Springer INdAM Series 34

Giulio Baù
Alessandra Celletti
Cătălin Bogdan Galeș
Giovanni Federico Gronchi *Editors*

Satellite Dynamics and Space Missions

Springer

Springer INdAM Series

Volume 34

Springer INdAM Series

This series will publish textbooks, multi-authors books, thesis and monographs in English language resulting from workshops, conferences, courses, schools, seminars, doctoral thesis, and research activities carried out at INDAM - Istituto Nazionale di Alta Matematica, http://www.altamatematica.it/en. The books in the series will discuss recent results and analyze new trends in mathematics and its applications.
THE SERIES IS INDEXED IN SCOPUS

More information about this series at http://www.springer.com/series/10283

Giulio Baù • Alessandra Celletti •
Cătălin Bogdan Galeş • Giovanni Federico Gronchi
Editors

Satellite Dynamics and Space Missions

Springer

Editors

Giulio Baù
Department of Mathematics
University of Pisa
Pisa, Italy

Alessandra Celletti
Department of Mathematics
University of Rome Tor Vergata
Roma, Italy

Cătălin Bogdan Galeş
Faculty of Mathematics
Alexandru Ioan Cuza University of Iasi
Iasi, Romania

Giovanni Federico Gronchi
Department of Mathematics
University of Pisa
Pisa, Italy

ISSN 2281-518X ISSN 2281-5198 (electronic)
Springer INdAM Series
ISBN 978-3-030-20632-1 ISBN 978-3-030-20633-8 (eBook)
https://doi.org/10.1007/978-3-030-20633-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

There is a long-lasting tradition of Celestial Mechanics training started with the school in Cortina D'Ampezzo (Italy) in 1981 and continued with other training events in more recent years. The contributions have usually been collected and published in books, thus providing researchers in this field with very useful reference texts.

This book is a collection of contributions given by internationally renowned scientists at the summer school SDSM 2017 "Satellite Dynamics and Space Missions: Theory and Applications of Celestial Mechanics", held in San Martino al Cimino, Viterbo (Italy) from August 28 to September 2, 2017. This school aimed to teach the latest theories, tools and methods for satellite dynamics and space missions. The contributions in this volume deal with a variety of important topics related to satellite dynamics and space mission design. A detailed description of the book contents is summarized below.

The contribution by Sylvio Ferraz-Mello concerns planetary tidal theories for a model with a homogeneous primary body rotating around a spin axis perpendicular to the orbital plane of the companion. It is assumed that the tidally deformed body has an ellipsoidal shape with a rotation delayed with respect to the motion of the companion. Different theories are presented: static tide, dynamic tide, the tidal evolution of the primary's rotation and of the orbital elements, Darwin's theory, constant time and phase lag models, and Mignard's theory.

In the contribution by Antonio Giorgilli the modern tools of Hamiltonian perturbation theory are reviewed. After a brief historical introduction, the problem of studying the dynamics of a small perturbation of an integrable Hamiltonian system (called the *general problem of dynamics* by Poincaré) is presented. This problem arises in a natural way by investigating the stability of the solar system. A short description of the properties of integrable Hamiltonian systems is given, with the Liouville-Arnold-Jost theorem, where the action-angle coordinates are used to describe portions of the phase space foliated in invariant tori. After that, Kolmogorov's contribution is explained, which gives a positive answer to the question of whether or not some of these tori survive small perturbations of the integrable system. This result gives rise to the so called KAM theory, after

Kolmogorov, Arnold, and Moser. The normal forms by Birkhoff and Poincaré are explained, to describe the dynamics in a neighborhood of a Kolmogorov invariant torus or of an elliptical equilibrium. The difficulties caused by the presence of the small divisors in the homological equation are explained in detail. Finally, Nekhoroshev's theorem is presented, addressing the stability problem for a very long time, whose proof requires a subtle geometric analysis of the resonances.

The contribution by Anne Lemaître presents basic techniques to study the dynamics of space debris and introduces a characterization of various dynamical phenomena revealed by applying modern tools of Celestial Mechanics. It discusses dynamical properties of these objects at several levels: gravitational resonances for MEO (medium Earth orbits) and GEO (geosynchronous Earth orbits), lunisolar resonances and secondary resonances involving the Sun, the stability of some regions by computing the chaos indicators (MEGNO, the frequency map), the effects of the solar radiation pressure with and without shadowing, the orbital decay of satellites in LEO (low Earth orbits) due to the atmospheric drag, and the Yarkovsky-Schachs effect. Several ideas describing the development of a synthetic population of virtual space debris are also presented.

The contribution by Josep-Maria Mondelo is about the computation of fixed points, periodic orbits, invariant tori of conservative dynamical systems, and their associated invariant manifolds. To this end, both numerical and semi-analytical methods are presented, discussing the advantages and the main differences between them. These techniques are meant to be applied to preliminary mission design of libration point missions. One of the goals is to select the orbit that best fulfills the requirements of a space mission; for this reason, it is important to compute families of trajectories and invariant manifolds, up to a certain precision. The author considers the restricted three-body problem (with primaries the Sun and a planet, or a planet and a natural satellite) as a model for numerical tests of the theory. Although these techniques are applied to this specific problem, they can be generalized to conservative dynamical systems, and many of them even to dissipative ones.

The contribution by Daniel Scheeres deals with the dynamics of a system of N spherical bodies that are resting on and orbiting about each other. This study is particularly relevant for understanding the motion of *rubble-pile* asteroids, which are composed of small pieces of rock attracted by their mutual gravity. The equations of motion for the N bodies are written with the Lagrangian formalism and include the non-holonomic constraints which arise when the bodies come into contact, always assuming a no-slip condition. Using the conservation of the angular momentum and Routh's reduction, the motion is referred to a suitable rotating frame. This step allows us to introduce the amended potential, which is shown to play an essential role in the determination of the relative equilibria of the system and in the discussion of their stability. The theory is applied to the case of collinear bodies resting on each other (*Euler resting configuration*). Moreover, the influence of the number of bodies and their dimension on the stability of the system is analyzed.

The two contributions by Massimiliano Vasile are about multi-objective optimal control and uncertainty quantification. The first contribution deals with optimal control problems where there is more than one single scalar cost function. The problem is transformed into a finite-dimensional non linear programming (NLP) problem. Pareto optimality is applied so that a solution is considered optimal if none of the cost functions can be improved without making worse the value of some of the others. A scalarization technique is used to transform the multi-objective problem into a single-objective one. Then, the problem is solved by a numerical integration scheme for ordinary differential equations (finite elements in time) and a memetic algorithm. A test case is also shown, with Goddard's rocket example, which has an analytical solution. The techniques presented in the second contribution apply to a wide range of practical problems. Specific examples of orbital mechanics, from orbit determination to collision avoidance, are included. Classification of uncertainty and quantification methods are addressed. Sampling-based methods, which are non-intrusive techniques (e.g., Monte Carlo method, Chebyshev polynomial expansions) are reviewed along with intrusive ones (e.g., description of state transition matrix, polynomial chaos expansion, interval arithmetic). Methods for capturing model uncertainty are also presented and a short description of evidence-based quantification is given.

