, C.: The frequency and intensity of comet showers from the Oort cloud. Icarus **70**, 269–288 (1987) 416, 423, 428
25. Heisler, J., Tremaine, S., Weissman, P., Greenberg, R.: Sky Distributions of Oort Cloud Comets During and Outside of Showers. In: Lunar and Planetary Institute Conference Abstracts, volume 22 of Lunar and Planetary Institute Conference Abstracts, p. 553 (March 1991)
26. Hills, J.G.: Comet showers and the steady-state infall of comets from the Oort cloud. Astron. J. **86**, 1730–1740 (November 1981) 402, 415, 416, 418, 421
27. Levison, H.F., Dones, L., Duncan, M.J.: The origin of Halley-type comets: Probing the inner Oort cloud. Astron. J. **121**, 2253–2267 (April 2001) 406, 416
28. Matese, J.J., Lissauer, J.J.: Characteristics and frequency of weak stellar impulses of the Oort cloud. Icarus **157**, 228–240 (May 2002) 416, 423
29. Matese, J.J., Whitman, P.G.: The Galactic disk tidal field and the nonrandom distribution of observed Oort cloud comets. Icarus **82**, 389–401 (1989) 403, 408, 410
30. Matese, J.J., Whitman, P.G.: A model of the galactic tidal interaction with the Oort comet cloud. Celest. Mech. Dynam. Astron. **54**, 13–35 (1992) 403, 408, 409, 410
31. Mazeeva, O.A., Emel'Yanenko, V.V.: Variations of the Oort cloud comet flux in the planetary region. In: Warmbein, B. (ed.), Asteroids, Comets, and Meteors: ACM 2002, volume 500 of ESA Special Publication, pp. 445–448 (November 2002) 416
32. Oort, J.H.: The structure of the cloud of comets surrounding the solar system and a hypothesis concerning its origin. Bull. Astron. Inst. Neth. **11**, 91–110 (January 1950) 401, 415
33. E. Öpik.: Note on stellar perturbations of nearby parabolic orbit. Proceedings of the American Academy of Arts and Science (1932) 415
34. Remy, F., Mignard, F.: Dynamical evolution of the Oort cloud. I. A Monte Carlo simulation. II. A theoretical approach. Icarus **63**, 1–30 (July 1985) 402, 416

35. Rickman, H.: Stellar perturbations of orbits of long-period comets and their significance for cometary capture. Bull. Astron. Inst. Czech. **27**, 92–105 (1976) 402, 416
36. Rickman, H., Fouchard, M., Froeschlé, Ch., Valsecchi, G.B.: Injection of Oort Cloud Comets: The Fundamental Role of Stellar Perturbations. To be published in Celestial Mechanics and Dynamical Astronomy (2008) 410, 426
37. Rickman, H., Fouchard, M., Valsecchi, G.B., Froeschlé, Ch.: Algorithms for stellar perturbation computations on Oort cloud comets. Earth Moon Planets **97**, 411–434 (December 2005) 416
38. Rickman, H., Froeschlé, Ch., Froeschlé, Cl., Valsecchi, G.B.: Stellar perturbations on the scattered disk. Astron. Astrophys. **428**, 673–681 (December 2004) 416, 418
39. Smoluchowski, R., Torbett, M.: The boundary of the solar system. Nature **311**, 38–+, (September 1984) 402
40. Weissman, P.R.: Physical and dynamical evolution of long-period comets. In: Duncombe, R.L. (ed.) Dynamics of the Solar System, volume 81 of IAU Symposium, pp. 277–282 (1979) 402, 416
41. Weissman, P.R.: Stellar perturbations of the cometary cloud. Nature **288**, 242–243 (1980) 416
42. Wiegert, P., Tremaine, S.: The evolution of long-period comets. Icarus **137**, 84–121 (January 1999) 403

An Introduction to Common Numerical Integration Codes Used in Dynamical Astronomy

S. Eggl and R. Dvorak

Abstract As the tree of numerical methods used to solve ordinary differential equations develops more and more branches, it may, despite great literature, become hard to find out which properties should be aimed for, given certain problems in celestial mechanics. With this chapter the authors intend to give an introduction to common, symplectic, and non-symplectic algorithms used to numerically solve the basic Newtonian gravitational *N*-body problem in dynamical astronomy. Six methods are being presented, including a Cash–Karp Runge–Kutta, Radau15, Lie Series, Bulirsch-Stoer, Candy, and a symplectic Hybrid integrator of Chambers (Mon. Not. R. Astron. Soc. **304**: 793–799, 1999). Their main properties, as for example, the handling of conserved quantities, will be discussed on the basis of the Kepler problem.

1 Introduction

For several thousands of years astronomers have been asked to predict positions of the Sun, the Moon, and the planets on the celestial sphere. Especially during the Babylonian and Egyptian eras, the "oldest of all sciences" had been essential for producing calendars used not only for agricultural demands but also for religious rituals. The ancient methods were rather descriptive in nature, nevertheless the predictions, e.g., for Solar and Lunar eclipses were quite precise (for often the astronomer's life was tied to them). Even Kepler's laws were derived from observations without an understanding of the underlying true "nature" of the related phenomena. It is interesting to note that Kepler already mentioned a meanwhile familiar concept, a power determining the elliptic motion of the planets, a so-called "force". The most renowned scientific minds tried to grasp its nature by using various tool-sets, be it Galileo Galilei's experiments or Isaac Newton's first comprehensive theoretical

S. Eggl (✉)
Institute of Astronomy, University of Vienna, Türkenschanzstr. 17, 1880 Vienna, Austria,
siegfried.eggl@univie.ac.at

R. Dvorak
Institute of Astronomy, University of Vienna, Türkenschanzstr. 17, 1880 Vienna, Austria,
dvorak@astro.univie.ac.at

Eggl, S., Dvorak, R.: *An Introduction to Common Numerical Integration Codes Used in Dynamical Astronomy*. Lect. Notes Phys. **790**, 431–480 (2010)
DOI 10.1007/978-3-642-04458-8_9

description by introducing his law of gravitation. Of course, our understanding of the basic principles of gravity as an interplay of space time with massed particles, as proclaimed by Einstein, has grown enormously since Newton's times, although it is far from being comprehensive, as the current problem of "dark matter" shows quite plainly. Still, the dominant influence of gravity on the motion of massed particles – at least in our Solar System – is very well modeled by Newton's ansatz.
Newton's Law of Universal Gravity describes the force $\boldsymbol{F}$ acting between two point masses m_ν and m_μ

$$\boldsymbol{F}_{\nu\mu} = k^2 \frac{m_\nu m_\mu}{\rho_{\nu\mu}^2} \frac{\boldsymbol{r_\nu} - \boldsymbol{r_\mu}}{\rho_{\nu\mu}},$$

where $\boldsymbol{r}_\nu$ and $\boldsymbol{r}_\mu$ are the position vectors in an inertial frame. The scalar distance between particles is called $\rho = \rho_{\nu\mu} = \|\boldsymbol{r_\nu} - \boldsymbol{r_\mu}\|$, and k denotes the Gaussian Gravitational Constant.[1] Unfortunately, the resulting system of differential equations is non-integrable for more than two bodies. So, in principle, one can use two different methods for treating the equations of motion:

1. Perturbation theory:
   Perturbation theory works with series expansions of the equations of motion, or their related Hamiltonian equations, often including thousands of terms, computing analytical approximations to the solutions for a whole bundle of initial conditions. The results produced are mostly retained in form of series (also Fourier series) in a continuous parameter being identified with time, so that inserting a certain date in the series immediately leads to the particles' positions in space and on the sky. The main problems are a rather constrained time interval, for which these series produce reliable results, and series-convergence issues.

2. The method of numerical integration:
   Even though the solution of the multi-body gravitational problem may not be manageably representable by means of analytical functions, it is possible to follow the system's development, through calculating evolution of the system step by step, instead of trying to achieve results, that are valid for all times. This discretization procedure constitutes the main difference to analytical approaches.[2] The price one has to pay, in turn, is that solutions gained in such a way are in fact approximations that stay close to the true system for a certain, limited time interval. This is due to truncation errors caused by algorithms used for integration, and roundoff errors, that are a consequence of a finite number representation

[1] The Gaussian Gravitational Constant is used to do calculations in problem-specific units, which is especially important for numeric treatments, avoiding the appearance of very small and very large numbers that contribute to roundoff errors. $k \simeq 0.01720209895 AU^{\frac{3}{2}} D^{-1} M_\odot^{-\frac{1}{2}}$.
[2] Appendix offers a short introduction to the main concepts and terms related to numeric algorithms that will be used in the following sections.

within today's computer architecture. For comets and near-Earth asteroids this time span is in the order of years, for the planets in the order of ten thousands of years. In contrast to perturbation theory, the solution gained is representing just *a single trajectory* in phase space for a whole system of equations of motion.

The Nautical Almanac Service at the Naval Observatory in Washington publishes the ephemerides using numerical integrations provided by the JPL in Pasadena. In generating these ephemerides even the perturbations of the major asteroids are taken into account and corrections arising from General Relativity are included.

Due to the enormous progress in the field of CPU processing power, numerics have become very popular over the last decades. Large-scale numerical experiments are feasible today, thus feeding the desire for potent algorithms. In the next sections, we will treat common numerical methods that are used to solve the gravitational multi-body problem. Though, it is always recommendable to see, if one can derive some basic information on the main properties of the equations governing a system's behavior, *before* focusing on the computational aspects. The force acting from N bodies with masses $m_\mu (\mu = 0, ..., N, \mu \neq \nu)$ on m_ν can be written as follows:

$$\boldsymbol{F}_\nu = m_\nu \ddot{\boldsymbol{r}}_\nu = k^2 m_\nu \sum_{\mu=0, \mu\neq\nu}^{N} \frac{m_\mu}{\rho_{\nu\mu}^3} (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu). \tag{1}$$

Summing all forces we derive

$$\sum_{\nu=0}^{N} m_\nu \ddot{\boldsymbol{r}}_\nu = k^2 \sum_{\nu=0}^{N} \sum_{\mu=0, \mu\neq\nu}^{N} \frac{m_\nu m_\mu}{\rho_{\nu\mu}^3} (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu) = 0, \tag{2}$$

which is zero, simply because each vector $\boldsymbol{r}_\mu - \boldsymbol{r}_\nu$ is canceled by its inverse vector $\boldsymbol{r}_\nu - \boldsymbol{r}_\mu$. A double integration with respect to time leads to the linear motion of the barycenter of the dynamical system, which can be described by the vector $\boldsymbol{s} = \boldsymbol{a}t + \boldsymbol{b}$. This result can be used to reduce the number of equations of the system due to the fact that $\boldsymbol{a}$ and $\boldsymbol{b}$ are constants; it is also obvious that the barycenter qualifies as the origin of an inertial coordinate system. In addition to these six constants of motion—the components of $\boldsymbol{a}$ and $\boldsymbol{b}$—another four integrals of motion exist, namely the angular momentum integral (three constants) and the energy integral $E_{\text{kin}} + E_{\text{pot}} = \text{const}$.

These 10 constants of motion are the classical integrals of the Newtonian N-body problem. In order to describe the motion of the planets of our Solar System, one may keep in mind, that more than 99.9% of the mass is accumulated in the Sun (m_0). As a consequence one often transforms the equations of motion of the planets to a heliocentric relative coordinate system; the relative (heliocentric) vectors will

be denoted by $\boldsymbol{q}_\nu = \boldsymbol{r}_\nu - \boldsymbol{r}_0$.[3] This transformation can be achieved by separating the attraction of the Sun on each planet m_ν from the other terms in (2),

$$\ddot{\boldsymbol{r}}_\nu = k^2 \left[\frac{m_0}{\rho_{0\nu}^3}(\boldsymbol{r}_0 - \boldsymbol{r}_\nu) + \sum_{\mu=1,\mu\neq\nu}^{N} \frac{m_\mu}{\rho_{\nu\mu}^3}(\boldsymbol{r}_\mu - \boldsymbol{r}_\nu) \right]. \tag{3}$$

Considering the Sun's equation of motion in the same form yields (4).

$$\ddot{\boldsymbol{r}}_0 = k^2 \left[\frac{m_\nu}{\rho_{0\nu}^3}(\boldsymbol{r}_\nu - \boldsymbol{r}_0) + \sum_{\mu=1,\mu\neq\nu}^{N} \frac{m_\mu}{\rho_{\mu 0}^3}(\boldsymbol{r}_\mu - \boldsymbol{r}_0) \right]. \tag{4}$$

Subtracting (4) from (3) and separating the vector $\boldsymbol{r}_\nu - \boldsymbol{r}_0$, we get the following—now heliocentric—equations of motion of the planet m_ν

$$\begin{aligned} \ddot{\boldsymbol{q}}_\nu = k^2 &\left[-\frac{m_0}{\rho_{0\nu}^3}(\boldsymbol{r}_\nu - \boldsymbol{r}_0) + \frac{m_\nu}{\rho_{0\nu}^3}(\boldsymbol{r}_\nu - \boldsymbol{r}_0) \right] + \\ &+ \sum_{\mu=1,\mu\neq\nu}^{N} \left[\frac{m_\mu}{\rho_{\mu\nu}^3}(\boldsymbol{r}_\mu - \boldsymbol{r}_\nu) - \frac{m_\mu}{\rho_{\mu 0}^3}(\boldsymbol{r}_\mu - \boldsymbol{r}_0) \right] \end{aligned} \tag{5}$$

which lead to

$$\ddot{\boldsymbol{q}}_\nu = k^2 \left[-\frac{m_0 + m_\nu}{\rho_{0\nu}^3}\boldsymbol{q}_\nu + \sum_{\mu=1,\mu\neq\nu}^{N} m_\mu \left(\frac{\boldsymbol{q}_\mu - \boldsymbol{q}_\nu}{\rho_{\mu\nu}^3} - \frac{\boldsymbol{q}_\mu}{\rho_{\mu 0}^3} \right) \right]. \tag{6}$$

The first term (6 r.h.s) represents just the unperturbed two-body motion of the respective particle around the Sun. As the solution of the gravitational two-body problem can be expressed in terms of conic sections, we can see why the bound orbits of the planets are—in a first approximation—ellipses. Size and orientation of these ellipses can be represented by five constants: semi-major axis (a) and (numeric) eccentricity (e) fixing the form of the conic section, plus three angles, consisting of the inclination (i) to a fixed plane, the position-angle of the perihelion (ω), as well as the argument of the ascending node (Ω), which gives the position of the intersection point of the orbit with the fixed plane. Together with some parameter pointing out the current position of the planet on the conic section, named mean anomaly (M), those six expressions constitute the Keplerian orbital elements. Keplerian elements are more appropriate to investigate the behavior of systems in celestial mechanics than the ever-changing position and velocity vectors

[3] $\boldsymbol{r}_0$ designates the vector from the artistically chosen origin of the coordinate system. This vector equation stays valid for its time derivatives, which means it is also true for the velocities and the accelerations.

and thus are the output of choice, when it comes to facilitating evaluation processes. Yet they are cumbersome to be dealt with in a numeric way, that is why practically all algorithms are derived from equations containing positions and velocities. The rest of the terms in (6) contain a factor including the planets' masses, which are only in the order of 10^{-1} to 10^{-4} percent of the Sun's mass as far as our Solar System is concerned. Therefore the equations of motion of the planets may be written as

$$\ddot{\boldsymbol{q}}_\nu = -k^2 \frac{m_0 + m_\nu}{\rho_{0\nu}^3} \boldsymbol{q}_\nu + \boldsymbol{P}_\nu, \tag{7}$$

where the perturbing vector $\boldsymbol{P}_\nu$ is

$$\boldsymbol{P}_\nu = k^2 \sum_{\mu=1,\mu\neq\nu}^{N} m_\mu \left(\frac{\boldsymbol{q}_\mu - \boldsymbol{q}_\nu}{\rho_{\mu\nu}^3} - \frac{\boldsymbol{q}_\mu}{\rho_{\mu 0}^3} \right). \tag{8}$$

Most of the classical numerical algorithms use a barycentric coordinate system as a reference frame, due to the fact that it is an inertial system and will not produce extra-pseudoforces. However, it may not always be the best choice, since the planets' movements are usually tied to Keplerian orbits, meaning that this information on the planets' motion is simply being neglected. The symplectic mappings of Wistom and Holman [35] also included in the hybrid algorithm of Chambers [4] contain an ansatz similar to that of (7), leading to smaller errors in standard planetary motion compared to, e.g., the method of Candy and Rozmus [1]. Of course, choosing a special set of coordinates has also disadvantages. Consider binary systems, for example, where the perturbation terms will be in the order of the terms describing the Keplerian motion. In this case, heliocentric types of coordinate systems will have to deal with non-negligible pseudo forces, slowing the integration process or elevating the integration errors. Generally speaking, the N second-order differential equations (7) may be reasonably used for planets, asteroids, and also comets, just taking Newtonian forces into account. For satellites and the Moon, it is preferable to use a coordinate system centered at the Earth or at the satellite's hosting planet. Considering the perturbing vector $\boldsymbol{P}_\nu$, the appearance of the third power of the distances $\mid \boldsymbol{q}_\mu - \boldsymbol{q}_\nu \mid = \rho_{\mu\nu}$ in the denominator may lead to large accelerations, when these distances become small. In planetary theories this fact is not a problem, because planets move on well-separated orbits. Treating comets is a different story, because they approach planets (especially Jupiter) quite often, which will change their orbits significantly. In addition to that, the equations of motion given above are not really suited for near-Earth asteroids (NEAs) like the Atens, Apollos, and Amors which also have frequent close encounters with a planet. Allowing for a reasonable description of close encounters is the main reason, why an *adaptive choice of step-size* is a requirement for any algorithm intended to prevail in the field of celestial mechanics. As the polar nature of the gravitational N-body problem also tends to enlarge roundoff errors, special regularization methods have been developed

(see, e.g., Mikkola and Aarseth [24]) that are meant to prevent divisions by small quantities during close encounters.

In the following, we will give a glimpse into the ideas behind six different methods often used to solve the gravitational N-body problem[4]. We will briefly compare their performance concerning conservational properties in the two body problem. With two exceptions,[5] we will restrict our introduction to explicit one-step methods with adaptive step-size control.

2 Classic Explicit Runge–Kutta-Type Integrators

2.1 Introduction

Explicit Runge–Kutta (RK)-type integrators are among the most popular algorithms concerning numerical analysis of initial value problems. This popularity may be due to a history dating back over a century, thus resulting in a well understood, elaborated theory. Also the possibility of a straightforward implementation and the relative ease of error control add to their appealing aura.

Nevertheless, their need for a relatively high number of right-hand side function evaluations and unfavorable energy conservation properties in their classic, non-symplectic forms are downsides, that will have to be taken into account, if an application to the field of Celestial Mechanics is intended.

2.2 Formalities

As a representative for the grand family of classic RK algorithms, the Cash-Karp version [3] has been chosen for the following reason: Although non-symplectic, the coefficients derived by Cash and Karp [3] have shown to work quite well-solving, non-stiff, ordinary differential equations, because they contain embedded formulas for lower order RK algorithms, which allow for a quick change in order, resulting in fewer function evaluations and a high-quality step-size control.

The $3N$ second-order differential equations describing Newtonian gravitational forces (9) of $\nu = 1, ..., N$ interacting bodies can be rewritten in $6N$ differential equations of first order (10) and (11).[6]

---

[4] The algorithms presented may not be the methods of choice when N tends to be very large.

[5] The Bulirsch–Stoer algorithm being an extrapolation method has been included for its popularity, and the Candy symplectic mapping, that does not support step-size control, is used for demonstrational reasons.

[6] Actually, relations (9), (10), and (11) describe the gravitational influence of N bodies acting on a particle with index ν, which are just three second-order differential equations. Though, since we would like to monitor the progress of the whole system, we have to calculate these equations for all particles with indices $\nu = 1, ..., N$, ending up with $3N$ second-order equations.

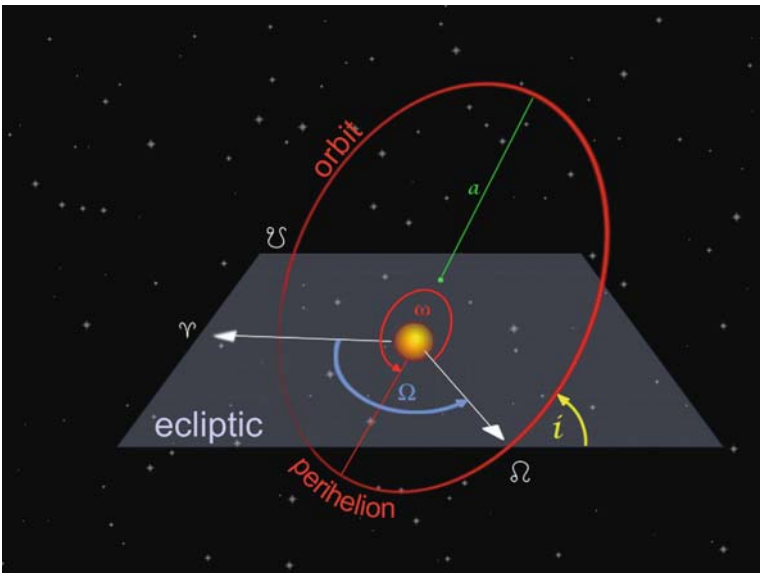

Fig. 1 The five Keplerian elements determining size and orientation of the orbit: semi-major axis (a), eccentricity (e), inclination (i), argument of perihelion (ω), and the argument of the ascending node (Ω)

$$m_\nu \ddot{\boldsymbol{r}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\nu m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3}, \tag{9}$$

$$\dot{\boldsymbol{r}}_\nu = \boldsymbol{v}_\nu, \tag{10}$$

$$\dot{\boldsymbol{v}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3}. \tag{11}$$

Let us combine $\boldsymbol{r}$ and $\boldsymbol{v}$ in order to gain a six-dimensional vector $\boldsymbol{y}$,

$$\boldsymbol{y}_\nu = \begin{pmatrix} \boldsymbol{r}_\nu \\ \boldsymbol{v}_\nu \end{pmatrix}. \tag{12}$$

Adding index n, that will denote the current time t_n, (10) and (11) can be reformulated accordingly, with their right-hand sides being put into an *evaluation function* $\boldsymbol{f}$:

$$\dot{\boldsymbol{y}}_{n,\nu} = \boldsymbol{f}(t_n, \boldsymbol{y}_{n,\nu}). \tag{13}$$

The essence of the RK method is the combination of intermediate results for the slopes between the old and new $\boldsymbol{y}$ values gained from function evaluations at

different points within the next time-step τ.[7] This procedure eliminates terms of lower orders in τ, as can be proven by Taylor expansion [33]. So the RK algorithm can be written in the following way:

$$\boldsymbol{y}_{n+1,v} = \boldsymbol{y}_{n,v} + \sum_{l=1}^{s} b_l \boldsymbol{k}_{l,v}, \tag{14}$$

with $\boldsymbol{k}_s$, the "intermediate slopes" being defined as

$$\begin{aligned}
\boldsymbol{k}_{1,v} &= \tau \boldsymbol{f}(t_n, \boldsymbol{y}_{n,v}) \\
\boldsymbol{k}_{2,v} &= \tau \boldsymbol{f}(t_n + c_2\tau, \boldsymbol{y}_{n,v} + a_{21}\boldsymbol{k}) \\
&\vdots \\
\boldsymbol{k}_{s,v} &= \tau \boldsymbol{f}(t_n + c_s\tau, \boldsymbol{y}_{n,v} + \sum_{m=1}^{s-1} a_{sm}\boldsymbol{k}_{m,v})
\end{aligned},$$

b_l, c_m, and a_{sm} are constants that can be chosen according to the corresponding Butcher tableau (Table 1).

Usually, when constants a_{sm} and b_l are being derived for different orders of RK algorithms without additional constraints, the spacings $c_m\tau$ where the functions $\boldsymbol{f}$

Table 1 Butcher tableau for Cash–Karp Runge–Kutta coefficients [3]

0	0						
c_2	a_{21}	0					
c_3	a_{31}	a_{32}	0				
c_4	a_{41}	a_{42}	a_{43}	0			
c_5	a_{51}	a_{52}	a_{53}	a_{54}	0		
c_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	0	
0	0						
$\frac{1}{5}$	$\frac{1}{5}$	0					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0				
$\frac{3}{5}$	$\frac{3}{10}$	$-\frac{9}{10}$	$\frac{6}{5}$	0			
1	$-\frac{11}{54}$	$\frac{5}{2}$	$-\frac{70}{27}$	$\frac{35}{27}$	0		
$\frac{7}{8}$	$\frac{1631}{55296}$	$\frac{175}{512}$	$\frac{575}{13824}$	$\frac{44275}{110592}$	$\frac{253}{4096}$	0	
	b_1	b_2	b_3	b_4	b_5	b_6	
	$\frac{37}{378}$	0	$\frac{250}{621}$	$\frac{125}{594}$	0	$\frac{512}{1771}$	order 5
	$\frac{2825}{27648}$	0	$\frac{18575}{48384}$	$\frac{13525}{55296}$	$\frac{277}{14336}$	$\frac{1}{4}$	order 4
	$\frac{19}{54}$	0	$-\frac{10}{27}$	$\frac{55}{54}$	0	0	order 3
	$-\frac{3}{2}$	$\frac{5}{2}$	0	0	0	0	order 2
	1	0	0	0	0	0	order 1

[7] A time-step τ is defined as follows: $\tau = t_n - t_{n-1}$.

are evaluated are unrelated. That is leading to an enormous amount of function evaluations necessary to calculate different orders.[8] The great advantage of embedded formulae used in the Cash–Karp RK is the fact, that all previous orders are contained in an $(s-1)$th order method. So one does not have to recalculate all function evaluations for each desired, new order. Instead, one uses the results of previous orders, resulting in a vast reduction of computational time required.

2.3 Variable Step-Size Determination

The choice of step-size within the RK algorithm depends on the local difference of two results $\boldsymbol{y}_n$ of varying orders. Let $\mathrm{ERR}(n,i)$ be the local difference of two results of orders i and $i+1$ at time t_n, and $E(n,i)$ the error relative to a user-specified tolerance ϵ, then the next step size will be

$$\mathrm{ERR}(n,i) = \|\boldsymbol{y}_n^{i+1} - \boldsymbol{y}_n^{i}\|^{1/(i+1)} \qquad i \in 1,2,4, \tag{15}$$

$$E(n,i) = \frac{\mathrm{ERR}(n,i)}{\epsilon^{1/(i+1)}}, \tag{16}$$

$$\tau_{n+1} = \frac{\mathrm{SF}\cdot\tau_n}{E(n,i)}. \tag{17}$$

SF is a safety factor, often taken to be $\mathrm{SF} = 0.9$. Here the order sequence $3-4$ is excluded from the step-size control for performance reasons. For further details see Cash and Karp [3].

2.4 Improvements

The relatively high number of right-hand side function evaluations necessary to solve a given differential equation, and their poor energy conservation behavior have made standard Runge–Kutta methods rather unappealing for problems concerning celestial mechanics. Even though continuing research on these algorithms brought up interesting results. Among others, Lasagni [21], Sanz-Serna [30], and Suris [31] started a detailed study on the construction of symplectic Runge–Kutta methods. Okunbor and Skeel [27] were able to construct explicit symplectic algorithms by putting up special relations for the Butcher coefficients, Cash and Gridlestone [2] even derived symplectic, variable step Runge–Kutta Nyström codes. This ongoing progress may well lead to a renaissance of Runge–Kutta-type integrators in dynamical astronomy.

---

[8] As the evaluation functions in the gravitational problem include the pair distances of each body to every other particle, the number of calculations having to be performed to evaluate $\boldsymbol{f}$ scales with N^2. Thus, viable measures of computational time required to solve the gravitational N-body problem can only be achieved by minimizing right-hand side function evaluations.

3 Gauss–Radau Quadratures

3.1 Introduction

The basic idea behind Gaussian quadratures is that the integral of a function can be approximated by the sum of functional evaluations taken at arbitrary points, respectively times, being multiplied by some weighting coefficients. Gauss–Radau quadrature uses one fixed abscissa point at the beginning of the integration interval, and chooses the other ones, as well as the weights, in order to maximize the degree of exactness of the quadrature rule. One of the first published applications of Gauss–Radau-based algorithms to astrodynamical problems was done by Everhart [12], even though they had already been used for nearly a decade back then. Positive attributes of Gauss–Radau algorithms are their affinity to implicit Runge–Kutta algorithms inheriting a very generous convergence behavior, their ability to integrate polynomials of an order that is related to the number of abscissa nodes up to machine precision, and their relatively moderate requirements concerning right-hand side function evaluations. A major drawback lies in the fact that all optimized coefficients as well as abscissa points have to be recalculated from scratch, if another order of the integration algorithm is needed. This makes order switching procedures very inefficient when compared to embedded schemes as the Cash–Karp Runge–Kutta method.