The school was attended by about 90 participants from all over the world and it was made possible, thanks to the support of the Departments of Mathematics of the University of Pisa and the University of Rome Tor Vergata, the ERC project COMPAT, the ERC project StableChaoticPlanetM, the European Space Agency, Gruppo Nazionale per la Fisica Matematica-INdAM, the Italian Space Agency, and Space Dynamics Services S.r.l. The School SDSM 2017 was held under the patronage of the IAU Commissions A4—Celestial Mechanics and Dynamical Astronomy and X2—Solar System Ephemerides and promoted by the Italian Society of Celestial Mechanics and Astrodynamics—SIMCA.

Pisa, Italy Giulio Baù
Rome, Italy Alessandra Celletti
Iasi, Romania Cătălin Bogdan Galeş
Pisa, Italy Giovanni Federico Gronchi
April 2019

Contents

About the Editors

Giulio Baù is a Researcher at the University of Pisa's Department of Mathematics. His research focuses on orbit propagation and determination methods for small celestial bodies, regularizations techniques in the N-body problem, dynamics of asteroids and space debris.

Alessandra Celletti received her PhD from ETH in Zurich in 1989 and she is currently a Full Professor of Mathematical Physics at the University of Rome Tor Vergata. Her research focuses on celestial mechanics and dynamical systems, especially KAM theory and stability problems.

Cătălin Bogdan Galeş is an Associate Professor at the Faculty of Mathematics, Al.I. Cuza University of Iaşi. His research interests include celestial mechanics, perturbation theories, and mechanics of deformable solids.

Giovanni Federico Gronchi is a Full Professor of Mathematical Physics at the University of Pisa. His research is on solar system body dynamics, perturbation theory, orbit determination, singularities, and periodic orbits of the N-body problem.

Planetary Tides: Theories

Sylvio Ferraz-Mello

Abstract Synthetic presentation of planetary tide theories in the simple case of a homogeneous primary rotating around an axis orthogonal to the orbital plane of the companion. The considered theories are founded on the dynamical equilibrium figure of the tidally deformed body, assumed as an ellipsoid whose rotation is delayed with respect to the motion of the companion. The orbital and rotational evolutions of the system are derived using standard physical laws. The main theory considered is the creep tide theory, a first-principles hydrodynamical theory where the dynamical tide is assimilated to a low-Reynolds-number flow and determined using a Newtonian creep law. The Darwin theories are also considered and are formally derived from the creep tide theory. The various rheologies used in Darwin theories are discussed, with emphasis on the CTL (constant time lag) and CPL (constant phase lag) theories. One introductory session is devoted to the main classical results on the hydrostatic figures of equilibrium of the celestial bodies (static tide).

Keywords Static tide · Dynamic tide · Creep theory · Darwin theory

1 Introduction

In these lectures, we consider the tidal evolution of a system of two homogeneous bodies close one to the other, the *primary* and its *companion*. The primary has mass m, mean radius R, rotational angular velocity $\Omega = |\mathbf{\Omega}|$ and its companion has mass M and is the source of the gravitational force that is tidally deforming the primary.

S. Ferraz-Mello (✉)
Instituto de Astronomia Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, Brazil
e-mail: sylvio@usp.br

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_1

Their instantaneous relative motion is a Keplerian orbit in the equatorial plane of the primary and the distance between the two bodies is $r(t)$. No hypotheses are done about the relative value of their masses. We consider the tide raised in the primary by the gravitational attraction of the companion. To know the consequences of the tide raised in the companion by the gravitational attraction of the primary, it is enough to invert the role played by the two bodies. In general, it is necessary to consider both cases and add their contributions to obtain the total variations of the orbital elements and energy of the system.

Since the pioneer work of Darwin, tides have been treated following many different approaches. In the hydrodynamical approaches considered here (the creep tide and the Maxwell model), the starting point is the hydrostatic equilibrium figure of the tidally deformed primary, the *static* tide. The actual deformation of the body due to the continuously changing gravitational attraction of the companion (the *dynamic* tide) is calculated using one approximate solution of the Navier-Stokes equation simplified by the assumption that the flow is laminar (low-Reynolds-number flow). In Darwinian models, the deformation (tide) is classically divided into two components studied separately. The main component is the *elastic* (or static) tide. The other, responsible for the dissipation of the energy of the system and for the torques acting on the primary body is the *anelastic* tide [24].[1] The approaches introducing the viscoelastic effect by means of a dissipation function (e.g. [17, 46]) were not considered in these lectures.

2 The Static Tide

The static tide is the deformation of the primary in the limit case where it does not offer any resistance to the deformation. The body behaves as a perfect fluid and takes instantaneously the shape corresponding to the equilibrium of the forces acting on it. The total force acting on the points on the surface of the body must be, in each point, perpendicular to the surface. These forces are the gravitational forces of the primary, the tidal forces due to the attraction of the primary by the companion, and, if the body is rotating, the inertial (centrifugal) forces due to the rotation, i.e.

$$\mathbf{F}_{\text{tot}} = -\nabla U_{\text{self}} + \mathbf{F}_{\text{tid}} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{d}) \tag{1}$$

(per unit mass).[2] $\mathbf{d}$ is the position vector of the considered surface point, U_{self} is the potential of the gravitational forces of the primary, $\mathbf{F}_{\text{tid}}$ is the tidal force per unit mass acting on the points of the primary.

[1] For the exact definition of the words *elastic* and *anelastic*, see the online supplement to [14]. One must keep in mind, however, that the involved restoring forces are gravitational, not elastic.

[2] The minus sign in the first term means that we are adopting the exact Physics convention: force is equal to minus the gradient of the potential.

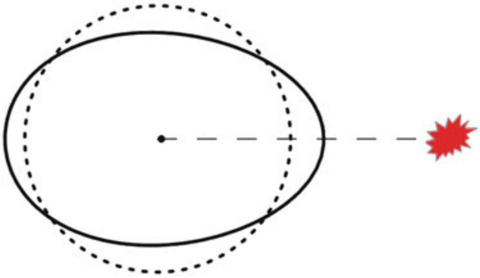


Fig. 1 Tidal deformation of the primary body under the attraction of a close companion

Figure 1 stresses the asymmetry of the tidal deformation. However, with a very few exceptions (e.g. [52]), theories of tidal evolution neglect all terms with $n > 2$; this is equivalent to consider that the shape of the deformed body is an ellipsoid. This ellipsoid is characterized by the relationships between its three axes: $a > b > c$ (c is directed along the rotation axis and a is directed towards the companion). They are the *equatorial prolateness*

$$\epsilon_\rho = \frac{a-b}{R_e}, \tag{2}$$

where $R_e = \sqrt{ab}$ is the mean equatorial radius of the body, and the *polar oblateness*

$$\epsilon_z = 1 - \frac{c}{R_e}. \tag{3}$$

The normal to the ellipsoidal surface in each point is given by $\nabla\mathcal{S}$ where $\mathcal{S}(\widehat{x}, \widehat{y}, \widehat{z}) = 0$ is the equation of the ellipsoid:

$$\mathcal{S} = \frac{\widehat{x}^2}{a^2} + \frac{\widehat{y}^2}{b^2} + \frac{\widehat{z}^2}{c^2} - 1 = 0, \tag{4}$$

and the condition that the force $\mathbf{F}_{\rm tot}$ is perpendicular to the ellipsoidal surface is expressed by the proportionality of the components of $\mathbf{F}_{\rm tot}$ and $\nabla\mathcal{S}$, that is,

$$\frac{\mathbf{F}_{\rm tot} \cdot \mathbf{i}}{\frac{\partial \mathcal{S}}{\partial \widehat{x}}} = \frac{\mathbf{F}_{\rm tot} \cdot \mathbf{j}}{\frac{\partial \mathcal{S}}{\partial \widehat{y}}} = \frac{\mathbf{F}_{\rm tot} \cdot \mathbf{k}}{\frac{\partial \mathcal{S}}{\partial \widehat{z}}}. \tag{5}$$

These proportionality relations give rise to two independent equations that may be solved to obtain the values of ϵ_ρ and ϵ_z.

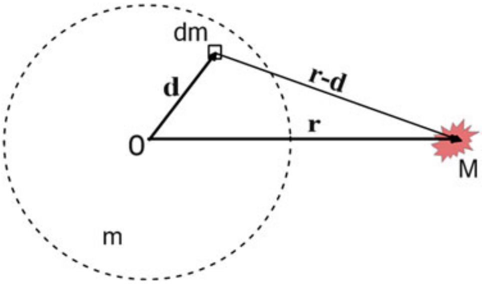


Fig. 2 The vectors **r** and **d**. The tidal force acting on dm is the difference between the gravitational attraction of dm by the companion and the resultant of the gravitational attraction of the whole primary by the companion. O is the center of gravity of the primary

The tidal forces are the forces acting in the interior of the body due to the gravitational attraction of the companion, referred to the resultant of the attraction forces acting on the whole primary, that is, the difference

$$\mathbf{F}_{\text{tid}} = -GM\nabla_{\mathbf{r}} \left(\frac{1}{\mid \mathbf{r} - \mathbf{d} \mid} - \frac{1}{m} \int_m \frac{dm}{\mid \mathbf{r} - \mathbf{d} \mid} \right), \tag{6}$$

where G is the gravitational constant (see Fig. 2). To be restricted to the ellipsoidal contribution, the parenthesis in the above expression may be reduced to its so-called quadrupole component. Hence,

$$\mathbf{F}_{\text{tid}} = -GM\nabla_{\mathbf{r}} \left(\frac{\mathbf{r} \cdot \mathbf{d}}{r^3} \right); \tag{7}$$

In the same order of approximation, the potential of the ellipsoid at the points of the surface (see [7, Chap. 3]; [42, Sec. 79]) is

$$U_{\text{self}} = \frac{3Gm}{4abc} \int_0^\infty \frac{dt}{\Delta} \left(\frac{\alpha \widehat{x}^2}{1+\alpha t} + \frac{\beta \widehat{y}^2}{1+\beta t} + \frac{\widehat{z}^2}{(1+t)} - c^2 \right), \tag{8}$$

where $\alpha = c^2/a^2$, $\beta = c^2/b^2$ and

$$\Delta = \sqrt{(1+\alpha t)(1+\beta t)(1+t)}.$$

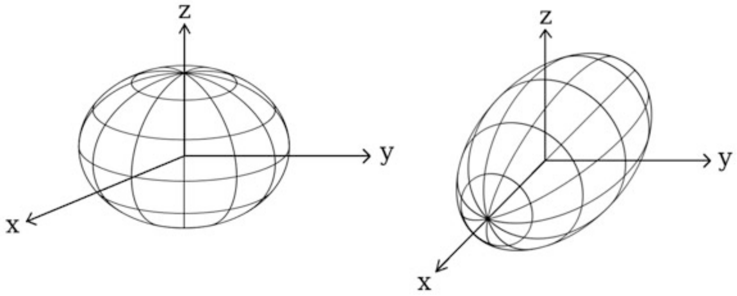


Fig. 3 (*Left*) Maclaurin oblate spheroid. (*Right*) Jeans prolate spheroid (reprinted from [29] with permission)

Two cases are of special importance in the study of the static tide, the spheroids of Maclaurin and Jeans (Fig. 3). They correspond to a body deformed either by the rotation or by the tide acting alone. If the body is not rotating, the ellipsoid becomes an ellipsoid of revolution (i.e. a spheroid) with a as axis of revolution. In this case, the name Jeans ellipsoid is often used [7, 53].

2.1 The Maclaurin Spheroid

This case corresponds to one isolated rotating body, not subjected to a tidal force. The rotation axis is a symmetry axis ($a = b$) and we have just one flattening to calculate. If we assume that the resulting polar oblateness is small, we may use approximations allowing to compute analytically the integrals in Eq. (8) to obtain

$$\epsilon_{\mathrm{M}} = \frac{5R^3\Omega^2}{4mG}. \tag{9}$$

2.2 The Jeans Spheroid

This case corresponds to one non-rotating body submitted to the gravitational attraction of one companion. The axis directed towards the companion (x) is a symmetry axis ($b = c$). As before, we have just one flattening to calculate and if we assume that the resulting equatorial prolateness is small, we may compute analytically the integrals in Eq. (8) to obtain the first-order approximation

$$\epsilon_{\mathrm{J}} = \frac{15}{4}\left(\frac{M}{m}\right)\left(\frac{R_e}{r}\right)^3 \tag{10}$$

(see Table 1).

Table 1 Jeans equatorial prolateness of some celestial bodies due to close companions

Primary	Companion	ϵ_J	$a-b$
Earth	Moon	2.1×10^{-7}	1.34 m
Earth	Sun	9.6×10^{-8}	0.6 m
Venus	Sun	2.6×10^{-7}	1.5 m
Jupiter	Sun	3.0×10^{-9}	0.2 m
Jupiter	Io	8.5×10^{-7}	61 m
Moon	Earth	2.8×10^{-3}	50 m
Io	Jupiter	4.9×10^{-3}	8.2 km
Titan	Saturn	1.5×10^{-4}	0.38 km
Planet CoRoT 7b	Star CoRoT 7	8×10^{-3}	85 km

2.3 *The General Ellipsoid*

If we must consider both, rotation and tide, the result is a composite of the two above ones. The equatorial prolateness is the same of the Jeans spheroid

$$\epsilon_\rho = \epsilon_J, \tag{11}$$

but the polar oblateness is a composite of the two spheroidal flattenings:

$$\epsilon_z = \epsilon_M + \frac{1}{2}\epsilon_J. \tag{12}$$

The main term in ϵ_z is the oblateness of a Maclaurin spheroid. However, the polar oblateness is also affected by the tidal deformation. Indeed, if the body is stretched along the axis a, the conservation of volume forces it to shrink in the directions orthogonal to that axis, thus decreasing both b and c and increasing the polar oblateness of the body.