3.2 Formalities

Stating once again the second-order differential equations of a classical gravity induced force exerted by N particles on a particle ν, and dividing these equations (18) by the mass m_ν will produce the acceleration $\ddot{\boldsymbol{r}}_\nu$ of particle ν:

$$m_\nu \ddot{\boldsymbol{r}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\nu m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3}, \tag{18}$$

$$\ddot{\boldsymbol{r}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3}. \tag{19}$$

For the sake of a compact notation we will refer to the right-hand side of (19) as being a function of all particle positions $\boldsymbol{F}_\nu(\boldsymbol{r}_\mu, \boldsymbol{r}_\nu)$,

$$\ddot{\boldsymbol{r}}_\nu = \boldsymbol{F}_\nu(\boldsymbol{r}_\mu, \boldsymbol{r}_\nu) \quad \mu = 1, ..., N \quad \mu \neq \nu \tag{20}$$

and omit the index ν denoting a certain particle, keeping in mind that

$$\boldsymbol{r} \equiv \boldsymbol{r}_\nu.$$

So the basic equation treated is of the form

$$\ddot{\boldsymbol{r}} = \boldsymbol{F}. \tag{21}$$

We will now perform a series expansion of (21) around time $t_1 = 0$ with initial conditions $\boldsymbol{r}_1$ and $\dot{\boldsymbol{r}}_1$,

$$\ddot{\boldsymbol{r}} = \boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{A}_1 t + \boldsymbol{A}_2 t^2 + \boldsymbol{A}_3 t^3 + \cdots + \boldsymbol{A}_n t^n. \tag{22}$$

Integrating (22) will result in

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{r}_1 + \dot{\boldsymbol{r}}_1 t + \boldsymbol{F}_1 t^2/2 + \boldsymbol{A}_1 t^3/6 + \cdots + \boldsymbol{A}_n t^{n+2}/((n+1)(n+2)), \\ \dot{\boldsymbol{r}} &= \dot{\boldsymbol{r}}_1 + \boldsymbol{F}_1 t + \boldsymbol{A}_1 t^2/2 + \boldsymbol{A}_2 t^3/3 + \cdots + \boldsymbol{A}_n t^{n+1}/(n+1). \end{aligned} \tag{23}$$

In contrast to a Taylor series, the coefficients $\boldsymbol{A}$ will not be chosen to represent $\boldsymbol{F}$ as well as possible for all times t. Instead, they are adapted to calculating $\boldsymbol{r}$, and $\dot{\boldsymbol{r}}$ as exactly as possible for a given time interval τ. The way to finding the weighting coefficients $\boldsymbol{A}$ is somehow similar to Runge–Kutta algorithms, because one explores the function $\boldsymbol{F}$ at several unequally spaced sub-steps $t_2, t_3, t_4, \ldots$ with $\tau > t_i > t_1 \quad i \in \mathbb{N}$. Let $\boldsymbol{F}_n$ be the function $\boldsymbol{F}$ evaluated at positions $\boldsymbol{r}_n(t_n)$, so we can perform an auxiliary expansion

$$\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{\alpha}_1 t + \boldsymbol{\alpha}_2 t(t - t_2) + \boldsymbol{\alpha}_3 t(t - t_2)(t - t_3) + \cdots . \tag{24}$$

Of course, this series development and all following terms must be consistent in order with the number of terms kept in (22).

As the functions $\boldsymbol{F}_n$ can easily be evaluated, truncations of (24) can be used to determine the coefficients $\boldsymbol{\alpha}_i$:

$$\begin{aligned} \boldsymbol{F}_2 &= \boldsymbol{F}_1 + \boldsymbol{\alpha}_1 t_2 \\ \boldsymbol{F}_3 &= \boldsymbol{F}_1 + \boldsymbol{\alpha}_1 t_3 + \boldsymbol{\alpha}_2 t_3 (t_3 - t_2) \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad . \end{aligned}$$

$$\begin{aligned} \boldsymbol{\alpha}_1 &= (\boldsymbol{F}_2 - \boldsymbol{F}_1)/t_2 \\ \boldsymbol{\alpha}_2 &= ((\boldsymbol{F}_3 - \boldsymbol{F}_1)/t_3 - \boldsymbol{\alpha}_1)/(t_3 - t2) \\ \boldsymbol{\alpha}_3 &= (((\boldsymbol{F}_4 - \boldsymbol{F}_1)/t_4 - \boldsymbol{\alpha}_1)/(t_4 - t2) - \boldsymbol{\alpha}_2)/(t_4 - t_3) \\ \boldsymbol{\alpha}_4 &= ((((\boldsymbol{F}_5 - \boldsymbol{F}_1)/t_5 - \boldsymbol{\alpha}_1)/(t_5 - t2) - \boldsymbol{\alpha}_2)/(t_5 - t_3) - \boldsymbol{\alpha}_3)/(t_5 - t_4) \end{aligned} \tag{25}$$

The relation between weighting coefficients $\boldsymbol{A}$ and $\boldsymbol{\alpha}$ can be established by comparing corresponding powers of t in (22) and (24):

$$\begin{aligned} \boldsymbol{A}_1 &= \boldsymbol{c}_{11}\boldsymbol{\alpha}_1 + \boldsymbol{c}_{21}\boldsymbol{\alpha}_2 + (t_2 t_3)\boldsymbol{\alpha}_3 + \cdots \\ \boldsymbol{A}_2 &= \boldsymbol{c}_{22}\boldsymbol{\alpha}_2 + \boldsymbol{c}_{32}\boldsymbol{\alpha}_3 + \cdots \\ \boldsymbol{A}_3 &= \boldsymbol{c}_{33}\boldsymbol{\alpha}_3 + \cdots \end{aligned} \quad . \tag{26}$$

With coefficients $\boldsymbol{c}_{ij}$ being defined as

$$\begin{aligned} \boldsymbol{c}_{ii} &= 1 \\ \boldsymbol{c}_{i1} &= -t_i \boldsymbol{c}_{i-1,1} \quad i > 1 \\ \boldsymbol{c}_{ij} &= \boldsymbol{c}_{i-1,j-1} - t_i \boldsymbol{c}_{i-1,j} \quad 1 < j < i \end{aligned} \quad . \tag{27}$$

The actual integration algorithm will work as follows: Let us assume, we integrate a sequence of three sub-steps at times t_2, t_3, t_4 that are not uniformly spaced. Actually t_4 does not even have to coincide with the end of the integration interval of length τ. At the starting point $t_1 = 0$ all initial conditions $\boldsymbol{r}_1, \dot{\boldsymbol{r}}_1$, and $\boldsymbol{F}_1$ are known. In this case we will have one (vector-valued) *predictor equation* per sub-step:

$$\begin{aligned} \boldsymbol{r}_2 &= \boldsymbol{r}_1 + \dot{\boldsymbol{r}}_1 t_2 + \boldsymbol{F}_1 t_2^2/2 + [\boldsymbol{A}_1 t_2^3/6 + \boldsymbol{A}_2 t_2^4/12 + \boldsymbol{A}_3 t_2^5/20] \\ \boldsymbol{r}_3 &= \boldsymbol{r}_1 + \dot{\boldsymbol{r}}_1 t_3 + \boldsymbol{F}_1 t_3^2/2 + \boldsymbol{A}_1 t_3^3/6 + [\boldsymbol{A}_2 t_3^4/12 + \boldsymbol{A}_3 t_3^5/20] \\ \boldsymbol{r}_4 &= \boldsymbol{r}_1 + \dot{\boldsymbol{r}}_1 t_4 + \boldsymbol{F}_1 t_4^2/2 + \boldsymbol{A}_1 t_4^3/6 + \boldsymbol{A}_2 t_4^4/12 + [\boldsymbol{A}_3 t_4^5/20] \end{aligned} \tag{28}$$

and two (vector-valued) *corrector equations* to find the positions and velocities at the end of the integration interval,

$$\begin{aligned} \boldsymbol{r}(\tau) &= \boldsymbol{r}_1 + \dot{\boldsymbol{r}}_1 \tau + \boldsymbol{F}_1 \tau^2/2 + \boldsymbol{A}_1 \tau^3/6 + \boldsymbol{A}_2 \tau^4/12 + \boldsymbol{A}_3 \tau^5/20 \\ \dot{\boldsymbol{r}}(\tau) &= \dot{\boldsymbol{r}}_1 + \boldsymbol{F}_1 \tau + \boldsymbol{A}_1 \tau^2/2 + \boldsymbol{A}_2 \tau^3/3 + \boldsymbol{A}_3 \tau^4/4 \end{aligned} \quad . \tag{29}$$

As the bracketed terms in the predictor equations (28) are not known in advance, the system is implicit. During the initial phases of an integration, where the bracketed terms are zero, it is thus recommended to make several passes through these equations improving estimates for $\boldsymbol{r}_n$ and corresponding $\boldsymbol{\alpha}$ and $\boldsymbol{A}$ values. If the integration is already in progress, it is possible to extrapolate current $\boldsymbol{\alpha}$ values from the previous steps and calculate new $\boldsymbol{A}$s, if the problem is "well behaved," meaning that it does not contain near discontinuous jumps.

So, if there are current $\boldsymbol{\alpha}$ values at hand, then a first prediction of $\boldsymbol{r}_2$ will be possible from (28). Evaluating $\boldsymbol{F}_2$ will allow for a correction of $\boldsymbol{\alpha}_1$ and consequently $\boldsymbol{A}_1$ still using all the previous $\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$, ... except for the renewed $\boldsymbol{\alpha}_1$. Using the second equation in (28) one finds $\boldsymbol{r}_3$, including the improved $\boldsymbol{A}_1$ and the old $\boldsymbol{A}_2, \boldsymbol{A}_3$ values. Evaluating $\boldsymbol{F}_3$, one obtains better estimates for $\boldsymbol{\alpha}_2$, and consequently $\boldsymbol{A}_2$ having just $\boldsymbol{\alpha}_3$ left in its old form. This procedure is continued until all values for $\boldsymbol{\alpha}$ and $\boldsymbol{A}$ have been updated and can be used in the corrector step.

3.3 Variable Step-Size Determination

In the previous section, we showed an example of a Gauss–Radau algorithm with three sub-steps, containing terms proportional to τ^5 in the corresponding corrector equations. For such a case, one can define a control parameter C that characterizes the desired size of the last term in (29) which is $\boldsymbol{A}_3\tau^5/20$. Let $H = (\|\boldsymbol{A}_3\|)/20$ then the new step-size $\tau' = (C/H)^{1/5}$. If the number of sequence terms is altered to achieve higher order integration algorithms, the corresponding, last terms of (29) and the exponent $1/5$ will have to be adapted, of course.

3.4 Gauss–Radau Spacings

As the algorithm presented contains terms up to order τ^5, one may initially assume that it is of order 5. One of the main advantages of Gauss–Radau quadrature is the possibility to choose the spacings between the sub-steps t_2, t_3, t_4 in such a way that the results for both $\boldsymbol{r}$ and $\dot{\boldsymbol{r}}$ are accurate to seventh order in τ *without* adding additional nodes. The essence of the method to find these spacings is that one adds two additional sub-steps t_5, t_6: watches the improvement of $\boldsymbol{r}$ and $\dot{\boldsymbol{r}}$ and chooses t_2, t_3, t_4 such that these improvements tend to zero. For further details on this topic, and a table of optimal spacings for different orders of the Gauss–Radau algorithm, the authors would like to refer the reader to Everhart [12].

4 Bulirsch–Stoer Method

4.1 Introduction

Though extrapolation methods were developed some time ago, they still rank among the most effective concerning high-accuracy solutions of ordinary differential equations. The so-called Bulirsch–Stoer method [32] is actually a clever combination of two separate algorithms.

Part I. In a first step, the treated differential equations are solved with a fast numerical integration method beginning at time T using a time-step of τ. Actually, there is no need for the integration algorithm to be of high order. The results R gained at time $T + \tau$ are *not* used as new starting points for the computation of the next time step. Instead the interval τ is split into n_m smaller intervals of size $\tau_m = \frac{\tau}{n_m}$, and the integration is now performed over all n_m intervals, up to the endpoint of the former one-step integration $T + n_m \cdot \tau_m = T + \tau$. Assuming that numerical rounding errors are negligible compared to the local error produced by the computing algorithm, the n_m integrations with the smaller step-size $\frac{\tau}{n_m}$ will give a better estimate of the result R. This interval splitting procedure is repeated, so that the step-sizes are reduced in a sequence proposed by Deuflhard [8]:

$$\begin{aligned} n_m &= 1 \qquad & m &= 0 \\ n_m &= 2 \cdot m \qquad & m &\in \mathbb{N} \end{aligned} \quad . \tag{30}$$

As a solver for the given differential equations the modified midpoint method is very popular.

Part II. The second part of the extrapolation method consists of an implementation of the idea to interpret the results R_m gained through the interval splitting procedure of Part I, as supporting points of a function depending on the step-size τ:

$$R_m = R(\frac{\tau}{n_m}). \tag{31}$$

This function R_m will then be extrapolated to its value R_∞. That result corresponds to the whole interval being calculated with a step size of $\tau = 0$. For this procedure, the Aitken–Neville scheme, based on polynomial interpolation, is used [14].

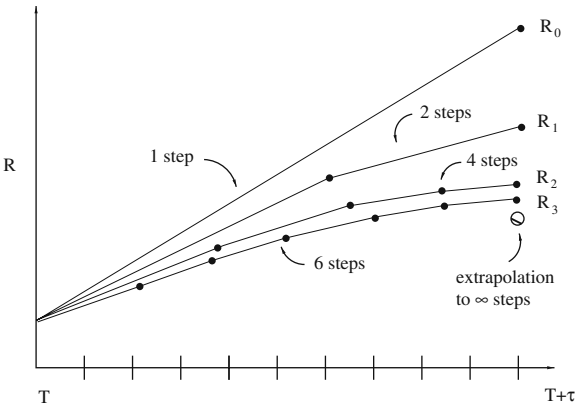

Fig. 2 Bulirsch–Stoer method. The results R_m after a time-step τ are sampled with different numbers of sub-steps $\frac{\tau}{n_m}$. These results are seen as a function of the number of sub-steps, and will finally be extrapolated to a value R_∞, that represents—in principle—the solution of a differential equation calculated with a (sub-) step size of $\tau_m = 0$

4.2 Formalities

Part I. **Modified Midpoint Method:**

Just as shown in Chap. 2, the main equations of the Newtonian gravitational problem (10) and (11) are reformulated as a set of first-order differential equations with $\boldsymbol{y}_{n,\nu}$ denoting the current phase space vector for the νth particle at time t_n, and $\boldsymbol{f}$ being the evaluating function (13):

$$\boldsymbol{y}_{n,\nu} = \begin{pmatrix} \boldsymbol{r}_{n,\nu} \\ \boldsymbol{v}_{n,\nu} \end{pmatrix}, \tag{32}$$

$$\dot{\boldsymbol{y}}_{n,\nu} = \boldsymbol{f}(t_n, \boldsymbol{y}_{n,\nu}). \tag{33}$$

Rooting in the so-called midpoint approximation for first derivatives, the midpoint method can be derived as follows[9]:

$$\begin{aligned} \dot{\boldsymbol{y}}_{n+1} &\simeq \frac{\boldsymbol{y}_{n+2} - \boldsymbol{y}_n}{2\tau}, \\ \boldsymbol{y}_{n+2} &= \boldsymbol{y}_n + 2\tau\, \dot{\boldsymbol{y}}_{n+1}. \end{aligned} \tag{34}$$

τ denoting the current time-step: $t_{n+1} = t_n + \tau$. Substituting (33) in (34) will produce

$$\boldsymbol{y}_{n+2} = \boldsymbol{y}_n + 2\tau\, \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}). \tag{35}$$

As we do not know the value of $\boldsymbol{y}_{n+1}$, we will perform a Taylor expansion up to first-order terms in τ:

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}(t_n + \tau) \simeq \boldsymbol{y}(t_n) + \tau\, \dot{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n + \tau\, \boldsymbol{f}(t_n, \boldsymbol{y}_n). \tag{36}$$

Again substituting the approximation (36) in (35) will result in

$$\boldsymbol{y}_{n+2} \simeq \boldsymbol{y}_n + 2\tau\, \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_n + \tau\, \boldsymbol{f}(t_n, \boldsymbol{y}_n)). \tag{37}$$

The final equation (37) approximates the true solution of (33) up to second-order terms in τ, but the main reason for its choice as a core algorithm in the Bulirsch–Stoer method is its cost-effectiveness concerning CPU time. As can be seen from (37) the midpoint method, once running, uses two evaluations of $\boldsymbol{f}$ performing two steps with step-size τ. This results in one evaluation per step, which is quite "cheap" compared to most of the second-order algorithms.

The most widespread formulation of the midpoint method is the "modified-midpoint" version, that can be found, e.g., in Numerical Recipes [28], that differs from the ordinary midpoint method in the way of calculating the first and the last points. In order to propagate a system over one relatively large time-step τ, having initial conditions $\boldsymbol{y}_0$ at time T, a sequence of sub-steps of size $\tau_m = \frac{\tau}{n}$ can be introduced with vectors $\boldsymbol{w}$ being intermediate approximations for $\boldsymbol{y}$

[9] The particle index ν will be omitted during the derivation section.

$$\begin{aligned} \boldsymbol{w}_0 &= \boldsymbol{y}_0 \\ \boldsymbol{w}_1 &= \boldsymbol{w}_0 + \tau_m \boldsymbol{f}(T, w_0) \\ \boldsymbol{w}_{l+2} &= \boldsymbol{w}_l + 2\tau_m \boldsymbol{f}(T + (l+1)\tau_m, w_{l+1}) \\ l &= 0, 1, 2, \ldots n-2 \\ \boldsymbol{y}(T+\tau) &\simeq \frac{1}{2}[\boldsymbol{w}_n + \boldsymbol{w}_{n-1} + \tau_m \boldsymbol{f}(T+\tau, \boldsymbol{w}_n)] \end{aligned} \quad . \tag{38}$$

The reason for introducing sub-steps is that as their number increases, the result $\boldsymbol{y}(T+\tau)$ will improve drastically, due to the reduced fractional time-step τ_m. This behavior will be used to extrapolate virtually perfect values for $\boldsymbol{y}(T+\tau)$ in Part II of this section.

Counting the number of necessary function evaluations for the modified scheme will result in $n+1$, meaning that it is still quite efficient compared to, e.g., a second-order Runge–Kutta, that would need $2 \cdot n$.

Part II. **Polynomial Interpolation:**

The problem of gaining a function, that may be able to extrapolate the results[10] R_m to the ideal result R_∞, theoretically achieved through numerical integration of an interval τ using a vanishing step-size is targeted by a polynomial interpolation algorithm. The Aitken–Neville method will fulfill that request without too much additional cost, because the results R_m are sufficient to calculate a polynomial of order q recursively. An error estimate is also possible, which will play an essential role in the choice of a fitting step-size.

Let q be the order of a polynomial, fit to m supporting points, that will give a vector of results $\boldsymbol{R}_q(0)$ for a supposed step-size equal to zero. Introducing the tabling index i, the Aitken-Neville recursive scheme can be constructed as follows [14]:

$$i = m - q \qquad i \geq 0, \tag{39}$$

$$\boldsymbol{R}_q^i(0) = \frac{\tau_{i+q}\boldsymbol{R}_{q-1}^i(0) + \tau_i \boldsymbol{R}_{q-1}^{i+1}(0)}{\tau_i - \tau_{i+q}}. \tag{40}$$

Consequently τ_{i+q} correspond to $\tau_m = \frac{\tau}{n_m}$ mentioned in the introduction (Sect. 4.1).

As an example, the Aitken–Neville table for a polynomial of third order improving results will be presented in Table 2.

[10] The results R_m correspond to $\boldsymbol{y}(T+\tau)$ in the former section, with index m denoting different subdivisions of the interval τ.

Table 2 Recursion for results-vector $\boldsymbol{R}_q^i(0)$

Time-step:	τ	$\frac{\tau}{2}$	$\frac{\tau}{4}$	$\frac{\tau}{6}$
Results for the corresponding time-step	$\boldsymbol{R}_0^0$	$\boldsymbol{R}_0^1$	$\boldsymbol{R}_0^2$	$\boldsymbol{R}_0^3$
First step of recursion	$\boldsymbol{R}_1^0$	$\boldsymbol{R}_1^1$	$\boldsymbol{R}_1^2$	
Second step of recursion	$\boldsymbol{R}_2^0$	$\boldsymbol{R}_2^1$		
Third step of recursion	$\boldsymbol{R}_3^0$			

4.3 Variable Step-Size Determination

The implementation of an efficient step-size choosing algorithm can be done following a method proposed by Deuflhard, which can be found in the Recipes [27].

The goal is to find the right trade-off between a further separation of the global time-step τ, thus increasing the order of the extrapolation polynomial which will allow for larger jumps, and its corresponding costs in function evaluations growing drastically in this process.

Assuming that the error estimate of the extrapolation algorithm in the ith column of the Aitken–Neville scheme is given by ϵ_i, the following relation holds

$$\epsilon_i < \epsilon. \tag{41}$$

Deufelhard's interval splitting sequence n_i (30) is correlated with ϵ_i through

$$\epsilon_i \sim \tau^{2i+1}. \tag{42}$$

Of course, this equation has to be adapted to the order of the core algorithm (see Sect. 4.2). From this relation, one can estimate a first approximation for the step-size aiming for convergence in the ith column of the extrapolation scheme:

$$\tau_i = \tau \cdot (\frac{\epsilon}{\epsilon_i})^{\frac{1}{2i+1}}. \tag{43}$$

But which column should be targeted to achieve convergence in?

A simple answer to this question is found in watching the work that has to be performed in order to build the Aitken–Neville scheme. Essentially, the costs will be defined via the number of interval parts n_i, where a full turn of the "Modified Midpoint" algorithm is defined to have a working expenditure of $a = 1$. So the total amount of work of an interval splitting procedure following Deuflhard [8] will be.

$$\begin{aligned} a_0 &= 1 \\ a_{i+1} &= a_i + n_{i+1} \quad . \end{aligned} \tag{44}$$

Consequently, one introduces the dimensionless work per unit-step W:

$$\begin{aligned} W_i &= \frac{a_i \tau}{\tau_i} \\ &= a_i \left(\frac{\epsilon_i}{\epsilon}\right)^{\frac{1}{2i+1}}. \end{aligned} \tag{45}$$

The best column for convergence is, of course, the one achieved by a minimum of W_i:

$$W_{i_{\text{opt}}} = \min\ W_i. \tag{46}$$

For the sake of consequent notation, the index i_{opt} is renamed to q.

While performing the numerical integration, this method will work perfectly well. For the initial step, a method of information theory is used to determine the best column of convergence q, namely the *mean estimated convergence behavior* $\alpha(i, q)$:

$$\alpha(i, q) = \epsilon^{\frac{a_{i+1} - a_{q+1}}{(2i+1)\cdot(a_{q+1} - a_1 + 1)}} \qquad i < q. \tag{47}$$

The *mean estimated convergence behavior* allows for an assessment of the development of the step-size τ:

$$\tau_{i+1} = \tau_i \alpha(i, i+1). \tag{48}$$

Finding the optimal column of convergence q will now be possible via

$$W_i > W_{i+1} \qquad i \leq q. \tag{49}$$

Using (45) and (48) one can deduce the inequality condition:

$$a_i \cdot \alpha(i, i+1) > a_{i+1} \qquad i \in \mathbb{N}. \tag{50}$$

If the inequality is being breached, the minimum of W_i and with it, the optimal column q_{start} has been found. Another spinoff is the upper boundary for i, namely $i_{\max}$, because a further augmentation of i will no longer cause a gain in efficiency (see relation (49)).

The introduction of a so-called order window will prevent the program to degenerate to low orders i and small step-sizes. This is achieved by allowing a change of step-size just in a certain interval centered around i_{opt}.

4.4 Improvements

Apart from the ansatz of Deuflhard to improve the step-size adaption by means of information theory, Fukushima [15] successfully implemented the idea of extrapolating increments instead of results, thus reducing round-off-errors.

5 Lie Series Integrator

5.1 Introduction

Lie Series, named after the famous Norwegian mathematician Sophus Lie, were used in some papers concerning the first lunar missions of NASA as analytical means to approximate solutions of non-trivial differential equations.[11] Their application as a numerical tool for gravitative N-body simulations was first investigated by Hanslmeier and Dvorak [19], Delva [7], and Lichtenegger [23]. Similar to symplectic algorithms, the ansatz for achieving a solution for the equations of motion with Lie Series is an infinitesimal transformation of a Hamiltonian system with respect to time. The striking difference between a Lie Series integration algorithm and its symplectic counterparts is the fact that the exponential operator of the time transformation will be expanded into a series, instead of being split up into separate mappings. Known downsides of Lie Series-based algorithms are their poor portability (Lie Series algorithms have to be completely redesigned if applied to other problems), the need to find recurrence relations between consecutive derivatives, and the lack of symplecticity in its current form. A major advantage is their excellent performance, especially when used with an adaptive choice of step-size.

5.2 Formalities

Once again, the three non-relativistic differential equations describing the gravitationally derived motion of particle ν are of the following form:

$$m_\nu \ddot{\boldsymbol{r}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\nu m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3},$$

k hereby denotes the Gaussian gravitational constant, $\boldsymbol{r}_\nu$ the position vector of particle ν with m_ν being its corresponding mass.

Dividing these equations by m_ν and splitting them into six first-order differential equations will produce (51) and (52)

$$\dot{\boldsymbol{r}}_\nu = \boldsymbol{v}_\nu, \tag{51}$$

$$\dot{\boldsymbol{v}}_\nu = k^2 \sum_{\mu=1,\nu\neq\mu}^{N} \frac{m_\mu (\boldsymbol{r}_\mu - \boldsymbol{r}_\nu)}{\|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\|^3}. \tag{52}$$

For the sake of not having to mention the Gaussian gravitational constant k explicitly all the way, we propose the following transformation of time T:

[11] For a detailed treatment of Lie Series, see Gröbner [17].

$$t = k \cdot T$$
$$t_j = t_{j-1} + \tau \quad j \in \mathbb{Z}. \tag{53}$$

Here, τ denotes the current step-size. The discrete solutions of (51) and (52) for a given time-step τ may be written in the following way:

$$\boldsymbol{r}_\nu(t_j) = e^{\tau D}\boldsymbol{r}_\nu(t_{j-1}) \tag{54}$$
$$\boldsymbol{v}_\nu(t_j) = e^{\tau D}\boldsymbol{v}_\nu(t_{j-1}). \tag{55}$$

Without any loss in generality, one can substitute the time relation (53) in (54) and (55) and choose a starting point $t_{j-1} \equiv 0$, ending up with

$$\boldsymbol{r}_\nu(\tau) = e^{\tau D}\boldsymbol{r}_\nu(0) \tag{56}$$
$$\boldsymbol{v}_\nu(\tau) = e^{\tau D}\boldsymbol{v}_\nu(0). \tag{57}$$

D is denoting the Lie operator [19]:

$$D = \sum_{i=1}^{3}\sum_{\nu=1}^{N}\left(v_\nu^i \frac{\partial}{\partial r_\nu^i} + \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu r_{\mu\nu}^i \rho_{\nu\mu}^{-3}\frac{\partial}{\partial v_\nu^i}\right). \tag{58}$$

The index i corresponds to the ith component of the respective vectors. We will define the Lie operator to act on one argument only, so the expression $D\boldsymbol{r}_\nu\boldsymbol{v}_\nu$ really means $(D\boldsymbol{r}_\nu)\boldsymbol{v}_\nu$, and we will use the notation $D^n\boldsymbol{r}_\nu$ instead of $D(D(D\ldots\boldsymbol{r}_\nu))$.