2.4 *Roche Ellipsoids*

The general ellipsoids are sometimes called Roche ellipsoids, however, this designation more strictly refers to the case in which the motion of the companion around the primary is circular and synchronous with the rotation of the primary ($\varOmega = n$). In such case, the third Kepler law may be used to obtain $\varOmega^2 = G(M+m)/a^3$, and then

$$\epsilon_{\rm M} = \frac{5R^3(M+m)}{4ma^3}. \tag{13}$$

If the primary is much smaller than the companion (one satellite of a big planet, or one hot planet orbiting a normal star), we may assume $m << M$ and so, $\epsilon_J \simeq$

$3\epsilon_M$ (see [53, t.2, Chap. 8]). In this case, several interesting relations may be easily derived, e.g.

$$\frac{a-c}{b-c} \simeq 4. \tag{14}$$

When the primary is one Roche ellipsoid, an important relation may be found between the quadrupole coefficients of the primary's potential. We may remember the general expressions[3]:

$$\begin{aligned} J_2 \equiv -C_{20} &= -\frac{1}{2mR_e^2}(A+B-2C) = \frac{1}{10R_e^2}(a^2+b^2-2c^2) \simeq \frac{2}{5}\epsilon_z \\ C_{22} &= \frac{1}{4mR_e^2}(B-A) = \frac{1}{20R_e^2}(a^2-b^2) \simeq \frac{1}{10}\epsilon_\rho \end{aligned}$$

(see [3, v.1, Sec. 3.4]) where A, B, C are the moments of inertia along the principal axis of the ellipsoid. Hence, if $\epsilon_J \simeq 3\epsilon_M$, we have

$$\frac{C_{22}}{J_2} \simeq \frac{3}{10}. \tag{15}$$

3 The Tide Harmonics

The equation of the surface of the equilibrium ellipsoid, in the general case, is

$$\rho = R_e\big(1 + \frac{1}{2}\epsilon_\rho \sin^2\widehat{\theta}\cos(2\widehat{\varphi} - 2\omega - 2v) - \epsilon_z \cos^2\widehat{\theta}\big), \tag{16}$$

where $\rho, \widehat{\varphi}, \widehat{\theta}$ are the radius-vector, longitude and co-latitude of one generic surface point, and $\omega + v$ is the true longitude of the companion in its equatorial orbit around the primary (ω is the argument of the pericenter and v is the true anomaly). These angles are reckoned from a fixed virtual node N and are such that the major axis is always oriented towards the companion, i.e. to the surface equatorial point whose longitude is $\widehat{\varphi} = \omega + v$. The dependent variables of this equation are the longitude of the generic point $\widehat{\varphi} = \varOmega(t - t_0)$, the radius vector r and the true anomaly v of

[3] Auxiliary first-order relations:

$$\begin{aligned} a &= R_e(1+\epsilon_\rho/2) \\ b &= R_e(1-\epsilon_\rho/2) \\ c &= R_e(1-\epsilon_z). \end{aligned}$$

Table 2 Some low-order Cayley expansions

$E_{2,-2} = \frac{17}{2}e^2 - \frac{115}{6}e^4 + \frac{601}{48}e^6$
$E_{2,-1} = \frac{7}{2}e - \frac{123}{16}e^3 + \frac{489}{128}e^5 - \frac{1763}{2048}e^7$
$E_{2,0} = 1 - \frac{5}{2}e^2 + \frac{13}{16}e^4 - \frac{35}{288}e^6$
$E_{2,1} = -\frac{1}{2}e + \frac{1}{16}e^3 - \frac{5}{384}e^5 - \frac{143}{18432}e^7$
$E_{2,2} = 0$
$E_{0,0} = 1 + \frac{3}{2}e^2 + \frac{15}{8}e^4 + \frac{35}{16}e^6$
$E_{0,1} = \frac{3}{2}e + \frac{27}{16}e^3 + \frac{261}{128}e^5 + \frac{14309}{6144}e^7$
$E_{0,2} = \frac{9}{4}e^2 + \frac{7}{4}e^4 + \frac{141}{64}e^6$

N.B. $E_{0,-k} = E_{0,k}$

the companion. The true anomaly v appears explicitly in the equation. The radius vector r is included in the equatorial prolateness ϵ_ρ, and in the polar oblateness $\epsilon_z = \epsilon_M + \frac{1}{2}\epsilon_\rho$, through the definition of ϵ_J. The equatorial radius is also variable since it is related to the constant mean radius R of the body through $R \simeq R_e(1 - \frac{1}{3}\epsilon_z)$.

We may expand Eq. (16) assuming that the functions $r(t)$, $v(t)$ are given by the two-body (Keplerian) approximation. The resulting equation is

$$\rho = R\Bigg(1 + \frac{1}{2}\overline{\epsilon}_\rho \sin^2\widehat{\theta} \sum_{k\in\mathbb{Z}} E_{2,k} \cos\big(2\widehat{\varphi} + (k-2)\ell - 2\omega\big)$$

$$-(\cos^2\widehat{\theta} - \frac{1}{3})\Big(\overline{\epsilon}_z + \frac{1}{2}\overline{\epsilon}_\rho \sum_{k\in\mathbb{Z}} E_{0,k} \cos k\ell\Big)\Bigg), \tag{17}$$

where ℓ is the mean anomaly, $E_{q,p}$ are the Cayley functions[4] ([6]; see [20] online supplement; see Table 2):

$$E_{q,p}(e) = \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{a}{r}\right)^3 \cos\big(qv + (p-q)\ell\big)\, d\ell, \tag{18}$$

$$\overline{\epsilon}_\rho = \frac{15}{4}\left(\frac{M}{m}\right)\left(\frac{R_e}{a}\right)^3, \tag{19}$$

and

$$\overline{\epsilon}_z = \epsilon_z - \frac{1}{2}\epsilon_\rho. \tag{20}$$

[4] The Cayley functions introduced here correspond to the degree 3 in a/r—since $\epsilon_\rho \propto (a/r)^3$. These functions are equivalent to the Hansen coefficients preferred by other authors and the equivalence is given by $E^{(n)}_{q,p} = X^{-n,q}_{2-p}$ (see [8]).