In order to simplify the description of gravitational interactions we further introduce the connecting position vector of particles ν and μ ($\boldsymbol{r}_{\nu\mu}$), its norm ($\rho_{\nu\mu}$) denoting the scalar distance between particles ν and μ and their mutual velocity ($\boldsymbol{v}_{\nu\mu}$):

$$\begin{aligned}\boldsymbol{r}_{\nu\mu} &= \boldsymbol{r}_\mu - \boldsymbol{r}_\nu = -\boldsymbol{r}_{\mu\nu}\\ \rho_{\nu\mu} &= \|\boldsymbol{r}_\mu - \boldsymbol{r}_\nu\| = \|\boldsymbol{r}_{\nu\mu}\| \quad .\\ \boldsymbol{v}_{\nu\mu} &= \boldsymbol{v}_\mu - \boldsymbol{v}_\nu = -\boldsymbol{v}_{\mu\nu}\end{aligned}$$

Up to this point, the procedure has not been too different from the initial stages of symplectic integrators, though, the next steps contain the essential difference of the Lie Series integration to symplectic methods. Instead of separating $e^{\tau D}$ into multiple symplectic mappings, we will perform a series expansion of the whole exponential:

$$\begin{aligned}\boldsymbol{r}_\nu(\tau) = e^{\tau D}\boldsymbol{r}_\nu(0) &= \Big(\sum_{n=0}^{\infty}\frac{(\tau D)^n}{n!}\Big)\boldsymbol{r}_\nu(0)\\ &= \left(1 + \tau D + \frac{(\tau)^2}{2!}D^2 + \frac{(\tau)^3}{3!}D^3 + \ldots + O\left(\frac{(\tau)^n}{n!}D^n\right)\right)\boldsymbol{r}_\nu(0) \quad n \in \mathbb{N}_0.\end{aligned}$$

It is fairly obvious, that computing the $D^n \boldsymbol{r}(0)$ up to the required order will constitute the main challenge.[12] Let us have a look at the first sequential applications of the Lie operator to given initial conditions $\boldsymbol{r}(0)$ and $\boldsymbol{v}(0)$:

$$D^0 \boldsymbol{r}_\nu = \boldsymbol{r}_\nu \qquad \text{position}$$

$$D^1 \boldsymbol{r}_\nu = \boldsymbol{v}_\nu \qquad \text{velocities}$$

$$D^2 \boldsymbol{r}_\nu = \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3}) \qquad \text{acceleration}$$

$$\begin{aligned} D^3 \boldsymbol{r}_\nu &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3} + \boldsymbol{r}_{\mu\nu} D\rho_{\nu\mu}^{-3}) \\ &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3} - 3\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-4} D\rho_{\nu\mu}) \\ &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3} - 3\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-4} \rho_{\nu\mu}^{-1} (\boldsymbol{r}_{\mu\nu} \cdot \boldsymbol{v}_{\mu\nu})) \\ &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3} - 3\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-5} (\boldsymbol{r}_{\mu\nu} \cdot \boldsymbol{v}_{\mu\nu})). \end{aligned}$$

In order to make things more transparent for higher derivatives, we will introduce the new variables ϕ and Λ [19]:

$$\begin{aligned} \phi_{\nu\mu} &= \rho_{\nu\mu}^{-3} \\ \Lambda_{\mu\nu} &= \boldsymbol{r}_{\mu\nu} \cdot \boldsymbol{v}_{\mu\nu} = \boldsymbol{r}_{\mu\nu} \cdot D\boldsymbol{r}_{\mu\nu} \end{aligned}$$

Another glance at $D^3 \boldsymbol{r}_\nu$ provides:

$$\begin{aligned} D^3 \boldsymbol{r}_\nu &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-3} + \boldsymbol{r}_{\mu\nu} D\rho_{\nu\mu}^{-3}) \\ &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (\phi_{\nu\mu} D\boldsymbol{r}_{\mu\nu} + D\phi_{\nu\mu} \boldsymbol{r}_{\mu\nu}) \\ &= \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu (D\boldsymbol{r}_{\mu\nu} \phi_{\nu\mu} - 3\boldsymbol{r}_{\mu\nu} \rho_{\nu\mu}^{-2} \phi_{\nu\mu} \Lambda_{\mu\nu}), \end{aligned}$$

[12] As a further convention will refer to all variables related to $\boldsymbol{r}$ and $\boldsymbol{v}$ as being initial conditions $\boldsymbol{r}(0)$, $\boldsymbol{v}(0)$, if not explicitly stated otherwise.

with $D\phi_{\nu\mu} = (-3)\rho_{\nu\mu}^{-2}\phi_{\nu\mu}\Lambda_{\mu\nu}$. Admittedly, it seems like we have not won a lot by throwing in ϕ and Λ, but it can be shown that, using these two variables, it is possible to find recurrence relations for higher derivatives. As the order of the derivative D^n is proportional to the order of the time-step τ, such recurrence relations allow — in theory[13] — to calculate solutions that are exact to arbitrary orders of τ very efficiently.

One may be able to grasp that idea by watching the behavior of the following derivatives $D^4\boldsymbol{r}_\nu$ to $D^6\boldsymbol{r}_\nu$:

$$D^3\boldsymbol{r}_\nu = \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu(\phi_{\nu\mu}D\boldsymbol{r}_{\mu\nu} + D\phi_{\nu\mu}\boldsymbol{r}_{\mu\nu})$$

$$D^4\boldsymbol{r}_\nu = \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu(\phi_{\nu\mu}D^2\boldsymbol{r}_{\mu\nu} + 2D\phi_{\nu\mu}D\boldsymbol{r}_{\mu\nu} + D^2\phi_{\nu\mu}\boldsymbol{r}_{\mu\nu})$$

$$D^5\boldsymbol{r}_\nu = \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu(\phi_{\nu\mu}D^3\boldsymbol{r}_{\mu\nu} + 3D\phi_{\nu\mu}D^2\boldsymbol{r}_{\mu\nu} + 3D^2\phi_{\nu\mu}D\boldsymbol{r}_{\mu\nu} + D^3\phi_{\nu\mu}\boldsymbol{r}_{\mu\nu})$$

$$D^6\boldsymbol{r}_\nu = \sum_{\mu=1,\nu\neq\mu}^{N} m_\mu(\phi_{\nu\mu}D^4\boldsymbol{r}_{\mu\nu} + 4D\phi_{\nu\mu}D^3\boldsymbol{r}_{\mu\nu} + 6D^2\phi_{\nu\mu}D^2\boldsymbol{r}_{\mu\nu} + 4D^3\phi_{\nu\mu}D\boldsymbol{r}_{\mu\nu} + D^4\phi_{\nu\mu}\boldsymbol{r}_{\mu\nu}).$$

The boxed recurrence formulae picture the backbone of the Lie Series integration algorithm:

$$\boxed{D^n\boldsymbol{r}_\nu = \sum_{k=1,l\neq k}^{N} m_\mu \sum_{l=0}^{n-2}\binom{n-2}{l} D^l\phi_{\nu\mu}D^{n-2-l}\boldsymbol{r}_{\mu\nu}} \tag{59}$$

$$\boxed{D^n\phi_{\nu\mu} = \rho_{\nu\mu}^{-2}\Big(\sum_{l=0}^{n-1} a_{n,l+1}D^{n-1-l}\phi_{\nu\mu}D^l\Lambda_{\mu\nu}\Big)} \tag{60}$$

with coefficients

$$a_{n,n} = -3 \qquad n \geq 0$$
$$a_{n,1} = a_{n-1,1} - 2 \qquad n \geq 1$$
$$a_{n,l} = a_{n-1,l-1} + a_{n-1,l} \qquad 1 \leq l < n$$

[13] As can be seen from the series expansion of the exponential, factorial numbers have to be evaluated for this process. Their rapid growth, as well as the weak convergence behavior at high orders, constitutes the major limitations of the accuracy of Lie Series expansions.

$$\boxed{D^n \Lambda_{\mu\nu} = \sum_{i=1}^{3} \sum_{l=0}^{nint(\frac{n}{2})} b_{n,l} D^l \boldsymbol{r}_{\mu\nu} D^{n+1-l} \boldsymbol{r}_{\mu\nu}} \tag{61}$$

with

$$\begin{aligned}
b_{n,0} &= 1 && n \geq 0 \\
b_{n,l} &= b_{n-1,l-1} + b_{n-1,l} && 1 < l < nint(\tfrac{n}{2}) \\
b_{n,nint(\frac{n}{2})} &= b_{n-1,nint(\frac{n}{2})-1} && n \text{ uneven} \\
b_{n,nint(\frac{n}{2})} &= 2b_{n-1,nint(\frac{n}{2})} + b_{n-1,nint(\frac{n}{2})-1} && n \text{ even}
\end{aligned}$$

The final nth order approximation to the solution of (51) and (52) for one time-step τ are given by

$$\boldsymbol{r}_\nu(\tau) = \left(1 + \tau D + \frac{(\tau)^2}{2!} D^2 + \frac{(\tau)^3}{3!} D^3 + \cdots + O\left(\frac{(\tau)^n}{n!} D^n\right)\right) \boldsymbol{r}_\nu(0), \tag{62}$$

$$\begin{aligned}
\boldsymbol{v}_\nu(\tau) &= \left(1 + \tau D + \frac{(\tau)^2}{2!} D^2 + \frac{(\tau)^3}{3!} D^3 + \cdots + O\left(\frac{(\tau)^n}{n!} D^n\right)\right) \boldsymbol{v}_\nu(0) \qquad (63) \\
&= \left(D + \tau D^2 + \frac{(\tau)^2}{2!} D^3 + \frac{(\tau)^3}{3!} D^4 + \cdots + O\left(\frac{(\tau)^n}{n!} D^{(n+1)}\right)\right) \boldsymbol{r}_\nu(0),
\end{aligned}$$

where *nint(x)* is a function giving the nearest integer (e.g., $nint(2.345) = 2$, $nint(2.5) = 3$, $nint(2.876) = 3$). A major point concerning the performance of the Lie Series integrator is the fact that the distances between bodies have to be evaluated *only once*, namely for $D^2 \boldsymbol{r}_\nu$. As this operation scales with the number of bodies squared (N^2), it is the most resource demanding part in every N-body algorithm. The rest of the calculations performed will scale with the number of terms of the Lie Series times the number of bodies ($n \cdot N$). Due to the fact that the best trade off between truncation error and CPU time was found to be around $n = 12$ [19] the Lie Series algorithm will really boost its performance for $N > 12$.

5.3 Variable Step-Size Determination

The fact that the Lie Series integration algorithm is based on a series expansion of an exponential function ((56), (57)) explains its missing symplecticity, due to the fact, that there is no nearby Hamiltonian solved exactly, as is the case with a symplectic partitioning of $e^{\tau D}$. On the other hand it does allow for a fairly easy implementation of a step-size choosing mechanism, through a simple estimate of the remainder of the exponential series.

$$e^{\tau D} = \sum_{l=0}^{n} \frac{(\tau D)^l}{l!} + R_{n+1} \tag{64}$$

$$\|R_{n+1}\| \leqq 2\frac{\|\tau D\|^{n+1}}{(n+1)!}, \tag{65}$$

which holds if $\|\tau D\| \leqq \frac{n}{2} + 1$.

$n \in \mathbb{N}$ is, as stated above, the number of included terms in the exponential series. This estimate was achieved using the geometric series in Forster [13]. Assuming the error of neglecting the remainder R_{n+1} to be less than a certain boundary value ϵ, one can derive the following relations using (65):

$$\begin{aligned} &\|R_{n+1}\| \leqq \epsilon, \\ &\tau \leqq \left(\epsilon \frac{(n+1)!}{2\|D^{n+1}\|}\right)^{\frac{1}{n+1}}. \end{aligned} \tag{66}$$

As the argument of the exponential function contains an operator part (64), it is essential to know, if all of the arguments are commutating. If this was not the case, then partitioning $\|\tau D\|^{n+1} \leqq \|(\tau)\|^{n+1} \cdot \|D^{n+1}\|$ would not work without taking the Baker–Campbell–Hausdorff development into account. Though, as the Lie operator D does not depend explicitly on time, both arguments of the exponential function commutate indeed.

For practical reasons, the authors would recommend to use the maximum value of all components of $D^{n+1}\boldsymbol{r}_v$ for the choice of the next-time step, because this term will also yield a maximum influence on the remainder (see (65)).

Strictly speaking, relation (66) gives the right step-size for the *current* values of the derivatives, which means for the *current* step. As it is not very efficient to compute everything twice per step, the prime assumption is made, that the problem is so "well behaved" that the best fitting current τ will also be valid for the next time-step.

5.4 Order Control

An appealing alternative to the common step-size control mechanism stated in the previous section is the notion of Order Control. The basic idea is to use a fixed step-size but adapting the order of the integration algorithm to keep errors below a certain boundary. Let us look once more at (65). Assuming a fixed time-step τ it is very easy to compute the remainder at the end of an integration step. If the value of the remainder falls below the desired error ϵ, then the algorithm may proceed with the next step. If not, then the number of terms in the series is raised by one remaining at the same point in time. At the end of this turn, the remainder is again compared to the desired error, and so on. Consequently, the Lie Series algorithm chooses the optimal order for a given time-step throughout the whole integration.

As the order of the Lie Series expansion is actually limited by a computer's ability to calculate and store greater factorial numbers, and as the global convergence weakens with growing order, the adaptive potential of an order-controlled algorithm is rather limited. Nevertheless, Order Control is a possible way to decrease the average error of fixed step Lie Series integration in "not-too-badly behaved" problems.

6 Symplectic Integrators

6.1 Introduction

The theory of symplectic integrators is rooted in the fact that certain geometrical properties of the phase space of a conservative mechanical system are time-invariant. In the case of a phase space being spanned by generalized coordinates q and their conjugated momenta p, the conservation of symplectic geometry can be defined via the invariant differential form ω with

$$\omega = \sum_{i=1}^{n} dp_i \wedge dq_i, \tag{67}$$

where n defines the dimension in configuration space, and $2n$ the dimension of the associated phase space. For the current non-relativistic gravitational N-body problem, n equals $3N$.

The link of symplecticity to classical descriptions of mechanical systems is given through Hamilton's equations:

$$\begin{aligned} \dot{\boldsymbol{p}} &= -\nabla_{\boldsymbol{q}} H \\ \dot{\boldsymbol{q}} &= \nabla_{\boldsymbol{p}} H, \end{aligned} \tag{68}$$

which are symplectomorphic, i.e., they support the conservation of the symplectic differential form (67).

By their special way of construction, symplectic algorithms are "aware" of these geometrical properties and thus very potent in modeling the propagation of dynamical systems of this kind.

For systems supporting a Hamiltonian of the form:

$$H = T(\boldsymbol{p}) + V(\boldsymbol{q}), \tag{69}$$

Candy and Rozmus [1] were able to construct a symplectic integration algorithm of fourth order, which will be described in the following section.

6.2 Formalities

Following Neri [26] we introduce $\boldsymbol{q}$ and $\boldsymbol{p}$ which are three-dimensional generalized coordinates and conjugate impulses of the point masses numbered $\nu = 1, ..., N$. The Hamiltonian be separable, which means it has to be of the form:

$$H = T(\boldsymbol{p}) + V(\boldsymbol{q}). \tag{70}$$

The Hamiltonian equations (68) can be rewritten using the phase-space vector $\boldsymbol{z}$ and Poisson's differential operator D_H:

$$\begin{aligned} \dot{z} &= \{z, H(z)\} \\ \dot{z} &= D_H z \end{aligned} \tag{71}$$

with

$$\begin{aligned} z &= \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \end{pmatrix} \\ D_H &= \{\_, H\}. \end{aligned} \tag{72}$$

Here the braces $\{\_, \_\}$ denote the Poisson brackets

$$\{F, G\} = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial q_i} \frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial G}{\partial q_i} \right). \tag{73}$$

The formal solution of (71) is given by

$$z(\tau) = e^{\tau D_H} z(0) \tag{74}$$

$$z(\tau) = e^{\tau (D_T + D_V)} z(0). \tag{75}$$

In a next step $e^{\tau(D_T+D_V)}$ will be decomposed into two independent mappings, one in direction of D_T and a second one in direction of D_V.

$$e^{\tau(D_T+D_V)} \rightarrow e^{\tau D_T} e^{\tau D_V}.$$

As D_T and D_V will not commutate, one has to be careful about splitting the exponential function $e^{\tau(D_T+D_V)}$. A quick look at the Baker–Campbell–Hausdorff identity (76) reveals, that the intended decomposition matches up to terms of commutators $[\_, \_]$ in D_T and D_V:

$$e^{\tau D_T} e^{\tau D_V} = e^{(\tau(D_T+D_V) + \frac{\tau^2}{2}[D_T, D_V] + \frac{\tau^3}{12}([D_T,[D_T,D_V]] - [D_V,[D_T,D_V]]) + ...)} \tag{76}$$

with

$$[D_T, D_V] = D_T D_V - D_V D_T.$$

For this reason an expansion with coefficients a^i and b^i is being performed, in order to cancel out unwanted terms containing commutators in (76) up to $O(\tau^{k+1})$:

$$e^{\tau(D_T+D_V)} \overset{O(\tau^{k+1})}{=} \prod_{i=1}^{k} e^{a^i \tau D_T} e^{b^i \tau D_V}. \tag{77}$$

Both exponentials $e^{a^i \tau D_T}$ and $e^{b^i \tau D_V}$ are symplectic mappings in their own right, and through properties of symplectic mappings, their product will be symplectic, too [18].

Consider a mapping $e^{a^i \tau D_T}$:

$$\begin{pmatrix} \boldsymbol{q}^{\,i-1} \\ \boldsymbol{p}^{\,i-1} \end{pmatrix} \to \begin{pmatrix} \boldsymbol{q}^{\,i} \\ \boldsymbol{p}^{\,i-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{q}^{i-1} + \tau a^i \nabla_{\boldsymbol{p}^{i-1}} T \\ \boldsymbol{p}^{i-1} \end{pmatrix}. \tag{78}$$

This changes just the coordinates $\boldsymbol{q}$. $\nabla_{\boldsymbol{p}^{i-1}}$ states the directional derivative with respect to our generalized impulses $\boldsymbol{p}^{i-1}$. The second mapping $e^{b^i \tau D_V}$ completes the transformation

$$\begin{pmatrix} \boldsymbol{q}^{\,i} \\ \boldsymbol{p}^{\,i-1} \end{pmatrix} \to \begin{pmatrix} \boldsymbol{q}^{\,i} \\ \boldsymbol{p}^{\,i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{q}^{i} \\ \boldsymbol{p}^{i-1} - \tau b^i \nabla_{\boldsymbol{q}^i} V \end{pmatrix}. \tag{79}$$

$\nabla_{\boldsymbol{q}^i}$ again states the directional derivative, but this time with respect to the generalized coordinates. The resulting Candy algorithm can be set up as follows [34].

Initial conditions $\boldsymbol{q}_{t,\nu}^{\,0}$ and $\boldsymbol{p}_{t,\nu}^{\,0}$ for a particle ν are taken at time t. Put up a loop with counting index $i = 1, ..., 4$ containing

$$\begin{aligned} \boldsymbol{q}_{t,\nu}^{\,i} &= \boldsymbol{q}_{t,\nu}^{\,i-1} + \tau a^{\,i} \nabla_{\boldsymbol{p}^{i-1}} T_\nu \\ \boldsymbol{p}_{t,\nu}^{\,i} &= \boldsymbol{p}_{t,\nu}^{\,i-1} - \tau b^{\,i} \nabla_{\boldsymbol{q}^i} V_\nu, \end{aligned} \tag{80}$$

with the coefficients

$$\begin{aligned} a^1 = a^4 &= \frac{1}{2(2-2^{1/3})} \qquad a^2 = a^3 = \frac{1-2^{1/3}}{2(2-2^{1/3})}. \\ b^1 = b^3 &= \frac{1}{2-2^{1/3}} \qquad b^2 = \frac{-2^{1/3}}{2-2^{1/3}} \qquad b^4 = 0 \end{aligned}$$

The results of this time-step will be the initial conditions of the next one:

$$\boldsymbol{p}_{t+1,\nu}^{\,0} = \boldsymbol{p}_{t,\nu}^{\,4} \qquad \boldsymbol{q}_{t+1,\nu}^{\,0} = \boldsymbol{q}_{t,\nu}^{\,4}. \tag{81}$$

The Hamiltonian of the N-body gravitational problem is given by

$$H = \sum_{\nu=1}^{N} \left(\frac{\boldsymbol{p}_\nu{}^2}{2m_\nu} - k^2 m_\nu \sum_{\mu \neq \nu}^{N} \frac{m_\mu}{|\boldsymbol{q}_\mu - \boldsymbol{q}_\nu|} \right). \tag{82}$$

The necessary derivatives for a particle ν are calculated as follows:

$$\begin{aligned} -\nabla_{\boldsymbol{q}} H_\nu &= k^2 m_\nu \sum_{\mu \neq \nu}^{N} \frac{m_\mu}{|\boldsymbol{q}_\mu - \boldsymbol{q}_\nu|^3} \left(\boldsymbol{q}_\mu - \boldsymbol{q}_\nu\right), \\ \nabla_{\boldsymbol{p}} H_\nu &= \nabla_{\boldsymbol{p}} \frac{p_\nu{}^2}{2\, m_\nu} = \frac{1}{2\, m_\nu} \left(2\, \boldsymbol{p}_\nu \nabla_{\boldsymbol{p}} \boldsymbol{p}_\nu\right) = \frac{\boldsymbol{p}_\nu}{m_\nu}. \end{aligned} \tag{83}$$

As stated earlier, (80) constitute two symplectic mappings, one performing the transformation $\binom{\boldsymbol{p}^{i-1}}{\boldsymbol{q}^{i-1}} \rightarrow \binom{\boldsymbol{p}^{i}}{\boldsymbol{q}^{i-1}}$ and the other one producing $\binom{\boldsymbol{p}^{i}}{\boldsymbol{q}^{i-1}} \rightarrow \binom{\boldsymbol{p}^{i}}{\boldsymbol{q}^{i}}$. The combined mapping is again symplectic [20], so the phase space structure remains intact throughout the integration process.

Equation (77) does not mean that the method constructed is equivalent to an arbitrary algorithm, which simply conserves the Hamiltonian up to order $O(\tau^k)$. Yoshida [36] was able to prove that a symplectic algorithm will conserve *a nearby* Hamiltonian $\tilde{H}$ *exactly*, which explains why there is no secular increase in the deviation of energy. The conserved nearby Hamiltonian $\tilde{H}$ differs from the original Hamiltonian H because of (76):

$$\begin{aligned} &e^{\tau D_H} = e^{\tau (D_T + D_V)} \\ &H = T + V \\ &e^{\tau D_T} e^{\tau D_V} = e^{\tau D_{\tilde{H}}}, \end{aligned}$$

$$\tilde{H} = T + V + \frac{\tau}{2}[T, V] + \frac{\tau^2}{12}([T, [T, V]] - [V, [T, V]]) + ... \tag{84}$$

If, of course, commutating terms are eliminated, as was done in (77) the integrated Hamiltonian $\tilde{H}$ will be closer to the original Hamiltonian H.

6.3 Improvements

For applications in the field of Dynamical Astronomy, Wisdom and Holman [35] found that splitting the Hamiltonian in terms of kinetic and potential energy is not the only possibility. In fact, separating the Hamiltonian into a part representing the Keplerian motion of each planet (H_{Kep}) and a second part, containing perturbation terms due to mutual interactions with other planets (H_{Int}), leads to a decrease in

longterm integration error proportional to $O(\epsilon\tau^2)$ instead of $O(\tau^2)$ for a second-order symplectic scheme, assumed that ($H_{\rm Int} \simeq \epsilon H_{\rm Kep}$) with ϵ denoting the planetary to stellar mass ratio. A very natural way to achieve this separation of the Hamiltonian arises when using Jacobi coordinates, where the position of the innermost planet is defined with respect to the central star, the other planets are added one by one and their positions are measured with respect to the consequently updated barycenter of the system. Another advantage of this method is the fact that the Keplerian part of the Hamiltonian can be advanced through Gauss's f and g functions [6] very efficiently, and as $H_{\rm Int}$ is a function of coordinates only within a Jacobian system, it may also be solved analytically.

Unfortunately this special separation of the Hamiltonian will loose its advantages when the condition $H_{\rm Int} \simeq \epsilon H_{\rm Kep}$ is violated, which will happen, when members of the system come close to each other (close encounters). Apart from that, the choice of Jacobi coordinates is not quite suited, if the radial ordering of the planets' orbits will not remain constant, which may be due to large eccentricities, for example.

6.4 Variable Step-Size

It was found (see, e.g., Saha and Tremaine [29], Yoshida [37]) that usual step-size-choosing techniques will destroy the favorable properties of symplectic algorithms. The reason for this behavior can be found in (84). As stated in the previous section, the Hamiltonian, that is conserved exactly, is $\tilde{H}$. This Hamiltonian $\tilde{H}$ *depends on the step-size* τ, even though some of the lower order dependencies may have been eliminated, as in (77). Consequently, if the step-size is changed, the integrated Hamiltonian also will, and a secular increase in the energy error will arise.

Different ansaetze have evolved to counter this dilemma, e.g., to use different timescales during the integration [22], symplectic implicit Runge–Kutta methods with symmetric determination of step-sizes [2], or hybrid solutions [4], the latter one being investigated in the following section.

7 Hybrid Integrators

7.1 Introduction

The main problem with symplectic integration algorithms is their inherent inability to adapt step-sizes during an ongoing computation. This is due to the fact, that the Hamiltonian actually solved by numerical integration $H_{\rm num}$ differs from the analytic Hamiltonian $H_{\rm an}$ by an expression proportional to all orders of the step-size τ, which means that changing the step-size automatically alters the integrated Hamiltonian, and will thus destroy the algorithm's energy conserving properties [29].

Given the need to calculate so-called close encounters (CE)[14] in celestial mechanics, this situation leaves users of symplectic algorithms with the utterly displeasing possibilities of choosing a tiny step-size right from the start, which will radically increase computational resource demands and round-off errors,[15] stopping calculations whenever a CE occurs, or simply ignoring CE, admitting that from this point onward the calculation has statistical significance at best.

An intriguing approach to circumnavigate this dilemma has been brought up by Chambers [4]. He combined a second-order mixed variable symplectic integrator [35] with fixed step-size and a Bulirsch–Stoer type extrapolation algorithm [32], using the symplectic part plus analytical advancing via Gauss's f and g functions [6], while close encounters that require changes in step-size are performed by the Bulirsch–Stoer method.

As there are some subtleties involved, the following section will just give a rough outline of the ideas involved.

7.2 Formalities

Let us again have a look at the time propagation equation (74), found in Chap.6

$$z(\tau) = e^{\tau D_H} z(0),$$

with z being the phase space vector of the observed system

$$z = \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \end{pmatrix}$$
$$D_H = \{\_, H\}$$

and $\{\_, \_\}$ denoting the Poisson brackets defined in (73). As the Hamiltonian H is separable in terms of coordinates and momenta, (74) can be rewritten as (75)

$$H = T(\boldsymbol{p}) + V(\boldsymbol{q})$$
$$z(\tau) = e^{\tau(D_T + D_V)} z(0).$$

Actually, there is no need to separate the Hamiltonian in this special way. Let us instead split the Hamiltonian in a part representing the unperturbed Keplerian

[14] "Close encounters" are very close approaches of two members of an N-body system for a finite time span, where the two-body interactions become the dominant forces.

[15] Round-off errors result from finite precision number representation and calculations in computers. Roughly speaking the round-off error grows with $\sqrt{N}$, where N is the number of calculations performed. For a more detailed review of round-off errors in symplectic integration algorithms see Gladman et al. [16].

motion of each planet around its guiding center (H_{Kep}) and the mutual perturbations (H_{Int}):

$$H = H_{\text{Kep}}(z) + H_{\text{Int}}(z) \qquad z(\tau) = e^{\tau(D_{H_{\text{Kep}}} + D_{H_{\text{Int}}})} z(0). \tag{85}$$

Wisdom and Holman [35] found that for such a separation of the Hamiltonian, it is rather convenient to use Jacobi coordinates. The problem with Jacobi coordinates is that they assume the radial ordering of planets in the system to be fixed, which will not be the case in general when notable eccentricities are being introduced. Consequently, a new set of coordinates has to be found that include the benefits of Jacobi coordinates, such as three spatial coordinates for the center of mass of the system, but treat them without making any assumptions on their orbits. In order to fulfill these requirements, we perform a change from barycentric to "democratic heliocentric" variables (DHV) derived in Duncan et al. [10], where the position of each planet is measured with respect to the central body, but the coordinates of the central body are replaced by those of the barycenter:

$$\boldsymbol{Q}_0 = \frac{m_0 \boldsymbol{q}_0 + \sum_{\mu=1}^{N} m_\mu \boldsymbol{q}_\mu}{m_{tot}}, \tag{86}$$

$$\boldsymbol{Q}_\nu = \boldsymbol{q}_\nu - \boldsymbol{q}_0. \tag{87}$$

$\boldsymbol{Q}_\nu$ denotes the new DHV, $\boldsymbol{q}_\nu$ the old barycentric coordinates, and m_ν the mass of body i. $N+1$ is the total number of bodies involved, and index 0 refers to the central body.

The canonically conjugate momenta are defined as

$$\boldsymbol{P}_0 = \boldsymbol{p}_0 + \sum_{\mu=1}^{N} \boldsymbol{p}_\mu, \tag{88}$$

$$\boldsymbol{P}_\nu = \boldsymbol{p}_\nu - \frac{m_\nu}{m_{tot}} \left(\boldsymbol{p}_0 + \sum_{\mu=1}^{N} \boldsymbol{p}_\mu \right). \tag{89}$$

Again, the momenta of the central body are replaced by the total momentum of the system, whereas the others simply correspond to barycentric momenta.

Even though the DHV are canonical, they produce some extra terms in the separation of our Hamiltonian (H), which we will denote H_{Jump}. More precisely the separated parts of the Hamiltonian will be

$$H = H_{\text{Kep}}(\boldsymbol{Z}) + H_{\text{Int}}(\boldsymbol{Z}) + H_{\text{Jump}}(\boldsymbol{Z}) \tag{90}$$

$$\boldsymbol{Z}(\tau) = e^{\tau(D_{H_{\text{Kep}}} + D_{H_{\text{Int}}} + D_{H_{\text{Jump}}})} \boldsymbol{Z}(0). \tag{91}$$

For the sake of clarity the compact notation in phase space coordinates $\boldsymbol{Z}$ will be interrupted in the next set of equations. Instead we reintroduce the momenta $\boldsymbol{P}_\nu$ and mutual distances $R_{\nu\mu} = \| \boldsymbol{Q}_\mu - \boldsymbol{Q}_\nu \|$ of bodies $\nu, \mu \in N$:

$$H_{\mathrm{Kep}} = \sum_{\nu=1}^{N} \left(\frac{P_\nu^2}{2m_\nu} - k^2 \frac{m_0 m_\nu}{R_\nu} \right), \tag{92}$$

$$H_{\mathrm{Int}} = -k^2 \sum_{\nu=1}^{N} \sum_{\mu>\nu}^{N} \frac{m_\nu m_\mu}{R_{\nu\mu}}, \tag{93}$$

$$H_{\mathrm{Jump}} = \frac{1}{2m_0} \left(\sum_{\nu=1}^{N} \boldsymbol{P}_\nu \right)^2. \tag{94}$$

Here, a term $\frac{P_0^2}{2m_{\mathrm{tot}}}$ has been omitted in the previous equations, because it would only account for a constant motion of the barycenter. As one can see from (93) and (94), H_{Int} *and* H_{Jump} are small corrections to the Keplerian motion, assuming that the mutual distances $R_{\nu\mu}$ are great and the momenta P_ν are small compared to the momentum of the central body.

We may gain a second-order symplectic integration algorithm by a symmetrized reformulation of (91), keeping in mind that the Baker–Campbell–Hausdorff formula holds (see [4]).

$$\boldsymbol{Z}(\tau) = e^{(\frac{\tau}{2} D_{H_{\mathrm{Int}}})} e^{(\frac{\tau}{2} D_{H_{\mathrm{Jump}}})} e^{(\tau D_{H_{\mathrm{Kep}}})} e^{(\frac{\tau}{2} D_{H_{\mathrm{Jump}}})} e^{(\frac{\tau}{2} D_{H_{\mathrm{Int}}})} \boldsymbol{Z}(0) \tag{95}$$

Just as with the mapping of Wisdom and Holman [35], the Keplerian part of the motion can be advanced through Gauss's f and g functions [6]. The fact that H_{Int} and H_{Jump} — small adjustments by definition — contain direct terms only is another advantage of DHV compared to Jacobi coordinates.

This algorithm works fine, as long as

$$H_{\mathrm{Int}} << H_{\mathrm{Kep}}. \tag{96}$$

Allowing for CEs will, of course, violate this relation, because the mutual influence of the participating bodies becomes comparable to Keplerian forces ($H_{\mathrm{Kep}} \simeq H_{\mathrm{Int}}$). As the reason for separating the Hamiltonian in a Keplerian and an interaction part is the improvement of the integration method's error properties,[16] a failure to comply to relation (96) will result in a substantial loss of accuracy.

Following Chambers [4], this problem can be remedied by ensuring that condition (96) remains satisfied even during CEs. This is done by transferring the increasing terms of H_{Int} to H_{Kep} via some partition function Γ that has the following properties:

[16] If $H_{\mathrm{Int}} \simeq \epsilon H_{\mathrm{Kep}}$ holds, then the approximation of the algorithm will be $O(\epsilon\tau^2)$ instead of $O(\tau^2)$, with ϵ denoting the planetary to stellar mass ratio [35].

$$\text{if } R \text{ is large, then } \Gamma(R) = 1 \qquad \text{if } R \to 0, \text{ then } \Gamma(R) \to 0.$$

This will result in a new separation of the Hamiltonian:

$$H_{\text{Large}} = \sum_{\nu=1}^{N} \left(\frac{P_\nu^2}{2m_\nu} - k^2 \frac{m_0 m_\nu}{R_\nu} \right) - k^2 \sum_{\nu=1}^{N} \sum_{\mu>\nu}^{N} \frac{m_\nu m_\mu}{R_{\nu\mu}} [1 - \Gamma(R_{\nu\mu})], \tag{97}$$

$$H_{\text{Small}} = -k^2 \sum_{\nu=1}^{N} \sum_{\mu>\nu}^{N} \frac{m_\nu m_\mu}{R_{\nu\mu}} \Gamma(R_{\nu\mu}), \tag{98}$$

$$H_{\text{Jump}} = \frac{1}{2m_0} \left(\sum_{\nu=1}^{N} \boldsymbol{P_\nu} \right)^2. \tag{99}$$

The new mapping will be constructed the same way as (95):

$$\boldsymbol{Z}(\tau) = e^{(\frac{\tau}{2} D_{H_{\text{Small}}})} e^{(\frac{\tau}{2} D_{H_{\text{Jump}}})} e^{(\tau D_{H_{\text{Large}}})} e^{(\frac{\tau}{2} D_{H_{\text{Jump}}})} e^{(\frac{\tau}{2} D_{H_{\text{Small}}})} \boldsymbol{Z}(0). \tag{100}$$

Unlike H_{Kep}, H_{Large} contains terms describing three-body motion and can therefore no longer be advanced analytically. Chambers [4] decided to evolve the system through such phases, by using a Bulirsch–Stoer algorithm with adaptive step-size, that calculates the advancement through H_{Large} to machine precision, so that — in theory — there should be no difference to analytically derived solutions. As the introduction of sequences of Bulirsch–Stoer integration will result in a significant loss of computation time, the non-symplectic integration will be performed for the interacting bodies only, while H_{Large} of all unperturbed particles is still advanced using Gauss's functions.

In summary, the Hybrid algorithm proposed by Chambers [4], reformulates the mixed variable symplectic algorithm of Wisdom and Holman [35] by introducing a new set of canonical coordinates (DHV) and includes a common non-symplectic algorithm with adaptive step-size. This enables the Hybrid integrator, even though having still a fixed step-size for its symplectic mapping part, to deal with CEs quite efficiently.

7.3 Improvements

The Hybrid algorithm described in the previous section is known to work for systems with one dominating central mass. When applied to binary systems, one will encounter some difficulties, as the second star will influence H_{Jump} heavily due to its enormous momentum, so that the condition $H_{\text{Jump}} << H_{\text{Kep}}$ will not be satisfied. Consequently the error per step becomes large compared to central mass systems. Solutions to this problem are discussed by Chambers [4]. Similar to the fact that the use of DHV instead of Jacobi coordinates made it possible to treat problems with no

constant radial order of orbits, the change from DHV to specially designed binary variables, and later on "Yosemite coordinates" in connection with a scheme derived by Duncan and Levison [11] solved most of the difficulties.

8 Comparison

In order to get an impression of the governing properties of the six algorithms presented, we will briefly compare the methods to analytically predicted solutions. Since the two-body problem is the only multibody gravitating system, that is perfectly integrable, it is an obvious choice in this respect. For testing, the system of the Sun and Jupiter has been chosen. Initial conditions for the equinox J2000 are readily available at Solar System Dynamics of JPL [25]. All numeric calculations were performed using the *mercury6* package [4] on the one hand containing Bulirsch–Stoer, Radau15, and the Hybrid algorithms, and the author-developed *nie* package on the other hand containing, among others, a Cash–Karp Runge–Kutta, a Lie Series and a fourth-order Candy integrator.

8.1 The Kepler Problem

As all the Keplerian elements, except for the mean anomaly (M), describe the size and orientation of a given orbit, they are to be viewed as constants of motion. This means, in principle, that the *semi-major axis (a), numeric eccentricity (e), inclination (i), argument of perihelion (ω)*, and *argument of the ascending node (Ω)* are meant to remain unchanged for all times, if Jupiter alone orbits the Sun in an empty universe that is ignoring relativity. The fact that the only force acting in this setup is Newtonian gravitation leads to *conservation of total energy (E)* as well as *total angular momentum (L)*[17] in our system.[18]

So, the easiest way to check on the reliability of the algorithms presented, is to watch the behavior of these conserved quantities:

$$E = E_{kin} + E_{pot} = \sum_{\nu=1}^{N} \frac{\boldsymbol{p}_\nu^2}{2m_\nu} - k^2 \sum_{\nu=1,\mu\neq\nu}^{N} \frac{m_\nu m_\mu}{\|\boldsymbol{q}_{\nu\mu}\|}, \tag{101}$$

$$L = \| \sum_{\nu=1}^{N} (\boldsymbol{q}_\nu \times \boldsymbol{p}_\nu) \|. \tag{102}$$

[17] In fact as the angular momentum is a vectorized quantity, not only its length but also its direction is conserved. In the following we will just check its length directly.

[18] Actually, more conserved quantities do exist in the two body problem, e.g. the Laplace-Runge-Lenz vector, pointing towards the pericenter. Yet, there is no analogon for $N > 2$, therefore we omit its calculation for the lack of comparability to general N-body results.

In order to being able to compare six numerical algorithms with entirely different orders and step choosing mechanisms, the Jupiter–Sun system was integrated to have *roughly the same overall deviation of total energy from its initial value*[19] for each method at the end of a time span of 10^7 ephemerical days [D]. That way, all the solutions are of comparable quality at the end of this interval.

The final error in energy achieved here, is by far not the best each algorithm is able to manage, though, as we would like to see the actual behaviour of these methods, there is no point in choosing stricter error constraints, that would lead to a major influence of round-off errors. As Fig. 3 demonstrates, this cannot be done perfectly, so we were content to ensure that this requirement was fulfilled to the same order of magnitude. Having a closer look at Fig. 3 reveals that the main desired property of symplectic algorithms, namely a linear growth in total energy error, can indeed be detected, whereas all non-symplectic methods show a quadratic total deviation from the initial energy.

In the picture presenting the total angular momentum error, the symplectic methods are so low valued that they do not even register.

The reason for this becomes clear in Fig. 4. These pictures show the logarithmic, normalized values of momentary deviation from initial energy and initial angular momentum. It is quite obvious that the momentary deviations in energy and angular momentum of non-symplectic algorithms follow the same, roughly linear growth. In contrast to this, the angular momentum conservation of symplectic algorithms is practically flawless causing the random walk pattern (Fig. 4, below), that is typical for calculations close to machine precision.[20] Another intriguing topic is the form of the logarithm of the deviation from initial energy of the Candy and Hybrid methods (Fig. 4, middle picture). First of all, the maximum deviation from initial energy stays bounded during the whole integration interval, thus, when summed, leading to the celebrated linear slope of the total energy error. The peculiar, seemingly osculating appearance is caused by the fact that a *nearby* Hamiltonian is exactly integrated. During its evolution, this nearby system comes close to the original system, explaining the constant drop in local energy errors also seen in Fig. 4.

Until now, the symplectic integrators seem to be the obvious choice for solving astrodynamical problems. Let us have a look at the development of the Keplerian orbital elements. As every element except for the mean anomaly should theoretically remain constant, we can apply the same procedure, as for energy and angular momentum, summing over the difference between initial values and numerically evaluated results. The mean anomaly will be compared to its analytically derived value.

---

[19] The overall deviation of energy is defined as the sum of all deviations measured every time a snapshot of the problem was taken: $\Sigma\Delta E = \sum_{t_s}(E_{t_s} - E_0)$. In other words, after each output time interval, the deviation from the initial energy value was calculated and summed. The output time interval has been the same for each algorithm.

[20] The fact that symplectic integrators actually conserve some nearby Hamiltonian flawlessly also leads to the conservation of all functions $f(q, p)$ whose Poisson bracket with the Hamiltonian vanishes. As this is the case for the angular momentum vector, it is perfectly conserved in symplectic integrators.

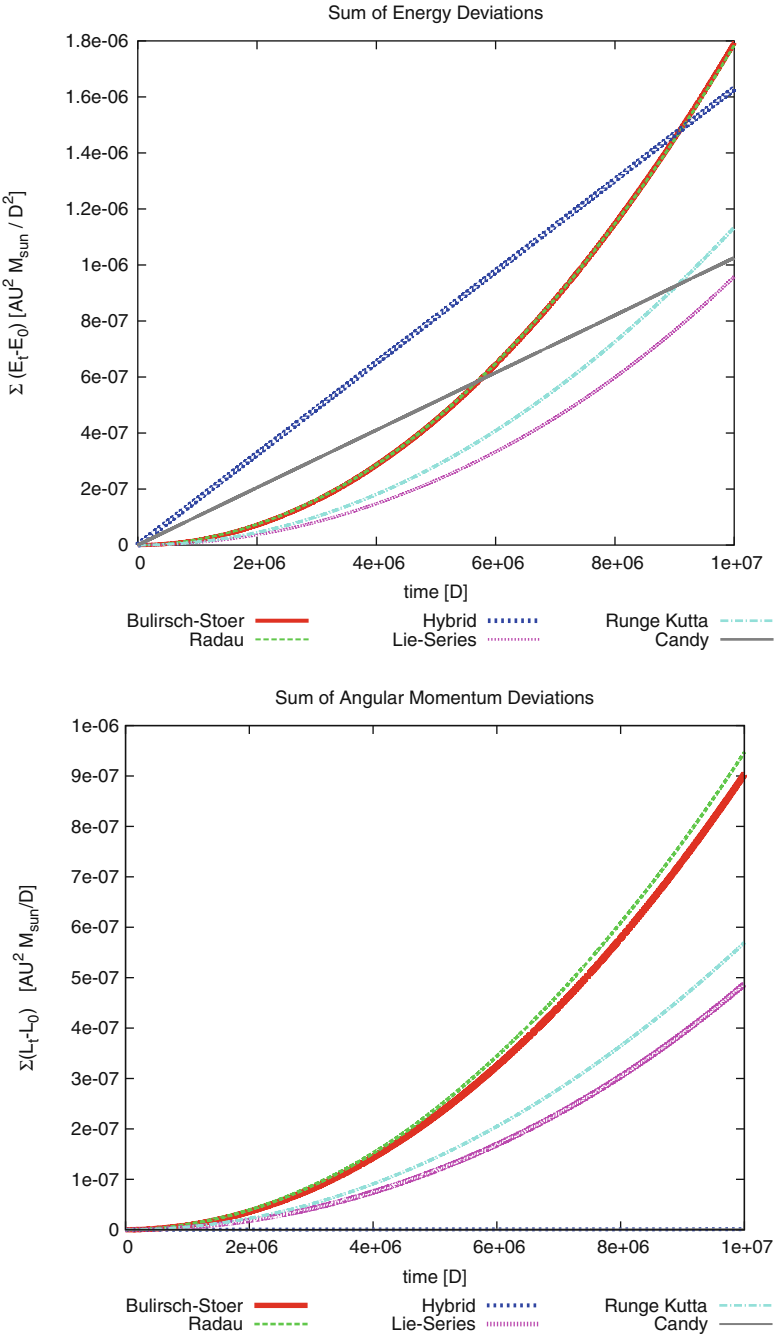

Fig. 3 Sum of deviations of total energy from its initial value for all integration algorithms (*above*); sum of deviations of total angular momentum from its initial value for all integrators (*below*); the accuracy parameters of all algorithms were chosen, such that the total sum of deviations of energy at time 10^7 [D] are roughly in the same order of magnitude. The symplectic algorithms show a linear trend in the deviation of total energy, whereas the non-symplectic show a quadratic trend

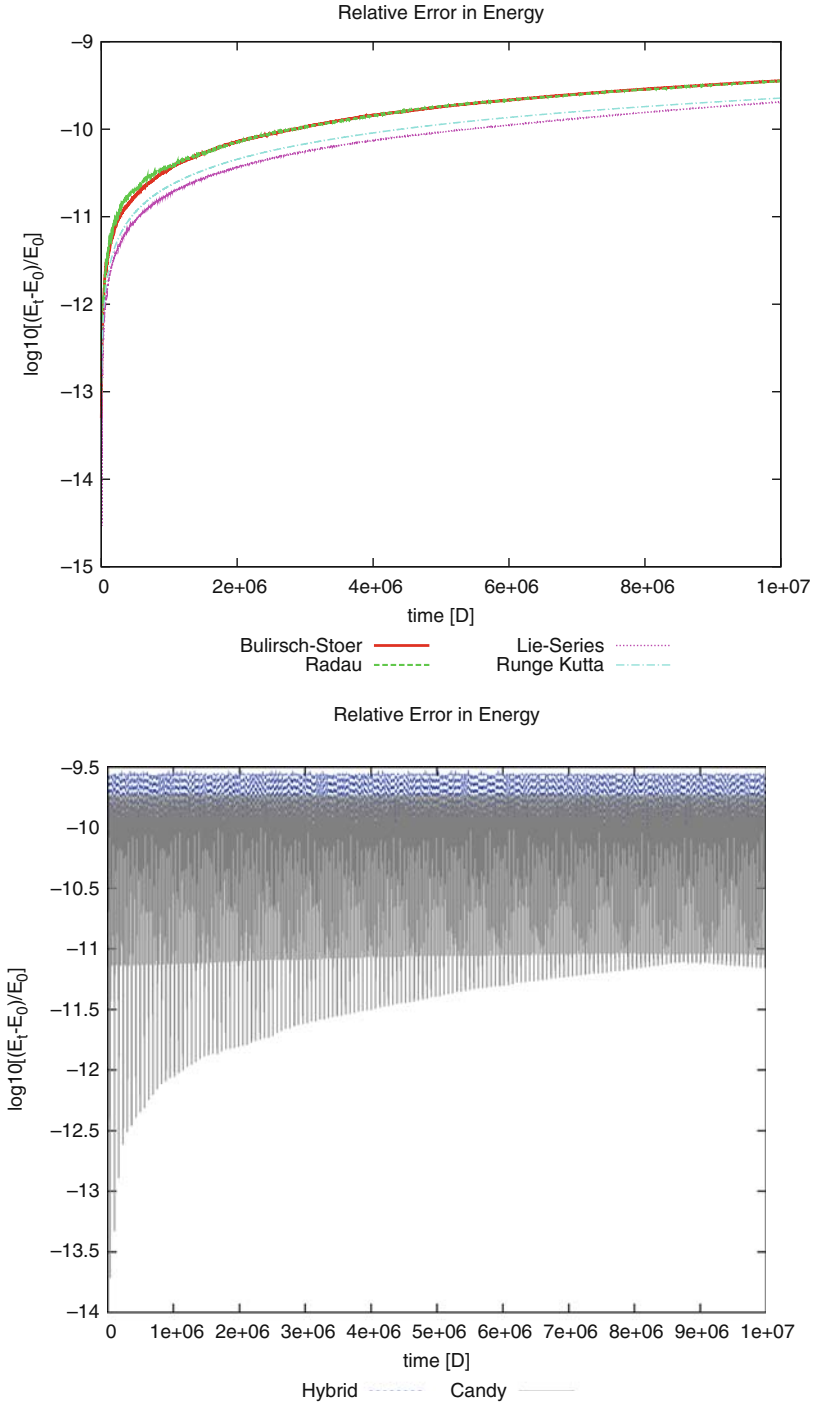

Fig. 4 (caption on next page)

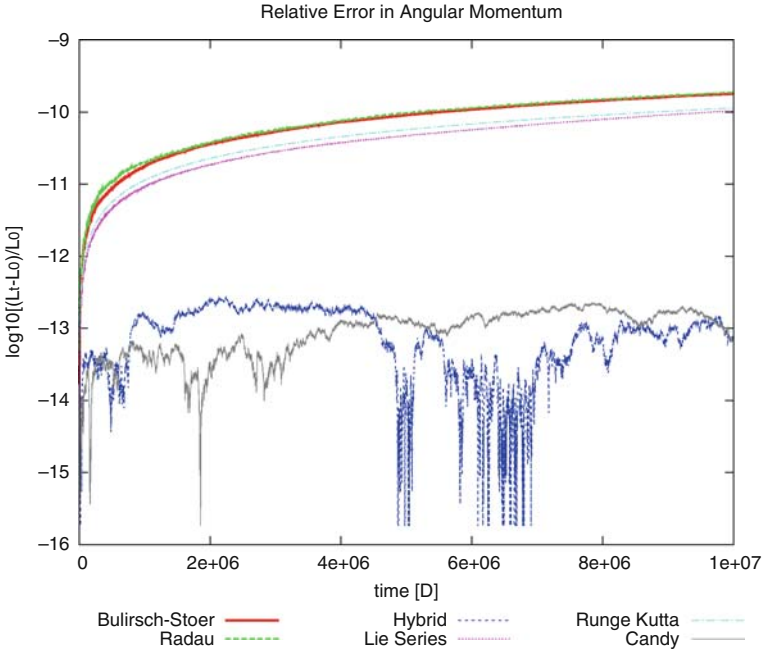

Fig. 4 (continued) Logarithm of the momentary deviation of total energy (E) against time for all non-symplectic algorithms (*above*); logarithm of the momentary deviation of total energy (E) against time of all symplectic algorithms (*middle*); the peculiar form in the deviation of energy is a trademark of symplectic integrators; logarithm of the momentary deviation of angular momentum (L) from its initial value; the tested symplectic algorithms show excellent angular momentum conservation properties (*below*)

The development of the semi-major axis (a) and the numeric eccentricity (e) in Fig. 5 mirrors the energy behaviour already encountered in Fig. 3, which means linear growth in errors for all symplectic and quadratic growth for all non-symplectic methods. It is quite interesting to see, that, even though Candy and Hybrid methods are both symplectic, they show a complementary behaviour.

Watching the development of numeric eccentricity (e), it can be seen that behaviour there is a quadratic error growth with non-symplectic integrators, yet the difference to both symplectic algorithms is still considerable at the end of the integration interval.

None of the methods explored shows any major deviations concerning inclination (i) or argument of the ascending node (Ω), meaning that the direction of the angular momentum, which has not been checked until now, actually is conserved by all integrators.

A weak spot of symplectic algorithms is their so-called phase error denoting an artificially induced circulation of the argument of pericenter (ω). This has already been found, e.g., in Gladman et al. [16] and will be treated more thoroughly in the next section.

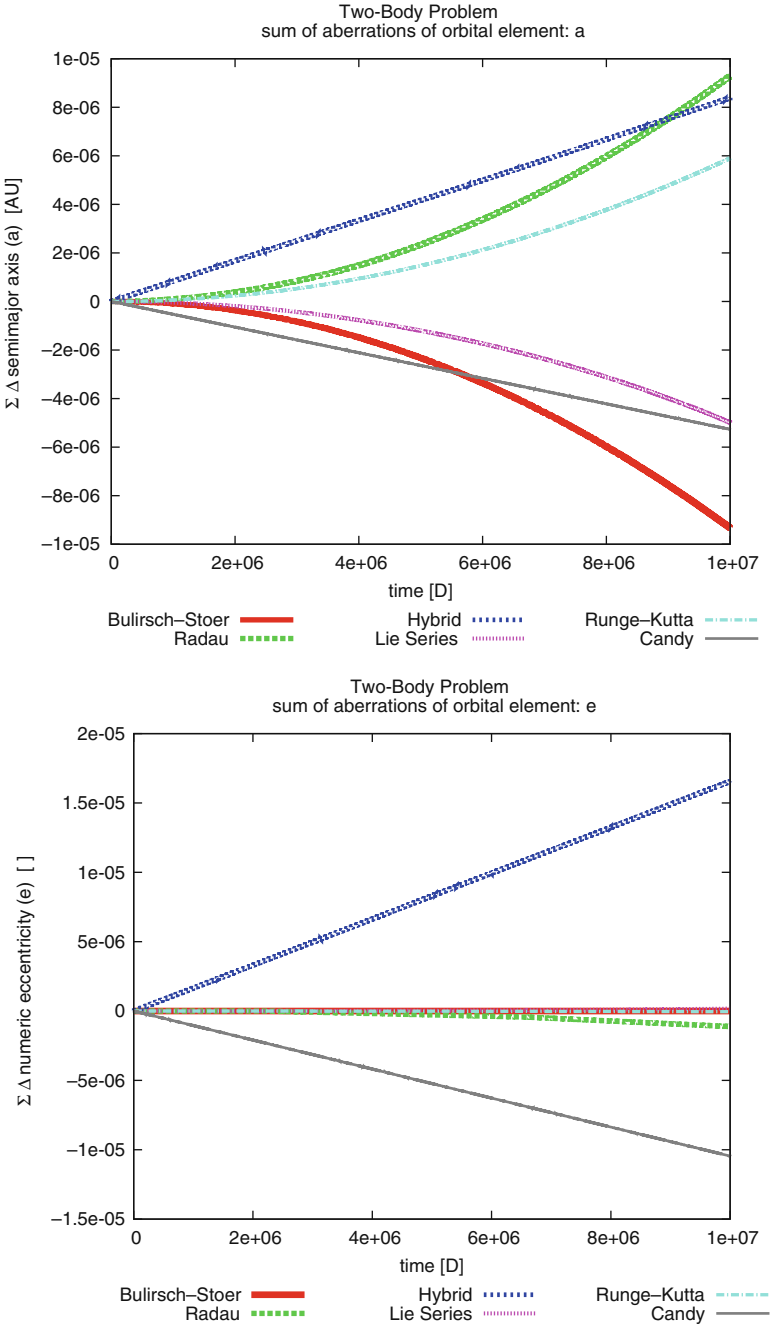

Fig. 5 Sum of deviations of orbital elements a and e from their initial values for all integration algorithms. The linear error growth of symplectic methods contrasts the quadratic one of non-symplectic integrations

The fact that the error in mean anomaly of the symplectic Candy and Hybrid algorithms seems to grow quadratically comes as a bit of a surprise.[21] Yet, finding an explanation for this result is relatively easy. As the momentary deviations of symplectic integrators in mean anomaly are known to grow linearly [16], the sum function of these momentary deviations will be of quadratic form. Non-symplectic algorithms will have a quadratic growth right from the start, resulting in a third-order polynomial, when summed. This can also be seen in Fig. 6, keeping an eye on the steepness of ascent of symplectic and non-symplectic curves.

In summary, the error in mean anomaly in Fig. 6 is bigger for symplectic methods, because the linear growth in error is a slight disadvantage in our setup. The overall integration time is chosen in such a way that the linear error function of symplectic and the quadratic error function of non-symplectic algorithms will intersect just at the end of the time interval. This means that the error of the non-symplectic methods will be smaller up to the moment of intersection, but from that point onward, this effect will turn tide.

8.2 Symplectic Phase Error

A distinct disadvantage of symplectic integration algorithms is their artificially introduced circulation of the argument of pericenter (ω).[22] This can be seen in Fig. 7 when one compares the symplectic to the non-symplectic error curves. Compared to the Candy algorithm, the Hybrid mapping exhibits a much smaller trend, that is nevertheless larger than that of all non-symplectic methods. In order to plainly visualize this effect, we manipulated the Sun–Jupiter problem, enlarging Jupiter's eccentricity by a factor of 10 and taking rather large time-steps regarding the order and fixed step-size of the Candy algorithm. The results in configuration space can be seen in Fig. 8. Such harsh initial conditions may seem a bit far fetched, but as more and more exoplanets are found, having highly eccentric orbits and knowing that such eccentricities are pretty common for comets, one has to be fully aware of these effects when using symplectic integrators. Taking a look at Fig. 7, it is clear that the way of construction of symplectic algorithms influences the extent of such a behaviour.

The reasons for the Hybrid integrator to show a better phase-error behaviour than the Candy algorithm (Fig. 7) are twofold. First of all, the Hybrid is able to reduce its current step-size when the planet approaches perihelion, which will greatly reduce truncation errors during this period. Secondly, the Hybrid integrator is based on the second-order mapping of Wisdom and Holman [35], which uses a set of coordinates

[21] The authors would like to stress that the results for the sum of deviations of the argument of pericenter (ω) and the mean anomaly (M) are *not* to be directly compared to the results in, e.g., Gladman et al. [16], as they have plotted the *momentary* deviation of ω and M which are both linear, of course. In contrast, we calculated the *sum of deviations* which is rather an integral measure, and therefore linear for constant deviations and quadratic for linear deviations.

[22] See, e.g., Kinoshita et al. [20].