To interpret the harmonic components of the static tide, we may consider one point fixed to the surface of the body and determine the time variation of the component amplitude at that point. Each term of ρ depending on $\widehat{\varphi}$ corresponds to a tidal harmonic traveling on the body with given direction and velocity. All harmonics contribute to the formation and evolution of the tidal bulge on the body. In the orthogonal model studied in these lectures, the terms of ρ appears in two groups:

1. *Sectorial* components having the argument $\big(2\widehat{\varphi}+(k-2)\ell-2\omega\big)$. The amplitudes of these terms are maximum at the equator ($\widehat{\theta}=\pi/2$) and decrease towards the poles. On the equator, they are maximum when $\widehat{\varphi}=-(k/2-1)\ell+\omega$. The main term ($k=0$) is an oscillation with period $\pi/(\Omega-n)$ (i.e. half the synodic rotation period), with two maxima located one on the sub-M point and the other on its antipodal. If $n\ll\Omega$, the period is nearly half of the rotation period.

   The next harmonic, ($k=1$) has two opposed maxima. One of them lies on the sub-M point when the tide generating body is at the periapsis (i.e., $\ell=0$) and the other when the tide generating body is at the apoapsis (i.e. $\ell=\pi$). The high tide moves, in this case, more slowly than the sub-M point. The harmonic $k=-1$ has a similar behavior, but the point of maximum amplitude on the equator propagates backward. Similar analyses can be done for the other harmonics.

   When $\Omega\gg n$ these terms have frequencies close to the semi-diurnal frequency

$$\nu=2\Omega-2n. \tag{21}$$

   They are called *semi-diurnal tides*. On the Earth, they have periods close to 12 h. They are shown in Table 3, which also summarizes the interpretation to be given in the other cases ($\Omega\ll n$ and $\Omega\simeq n$).

Table 3 The main tide harmonics

		Type 1	Type 2	Type 3
k	Frequency	$\Omega\gg n$	$\Omega\simeq n$	$\Omega\ll n$
Sectorial terms				
0	$2\Omega-2n$	Semi-diurnal	–	Semi-annual
-1	$2\Omega-3n$	Semi-diurnal	Monthly	3rd of annual
+1	$2\Omega-n$	Semi-diurnal	Monthly	Annual
-2	$2\Omega-4n$	Semi-diurnal	Semi-monthly	4th of annual
+2	2Ω	Semi-diurnal	Semi-monthly	“Semi-diurnal”
Radial terms				
1	n	Monthly	Monthly	Annual
2	$2n$	Semi-monthly	Semi-monthly	Semi-annual

The frequencies and corresponding names refer to how the tidal disturbance is felt on a given (fixed) point of the body. For the type 2 tides, the paradigm is the Moon, but when the synchronous companion is an exoplanet, the names annual and semi-annual would be more appropriate

2. *Zonal* components independent of the longitude. These harmonics do not depend on the longitude of the considered points. ρ oscillates all over the body with amplitudes depending on the latitude of the points and on the mean longitude of the tide generating body. These terms are often called *radial tides* because there is no propagation of a crest on the surface of the body.

In Table 3, we summarize the names of the tidal harmonics for three different cases depending on the rotation speed of the primary. Type 1 corresponds to a body rotating with angular velocity much larger than the orbital mean motion ($\Omega \gg n$). It is the case of the Earth-Moon system, with the Earth as primary and the Moon as companion. Type 2 corresponds to *synchronous* or *almost synchronous* motions and, again, the Earth-Moon system serves as an example, but now the Moon is the primary and the Earth is the companion. Looking at Table 3, we see that synchronization gives rise to harmonics whose periods are related to the rotation period of the companion, and they are called monthly, semi-monthly, etc. The names come from the tidal action of the Earth on the Moon (for this reason, the semi-monthly tide is often called *fortnightly*). In the case of tidally locked exoplanets, the main period is the orbital period of the planet and the names annual, semi-annual, etc. are more appropriated. Type 3 corresponds to a slow rotating body ($\Omega \ll n$); it is the case of the tides on a typical main sequence star due to a close-in planet (hot Jupiter). Using names like those used in the other cases and considering that the main period is the planet's orbital period (or "year"), we will call them, respectively, annual, semi-annual, tierce-annual and so on.[5]

It is worth emphasizing that the given tidal frequencies and corresponding names refer to how the tidal disturbance is felt on a given (fixed) point of the body. The propagation of the tidal harmonic in the body must be analyzed separately. For instance, on the Earth, the tidal bulges of both diurnal and semi-diurnal tides circulate around the Earth with the synodic rotation speed. The names and frequencies given in Table 3 refer rather to the shape of the tidal harmonic.

4 Tidal Evolution Due to the Static Tide

The deformation of the body will modify the gravitational potential in its neighborhood. The simplest form of this potential is obtained when we use a reference system whose axes are the principal axes of inertia of the ellipsoid. It is

$$U = -\frac{Gm}{r} - \frac{G}{2r^3}(A + B + C) + \frac{3G}{2r^5}(AX^2 + BY^2 + CZ^2) + \cdots, \qquad (22)$$

[5] We have, however, to keep in mind that these "years" are very short. In type 3 tides, "diurnal" is slower than "annual".

where A, B, C are the moments of inertia w.r.t. to the three axes, X, Y, Z are the coordinates of one generic point and $r = \sqrt{X^2 + Y^2 + Z^2}$ (see [3, v.1 Sec. 3.3.5]):

$$\begin{aligned} X &= r \sin\theta \cos(\varphi - \omega - v) \\ Y &= r \sin\theta \sin(\varphi - \omega - v) \\ Z &= r \cos\theta. \end{aligned} \tag{23}$$

where θ, φ are the co-latitude and longitude of the point in a system of reference whose fundamental plane lies in the equator of the body, but whose axes are fixed (i.e., not rotating with the body), and $\omega + v$ is the true longitude of the companion.

The acceleration of the considered point is minus the gradient of U. In a right-handed orthogonal set of unit vectors along the positive direction of the increments of (r, θ, φ), the components of the acceleration are

$$\begin{aligned} a_1 &= -\frac{\partial U}{\partial r} &&= -\frac{Gm}{r^2} - \frac{3G}{2r^4}(A + B + C) + \frac{9G}{2r^6}(AX^2 + BY^2 + CZ^2) \\ a_2 &= -\frac{1}{r}\frac{\partial U}{\partial \theta} &&= -\frac{3G}{r^6}(AX^2 \cot\theta + BY^2 \cot\theta - CZ^2 \tan\theta) \\ a_3 &= -\frac{1}{r\sin\theta}\frac{\partial U}{\partial \varphi} &&= \frac{3G}{r^6 \sin\theta}(A - B)XY. \end{aligned} \tag{24}$$

In particular, the components of the force acting on the companion whose coordinates are $X = r, Y = 0, Z = 0$ ($\theta = \frac{\pi}{2}, \varphi = \varpi + v$), are

$$\begin{aligned} F_1 &= -\frac{GmM}{r^2} - \frac{3GM}{2r^4}(A + B + C) + \frac{9GMA}{2r^4} \\ F_2 &= 0 \\ F_3 &= 0. \end{aligned} \tag{25}$$

The force is radial, its torque is equal to zero and, therefore, the rotation of the body is not affected by the static tide. There is no exchange of angular momentum between the rotation of the body and the orbital motion. The rate of work done by the static tide force (its power) is $\mathbf{F} \cdot \mathbf{v}$ where $\mathbf{v}$ is the velocity of the companion. The sequence $\dot{W} = \mathbf{F} \cdot \mathbf{v} = \mathcal{F}(r)\mathbf{r} \cdot \mathbf{v} = \frac{1}{2}\mathcal{F}(r)d(\mathbf{r}^2)/dt = \frac{1}{2}\mathcal{F}(r)d(r^2)/dt = \mathcal{F}(r)r dr/dt$, where $\mathcal{F} = F_1/r$, shows that the work is an exact differential and, therefore, the total mechanical energy of the system remains constant in one cycle. There is no dissipation of energy due to the static tide.