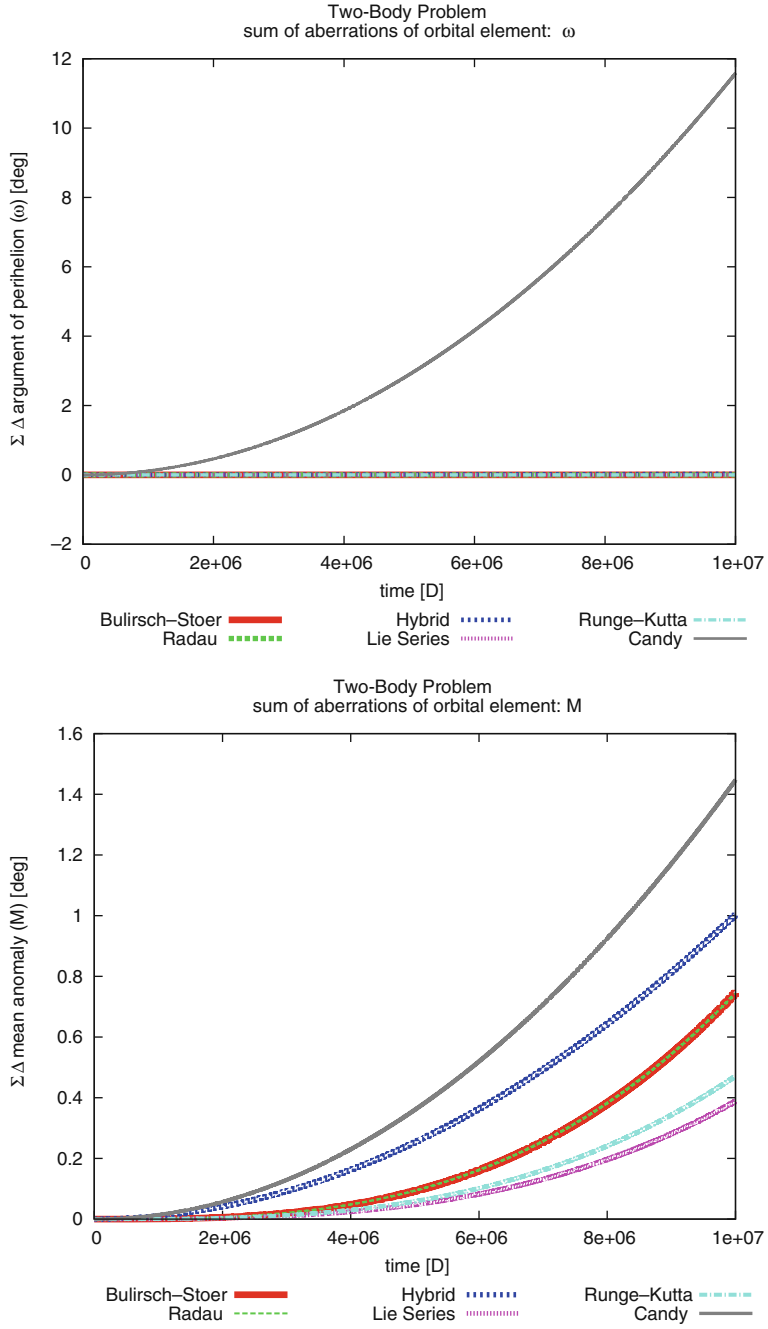

Fig. 6 Sum of deviations of orbital elements ω and M from their initial values for all integration algorithms

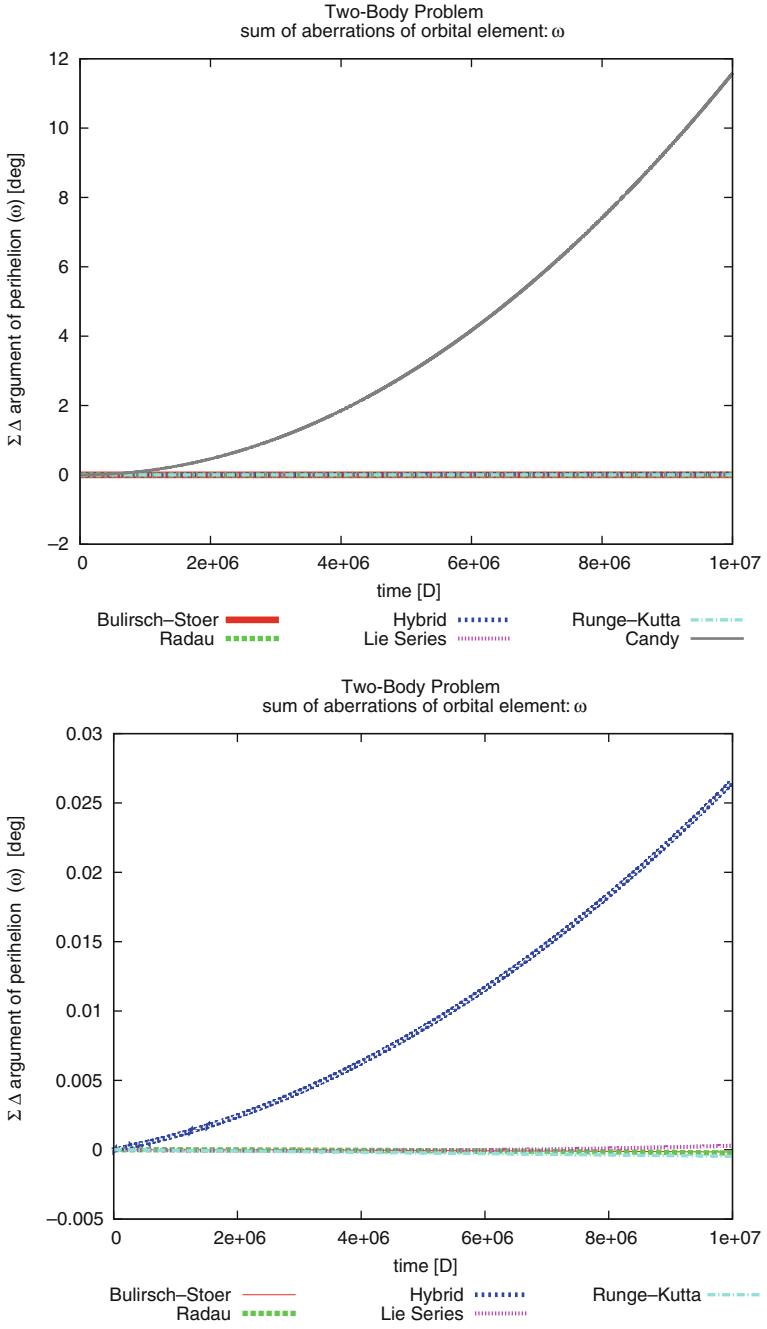

Fig. 7 Error in the argument of pericenter (ω) of all integration algorithms (*above*) and without Candy (*below*); even though the error in ω of the Hybrid-type integrator is lower than that of the Candy algorithm, it is still large compared to its non-symplectic competitors

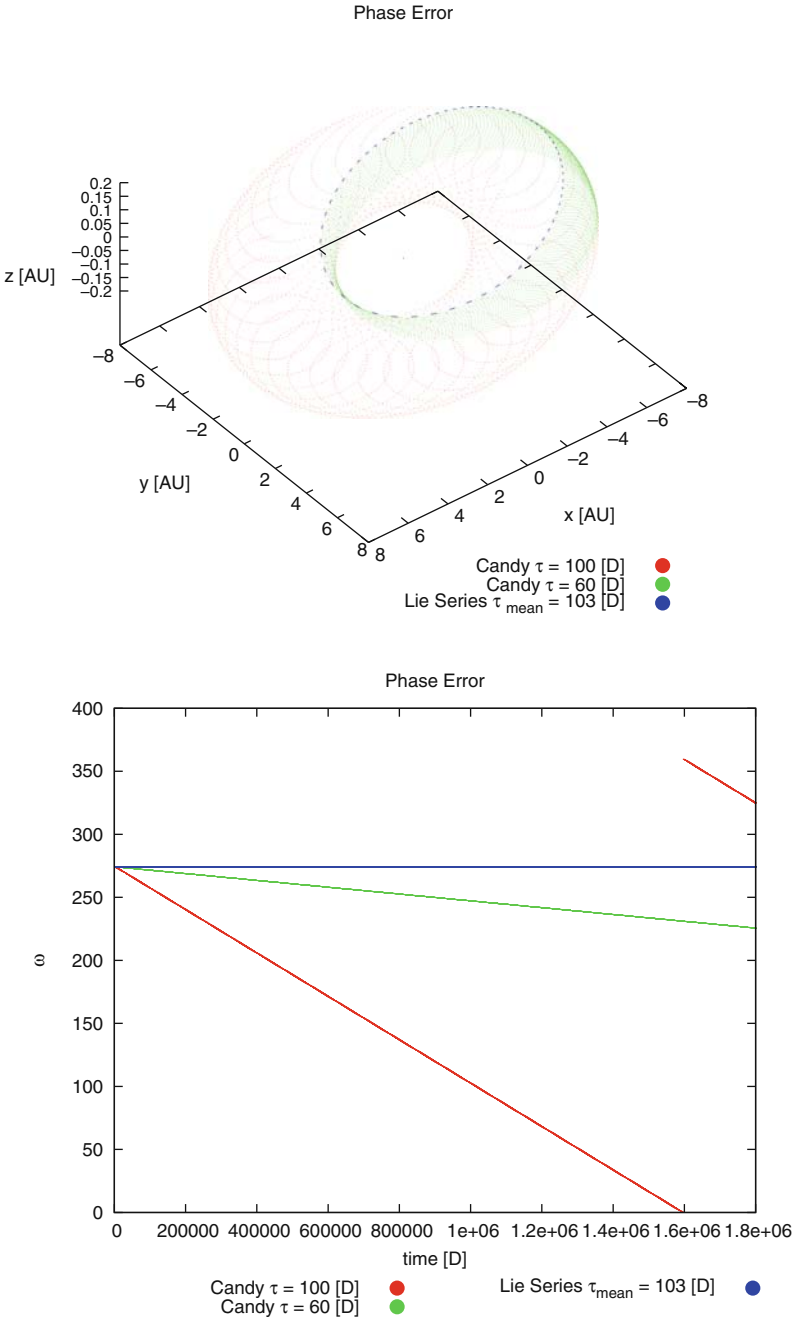

Fig. 8 The phase error of symplectic integrators (here Candy) changes the orientation of the orbital ellipse. This effect is related to the step-size used

that take Keplerian motion into account, and therefore shows a better behaviour in the conservation of orbital elements.

8.3 Performance

Conservational properties are not the only indicators for the quality of integrators. The amount of computational resources consumed during the calculation process is equally important, as any algorithm can be trimmed to produce highly accurate results. Yet, methods will become rather unappealing, when the timescales involved in gaining usable data start to surpass weeks.

Of course, comparing algorithms contained in different package environments is rather tricky and cannot be entirely fair.[23] This is the reason, why the authors chose to split the results of CPU-time measurements to be seen in Figs. 9 and 10 according to package memberships. As the quality of the results was set to be comparable with

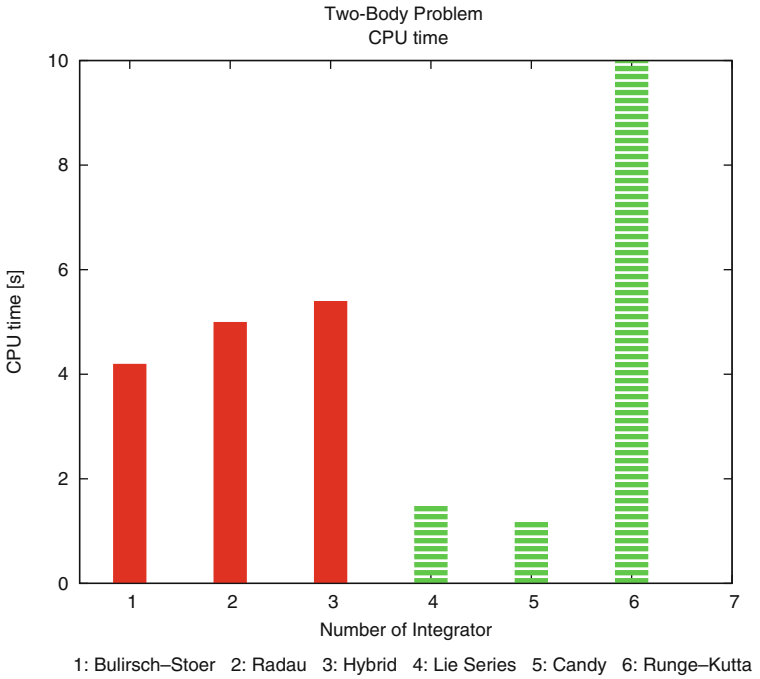

Fig. 9 CPU time needed to calculate the Sun–Jupiter system up to 10^7 [D]. *Full histogram*: CPU-time consumption of algorithms contained in the *mercury6* package; *Dashed histogram*: CPU-time consumption of algorithms contained in the *nie* package

[23] Different values of integration time may be due to internal transformations and output routines that are not necessarily connected to the integration algorithms themselves.

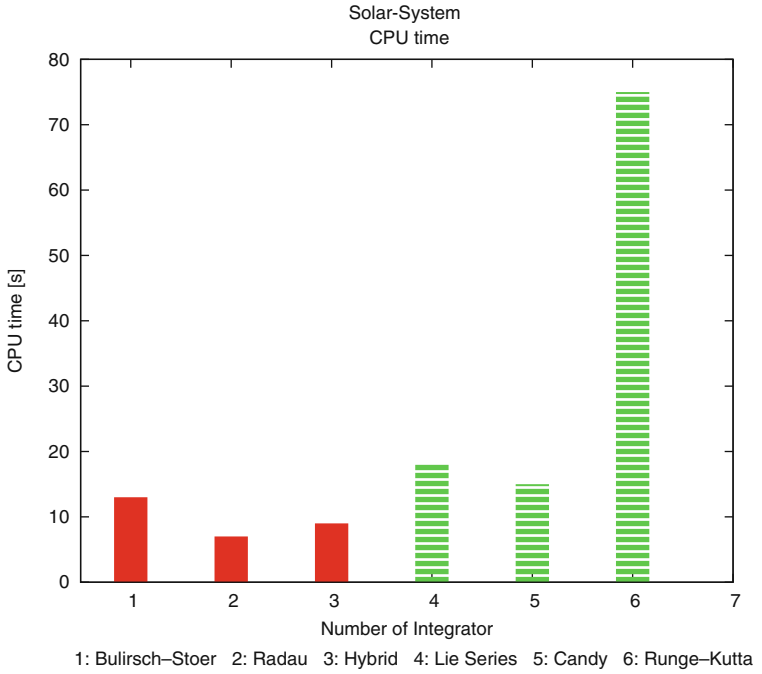

Fig. 10 CPU time needed to calculate the Solar System up to 10^6 [D]. *Full histogram*: CPU time consumption of algorithms contained in the *mercury6* package; *Dashed histogram*: CPU time consumption of algorithms contained in the *nie* package

respect to total energy conservation, one just has to time the integration. The interference of the operating system was tracked and taken into account in the following figures.[24] These measurements have been done for two configurations. Figure 9 shows the results for the Sun–Jupiter system that was integrated up to $10^7[D]$. As one can see quite clearly, the algorithms within the *mercury6* package are comparable for such short integration periods. As the Hybrid algorithm accomplishes a linear error growth in the same time span, it is to be considered the most efficient. The performance of the Lie Series integrator is in the same order of magnitude as the Candy algorithm. As to be expected, Runge–Kutta takes longest, which is due to its high number of right-hand side function evaluations, which are not only taking time by themselves, they also scale with the number of bodies squared. Figure 10 depicts the CPU times for a run including all the planets of the Solar System over $10^6[D]$. Certainly, the algorithms have been set to produce the same quality of output data, in order to make their performance comparable within their respective packages.

[24] Figs. 9 and 10 show the statistical means of a series of repeated calculations timed with the Unix-based "time" command. The number of calculation runs was chosen such that the standard deviation of all measured values was below 1%.

Table 3 A rough summary of the main properties of the six common N-body integration algorithms described in this chapter

Integrator	Symplectic	Variable step-size	Pros	Cons
Cash–Karp Runge–Kutta	No	Yes	Easy implementation adaptability	Performance
Gauss Radau	No	Yes	Accuracy above order, stability	Adaptability
Bulirsch–Stoer	No	Yes	Results for $\tau = 0$	Weak performance when many substeps are required
Lie Series	No	Yes	High performance	Specifically designed for a given problem
Candy	Yes	No	Energy and angular momentum conservation	Fixed step-size, symplectic phase-error
Hybrid	Yes	Yes	Energy and angular momentum conservation	Specifically designed for a given problem, symplectic phase error

The global picture did not change too much. Still the *mercury6* routines are almost on the same performance level, just like the Lie Series and the Candy mapping. Far off resides the Runge Kutta algorithm, which suffers most from the increased number of particles. Another important point, that can be taken out of Figs. 9 and 10 is that inter-package comparison is not easy a task, and should be avoided where possible. Obviously, the *mercury6* integrators were less influenced by the number of bodies involved. This is at least partly caused by its clever output policy (for details see [4]).

9 Conclusions

In this chapter the authors have given a short overview concerning six common integration algorithms used to solve the gravitational N-body problem in dynamical astronomy. A listing of integration methods together with a short overview concerning advantages and disadvantages can be found in Table 3.

The main dividing lines between the algorithms presented are symplecticity on the one hand, and adaptive step-size control on the other. For long-time integrations, where the orientation of single orbits is not as important as the overall energetic behaviour, symplectic algorithms are probably the better choice, due to their favourable energy and angular momentum conservation properties. If one is interested in short-term, high-accuracy calculations, non-symplectic methods may be more effective. As every integrator mentioned in this chapter, except for the Candy algorithm, contains step-size control mechanisms, close encounters during

calculation-runs should - in theory - not pose any major problems, although this is still an ongoing field of research. Binary systems require an updated version of the hybrid algorithm contained in the *mercury6* package [5]. The differences between non-symplectic algorithms are basically restricted to performance issues, and directions of energy-drifts. Concerning performance, the inner-package competitions showed that the only algorithm that is too far off to be recommended is the Cash–Karp Runge–Kutta, simply because the ratio of step-size to righthand side function evaluations is rather low compared to its competitors.

For a more detailed analysis of every method presented, the authors would like to refer the reader to the respective papers:

Cash–Karp Runge–Kutta:	Cash and Karp [3]
Gauss Radau:	Everhart [12]
Bulirsch–Stoer:	Deuflhard [8] and references therein [8]
Lie Series:	Hanslmeien and Dvorak [19]
Candy:	Candy and Rozmus [1]
Hybrid:	Chambers [4]

Acknowledgments S. Eggl would like to acknowledge the support from Austrian FWF Project P-20216. R. Dvorak would like to acknowledge the support from Austrian FWF Project P-18930 - N16

References

1. Candy, J., Rozmus, W.: A symplectic integration algorithm for separable Hamiltonian functions. J. Comp. Phys. **92**, 230–256 (1991) 435, 455, 477
2. Cash, J.R., Gridlestone, S.: Variable Step Runge-Kutta-Nyström methods for the numerical solution of reversible systems. J. Num. Anal. Ind. Appl. Math. 59–80 (2006) 439, 459
3. Cash, J.R., Karp, A.H.: A variable order Runge-Kutta Method for initial value problems with rapidly varying right-hand sides. ACM Transac Math. Softw. **16**(3), 201–222 (1990) 436, 438, 439, 477
4. Chambers, J.E.: A hybrid symplectic integrator that permits close encounters between massive bodies. Mon. Not. R. Astron. Soc. **304**, 793–799 (1999) 435, 459, 460, 462, 463, 464, 476, 477
5. Chambers, J.E.: N-body integrators for planets in binary star systems http://arxiv.org/abs/0705.3223v1, Cornell University Library, Cornell (2007) 477
6. Danby, J.M.A.: Fundamental of Celestial Mechanics Atlantic Books. Willmann-Bell, Richmond (1988) 459, 460, 462
7. Delva, M.: Integration of the elliptic restricted three-body problem with Lie series. Celestial Mech. **34**, 145–154 (1984) 449
8. Deuflhard, P.: Order and stepsize control in extrapolation methods. Num. Math. **41**, 399–422 (1983) 443, 447, 477
9. Deuflhard, P., Bornemann, F.: Scientific Computing with Ordinary Differential Equations. Springer, New York (2002)
10. Duncan, M.J., Levison, H.F., Lee, M.H.: A multiple time step symplectic algorithm for integrating close encounters. Astron. J. **116**, 2067–2077 (1998) 461
11. Duncan, M., Levison, H.F.: Symplectically integrating close encounters with the Sun. Astron. J. **120**, 2117–2123 (2000) 464

12. Everhart, E.: Implicit single-sequence methods for integrating orbits Celestial Mech. **10**, 35–55 (1974) 440, 443, 477
13. Forster, O.: Analysis I, p.73 et sqq. Vieweg Verlag, Wiesbaden (2006) 454
14. Flaherty, J.E.: Course notes – ODE4 http://www.cs.rpi.edu/∼flaherje/ (2007) 444, 446
15. Fukushima, T.: Reduction of round-off errors in the extrapolation methods and its application to the integration of orbital motion. Astron. J. **112**, 1298 (1996) 448
16. Gladman, B., Duncan, M, Candy, J.: Symplectic integrators for long-term integrations in celestial mechanics. Celestial Mech. Dynam. Astron. **52**, 229 (1991) 460, 468, 470
17. Gröbner, W.: Die Lie-Reihen und ihre Anwendungen. Deutscher Verlag, der Wissenschaften, 1967 – VI 449
18. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration illustrated by the Størmer Verlet method. Cambridge University Press, Acta Numerica 1–51 (2003) 457
19. Hanslmeier, A., Dvorak, R.: Numerical integration with Lie-series. Astron. Astrophys. **132**, 203–207 (1984) 449, 450, 451, 453, 477
20. Kinoshita, H., Yoshida, H., Nakai, H.: Symplectic integrators and their application to dynamical astronomy. Celestial Mech. Dynam. Astron. **50**, 59–71 (1990) 458, 470
21. Lasagni, F.M.: Canonical Runge-Kutta methods. ZAMP **39**, 952–953 (1988) 439
22. Lee, M.H., Duncan, M.J., Levison, H.F.: Variable time step integrators for long-term orbital integrations computational astrophysics In: Clarke, D.A. and West, M.J. (eds.) 12th Kingston Meeting on Theoretical Astrophysics; Proceedings of meeting held in Halifax; Nova Scotia, Canada October 17–19, 1996, ASP Conference Series #123, p. 32 (1997) 459
23. Lichtenegger, H.: The dynamics of bodies with variable masses, Celestial Mech. **34**: 357–368 (1984) 449
24. Mikkola, S., Aarseth, S. J.: An implementation of n-body chain regularization. Celestial Mech. Dynam. Astron. **57**, 439 et sqq. (1993) 436
25. NASA – JPL http://ssd.jpl.nasa.gov/txt/p_elem_t1.txt(2008) 464
26. Neri, F.: Lie Algebras and Canonical Integration. Department of Physics, University of Maryland, Maryland, preprint (1987) 456
27. Okunbor, D.I., Skeel, R.D.: Canonical Runge-Kutta-Nyström methods of orders five and six. J. Comput. Appl. Math. **51**:375–382 (1994) 439, 447
28. Press, W.H., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in Fortran 77, p.718 et sqq. Cambridge University Press, Cambridge (1992) 445
29. Saha, P., Tremaine, S.: Symplectic integrators for solar system dynamics. Astron. J. **104**, 1633–1640 (1992) 459
30. Sanz-Serna, J.M.: Runge-Kutta schemes for Hamiltonian systems. BIT **28**, 877–883 (1988). 439
31. Suris, Y.B.: On the Conservation of the Symplectic Structure in the Numerical Solution of Hamiltonian Systems, pp. 148–160. USSR Academy of Sciences, Moscow (1988) 439
32. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. Springer, New York (1980) 443, 460
33. Stoer, J., Bulirsch, R.: Numerische Mathematik 2 Springer, 4. Auflage (2000) 438, 480
34. Vesely, F.: Computational Physics – An Introduction, p. 105 et sqq. Springer, New York (2001) 457, 480
35. Wisdom, J., Holman, M.: Symplectic maps for the n-body problem. Astron. J. **102**. 1528–1538 (1992) 435, 458, 460, 461, 462, 463, 470
36. Yoshida, H.: Conserved quantities of symplectic integrators 23. Symposium on Celestial Mechanics, pp. 16–19 (1990) 458
37. Yoshida, H.: Recent progress in the theory and application of symplectic integrators. Celestial Mech. Dynam. Astron. **56**: 27–43 (1993) 459

Appendix

This chapter contains a short introduction to the most familiar terms concerning algorithms used to solve ordinary, first-order differential equations.

A simple example, how a transition from analytics to discrete mathematics may be used to solve an ordinary first-order differential equation can be constructed from the definition of a functional derivative of a function f with argument t:

$$\lim_{\tau \to 0} \frac{f(t+\tau) - f(t)}{\tau} = \dot{f}(t). \tag{103}$$

The discrete equivalent of this equation simply allows τ to be a non-vanishing number called *step-size*, or, in this case *time-step*.

$$\begin{aligned} \frac{f(t+\tau) - f(t)}{\tau} &= \dot{f}(t), \\ f(t+\tau) &= f(t) + \tau \dot{f}(t). \end{aligned} \tag{104}$$

Equation (104) states, that, given $f(t)$ and $\dot{f}(t)$ as initial values at time t, one can calculate the value of $f(t+\tau)$, which is the approximate result of the true integral $F|_t^{t+\tau} = \int_t^{t+\tau} \dot{f}(s)ds$ in the interval $[t, t+\tau]$. f is the *approximate* result, because we have taken τ as being a real number instead of tending towards zero, which will therefore produce some discretisation error. Consequently, when we down-size τ, our estimates of the true integral will improve, yet, more steps will have to be performed in order to cover a given integration interval. As equation (104) is linear in τ, so will be the approximation to F, and we will call this algorithm to be of *first order*.[25] Integration algorithms of first order lead to rather time consuming computations in practice, as many computational steps are required to gain viable solutions. Due to the finite number representation within today's computational architectures, each calculation step will result in *round-off errors*. Therefore, the more steps an algorithm requires, the larger will be the result's round-off error. We will have another look at equation (104), which reveals a resemblance to a Taylor series expansion of $f(t+\tau)$ up to the first order in τ. In order to decrease the *truncation error* of this expansion ansatz, a logical, next step would be to construct algorithms of higher order by simply enhancing the corresponding Taylor polynomial. This works, but it is rather inefficient, as for each added order, a higher functional derivative has to be calculated at each step. As a consequence, many different ideas have evolved to gain access to high orders with a minimum of additional computations, constituting the broad spectrum of solvers for differential equations.

Another aspect of (104) is the fact that the results for the next step are gained exclusively through values of the previous step. This is the reason for calling such algorithms *one-step methods*. If values of more previous steps are used, these algorithms are named *multi-step methods*.

$$f(t+\tau) = g\left(\tau, f(t), \dot{f}(t), f(t-\tau), \dot{f}(t-\tau), ...\right) \tag{105}$$

[25] The algorithm described in equation (104) is usually referred to as explicit Euler method.

Multi-step methods tend to be very efficient compared to one-step methods, as one can save and reuse the previous results leading to very few extra function evaluations (e.g., Predictor Corrector algorithms). Plus, especially if they are symmetric with respect to time inversion, they are very well behaved as far as conservational properties are concerned. Yet the computational overhead of an implemented variable step-choosing technique and certain problems with resonance induced errors often even the odds toward one-step algorithms [33].

There is still one major criterion, that separates different numerical methods for solving ordinary differential equations. Once again referring to (104), it occurs that the results at time $t + \tau$ are computed from initial values, that are *already known*. Let us call this type of method *explicit*. Though, the relation leading to results of the next time-step could actually be of the form:

$$f(t+\tau) = h\left(\tau, f(t+\tau), \dot{f}(t+\tau), f(t), \dot{f}(t), ...\right) \tag{106}$$

Such formulations are called *implicit*, as the function values $f(t + \tau)$, that are required to calculate the results for the next step, are not known when first needed. Mostly, implicit methods rely on additional interpolation steps to solve differential equations, which is granting them a large stability region, meaning that in theory there is no upper limit to the time-step τ, except for limits imposed by the truncation error. These interpolation steps are, on the other hand, responsible for the rather weak performance of implicit algorithms.

In order to be certain, that a chosen algorithm is suited for a given problem, it is also necessary, to check for *stability* and *convergence*. An algorithm is referred to as being *convergent*, if the solution of the discrete equations will tend towards the true integral for vanishing step-sizes.

$$\lim_{\tau \to 0} f = F \tag{107}$$

Checking for *stability* can be considered as the search for the largest step-size τ that will not permit errors to grow with time. Usually, stability is huge an issue with explicit, but not with implicit algorithms.

There is quite a lot of literature on the topic of numerical integration of ordinary differential equations, so the following references are simply due to the authors' preferences. For a detailed development and stability analysis of numerical methods in general, see, e.g., Stoer and Bulirsch [33], Deuflhard and Bornemann [9] or Vesely [34].

Dynamical Stability of Extra-Solar Planets

E. Pilat-Lohinger and B. Funk

Abstract In this chapter we discuss the orbital stability of extra-solar planetary systems. After a short general introduction into the very popular topic of extra-solar planets, we classify the more than 400 planets that have been detected so far according to dynamical aspects. We discuss planetary motion in (i) single-star single-planet, (ii) multi-planet systems, and (iii) binary systems. For the first group we show the application of a general stability study which helps to verify the dynamical behavior of an additional planet that may be discovered in the future. For the other two groups—that are more complicated from the dynamical point of view—we selected some interesting systems—like Jupiter–Saturn analogs and the close binary systems HD41004 AB, Gliese 86, and γ Cephei—for which we discuss the dynamical stability . Finally, we provide an insight into dynamical contributions to the interdisciplinary research of habitability.

1 Introduction

Planets outside the solar system are called *extra-solar planets* and have been detected numerously since 1989 [41]. By now (November 2009) we have knowledge of more than 400 extra-solar planets (see, e.g., the Web site of J. Schneider: http://www.exoplanet. eu) that were primarily discovered by radial velocity measurements. A very sensitive Doppler technique reveals perturbations of a star due to an accompanying planet. Since the star is forced to move around the center of mass, the spectrum is shifted periodically to the blue (the star is moving towards the Earth)

E. Pilat-Lohinger (✉)
Institute for Astronomy, University of Vienna, Türkenschanzstrasse 17, A-1180 Vienna, Austria,
elke.pilat-lohinger@univie.ac.at

B. Funk (✉)
Department of Astronomy, Eötvös Loránd Univesity, Pázmány Pèter Sétány 1/A, 1117 Budapest
funk@astro.univie.ac.at

Pilat-Lohinger, E., Funk, B.: *Dynamical Stability of Extra-Solar Planets*. Lect. Notes Phys. **790**, 481–510 (2010)
DOI 10.1007/978-3-642-04458-8_10

and to the red (the star is moving away from the Earth). From the radial velocity measurements we are able to determine the planet's mass—or more precisely the minimum mass[1]—its orbital period and its distance to the star with the aid of the Newtonian law.

The discoveries were quite surprising. In 1992 the first planet orbiting a pulsar was discovered (see [71]). For such a detection a precision of a few μs is needed, hence Earth-like planets can be discovered easily by pulsar timing. Three years later, the discovery of a planet orbiting the Sun-like star 51 Peg by [46] was certainly the breakthrough for the extra-solar planetary science. The analysis of the orbital parameters of 51 Peg b showed an astonishing result, since the Jupiter-like planet is orbiting its host star in a very close orbit (much closer than Mercury, the innermost planet of our solar system). Nowadays, many such close-in planets (also called hot-Jupiters) are known, which is probably a bias due to the observation technique. We are also faced with high-eccentric motions of the planets, so that the systems are quite different compared to our planetary system.

Another, very promising detection method is the *transit*, where we observe the planet passing in front of the star, which produces a drop in the star light. With this method we are able to determine the exact mass of a planet and it allows the discovery of (a) Jupiter-size planets with ground-based observations and (b) Earth-size planets from space. In fact, it is presently the only method to detect an exo-Earth. Moreover, this method is used by the first true exo-planetary mission CoRoT (Convection, Rotation, and planetary Transits; see [11]) and will gain more and more importance in the forthcoming years. Other observational methods are

- *Astrometrie measurements*: where a wobble in the star's motion can be observed that is caused by an accompanying planet. Nowadays the best astrometric measurements can reach a precision of about 100 μarcs which allows the detection of Jupiter-size planets.
- *Gravitational lensing*: where an Exo-planet can produce a gravitational amplification of the light of background stars for a certain time depending on its transverse velocity;
- *Direct imaging*: a direct observation of planets outside the solar system near a Sun-like star is up to now impossible,[2] even the Hubble Space Telescope would not be able to detect them at the expected distances from their stars, since Sun-like stars are about 1 billion times brighter than planets in the visible light. Even if four objects have been imaged so far, we have to point out that either they are very massive (most probably brown dwarfs), or they move around very low-massive M-dwarfs so that their imaging was easier. From the observations (until November 2009) the following systems are known:

[1] We cannot determine the exact mass with this method, since the inclination of the system with respect to the line of sight is not known.

[2] The announcement of GQ Lupi—the possibly first direct discovered extra-solar planet by Neuhäuser and coworkers in 2004 with the ESO VLT NACO—has still to be proven and is not accepted by all researchers yet.

321 *planets* have been detected by the *radial velocity method*
9 *planets* by *microlensing*
11 *planets* by *imaging*
8 *planets* by *Pulsar timing*.

From the dynamical point of view it is useful to distinguish between

(1) Single-star single-planet systems
(2) Single-star multi-planet systems
(3) Planets in double star systems.

The first group contains most of the detected systems and is with respect to the dynamics the simplest one. For such systems we are able to predict the stability of an additional planet via global stability studies (like the one by [68]). The latter two groups are more complicated from the dynamical point of view, since the stable planetary motion is restricted to certain regions of the phase space of a system due to the gravitational interactions between the celestial bodies. Up to now we have knowledge of about 30 multi-planet systems and about 31 binary systems that host one or more planets.

The necessity to verify the dynamical stability of multi-planet systems was demonstrated by Ferraz–Mello, who showed in a numerical simulation of the system HD82943 that the two planets might end in a catastrophe after about 50,000 years, using the orbital parameters given by the observations. Therefore, it is quite evident that the determination of the orbital parameters is quite a tricky task. Especially, when only few observations for a system are available, the errors of the data set are relatively high, particularly in the eccentricity.

In this chapter we discuss the stability of planetary motion for the three dynamical groups mentioned above. We present some selected general stability studies as well as the application to some real extra-solar planetary systems. And we will end with a brief discussion about the stability of terrestrial planets moving in the so-called habitable zone (HZ).

2 Single-Star Single-Planet Systems

The majority of the detected extra-solar planetary systems (EPSs) belong to this group, that consists of a star and a giant planet. However, we should rule out that other small planets—maybe Earth-like planets—may exist in these systems. Therefore, it is interesting to study the dynamical stability of such systems in order to determine the regions where other planets might exist. This can be done (i) by exploring the stable and unstable regions of the phase space of each EPS separately or (ii) by calculating general stability maps for a large set of orbital parameters as it was done by [68]. This second method has the advantage that the stability properties of a low-mass planet in an EPS can be easily established, when the orbital parameters of the giant planet of the system are modified due to new observational runs. In

that case, normally one has to re-explore the phase space of the individual EPS after each modification of the orbital parameters of the giant planet. This is not necessary in the case of the second method, since the stability properties of the investigated EPS can be re-established easily from the existing stability maps.

The so-called "Exocatalogue" consists of 92 stability maps covering 23 mass-ratios ($\mu = m_2/(m_1 + m_2)$) of the star and the giant planet (from 0.0001 to 0.05). The semi-major axis of the giant planet (a_{GP}) was set to unity and its eccentricity (e_{GP}) was varied from 0 to 0.5. Moreover, two starting positions were used for the giant planets defined by the variation of its mean anomaly M_{GP}. Using the elliptic restricted three-body problem (ERTBP)[3] as dynamical model we studied (i) the *inner region* (i.e., the region between the host-star and the giant planet), where the test-planets have starting positions between 0.1 and 0.9, and (ii) the *outer region* (i.e., outside the giant planet) from 1.1 to 4. To distinguish between regular and chaotic motion, three methods were used: (1) the relative Lyapunov indicator (RLI) [66, 67], (2) the fast Lyapunov indicator (FLI) [25, 28], and (3) the maximum eccentricity (max-e) (see, e.g., [19]).

The RLI measures the convergence of the finite time Lyapunov indicators to the maximal Lyapunov characteristic exponent of two very close orbits. This method is extremely fast to determine the dynamical behavior of individual orbits ([67], e.g., sticky orbits). When applying the FLI, one measures the growth of the largest tangent vector, which increases either linearly (in the regular case) or exponentially (in the chaotic case). For the calculation of the max-e one needs long-term orbital computations of the orbits (10^5 orbital periods in this chapter) and determines the maximum eccentricity of the whole time interval. All methods have been applied successfully to several EPSs to determine their dynamical stability properties: for the RLI see [21, 67], for the FLI see [55, 19, 10], and for the max-e see [19, 21].

A comparison of the results of the different methods, showed them in good agreement. Since more than 80,000 orbits were calculated for each stability map, it is obvious that the computation time was a crucial parameter to select the method for the huge amount of computations. According to different numerical experiments, the RLI needs only a few 100 periods of the giant planet to determine the motion, while the FLI showed good results after several tens of thousands orbital periods. So that the RLI seemed to be the appropriate tool for the whole computations of the stability map catalogue.

In Fig. 1 we show two example maps of the catalogue, where Fig. 1a (upper panel) displays the stability of the inner region for $\mu = 0.001$ and Fig. 1b (lower panel) shows the outer region for this mass ratio (which corresponds to the mass ratio of the Sun and Jupiter). In both panels the stable motion is given by the colors yellow to dark orange, which are connected to the lowest values of the RLI. Higher RLI values (see the color scale of Fig. 1a, b) label chaotic motion in this system, where dark blue and black indicate strong chaotic orbits.

[3] The elliptic restricted three body problem studies the motion of a massless body moving in the gravitational field of two massive bodies, which move in Keplerian orbits around their center of mass.

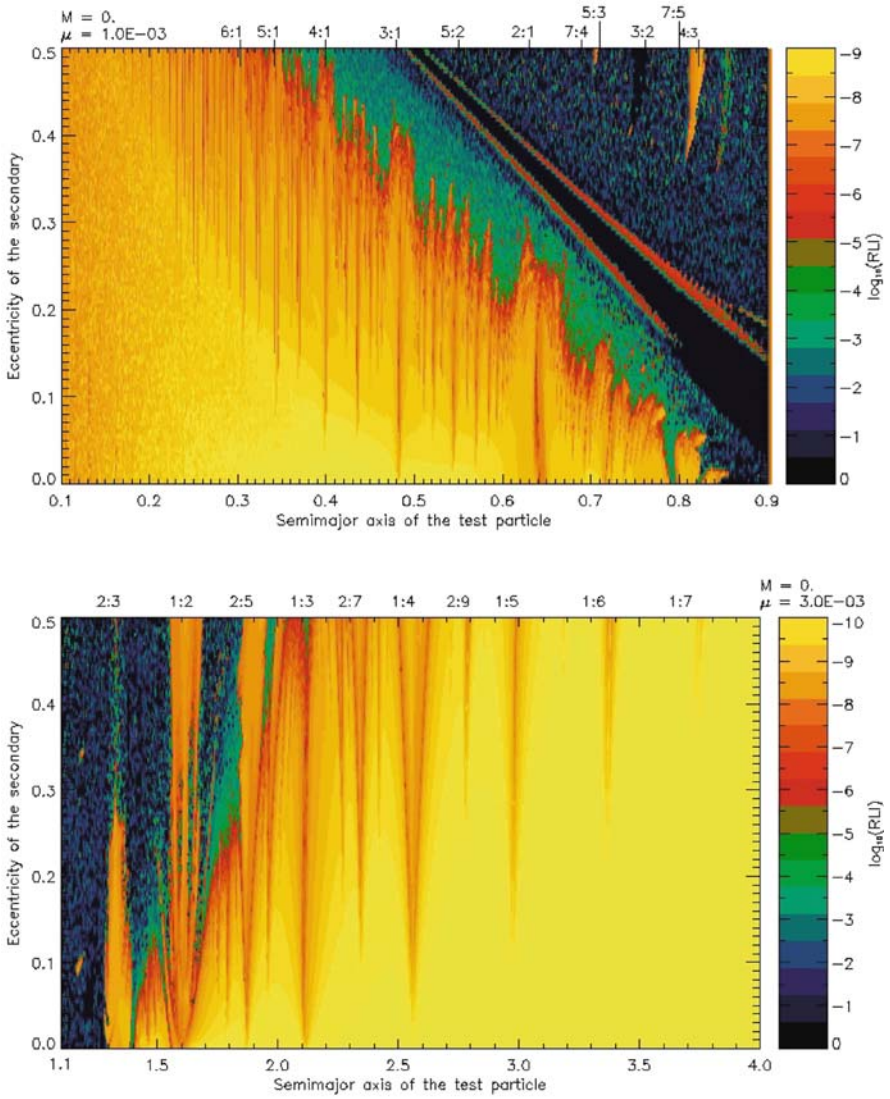

Fig. 1 RLI stability maps for $\mu = 0.001$ and $M_{GP} = 0^{\circ}$. *Yellow* to *orange* (representing small values of the RLIs) show the stable regions, *red* to *blue* indicate chaotic motion and *dark blue* and *black mark* the strong chaotic regions, where collisions and escapes might occur

It is clearly seen, that an increase of the planet's eccentricity (e_{GP}, i.e., the y-axis) decreases the stable region in this system. Perturbations indicated by vertical lines mark the mean motion resonances (MMRs) with respect to the giant planet. These MMRs were found very numerous in the RLI maps. A comparison with the corresponding FLI map shows the same overall structure but a complementary

character of the results was obtained by the two methods: The RLI detects the chaotic separatrices of the resonances (dark V-shaped stripes in Fig. 1a, b), while the FLI finds the stable resonant orbits inside the resonances. For details of this work we recommend the reader to the paper by Sándor et al. [68] or to the Web Site: `http://astro.elte.hu/exocatalogue.html` that provides all necessary information about this tool.

To get information about the stability of a new discovered planet, one needs to know (i) the mass ratio between the star and the giant planet and (ii) whether the new planet orbits the star inside or outside the giant planet, so it is possible to select the appropriate stability map. Then it is necessary to convert the system units into the units of the catalogue. With the new semi-major axis (i.e., x-axis of the stability map) and the eccentricity of the existing giant planet (e_{GP}, i.e., y-axis of the stability map) one is able to locate the position in the stability map, where the color defines the dynamical behavior of the orbital parameters of the new discovered planet. In this context we advice the reader of "ExoStab" (see http://www.univie.ac.at/adg/exostab)—an Internet tool that does the necessary conversion and displays the appropriate stability map of the Exocatalogue.

However, we have to point out that stability maps are mostly valid for small planets (e.g., terrestrial planets) and low-eccentric motion ($e < 0.2$) since the computations were performed in the ERTBP. If the new planet is very massive ($> M_{\text{Jupiter}}$) and moves in an eccentric orbit, it is advisable to proof the stability of using the three-body problem.

3 Multi-Planet Systems

Currently (November 2009) we know 40 multi-planet systems – most of them consist of only 2 planets, only 9 systems have 3 planets, 2 systems with 4 and 1 with 5 planets. From the dynamical point of view, we can distinguish four classes of multi-planet systems (according to [23]) that are based on the mutual distance between the planets and the orbital eccentricities:

Class Ia—Planets in mean motion resonance (MMR): Planet pairs with large masses moving in eccentric orbits that are relatively close to each other, so that strong gravitational interactions might occur. Such systems remain stable if the two planets are in mean motion resonance, i.e., if the ratio of the orbital periods of two planets is quite close to a ratio of two integers. A MMR can be written as $(p + q)/q$, where p and q are integers and the latter represents the order of the resonance. The critical angles of a MMR are defined as $\theta_i = (p+q)\lambda_2 - q\lambda_1 - q\varpi_i$, where $\lambda_i, i = 1, 2$ are the mean longitudes of the planets, and ϖ_i, $i = 1{,}2$ are the longitudes of perihelion. The behavior of these angles show whether a system is in resonance or not. If one of these angles oscillates then the system is inside the resonance. There are many examples of planet pairs in MMRs like, e.g., GJ 876, 55 Cnc, HD829422, HD202206, HD160691, ν And, and GJ873. Since these systems

Table 1 Discovered multi-planet systems (part I)

Star	M_{pl} [M_{Jup}]	a_{pl} [AU]	e_{pl}
GJ 876	1.935	0.20783	0.0249
	0.56	0.13	0.27
	0.018	0.020807	0.0
GJ 581	0.0492	0.041	0.02
	0.0158	0.073	0.16
	0.0243	0.25	0.2
HD 69830	0.0322	0.0789	0.1
	0.0374	0.187	0.13
	0.0573	0.633	0.07
55 Cnc	0.824	0.115	0.014
	0.169	0.24	0.086
	3.835	5.77	0.025
	0.034	0.038	0.07
	0.144	0.781	0.2
HD 82943	1.81	0.752	0.39
	1.74	1.19	0.02
47 Uma	2.6	2.11	0.049
	1.34	7.73	0.005
HD 128311	2.19	1.1	0.25
	3.22	1.76	0.17
HD 160691	1.675	1.501	0.132
	1.814	5.171	0.097
	0.033	0.091	0.172
	0.522	0.919	0.049
HD 190360	1.502	3.92	0.36
	0.057	0.128	0.01
υ And	3.95	2.51	0.242
	1.98	0.83	0.254
	0.69	0.059	0.029
HD 11964	0.09	0.2527	0.23
	0.213	1.132	0.63
	0.77	3.46	0.05
HD 12661	2.3	0.83	0.35
	1.57	2.56	0.2
HIP 14810	3.91	0.0692	0.147
	0.76	0.407	0.4091
OGLE-06-109L	0.71	2.3	-
	0.27	4.6	0.11

are of special interest in dynamical studies, the most interesting ones have been examined by many research groups, see, e.g., [43, 44, 29, 1, 2, 10].

Class 1b—Low-eccentricity near-resonant planet pairs: In this case a mean motion resonance of the planet pair is not needed to guarantee the long-term stability of the system. Therefore, the eccentricities of the planets have to be small to exclude a crossing of the orbits. Our solar system belongs to this group and the recently discovered OGLE-06-109L system [26]. Even if there is no other multi-planet system with similar characteristics known at the moment, we have to take in mind

Table 2 Discovered multi-planet systems (part II)

Star	M_{pl} [M_{Jup}]	a_{pl} [AU]	e_{pl}
HD 37124	0.624	0.519	0.079
	0.574	1.61	0.15
	0.695	3.142	0.297
HD 38529	0.852	0.131	0.248
	13.2	3.74	0.3506
HD 73526	2.9	0.66	0.19
	2.5	1.05	0.14
HD 74156	1.88	0.29	0.64
	8.03	3.86	0.43
	0.4	1.04	0.25
HD 108874	1.36	1.051	0.07
	1.018	2.68	0.25
HD 155358	0.89	0.628	0.112
	0.504	1.224	0.176
HD 168443	7.48	0.29	0.53
	16.87	2.84	0.222
HD 169830	2.88	0.81	0.31
	4.04	3.6	0.33
HD 187123	0.522	0.0426	0.0099
	1.95	4.8	0.249
HD 202206	17.4	0.83	0.435
	2.44	2.55	0.267
HD 217107	1.33	0.073	0.132
	2.5	4.41	0.537
HD 102272	5.9	0.614	0.05
	2.6	1.57	0.68
GJ 436	0.072	0.0287	0.15
	0.015	0.045	0.2
14 Her	4.975	2.864	0.359
	7.679	9.037	0.184
HD 82943	0.88	0.73	0.54
	1.63	1.16	0.41
PSR 1257+12	0.00007	0.19	0.0.0
	0.013	0.36	0.0186
	0.012	0.46	0.0252

that only six far away planetary systems have been detected so far. Out of these systems, one has similarities with our solar system, therefore, it can be assumed that many such systems may exist but have not been discovered yet.

In a numerical investigation by Pilat-Lohinger et al. [60] the stability of different fictitious Jupiter–Saturn configurations were studied. To obtain similar systems, the mass of Saturn (m_S) was increased by factors of 2—40 and its initial semi-major axis (a_S) was varied from 8 to 11 AU. To see the mutual perturbations of Jupiter and Saturn, we show in Fig. 2a summary of all fictitious Jupiter–Saturn configurations. The dynamical state for each (a_S, m_S) was determined by a method base d on the frequency analysis described in [65]: First Saturn's "proper mean motion"

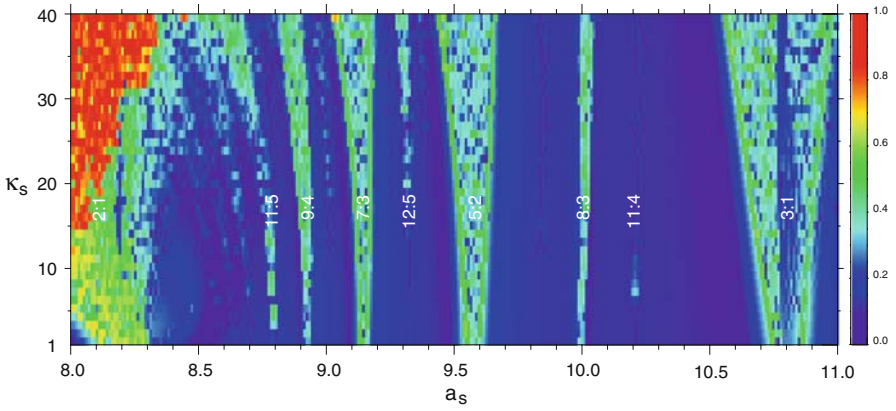

Fig. 2 MMRs between Jupiter and Saturn in the region between 8 and 11 AU (x-axis) for various masses of Saturn (y-axis). The colors show whether a trajectory can be considered as quasiperiodic one (i.e., *dark blue*) or as perturbed orbit (*blue* to *yellow*) for which the associated diffusion remains (for almost all initial condition) bounded or if they are in areas indicating that the planetary system is chaotic and its destruction is possible (*orange* and *red*) (taken from [60])

was determined by a quasiperiodic approximation of $a_S \exp(i\lambda_S)$—where λ_S is the mean longitude of Saturn. Then, for each Saturn-mass the second derivative of the proper mean motion with respect to Saturn's semi major axis was computed and is related to some diffusion rate in the frequency space [40]. Therefore, it defines the dynamical state of motion of the two giant planets, which is described by the index of quasiperiodicity in the interval $[0, 1]$, where 0 means that the trajectory is quasiperiodic, and 1 that the trajectory is no longer quasiperiodic at all. If the index in Fig. 2 is lower than 0.2 (dark blue), the orbit cannot be distinguished from a quasiperiodic one. When the index is between 0.2 and 0.7 (blue to yellow), the instability increases, but the associated diffusion remains (for almost all initial conditions) bound, which means that the disruption probably does not occur on a billion years timescale. For higher values of the index (orange to red), the corresponding planetary system is chaotic and its destruction is possible.

Class II—Non-resonant planets with significant secular dynamics: Planet pairs of this class can have strong gravitational interactions, where long-term variations are ascribed to secular perturbations, large variations of the eccentricities and dynamical effects like the alignment and anti-alignment for the apsidal lines [47]. For the long-term stability of such a system, it is not necessary that the planets are in MMR. Examples are, e.g., 55 Cnc (e and b), HD169830, and HD37124.

Class III—Hierarchical planet pairs: In this class one finds all planet pairs with a large ratio of their orbital periods—$P1/P2 > 10$. So that the gravitational interaction are not so strong like in class II and the probability of a capture in a MMR is negligible. The weaker interactions lead to stable motion in the numerical simulations, even if the orbits of the planets are not so determined. Examples are well HD168443, HD74156, and HD38529.

For a detailed description of the interesting topic of multi-planet systems we refer the reader to [23] or [48].

4 Binary Systems

The fact that more than 60% of the stars in the solar neighborhood build double- or multiple star systems (see [12]) underlines the necessity of stability studies for binaries. It is well known that in such systems the stable planetary motion is restricted to certain regions of the phase space due to the gravitational interactions between the celestial bodies. From the dynamical point of view, we distinguish three types of motion in double star systems [13]:

(i) *the satellite-type (or S-type) motion*, where the planet moves around one stellar component;
(ii) *the planet-type (or P-type) motion*, where the planet surrounds both stars in a very distant orbit and
(iii) *the libration-type (or L-type) motion*, where the planet moves in the same orbit as the secondary but 60° before or behind, furthermore, they are locked in 1:1 mean motion resonance.

Long before the first planet in a binary system has been discovered, astronomers, working in Dynamical Astronomy, carried out theoretical and numerical stability studies for the different types of motion (see, e.g., [13, 14, 62, 15, 16, 45, 32, 55, 56]) using the elliptic restricted three body problem (ER3BP)[4] for the numerical simulations. Between 1988 and 1998 Benest [3–8], studied in a series of papers several binaries. The discovery of planets in such systems encouraged other research groups to examine special double star systems (see, e.g., [31, 17, 19, 18, 58]).

Additionally, there are investigations, that used the general three body problem: see, e.g., [30, 27, 9], and more recently by [33] or [52].

At the moment the S-type motion is the most interesting one, since all detected extra-solar planets in binary systems orbit one of the stars (see Table 2).

The P-type motion will be more important as soon as planets will be discovered in very close binaries. In principle we know that the planetary motion around both stars is only stable for distances (from the mass center) $> 2\times$ the distance of the two stars. In the case of high-eccentric motion of the binary (around 0.7) the planet's distance has to be more than $4\times$ that of the two stars to be stable. For details see, e.g., [32, 56, 59].

The third type (L-type motion)—where the planet librates around one of the two Lagrangian triangular points of one of the stars is not so interesting for planetary

[4] The elliptic restricted three-body problem describes the motion of a massless body in the gravitational field of two massive bodies; the so-called primaries move in elliptic orbits (Keplerian motion) around their center of mass, without being influenced by the massless body.

motion in double stars due to a limitation in the mass ratio of the two stars:

$$\mu = m_2/(m_1 + m_2) < 1/26.$$

This motion is more interesting for single-star—giant planet—systems, where the limit of the mass ratio is easily fulfilled.

4.1 Discovered Planets in Binary Systems

From the more than 300 extra-solar planets discovered so far only a small part (around 30) were found to move in double star systems (see Table 3). According to Table 3 it is obvious that all discovered planets in binary systems move in S-type orbits due to the fact that most of these systems are wide binaries, where the distance between the two stars is more than 100 AU, except the systems *Gliese 86, HD41004 AB*, and γ *Cephei*. These are the most interesting ones from the dynamical point of view, therefore, we will discuss the orbital stability in these systems. First a short overview about general stability studies of S-type motion will be given.

4.2 S-Type Motion

Most of the general stability studies of S-type motion[5] in the planar ERTBP determined the stable region as a function of the binary's eccentricity, where the motion of the planet is initially circular (see, e.g., [62], 1988 [62] (= RD), or [32] (=HW)). Only the numerical investigation by [55] (=PLD) analyzed also the influence of the planet's eccentricity.

The three cited works determined the stable regions of planetary motion in a similar way. The host-star about which the planet (which is the massless body) moves is always m_1, then *the initial conditions of the binaries are* a fixed semi-major axis of 1 AU, a variation of the eccentricity between 0 and 0.9 with a step of 0.1 and two starting positions for the second star m_2 (the peri-center and the apo-center). *The initial conditions of the planets are* a semi-major axis between 0.1 and 0.7 AU with a various step δa and four starting positions were used for each orbit (i.e., mean anomaly $= 0^o, 90^o, 180^o, 270^o$). The initial eccentricity was zero in RD and HW and was varied between 0 and 0.5 with a step of 0.1 for all mass ratios and in some cases up to 0.9 in PLD.

In PLD the orbital behavior was determined by means of the FLI [25], which is quite a fast tool to distinguish between regular and chaotic motion. Chaotic orbits can be found very quickly because of the exponential growth of the tangent vector in the chaotic region. For most chaotic orbits only a few number of primary revolutions is needed to determine the orbital behavior. In order to distinguish between

[5] S-type motion is also called circumstellar motion.

Table 3 Planets in double stars [63]

Star	a_{binary} [AU]	a_{planet} [AU]	$M_{pl} \sin i$ [M_{Jup}]	e_{planet}
HD38529	12042	0.129	0.78	0.29
		3.68	12.7	0.36
HD40979	6394	0.811	3.32	0.23
HD222582	4746	1.35	5.11	0.76
HD147513	4451	1.26	1.00	0.52
HD213240	3909	2.03	4.5	0.45
Gl 777 A	2846	0.128	0.057	0.1
		3.92	1.502	0.36
HD89744	2456	0.89	7.99	0.67
GJ 893.2	2248	0.3	2.9	–
HD80606	1203	0.439	3.41	0.927
55 Cnc	1050	0.038	0.045	0.174
		0.115	0.784	0.02
		0.24	0.217	0.44
		5.25	3.92	0.327
GJ 81.1	1010	0.229	0.11	0.15
		3.167	0.7	0.3
16 Cyg B	860	1.66	1.69	0.67
HD142022	794	2.8	4.4	0.57
HD178911	789	0.32	6.292	0.124
Ups And	702	0.059	0.69	0.012
		0.83	1.89	0.28
		2.53	3.75	0.27
HD188015	684	1.19	1.26	0.15
HD178911	640	0.32	6.29	0.124
HD75289	621	0.046	0.42	0.054
GJ 429	515	0.119	0.122	0.05
HD196050	510	2.5	3.00	0.28
HD46375	314	0.041	0.249	0.04
HD114729	282	2.08	0.82	0.31
ϵ Ret	251	1.18	1.28	0.07
HD142	138	0.98	1.00	0.38
HD114762	132	0.3	11.02	0.25
HD195019	131	0.14	3.43	0.05
GJ 128	56	1.30	2.00	0.2
HD120136	45	0.05	4.13	0.01
γ Cep	20.3	2.03	1.59	0.2
Gl 86	21	0.11	4.01	0.046
HD41004 AB	23	1.7	2.64	0.5

stable and chaotic motion we defined a critical value for the FLIs depending on the computation time. In the general stability study of S-type motion the FLIs were computed for 1000 periods of the binary.[6] A comparison of the results of RD, HW,

[6] Even if the computation time seems to be quite short, one has to take into account that the results are valid for a much longer time due to the application of the FLI. Test computations of three selected mass ratios over a longer time (of 10^4, 10^5, and 10^6 primary periods) did not change the result significantly.