Another important consequence is that the variation of the eccentricity, which is a function of the variations of both the energy and the angular momentum, also averages to zero in a cycle.

The only effects not averaged to zero are the precessions of the pericenter and of the longitude at the epoch (the third Kepler law needs a correction). To obtain these variations, we may use the classical Lagrange (or Gauss) equations for the variation of these orbital elements. It is worth stressing that the disturbing potential $\mathcal{R}$ appearing in Lagrange equations is the potential of an *external* perturbation acting on the companion. However, in the present case, the disturbing forces acting on M are *internal* to the system of bodies. Therefore, as is usually done in the formulation of an N-body problem, the reactions must also be considered, that is, the disturbing function $\mathcal{R}$ in Lagrange variational equations must be substituted by $-(1+\frac{M}{m})\delta U$ to consider the reaction on the primary of its tidal action on the companion.[6] This correction can be neglected only when $M \ll m$. However, the neglect of the reaction in the general case leads to wrong equations and cannot be done.

For the sake of completeness, we recall that the reaction to the bulk gravitational attraction of m on M is duly considered in the current formulation of Gauss and Lagrange equations and we need to consider here only the reaction to the disturbing force.

5 The Dynamic Tide

Because of the forces acting on it (self-gravitation, tidal potential and centrifugal forces), one extended inviscid body immediately changes its shape to become an equilibrium ellipsoid (static tide). However, a real celestial body does not relax instantaneously to the equilibrium. It will offer some resistance to the change and will relax slowly towards the equilibrium. But since the relative position of the primary and the companion is continuously changing, the equilibrium is also changing, and the actual shape of the body will be continuously trying to adjust itself to it. To describe mathematically this process, we introduce two functions: $\zeta = \zeta(\widehat{\theta}, \widehat{\varphi}, t)$ and $\rho = \rho(\widehat{\theta}, \widehat{\varphi}, t)$, where the angles $\widehat{\theta}, \widehat{\varphi}$ are the co-latitude and longitude of one direction in a fixed reference system and ζ and ρ are the radii vectors of the corresponding points on the actual surface of the body and on the surface of the instantaneous equilibrium ellipsoid.

The creep tide theory of Ferraz-Mello [19] assumes that, at each instant, the actual surface tends to the equilibrium ellipsoid surface with a speed $\dot{\zeta}$ proportional to the distance between the two surfaces (see Fig. 4). The equation of the motion is

$$\dot{\zeta} = -\gamma(\zeta - \rho). \tag{26}$$

This is the equation of a *Newtonian creep* (see [44, chap. 5]) where the stress was considered as proportional to the distance to the equilibrium. It does not consider inertia or azimuthal motions and is linear.

This equation was used by Darwin in his first paper on the precession of a viscous Earth [9] to define its rate of adjustment to a new form of equilibrium and was described in a very pictorial way. We may paraphrase his statement by changing some words and symbols to make them correspond to the words and symbols used here:

> But because of the [primary's] viscosity, [ζ] always tends to approach [ρ]. The stresses introduced in the [primary] by the want of coincidence of [ζ] with [ρ] vary as [$\rho - \zeta$]. [...] Hence the linear velocity (on the map), with which [ζ] approaches [ρ], varies as

[6] Remember that in the considered case, the equations of the relative motion are $M\ddot{\mathbf{r}} = (1+\frac{M}{m})\mathbf{F}$.

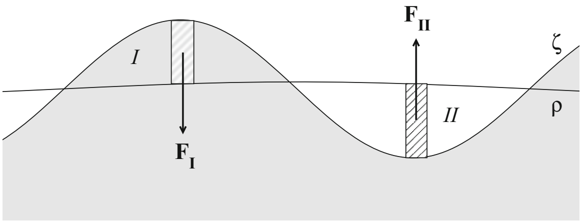


Fig. 4 The creep model: ζ is the actual surface of the body at the time t and ρ is the surface of the static tide or equilibrium ellipsoid at the same time. Adapted from [27] with permission

> [$\rho - \zeta$]. Let this velocity be [$\gamma(\rho - \zeta)$], where [γ] depends on the viscosity of the [primary], decreasing as the viscosity increases.

The relaxation factor γ is a radial deformation rate gradient and has dimension T^{-1}. $\gamma \rightarrow 0$ in the rigid body limit and $\gamma \rightarrow \infty$ in the inviscid fluid limit. Between these two extremes, we have the real bodies, which, under stress, relax towards the equilibrium, but not instantaneously.

5.1 The Navier-Stokes Equation

It is possible to show that Eq. (26) is an approximated solution of the Navier-Stokes equation of a radial flow across the two surfaces, with a very low Reynolds number (Stokes flow). In this case, the inertia terms can be neglected (see [31]) and the Navier-Stokes equation becomes simply [51]

$$\nabla p = \eta \Delta \mathbf{V}, \tag{27}$$

where p is the pressure, η is the uniform viscosity and $\mathbf{V}$ is the velocity. The additional external force (per volume unit) is omitted since we are studying the immediate neighborhood of the equilibrium surface and the stress in that neighborhood is already considered in the pressure term.

We notice that the symbol Δ is operating on a vector, contrarily to its usual definition. Actually, this pseudo-vectorial notation can be converted to a legitimate vector formula through [51]

$$\Delta \mathbf{V} = \frac{1}{2}\nabla(\mathbf{V}^2) - \mathbf{V} \times \nabla \times \mathbf{V}.$$

If the radial flow is assumed independent of the azimuthal variables, the vector Laplacian becomes

$$\Delta \mathbf{V} = \Delta V_r - \frac{2V_r}{\zeta^2}. \tag{28}$$

In a generic point of radius vector ζ, we then have

$$\frac{\partial^2 V_r}{\partial \zeta^2} + \frac{2}{\zeta}\frac{\partial V_r}{\partial \zeta} - \frac{2V_r}{\zeta^2} = -\frac{w}{\eta}. \tag{29}$$

where w is the local specific weight (N.B. $w = -\nabla p$). The pressure due to the body gravitation was approximated by the weight of the mass lying above (or is missing below[7]) the equilibrium surface, that is, $-w(\zeta - \rho)$; the modulus of the pressure gradient is the specific weight w. Terms of second order with respect to ϵ_ρ are neglected in this and in the following calculations.