Table 4 Stable zone (in units of length) of S-type motion for all computed mass ratios and eccentricities of the binary. The given size for each μ, e_{binary} pair is the lower value of the studies by HW and PLD

	Mass ratio (μ)								
e_{binary}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.45	0.38	0.37	0.30	0.26	0.23	0.20	0.16	0.13
0.1	0.37	0.32	0.29	0.27	0.24	0.20	0.18	0.15	0.11
0.2	0.32	0.27	0.25	0.22	0.19	0.18	0.16	0.13	0.10
0.3	0.28	0.24	0.21	0.18	0.16	0.15	0.13	0.11	0.09
0.4	0.21	0.20	0.18	0.16	0.15	0.12	0.11	0.10	0.07
0.5	0.17	0.16	0.13	0.12	0.12	0.09	0.09	0.07	0.06
0.6	0.13	0.12	0.11	0.10	0.08	0.08	0.07	0.06	0.045
0.7	0.09	0.08	0.07	0.07	0.05	0.05	0.05	0.045	0.035
0.8	0.05	0.05	0.04	0.04	0.03	0.035	0.03	0.025	0.02

and PLD show them in good agreement. Minor variations are caused by the different methods used to determine the stable region. In some cases the FLI results gave a slightly larger stable region due to the fact that only four starting positions were used whereas HW used eight. Table 4 shows the border of the stable region (i.e., the semi-major axis of the last stable orbit) for different mass ratios, where we took the lower value of the two studies by HW and PLD.

The variation in the size of the stable zone due to an increase of e_{binary} or e_{planet} is shown in Table 5 for all mass ratios. For each μ we show the extension of the stable zone for circular motion of the binary and the planet and for an eccentric motion of 0.5 for both bodies.

Table 5 Stable zone (in dimension less units) of S-type motion for different mass ratios

		Stable zone	
Mass ratio $\mu = m_2/(m_1 + m_2)$	e_{binary}	$e_{planet} = 0$	$e_{planet} = 0.5$
0.1	0	0.45	0.36
	0.5	0.18	0.13
0.2	0	0.40	0.31
	0.5	0.16	0.12
0.3	0	0.37	0.28
	0.5	0.14	0.11
0.4	0	0.30	0.25
	0.5	0.12	0.07
0.5	0	0.27	0.22
	0.5	0.12	0.07
0.6	0	0.23	0.21
	0.5	0.10	0.07
0.7	0	0.20	0.18
	0.5	0.09	0.07
0.8	0	0.16	0.16
	0.5	0.09	0.05
0.9	0	0.13	0.12
	0.5	0.06	0.04

It can be seen that the reduction of the stable zone due to an increase of the binary's eccentricity is between 0.07 AU (i.e., for the initially circular motion in a binary with $\mu = 0.9$) and 0.28 AU (i.e., for the initially circular motion in a binary with $\mu = 0.1$). Even if the size of the stable region does not show a strong dependence on the eccentricity of the planet, it is not negligible, especially if a planet is close to the border of chaotic motion and moves in a highly eccentric orbit.

Figure 3 shows a summary of this study for $\mu = 0.2$, where we see for each (e_{Binary}, e_{Planet}) pair on the (x,y) plane the respective extension of the stable zone (z-axis), which is defined by the semi-major axis of the last stable orbit (corresponding to the largest distance of the planet to its host-star). The grey plane represents the limiting plane for stable motion. Similar 3-D plots for all mass ratios (from 0.1 to 0.9) as well as a detailed discussion of this work is given in PLD. As an example we show that the results for $\mu = 0.2$ given in Fig. 3 can be applied to the binary γ Cephei that hosts a giant planet.

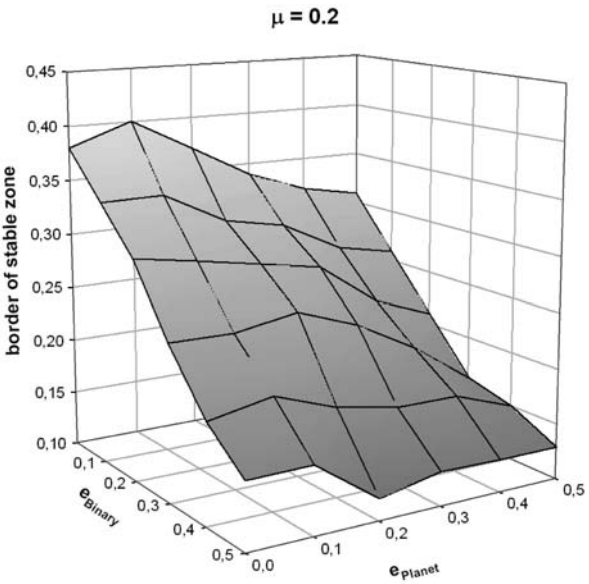

Fig. 3 The size of the stable zone of S-type motion in a binary with mass-ratio $\mu = 0.2$ depending on the eccentricity of the binary (x-axis) and of the planet (y-axis). It is clearly seen that the variation of e_{Binary} influences the extension of the stable zone stronger than the variation of e_{Planet}

4.3 γ Cephei

γ Cephei is one of the most interesting double star systems that hosts a planet. It is about 11 pc away from our solar system and consists of a K1 IV star (of 1.6 solar masses) and a M4 V star (of 0.4 solar masses). Thus the mass ratio $(m_2/(m_1 + m_2))$

of this system is 0.2. The detected planet of 1.76 Jupiter-masses orbits the K1 IV star at distance of 2.13 AU.

In Table 6 we show the extension of the stable zone (last column) for the old (upper part) and the new (lower part) orbital parameters using the results of PLD. In any case the stable region exceed 3 AU so that the detected giant planet at about 2 AU is clearly inside the stable zone.

Table 6 Stable zone derived from the study by [55]

Old orbital parameters			
	M4 V star	Giant planet	Border of stable zone
Mass	0.4 M_{Sun}	1.7 M_{Jup}	
Semi-major axis [AU]	$\sim$ 22 AU	$\sim$ 2 AU	
Eccentricity	0.44	0.21	
			$\sim$ 3.6 AU
New orbital parameters			
Mass	0.4 M_{Sun}	1.7 M_{Jup}	
Semi-major axis [AU]	$\sim$ 18.5 AU	$\sim$ 2.13 AU	
Eccentricity	0.36	0.12	
			$\sim$ 3.2 AU

Since the work by HW [32] is often used to confirm the stability of a detected planet in a binary, we show in Fig. 4a, b that one has to be careful, especially in the case of eccentric motion of the planet. As HW studied only circular planetary motion, they give a larger stable zone. This is well visible in Fig. 4b, where the two dashed lines show the zone for the stability borders defined by the relation given in HW. The plotted results of all eccentricities of a planet indicate that the majority of the stability borders is outside the given zone by HW. Figure 4a shows the results for circular orbits (full line with crosses) and low-eccentric motion ($e_{planet} = 0.1$, dashed line with stars), as well as the results of HW (dotted line with white squares). It is clearly seen that the results for the circular problem ($e_{binary} = 0$) are the same for the three cases; but for the elliptic problem (e_{binary} from 0.1 to 0.8) the stability borders determined by PLD are closer to m_1. Moreover, the results for $e_{planet} = 0$ and 0.1 are the same up to $e_{binary} = 0.3$ and again for $e_{binary} \geq 0.7$; and two cases ($e_{binary} = 0.1$ and 0.2) are outside the zone determined with the relation of HW. A higher eccentricity of the planetary motion shows that most of the stability borders are closer to the host-star. This could be important if the detected planet is quite close to the border of the stable zone.

4.3.1 Influence of the Secondary

To study the influence of the secondary, we examined in the system γ Cephei the region between the host-star and the detected planet by doing the calculations with (left panel of Fig. 5) and without (right panel of Fig. 5) secondary. A comparison of the two results shows significant differences. The presence of the perturbing star (see Fig. 5a) decreases the stable region (i.e., the faint region in the panels) and shows

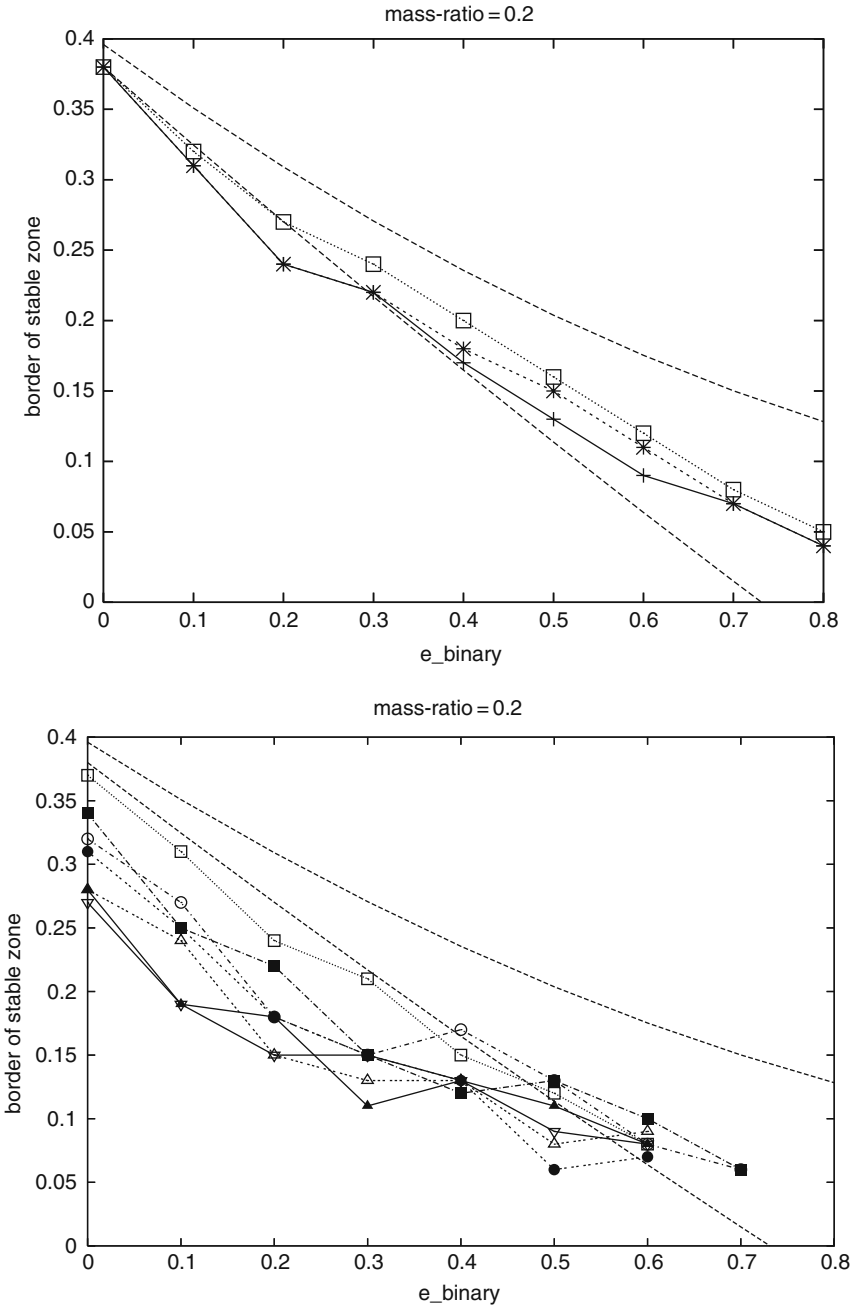

Fig. 4 A comparison of the results of PLD and HW. The area between the two *dashed lines* defines the zone for the stability border according to the relation given in HW. Panel (**a**) shows the results of PLD for $e_{planet} = 0$ (*full line with crosses*) and $e_{planet} = 0.1$ (*dashed line with stars*) and the result of HW (*dotted line with white squares*). Panel (**b**) shows the results for all e_{planet} (from 0 to 0.9) in comparison with the theoretical zone for the borderline of stability

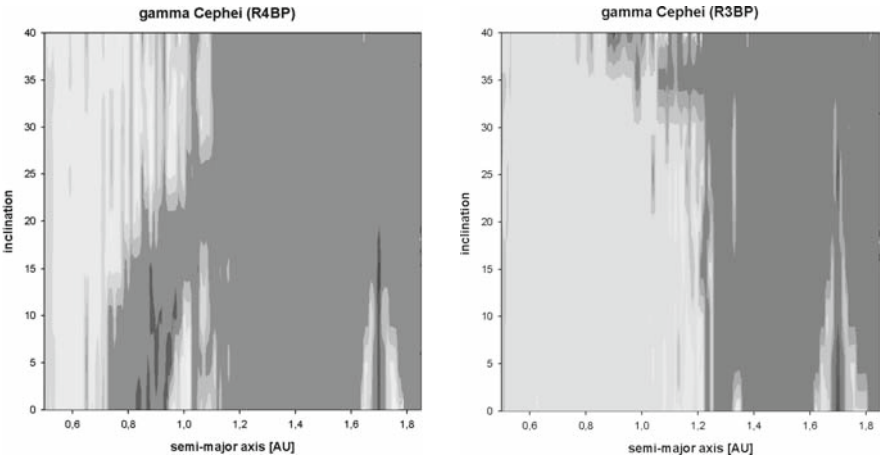

Fig. 5 FLI-stability maps for a fictitious planet in the vicinity of γ Cephei: *left panel* shows the result in the restricted four body problem (R4BP) (i.e., γ Cephei + secondary + detected planet + fictitious planet) and *right panel* shows the result in the restricted three body problem (R3BP) (i.e., γ Cephei + detected planet + fictitious planet). The *dark region* shows the chaotic zone and the *white area* the stable one

an arc-like chaotic path with a stable island around 1 AU (which corresponds to the 3:1 MMR). A first study about this significant difference is given in [57], where a variation of the semi-major axis of the detected giant planet shows the following.

When the giant planet is close enough to the host-star (e.g., around 1.3 AU) the region is mainly perturbed by MMRs with respect to the giant planet. An curved chaotic structure appears if the giant planet is shifted toward the secondary, so that a secular perturbation occurs, which means that the secondary causes a precession of the perihelion of the giant planet.[7]

4.4 Gliese 86

The binary Gliese 86 is about 11 pc away from the Sun in the constellation Eridanus and consists of a K1 main sequence star ($m_1 = 0.7 M_S$) and in all probability a white dwarf (with a minimum mass of $0.55 M_S$) at about 21 AU as proposed by [50] using NAOS-CONICA (NACO) and its new Simultaneous Differential Imager (SDI). The former detection by coronagraphic images using the ESO adaptive optic system ADONIS [20] identified a late brown dwarf (BD) of about 50 Jupiter-masses moving at a distance of at least 18.75 AU. But [20] could not explain the linear trend in the observation, which was possible with the new detection by [50] (=MN).

[7] Since all massive bodies were placed in the same plane a precession of the ascending node cannot be modeled.

However, the first who suggested a white dwarf (WD) companion for Gliese 86 was [35] in 2001.

A planet was found to be very close to the K1 V star, at 0.11 AU with an orbital period of less than 16 days [61]. Due to the CORALIE measurements a minimum mass of $4M_{Jupiter}$ was determined. It is evident that this close-in planet is not perturbed by the secondary star at 18.75 or 21 AU.

Since the eccentricity of the binary is not known, we examined numerically the dynamical behavior of fictitious low-mass planets orbiting Gliese 86. As we were interested in the size of the stable region for different eccentricities of the binary, we neglected the detected planet. The initial conditions of the massive bodies (m_1 and m_2 were taken from Table 7 and all angles (inclination (i), node (Ω), perihelion distance (ω), and mean anomaly (M)) were set to zero.

Table 7 Orbital parameters of the binary Gliese 86

	m_1 (K1 V star)	m_2 (BD)	m_2 (WD)	m_3 (planet)
Mass	0.79 M_S	50 M_J	0.55 M_S	4 M_J
Semi-major axis [AU]	0	18.75	21	0.11
Eccentricity	0.0–0.9	0.0–0.9	0.0–0.7	0.046

The FLIs were used to determine the dynamical behavior of the trajectories and the integration time was between 1000 and 100,000 periods of the binary, using the ER3BP as dynamical model. The massless bodies were started in the semi-major axes range between 0.3 and 12 AU with a step of 0.01 AU, and e, i, ω, ΩM were set to 0.

The results of the FLI computations are shown in Fig. 6 which splits the (semi-major axis e_{binary}) parameter space into three zones: (i) a *stable zone* whose border (dashed lines with black squares (for m_2 = WD)/solid line with black circles (for m_2 = BD)) is defined by the largest distance from Gliese 86 up to which we have found only regular motion; (ii) a *chaotic zone*, where no regular motion can be found—which is outside the dashed line with open circles. In between the two border lines one can see (iii) a *mixed zone* where both regular and chaotic motion can be found. It is clearly seen that for low-eccentric motion of the binary the two border-lines of stable motion coincide whereas those of the chaotic zone are not at the same position. Therefore, the new system (WD secondary) has a smaller mixed zone. For the new system we observe a linear decrease of the stable zone when increasing e_{binary} from 0.1 to 0.9. Moreover, the mixed zone shrinks significantly at $e_{binary} = 0.4$ and remains nearly constant for higher eccentricities (except for $e_{binary} = 0.6$). When $e_{binary} \geq 0.4$, one can see that both border lines of the new system lie inside the stable zone of the old system (BD secondary). This means that especially for eccentric motion of the binary the mass of the perturbing star plays an important role for the size of the stable zone. The result of the old system shows a quite constant extension of the mixed zone for $e_{binary} \leq 0.2$, while for higher eccentricities a linear decrease in the size can be observed.

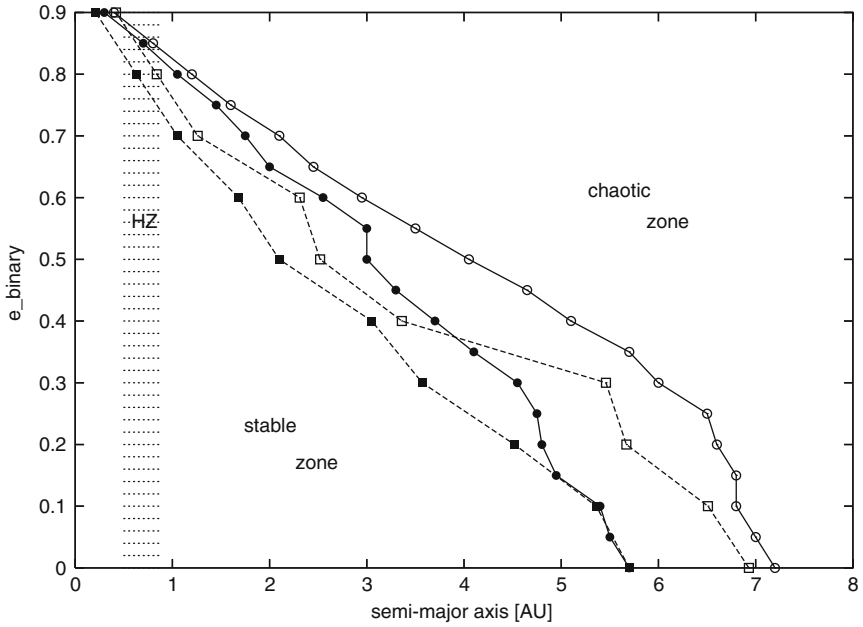

Fig. 6 The stability of a fictitious massless body in the gravitational field of the binary Gliese86 AB, where the detected planet at 0.11 AU was neglected. One can see three zones (stable, mixed, and chaotic) for both configurations: *dashed lines* with *open and full squares* for a WD and *full lines* with *open and full circles* for a BD as secondary

Remark. As a comparison we did some computations using the general three body problem with a planet's mass of about five Jupiter-masses. For such a system the stable zone shrinks significantly only for high-eccentricity motion of the fictitious planet ($e_{planet} \geq 0.3$).

To get a first picture about the gravitational influence of the secondary (i.e., the brown dwarf) and of the discovered planet on a fictitious planet moving in the region between these two bodies, we computed the MMR up to the order 20. Its representation is given in Fig. 7, where the lower part refers to the detected planet, the upper part refers to the secondary, and the dotted region labels the habitable zone. It is clearly seen that most of the resonances with respect to the detected planet are concentrated to distances < 0.3 AU from the K1V star and only a few, very high-order resonances were found in the habitable zone.

4.5 HD41004 AB

Another interesting binary system is HD41004 AB, which was in the sample of the Geneva extra-solar planet search program using the Coralie spectrograph at La Silla Observatory. For the first analysis, 86 radial velocity measurements were used [69], where they identified a brown dwarf orbiting HD41004 B with an observed period of

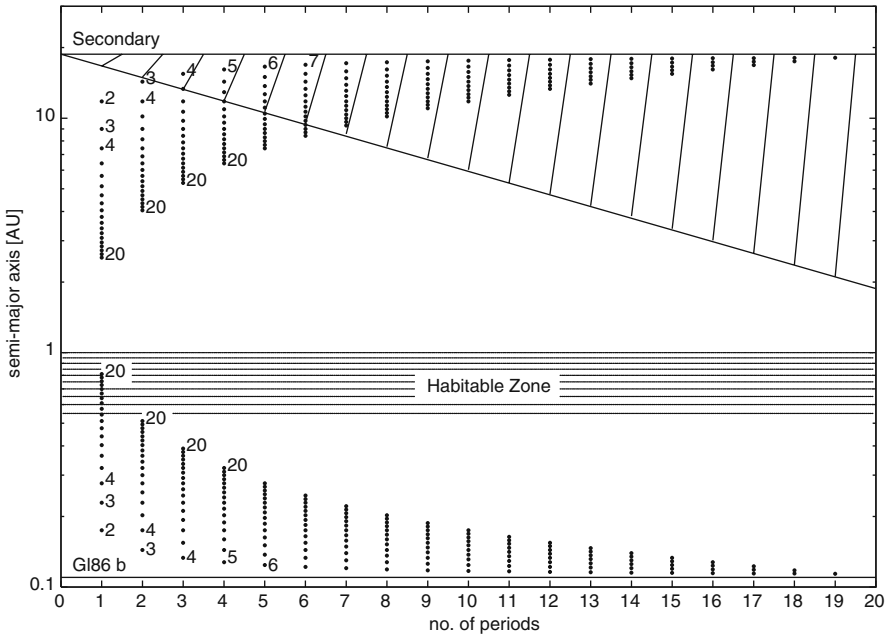

Fig. 7 The mean motion resonances (MMRs) up to the order 20 of an additional fictitious planet with respect to Gliese 86b (*lower part*) and with respect to the secondary (*upper part*). The x-axis denotes the number of periods either for a fictitious planet (*lower part*) or for the secondary—BD—(*upper part*), and the y-axis shows the position of the resonances (on a log scale). For a better understanding of the graphical presentation we give the following examples: e.g., "2" at x=1 and y=11.14 (*upper part of the figure*) means 2:1 resonance of a fictitious planet with the secondary at a=11.14 AU; and "2" at x=1 and y= 0.174 (*lower part*) means 1:2 resonance of a fictitious planet with Gliese 86b at a=0.174 AU. The hatching denotes the region which is occupied by the secondary, and since we do not have any knowledge about its eccentricity we marked this region for $e_{secondary}$ from 0 (*left border*) to 0.9 (*right border*). The *dotted region* labels the habitable zone of Gliese 86. Here it is clearly seen that the MMRs do not influence the HZ of Gliese 86 except the high-order MMRs with respect to the detected planet, which are not that important. We have to note that the MMR plot for the new system (m_2 = WD) is quite similar to Fig. 2 — so it is useless to show both

only 1.3 days. Moreover, the observed long-term linear trend in the radial velocities was ascribed primarily to the motion of HD41004 A around component B—but it could also be caused by an additional component, which was proofed by further observations and analyses of the system that showed a planetary companion near HD41004 A [73]. The system HD41004 AB can be divided into 2 subsystems, with a projected distance of the two stellar components between 20 and 23 AU according to the different observations. Both stars have a sub-stellar companion, whose orbital parameters are given in Table 8.

Since the different observational samples change the orbital parameters of this planetary system drastically, we examine the stability of planetary motion for the

Table 8 Orbital parameters of the binary HD41004 AB [74], where we divide the binary system into two subsystems, with a distance of at least 23 AU

	HD41004 A ($m = 0.7M_{Sun}$)			HD41004 B ($m = 0.4M_{Sun}$)
	Planet (IC1)	Planet (IC2)	Planet (IC3)	brown dwarf
$m_P[M_{Jup}]$	2.3	2.436 ± 0.098	2.6 ± 1.8	18.64 ± 0.26
a_P [AU]	1.31	1.64	1.7 ± 0.11	0.0007544
e_P	0.39 ± 0.17	0.5	0.74 ± 0.2	0.065 ± 0.014
ω_P [deg]	114 ± 10	71.7 ± 4.6	97 ± 31	171 ± 11
P [days]	655 ± 37	924 ± 25	963 ± 38	1.328199

different orbital parameters in the (a_P, e_B-plane)[8] fixing the eccentricity of the giant planet to the different observed values. So that we are able to determine a maximum eccentricity for the binary that assures long-term stability for the detected giant planet, since we have no knowledge about the eccentricity of the binary. In addition, we are faced with a large error in the eccentricity of the detected giant planet. Figure 8 shows clearly that—depending on the planet's eccentricity—e_B has to be < 0.6 in all cases, and for $e_P = 0.74$, e_B has to be < 0.15, otherwise the detected planet would not be in the stable region.

5 Planets in the Habitable Zone (HZ)

The HZ is the region around a star, where liquid water is stable on the surface of an Earth-like planet [37]. Another assumption for such a planet is the existence of an appropriate planetary atmosphere. The study of habitability is certainly an interdisciplinary venture including astrophysical, biological, geophysical, and chemical studies. From the astrophysical point-of-view studies of the stellar luminosity and its influence on the distance of the HZ, as well as the planet's mass (to maintain an atmosphere) and planetary composition (assuming a terrestrial planet), are important contributions to the science of habitability.

The evolution of a biosphere is a process over a long time, therefore, it is obvious that long-term orbital stability in the HZ represents one of the basic requirements for habitability. This emphasizes the importance of such numerical investigations for known and future extra-solar planetary systems.

To define the boundaries of this zone we used the work by [37], that is, based on a planet with a terrestrial ocean of superficial water, the carbonate–silicate cycle which controls the CO_2 level in the atmosphere and the surface temperature that is above freezing in the HZ. The conservative estimate of the outer boundary in the solar system is at 1.3 AU. For larger semi-major axes CO_2 condensates in the atmosphere producing CO_2 clouds that can affect the temperature-CO_2 coupling significantly. The inner boundary is at 0.93 AU. For $a < 0.93$ AU, H_2O becomes a major atmospheric compound and is rapidly lost to space after UV photolysis.

[8] a_P is the semi-major axis of the giant planet and e_B is the eccentricity of the binary.

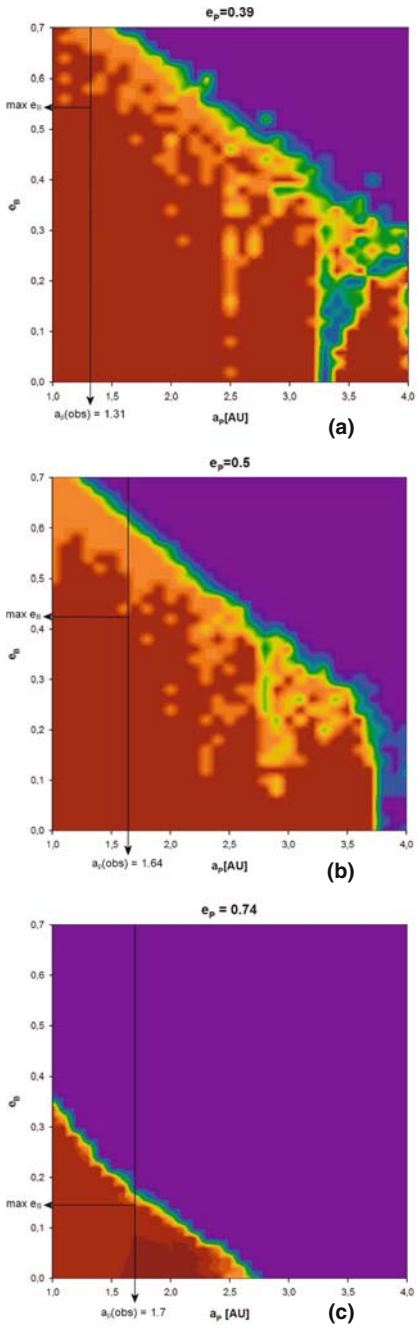

Fig. 8 FLI-stability maps for the different parameter sets: (**a**) the old HD41004 A system with $e_P = 0.39$, and the new HD41004 A system with (**b**) $e_P = 0.5$ and (**c**) $e_P = 0.74$. A variation of the planet's semi-major axis a_P (x-axis) and the binary's eccentricity e_B (y-axis) allows to determine the maximum e_B for which we have found long-term stability of the detected giant planet (see the *horizontal black line*). The *vertical line* in each panel labels the observed position of the detected planet for the different orbital parameters. Long-term stability can be expected in the *red region* while *dark blue mark* highly chaotic orbits

Further studies find a potentially larger HZ for a Sun-like star—see e.g., [24] or [49] (two groups that include CO_2 cloud effects).

The size of the HZ is limited to a small region, depending on the spectral type and the age of the host-star, therefore the planet's eccentricity has to be small enough if we require that the planet is always in the HZ. In dynamical studies we distinguish different types of HZ depending on the giant planet of the system:

(1) The *solar system type HZ*, where the HZ is between the host-star and the detected giant planet; this is the case for HD41004 A, which is studied in this chapter.
(2) The *hot-Jupiter type HZ*, where the HZ is outside the giant planet.
(3) The *giant planet habitable zone (GPHZ)* where the detected giant planet moves in the HZ. In this case we can only expect so-called habitable moons or habitable Trojan planets (see [42, 18, 22]).

If an Earth-like planet is found in a single-star single-planet system, it is possible to apply the Exocatalogue like it is described in Sect. 2 or more detailed in [68].

In this investigation we give only a brief introduction to the very interesting study of planets in the HZ — showing results for Jupiter–Saturn like-systems and for the two binary systems Gliese 86 and HD41004 AB. For a detailed information of this topic we recommend the reader to the book *Extrasolar planets* ed. Dvorak [48], where several articles provide a good overview about the problematic of the HZ.

5.1 EPS Similar to the Jupiter–Saturn Configuration

The influence of Jupiter and Saturn on the motion in the HZ of a Sun-like star was studied by the computation of test-planets (with negligible mass) between $a_{tp} = 0.6$ and 1.6 AU.[9] For the giant planets, we used the orbital parameters of Table 9 and for the test-planets, we varied the semi-major axis between 0.6 and 1.6 AU in steps of 0.02 AU and assumed Earth's orbital parameters for the eccentricity $e_{tp} = 0.0167$— the inclination $i_{tp} = 0.0008^{\circ}$—the argument of perihelion $\omega_{tp} = 103.946^{\circ}$—the node $\Omega_{tp} = 358.859^{\circ}$—and the mean anomaly $M_{tp} = 206.900^{\circ}$.

Table 9 Orbital elements of the gas giants

Planet	a [AU]	e	Inc. [deg]	ω [deg]	Ω [deg]	M [deg]	Mass [m_{Sun}]
Jupiter	5.2028	0.0483	1.3046	275.201	100.471	183.898	0.9547907e-3
Saturn	9.5300	0.0533	2.4864	339.520	113.669	238.293	0.2858776e-3

Calculating the test-planets for the different starting positions of Saturn (between 8 and 11 AU in steps of 0.1 AU), a stability map representing the (a_{tp}, a_S) plane

[9] The HZ defined by [37] (i.e., from 0.93 to 1.3 AU) was extended to get additional information for the positions of Venus and Mars for the solar system configuration (i.e., $[\kappa = 1]$), which is interesting from the dynamical point of view.

contains information for 1581 data pairs (see, e.g., Fig. 9). This figure shows the perturbations of the Jupiter–Saturn system via the maximum eccentricity (max-e) over the whole computation time, where red labels the region of lowest max-e and blue that of the highest one. Analyzing this result, we find that most of the HZ is not affected by Jupiter and Saturn (i.e., the red region). The stability map is dominated by an arched band that indicates higher eccentricities of the test-planet due to a stronger influence of the giant planets in this region. Within this band are several positions, where the eccentricity is about 0.5 (green regions) or even higher (see, e.g., the blue region near the position of Mars). Apart from this significant stripe some smaller regions show an increase in eccentricity, especially in the outer part of the HZ ($a_{tp} > 1.2$ AU), e.g., when Saturn is nearly in 2:1 MMR with Jupiter. In that case, a test-planet moving near the position of Mars would be influenced by the gas giants, causing an eccentricity of about 0.2 for the test-planet. An even higher eccentricity can be observed when Saturn is placed between 10 and 10.7 AU, especially for the region outside the orbit of Mars. If Saturn orbits the Sun at 11 AU perturbations in the area between 1.15 and 1.5 AU appear.

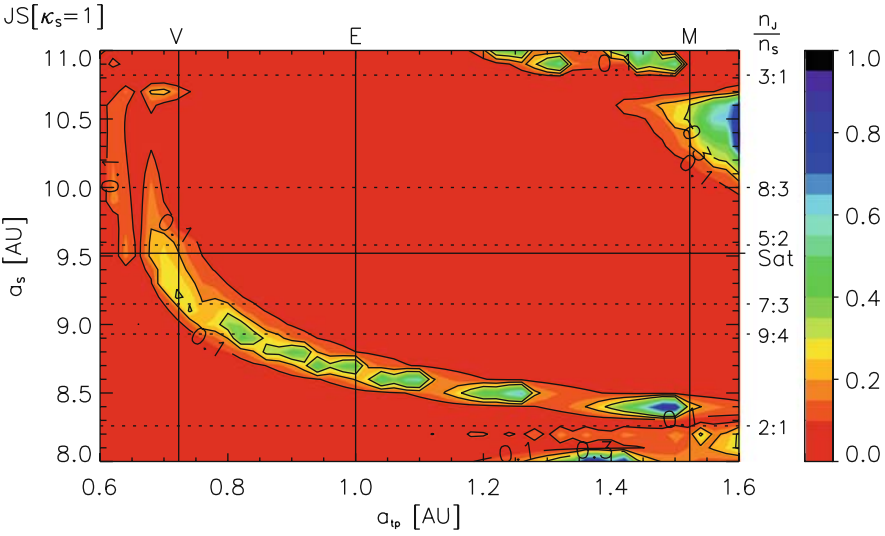

Fig. 9 Stability map for Earth-like planets under the influence of Jupiter and Saturn. Both axes show different initial semi-major axes either of the massless Earth-like planets (x-axis) or of Saturn (y-axis)—grid-size in x is 0.02 AU and in y 0.1 AU. The *vertical black full lines* indicate the positions of Venus (V), Earth (E), and Mars (M); and the *horizontal one* represents that of Saturn (Sat). The *dashed horizontal lines* label the positions of different mean motion resonances. Different colors belong to different values of max-e (see the color scaling)

Another interesting result is the heightened eccentricity for a test planet at the position of Venus, especially in the case, when Saturn is placed at its actual semi-major axis. The color scaling indicates a max-e around 0.2, in contrast Venus is moving in a nearly circular orbit in our solar system.

Plotting the eccentricity of the test-planet at 0.72 AU one can see a regular signal with large variations between nearly 0 and 0.22 (see the red line in Fig. 10). The same result was found when we replaced our initial orbital elements for the test-planets with those of Venus[10] or when we started with $e_{tp}, i_{tp}, \omega_{tp}, \Omega_{tp}, M_{tp} = 0$. Even the mass of Venus did not change the result significantly (see the blue line in Fig. 10). Only, the presence of the Earth in this system causes a significant decrease in the eccentricity of Venus' orbit: in the case of a massless body the eccentricity will not exceed 0.09 (green line in Fig. 10) and for a body with Venus-mass the eccentricity is always < 0.05 (see the black line in Fig. 10).

The increase in Venus' eccentricity caused by the absence of the Earth–Moon system has been already found by [34], when studying the dynamical stability of the inner solar system. In one of their models,[11] that can quasi be compared to our dynamical model (for $\kappa = 1$), they have found the maximum of e_V near 0.6 and a period of e_V' variation of about 8.1 Myrs. In our restricted four body model (Sun–Jupiter–Saturn + massless test-planets) the period of the variation in e_V is around 4

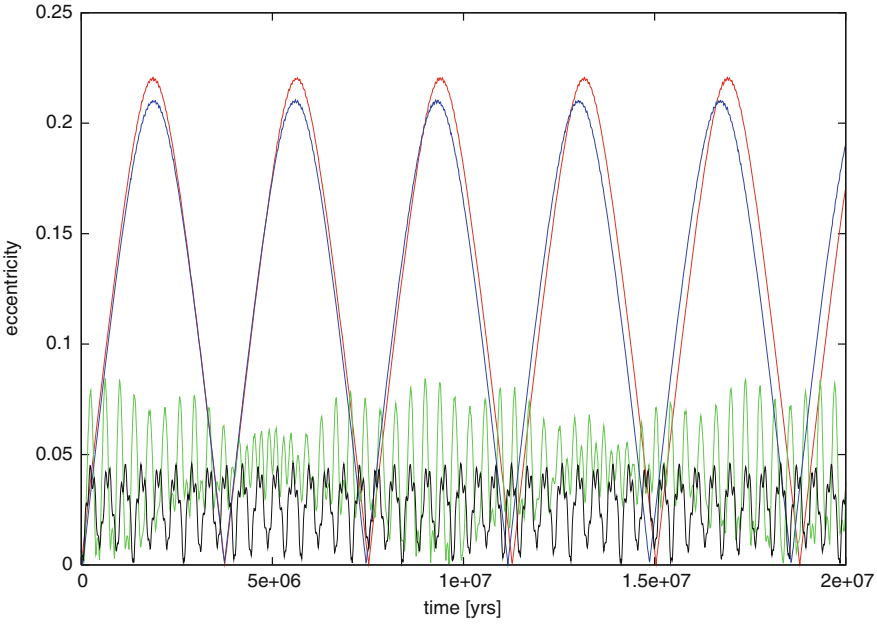

Fig. 10 The eccentricity (y-axis) of a (test-)planet placed at Venus' semi-major axis, computed over 2×10^7 years (x-axis). We compare the evolution of $e_V(t)$ in different systems: (i) Venus in the Jupiter–Saturn system as massless test-planet (*red line*) and as massive body (*blue line*); (ii) Venus in the Jupiter–Saturn–Earth system as massless test-planet (*green line*) and as massive body (*black line*)

[10] $a_V = 0.7233$ AU, $e_V = 0.007$, $i_V = 3.3947^o$, $\omega = 54.7176^o$, $\Omega = 76.6953^o$, $M = 254.37111^o$; note that in our calculation we took 0.72 AU as initial a_V.

[11] Consisting of Sun increased by the mass of Mercury and the planets Mars through Neptune.

Myr and the maximum of e_V is about 0.22. This disagreement results mainly from differences in the models used. For details of this study, we recommend the reader to the paper by Pilat-Lohinger et al. [60], where systems with higher Saturn-mass are also analyzed.

5.2 Planets in the HZ of Close Binary Systems

5.2.1 Gliese 86 A

In the case of Gliese86 A the HZ is—according to [37]—between 0.48 and 0.95 AU. As the detected gas giant moves at 0.11 AU, its gravitational influence on the HZ cannot be strong, as we have already seen in Fig. 7 that only high-order resonances can be found in this zone. Our stability study shows the HZ of Gliese 86 in a very stable state up to an eccentricity of the binary of 0.75 (for a BD secondary) an 0.7 (for a WD secondary). While for higher e_{binary} the HZ will be chaotic (see Fig. 6).

The most important question for the binary Gliese86 AB is, where was the planet built? If it was formed at a distance between 4 and 5 AU[12] and migrated toward the star through the HZ, an already existing terrestrial planet would have been ejected from the system. But if the gas giant was built closer to the star—maybe quite near to the region, where it was found (see [72]), then we can expect terrestrial planets in the HZ (which cannot be detected up to now). However, there are only a few studies that deal with the difficult problem of planetary formation in binaries (see, e.g., [39, 38, 53] or [54]), which needs still a lot of work.

In a paper by [64] results of simulations are discussed, where they have shown the formation of several Earth-mass planets in systems with a close-in planet. They claim that more than a third of the known systems might harbor Earth-like planets.

5.2.2 HD41004 A

Taking the old parameter of this system and using the R4BP as numerical method we calculated various stability maps of the HZ of HD41004A to have a global overview of the dynamical state of motion in this region.

Each map contains between 900 and 1500 orbits, which are analyzed using both methods: (i) the FLIs (for 10^4 periods of the binary) to cancel out the fast diffusion and (ii) the max-e (for 10^5 years) to determine those orbits, which are always in the HZ. In total we computed 44 stability maps (26 FLI-maps and 18 max-e maps). A selection of these results is given in figs.11(a-i). All stability maps show that

(i) the stable motion is limited to the inner region of the HZ ($a_{TP} < 0.7$ AU) and
(ii) the stability region is fragmented into several stable stripes, which reminds one of the distribution of the main-belt asteroids, where a similar resonant structure

[12] Before the discovery of extra-solar planets it was claimed by A. Boss that the formation of gas planets is at or outside 5 AU (which is called snow-line).

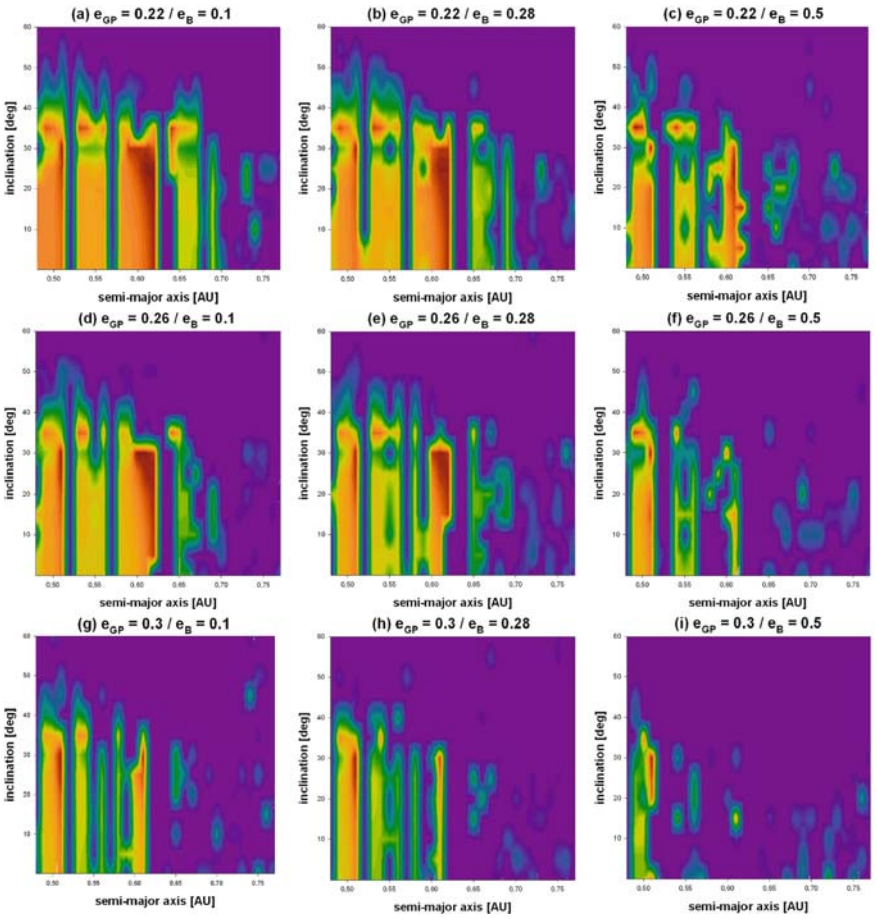

Fig. 11 Stability maps (max-e) in the a–i plane for different eccentricities of the binary and of the giant planet. The different colors indicate zones of different max-e: < 0.1 (*red*), ..., > 0.8 (*dark blue*). For details see the text

can be found due to the so-called Kirkwood gaps (see, e.g., Murray and Dermott [51] for a detailed explanation). It is well known that the Kirkwood gaps occur as a result of resonances with Jupiter. The similar structure in the HZ of HD41004 A can be explained by MMRs with the detected giant planet — where the **4:1** MMR is near 0.52 AU, the **7:2** MMR is near 0.57 AU, the **3:1** MMR is near 0.63 AU, and the **8:3** MMR is near 0.68 AU.

In the 9 panels of Fig. 4 we summarize our computations, and show how the stable regions depend on (i) the eccentricity of the binary (left: $e_B = 0.1$, middle: $e_B = 0.28$ and right: $e_B = 0.5$) and (ii) the eccentricity of the giant planet (top: $e_P = 0.22$, middle: $e_P = 0.26$, and bottom: $e_P = 0.3$). In all plots, the red region marks the orbits with the lowest e-max (< 0.15) and the dark blue area labels the orbits with the highest e-max (i.e., corresponding to unstable motion) that the test-planets

achieved during the computations. Besides the restriction of e_B to values < 0.7, one can see from Fig. 11a limitation in the inclination of the test-planets to $i_{TP} < 50^\circ$. Comparing the different panels, it is clearly seen that an increase of either e_P or e_B would lead to smaller stable areas.

The study of the existence of possible "dynamically habitable planets" (DHPs) in the HZ of HD41004 A yields the following result:

If we consider planetary motion with eccentricities $e_{TP} < 0.25$ to be dynamically habitable, then about 10% of the orbits are DHPs when $e_P = 0.22$ and $e_B = 0.1$ (Fig. 4a). An increase of either e_B to 0.5 (see Fig. 4c) or e_P to 0.3 (see Fig. 4g) would reduce the DHPs to 3 or 2, respectively. If both eccentricities are quite large (see Fig. 4i), only a small stable island near the inner border of the HZ remains. Another remarkable feature, that can be seen in most of the dynamical maps of Fig. 4, is an increase of the lowest e-max area around $a_{TP} = 0.61$ AU for higher inclinations (up to 30°—visible by the V-shape of the dark region).

The most appropriate regions for habitable planets are found around 0.51 AU, where the planetary motion has to be nearly circular ($e < 0.5$), to remain in the HZ, and around 0.61 AU, where eccentricities up to about 0.24 are allowed for DHPs. Here we should note, that the application of the HZ defined by [36]—i.e., between 0.36 and 0.71 AU—would allow higher eccentricities for DHPs near 0.51 AU (up to nearly 0.3) while in the region around 0.61 AU the DHPs should have an eccentricity < 0.17 to guarantee that the orbit of the terrestrial planet is confined to the HZ. Even if [70] claim that a habitable planet may leave the HZ in the peri-center/apo-center without loss of its habitability (to allow a higher eccentricity for habitable planets), we remind the reader that in the case of HD41004 A a high-eccentricity motion of a terrestrial planet would enter the planet into the chaotic zone outside 0.7 AU.

Acknowledgments The authors wish to acknowledge the support by the Austrian FWF: EP-L was financed by project P19569-N16 and BF by the Erwin Schrödinger grant no. J2892-N16.

References

1. Beaugé, C., Ferraz-Mello, S., Michtchenko, T.: Astron. Astrophys. J. **598**, 1124 (2003) 487
2. Beaugé, C., Ferraz-Mello, S., Michtchenko, T.: MNRAS. **341**, 760 (2003) 487
3. Benest, D.: Astron. Astrophys. **206**, 143 (1988) 490
4. Benest, D.: CMDA. **43**, 47 (1988) 490
5. Benest, D.: Astron. Astrophys. **223**, 361 (1989) 490
6. Benest, D.: CMDA. **56**, 45 (1993) 490
7. Benest, D.: Astron. Astrophys. **314**, 983 (1996) 490
8. Benest, D.: Astron. Astrophys. **332**, 1147 (1998) 490
9. Black, D.C.: Astrophys. J. **87**, 1333 (1982) 490
10. Bois, E., Kiseleva-Eggleton, L., Rambaux, N., Pilat-Lohinger, E.: Astron. Astrophys. J. **598**, 1312 (2003) 484, 487
11. Boisnard, L. Auvergne, M.: In: Fridlund, M., Baglin, A., Lochard, J., Conroy, L. (eds.) Proceedings of the CoRoT Mission Pre-Launch Status – Stellar Seismology and Planet Finding (ESA SP-1306). ISBN 92-9092-465-9, p. 19 (2006) 482
12. Duquennoy, A., Mayor, M.: Astron. Astrophys. **248**, 485 (1991) 490

13. Dvorak, R.: CMDA. **34**, 369 (1984) 490
14. Dvorak, R.: Astron. Astrophys. **167**, 379 (1986) 490
15. Dvorak, R., Froeschlé, Ch., Froeschlé, C.: Astron. Astrophys. **226**, 335 (1989) 490
16. Dvorak, R., Lohinger, E.: In: Roy, A.E. (ed.) Proceedings of the NATO ASI Series, vol. 439. Plenum Press, New York and London (1991) 490
17. Dvorak, R., Pilat-Lohinger, E., Funk, B., Freistetter, F.: Astron. Astrophys. **398**, L1 (2003) 490
18. Dvorak, R., Pilat-Lohinger, E., Schwarz, R., Freistetter, F.: Astron. Astrophys. **426**, L37 (2004) 490, 503
19. Dvorak, R., Pilat-Lohinger, E., Funk, B., Freistetter, F.: Astron. Astrophys. **410**, L13 (2003) 484, 490
20. Els, S.G., Sterzik, M.F., Marchis, F., Pantin, E., Endl, M., Kürster, M.: Astron. Astrophys. **370**, L1 (2001) 497
21. Érdi, B., Dvorak, R., Sándor, Zs., Pilat-Lohinger, E.: MNRAS. **351**, 1943 (2004) 484
22. Érdi, B., Sándor, Zs.: CMDA. **92**, 113 (2005) 503
23. Ferraz-Mello, S., Michtchenko, T.A., Beaugé, C., Callegari, Jr.: LNP Proceedings. In: Dvorak, R. et al. (eds.) Springer, Heidelberg **683**, 219 (2005) 486, 490
24. Forget, F., Pierrehumbert, R.T.: Science. **278**, 1273 (1997) 503
25. Froeschlé, C., Lega, E., Gonczi, R.: CMDA. **67**, 41 (1997) 484, 491
26. Gaudi, B.S., Patterson, J., Spiegel, D.S., Krajci, T., Koff, R., Pojman'ski, G., Dong, S., Gould, A., Prieto, J.L., Blake, C.H., Roming, P.W.A., Bennett, D.P., Bloom, J.S., Boyd, D., Eyler, M.E., de Ponthire, P., Mirabal, N., Morgan, C.W., Remillard, R.R., Vanmunster, T., Wagner, R.M., Watson, L.C.: Astron. Astrophys. J. **677**, 1268 (2008) 487
27. Graziani, F., Black, D.C.: Astron. Astrophys. J. **251**, 337 (1981) 490
28. Guzzo, M., Lega, E., Froeschlé, C.: Physica D. **163**, 1 (2002) 484
29. Hadjidemetriou, J.: CMDA. **83**, 141 (2002) 487
30. Harrington, R.S.: Astrophys. J. **82**, 753 (1977) 490
31. Holman, M.J., Touma, J., Tremaine, S.: Nature. **386**, 254 (1997) 490
32. Holman, M.J., Wiegert, P.A.: Astrophys. J. **117**, 621 (1999) 490, 491, 495
33. Innanen, K.A., Zheng, J.Q., Mikkola, S., Valtonen, M.J.: Astrophys. J. **113**, 1915 (1997) 490
34. Innanen, K., Mikkola, S., Wiegert, P.: Astrophys. J. **116**, 2055 (1998) 505
35. Jahreiß, H.: Astronomische Gesellschaft Abstract Series. **18**, 110 (2001) 498
36. Jones, B.W., Underwood, D.R., Sleep, P.N.: Astron. Astrophys. J. **622**, 1091 (2005) 508
37. Kasting, J.F., Whitmire, D.P., Reynolds, R.T.: Icarus. **101**, 108 (1993) 501, 503, 506
38. Kley, W., Burkert, A.: Proceedings of Conference: Disks, Planetesimals and Planets, Tenerife, January (2000) 506
39. Kley, W.: The Formation of Binary Stars, Proceedings of IAU Symp. 200, held 10–15 April 2000, in Potsdam, Germany, Edited by Hans Zinnecker and Robert D. Mathieu, p. 511 (2001) 506
40. Laskar, J.: Instability, Chaos and Predictability in Celestial Mechanics and Stellar Dynamics. Proceedings of the International Astronomical Union Colloquium 132, held in Delhi, India, October 10–13, 1990 [i.e. 1991] Editor, K.B. Bhatnagar; Publisher, Nova Science Publishers, Commack, N.Y., 1993. LC # QB349. I567 1990. ISBN # 1560720549., p. 21 (1993) 489
41. Latham, D.W., Stefanik, R.P., Mazeh, T., Mayor, M., Burki, G.: Nature. **339**, 38 (1989) 481
42. Laughlin, G., Chambers, J.E.: Astrophys. J. **124**, 592 (2002) 503
43. Lee, M.H., Peale, S.J.: Astron. Astrophys. J. **567**, 596 (2002) 487
44. Lee, M.H.: Astron. Astrophys. J. **611**, 517 (2004) 487
45. Lohinger, E., Dvorak, R.: Astron. Astrophys. **280**, 683 (1993) 490
46. Mayor, M., Queloz, D.: Nature. **378**, 355 (1995) 482
47. Michtchenko, T., Malhotra, R.: Icarus. **168**, 237 (2004) 489
48. Michtchenko, T.: Extrasolar Planets. Formation, Detection and Dynamics, Edited by Rudolf Dvorak. ISBN: 978-3-527-40671-5, Wiley-VCH, New York, p. 151 (2007) 490, 503
49. Mischna, M.A., Kasting, J.F., Pavlov, A., Freedman, R.: Icarus. **145**, 546 (2000) 503
50. Mugrauer, M., Neuhäuser, R.: MNRAS. **361**, L15 (2005) 497
51. Murray, C.D., Dermott, S.F. Solar system dynamics. Cambridge University Press, Cambridge (1999) 507

52. Musielak, Z.E., Cuntz, M., Marshall, E.A., Stuit, T.D.: Astron. Astrophys. **434**, 355 (2005) 490
53. Nelson, R.P.: Astronomische Gesellschaft Abstract Series, Vol. 18., Abstracts of Contributed Talks and Posters presented at the Annual Scientific Meeting of the Astronomische Gesellschaft at the Joint European and National Meeting JENAM 2001 of the European Astronomical Society and the Astronomische Gesellschaft at Munich, September 10–15, 2001, abstract # MS 04 13 (2001) 506
54. Nelson, R.P., Papaloizou, J.C.B.: In: Proceedings of the Conference on Towards Other Earths: DARWIN/TPF and the Search for Extrasolar Terrestrial Planets, 22–25 April 2003, Heidelberg, Germany. Edited by M. Fridlund, T. Henning, compiled by H. Lacoste. ESA SP-539, Noordwijk, Netherlands: ESA Publications Division, ISBN 92-9092-849-2, p. 175 (2003) 506
55. Pilat-Lohinger, E., Dvorak, R.: CMDA. **82**, 143 (2002) 484, 490, 491, 495
56. Pilat-Lohinger, E., Funk, B., Dvorak, R.: Astron. Astrophys. **400**, 1085 (2003) 490
57. Pilat-Lohinger, E.: In Dynamics of populations of planetary systems. In: Knezevic, Z., Milani, A. (eds.) Proceedings of IAU Coll. 197, Cambridge University Press, Cambridge, p. 71 (2005)** 497
58. Pilat-Lohinger, E., Funk, B.: The stability of exo-planets in the binary Gliese 86AB, Proceedings of the 4th Austro-Hungarian Workshop, in Budapest 2005, p. 103 (2006) 490
59. Pilat-Lohinger, E., Dvorak, R.: Extrasolar planets. Formation, detection and dynamics, Edited by Rudolf Dvorak. ISBN: 978-3-527-40671-5, Wiley-VCH, New York, p. 179 (2007) 490
60. Pilat-Lohinger, E., Süli, Á., Robutel, P., Freistetter, F.: Astron. Astrophys. J. **681**, 1639 (2008) 488, 489, 506
61. Queloz, D., Mayor, M., Weber, L., Blicha, A., Burnet, M., Confino, B., Naef, D., Pepe, F., Santos, N., Udry, S.: Astron. Astrophys. **354**, 99 (2000) 498
62. Rabl, G., Dvorak, R.: Astron. Astrophys. **191**, 385 (1988) 490, 491
63. Raghavan, D., Henry, T.J., Mason, B.D., Subasavage, J.P., Jao, W.C., Beaulieu, T.D., Hambly, N.C.: ApJ **646**, 523 (2006) 492
64. Raymond, S.N., Quinn, T., Lunine, J.I.: Icarus. **183**, 265 (2006) 506
65. Robutel, P., Laskar, J.: Icarus. **152**, 470 (2001) 488
66. Sándor, Zs., Érdi, B., Efthymiopoulos, C.: CMDA. **78**, 113 (2000) 484
67. Sándor, Z., Érdi, B., Széll, A., Funk, B.: CMDA. **90**, 127 (2004) 484
68. Sándor, Z., Suli, A., Erdi, E., Pilat-Lohinger, E., Dvorak, R.: MNRAS. **375**, 1495 (2007) 483, 486, 503
69. Santos, N.C., Mayor, M., Naef, D., Pepe, F., Queloz, D., Udry, S., Burnet, M., Clausen, J.V., Helt, B.E., Olsen, E.H., Pritchard, J.D.: Astron. Astrophys. **392**, 215 (2002) 499
70. Williams, D.M., Pollard, D.: Int. J. Astrobiol. **1**, 61 (2002) 508
71. Wolszczan, A., Frail, D.A.: Nature. **355**, 145 (1992) 482
72. Wuchterl, G., Guillot, T., Lissauer, J.J.: In: Mannings, V., Boss, A.P., Russell, S.S. (eds.) Protostars and Planets IV, p. 1081. University of Arizona Press, Tucson (2000) 506
73. Zucker, S., Mazeh, T., Santos, N.C., Udry, S., Mayor, M.: Astron. Astrophys. **404**, 775 (2003) 500
74. Zucker, S., Mazeh, T., Santos, N.C., Udry, S., Mayor, M.: Astron. Astrophys. **426**, 695 (2004) 501

Index

Souchay, J., Dvorak, R. (eds.): *Index*. Lect. Notes Phys. **790**, 511–517 (2010)
DOI 10.1007/978-3-642-04458-8