The solution of this differential equation is:

$$V_r(\zeta) = C_1\zeta + \frac{C_2}{\zeta^2} - \frac{w}{4\eta}\zeta^2, \tag{30}$$

where C_1 and C_2 are integration constants. These constants are determined by the boundary conditions

- $V_r(\rho) = 0$ i.e. the velocity vanishes when $\zeta = \rho$; and
- $V_r''(\rho) \equiv 0$ i.e. the approximation is linear.

Hence $C_1 = \frac{w\rho}{6\eta}$, $C_2 = \frac{w\rho^4}{12\eta}$ and, after linearization in the neighborhood of $\zeta = \rho$, we get $V_r(\zeta) = \gamma(\delta\rho - \delta\zeta) = \gamma(\rho - \zeta)$, showing that the basic creep equation adopted in Ferraz-Mello's theory is the linearized solution of an approximate version of the Navier-Stokes equation and that the relaxation factor γ is related to the uniform viscosity of the body through

$$\gamma \simeq \frac{wR}{2\eta} \simeq \frac{3gm}{8\pi R^2\eta}, \tag{31}$$

where g is the gravity at the surface of the body and R is its mean radius (Table 4). Darwin [10] also studied this equation, but using a different construction of the Navier-Stokes equations he obtained the numerical factor 3/38 instead of 3/8. His factor was determined using the spheroidal form of the tidal potential, but the spheroidal parameters do not appear in his results.

5.2 *The Creep Equation*

Equation (26) is a non-homogeneous ordinary differential equation of first order with constant coefficients. The right-hand side is a known time function depending

[7] This does not mean that a negative mass is being assigned to void spaces; it means just that forces included in the calculation of the equilibrium figure need to be subtracted when the masses creating them are no longer there.

Table 4 Typical values of the relaxation factor adopted in applications

Body	γ (s^{-1})	$2\pi/\gamma$	η (Pa s)
Moon	$2.0 \pm 0.3 \times 10^{-9}$	100 year	$2.3 \pm 0.3 \times 10^{18}$
Titan	$2.9 \pm 0.2 \times 10^{-8}$	6.8 year	$1.1 \pm 0.1 \times 10^{17}$
Solid Earth	$0.9 - 3.6 \times 10^{-7}$	200–800 d	$4.5 - 18 \times 10^{17}$
Io	$4.9 \pm 1.0 \times 10^{-7}$	150 d	$1.2 \pm 0.3 \times 10^{16}$
Europa	$1.8–8.0 \times 10^{-7}$	90–400 d	$4–18 \times 10^{15}$
Neptune	2.7–19	<2 s	$1.2–4.8 \times 10^{10}$
Saturn	>7.2	<0.9 s	$< 15 \times 10^{10}$
Jupiter	23 ± 4	~0.3 s	$4.7 \pm 0.9 \times 10^{10}$
Hot Jupiters	8–50	0.1–0.8 s	$5 \times 10^{10}–10^{12}$
Solar-type stars	>30	<0.2 s	$< 2 \times 10^{12}$

See [19, 20]

on the longitude $\widehat{\varphi}$ and on the coordinates of the companion, r and v. The radius vector of the companion, r, is introduced in the equation by the flattenings ϵ_ρ and ϵ_z. If the expression of the static equilibrium ellipsoid ρ is expanded in a Fourier series (cf. Eq. (17)) and introduced into Eq. (26), we obtain

$$\dot{\zeta} + \gamma\zeta = \gamma R + \gamma R \sum_{k\in\mathbb{Z}} \left(\mathcal{C}_k \sin^2\widehat{\theta} \cos\Theta_k + \mathcal{C}''_k (\cos^2\widehat{\theta} - \frac{1}{3}) \cos\Theta''_k \right), \tag{32}$$

where we have introduced the constants:

$$\mathcal{C}_k = \frac{1}{2}\overline{\epsilon}_\rho E_{2,k} \tag{33}$$

$$\mathcal{C}''_k = -\frac{1}{2}\overline{\epsilon}_\rho E_{0,k} - \delta_{0,k}\overline{\epsilon}_z \tag{34}$$

($\delta_{0,k}$ is the Kronecker delta), and the linear time functions

$$\Theta_k = 2\widehat{\varphi} + (k-2)\ell - 2\omega \tag{35}$$

$$\Theta''_k = k\ell. \tag{36}$$

After the integration, we obtain

$$\zeta = Ce^{-\gamma t} + R + \delta\zeta, \tag{37}$$

where $C = C(\widehat{\varphi}, \widehat{\theta})$ is an integration constant. The forced terms arising from the non-homogeneous part of the differential equation are

$$\delta\zeta = R \sum_{k\in\mathbb{Z}} \left(\mathcal{C}_k \sin^2\widehat{\theta} \cos\sigma_k \cos(\Theta_k - \sigma_k) + \mathcal{C}''_k (\cos^2\widehat{\theta} - \frac{1}{3}) \cos\sigma''_k \cos(\Theta''_k - \sigma''_k) \right), \tag{38}$$

where

$$\tan\sigma_k = \frac{\dot{\Theta}_k}{\gamma} \qquad \cos\sigma_k = \frac{\gamma}{\sqrt{\dot{\Theta}_k^2+\gamma^2}} \qquad \sin\sigma_k = \frac{\dot{\Theta}_k}{\sqrt{\dot{\Theta}_k^2+\gamma^2}} \tag{39}$$

$$\tan\sigma_k'' = \frac{\dot{\Theta}_k''}{\gamma} \qquad \cos\sigma_k'' = \frac{\gamma}{\sqrt{\dot{\Theta}_k''^2+\gamma^2}} \qquad \sin\sigma_k'' = \frac{\dot{\Theta}_k''}{\sqrt{\dot{\Theta}_k''^2+\gamma^2}}. \tag{40}$$

The subtracting constant phases σ_k and σ_k'' behave as lags, but they are not ad hoc plugged constants as in Darwinian theories. They are finite (i.e. not small) exact quantities resulting from the integration of the first-order linear differential equation. We note that, in the integration, the orbital elements a, e, the rotational velocity Ω and the pericenter precession $\dot{\omega}$ are taken as constants. In the actual problem, they are variable. However, their resulting variations are of the order $\mathcal{O}(\overline{\epsilon}_\rho, \overline{\epsilon}_z)$ and their contributions can be neglected, at least for limited times. Another warning to introduce concerns the Keplerian approximation adopted. In the case of some planetary satellites, as the Moon, the perturbations of the orbital motion and the precession of the pericenter must be necessarily included in the model.

The transient ($\zeta = Ce^{-\gamma t}$) is generally not considered. It is assumed that the past elapsed time is long enough allowing the transient to be fully damped.

6 Forces and Torques

The body surface is defined by $\zeta = R + \delta\zeta$ and it is simple to compute the force and torque that the primary exerts on the companion M because $\delta\zeta$ is composed by the bulges of a set of quadrics (which may give positive or negative contributions) superposed to one sphere. Since these bulges are very thin (they are proportional to the flattenings), we may calculate the attraction of M by the resulting composite, as the sum of the forces due to each ellipsoid bulge [19]. The errors of this superposition are of second order w.r.t the flattenings.

Alternatively, we may use a more direct approach [20]. We may substitute the bulges by a thin spherical shell of radius R and assume for the mass element at the shell coordinates $(\widehat{\theta}, \widehat{\varphi})$, the sum of the masses of the bulges at that point. The generic mass element in the shell is

$$dm(\widehat{\theta}, \widehat{\varphi}) = R^2 \mu_{\mathsf{m}} \sin\widehat{\theta} d\widehat{\varphi} d\widehat{\theta} \delta\zeta, \tag{41}$$

where μ_m is the density of the body. The contribution of the element dm to the potential in the external point $\mathsf{P}(r, \varphi, \theta)$ is

$$dU = -\frac{Gdm}{\varDelta}, \tag{42}$$

where G is the gravitational constant and $\varDelta$ is the distance from the element dm to the point $\mathsf{P}(r, \varphi, \theta)$; the potential created by the whole shell is given by

$$\delta U = -GR^2\mu_\mathsf{m} \int_0^\pi \sin\widehat{\theta} d\widehat{\theta} \int_0^{2\pi} \frac{\delta\zeta}{\varDelta} d\widehat{\varphi}. \tag{43}$$

The integration is simple and may be easily done either numerically or algebraically to the desired precision. The result is $\delta U = \sum_{k\in\mathbb{Z}}(\delta U_k + \delta U_k'')$, where we have considered separately the contributions of the sectorial and zonal components of $\delta\zeta$ and neglected terms of higher orders in R/r:

$$\delta U_k = -\frac{3GmR^2}{5r^3} \mathcal{C}_k \cos\sigma_k \sin^2\theta \cos(2\varphi - \beta_k), \tag{44}$$

$$\delta U_k'' = -\frac{GmR^2}{5r^3} \mathcal{C}_k'' \cos\sigma_k''(3\cos^2\theta - 1)\cos\beta_k''. \tag{45}$$

The β_k are the linear time functions:

$$\beta_k = (2-k)\ell + 2\omega + \sigma_k \tag{46}$$

$$\beta_k'' = k\ell - \sigma_k''. \tag{47}$$

6.1 Diana

The construction used in the calculation of the forces, since Darwin, is the following: The companion creates a deformation in the primary body and this deformation changes the attraction of the companion by the primary (see Fig. 5). This construction has some consequences. First, the two processes are physically separated. The disturbing potential δU considers the deformation of the primary, but it ignores how the deformation is produced. It just embeds a time variation associated with the relative motion of the companion.

Because of this construction, the coordinates of the companion enter in δU twice. One time as the coordinates of the body producing the deformation of the primary and, again, as the coordinates of the body whose motion is being disturbed by the deformation. To indicate the origin of each set of parameters, it is usual to give different symbols to them—e.g. marking one of them with asterisks. More yet, following Darwin's prose, the body creating the deformation of the primary is

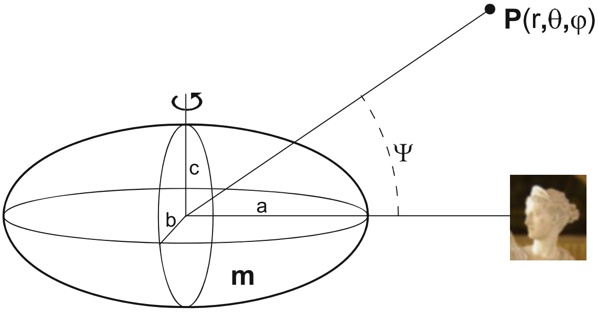


Fig. 5 Darwin's construction. The attraction of *Diana* deforms the primary and the deformed primary attracts the point P. Subsequently P and Diana are identified one with the other

called Diana, the Roman goddess of the Moon, while the other is the Moon (here, the *companion*). The identification of Diana with the companion cannot be done before the forces are calculated, because in the calculation of the gradient of δU, this operation must be done with respect to the coordinates of the generic point P, regardless of the time variation of the potential due to the motion of Diana. We remind that in the calculus of the force arising from the deformation of m via the gradient of δU, only the derivatives w.r.t. the components of $\mathbf{r}$ are considered (see [14, 34]).

When the integration shown in the previous sections is changed to give directly the components of the force, this kind of precaution is superfluous, as no gradient is calculated. The direct calculation of the force is more straightforward and prescinds of Darwin's discussion. The price to pay, however, is that we then must calculate the three components of the force separately, instead of just one potential.

6.2 *Forces and Torques Acting on the Companion*

To obtain the force acting on one mass located at one point, we must take the negative gradient of the potential at that point and multiply the result by the mass placed on the point. In this section, we do not use asterisks[8] to differentiate the parameters related to Diana from the parameters of P because they are well separated: r, θ, φ are the coordinates of P and ℓ, ω are the mean anomaly and the argument of the pericenter of Diana, respectively.

[8] We avoided to overcharge this text with asterisks. We will only use them in Sect. 8 to indicate the mean anomaly and the argument of the pericenter of Diana. If the other Lagrange variational equations are used, other elements of Diana also need to be indicated. In that case, it is convenient to use the asterisk for all orbital parameters of Diana, from the beginning, and drop the asterisks only after all derivatives of the disturbing potential are calculated. *Note added in proof:* The presentation of the creep tide theory becomes much simpler when the new Folonier equations are used. See [28].

In a right-handed orthogonal set of unit vectors along the positive directions of the increments of (r, θ, φ), the components of the force are

$$
\begin{aligned}
F_{1k} &= -\frac{3GMmR^2}{5r^4}\Big(3\mathcal{C}_k \cos\sigma_k \sin^2\theta \cos(2\varphi - \beta_k) + \mathcal{C}''_k \cos\sigma''_k (3\cos^2\theta - 1)\cos\beta''_k\Big)\\
F_{2k} &= \frac{3GMmR^2}{5r^4}\Big(\mathcal{C}_k \cos\sigma_k \sin 2\theta \cos(2\varphi - \beta_k) - \mathcal{C}''_k \cos\sigma''_k \sin 2\theta \cos\beta''_k\Big)\\
F_{3k} &= -\frac{6GMmR^2}{5r^4}\, \mathcal{C}_k \cos\sigma_k \sin\theta \sin(2\varphi - \beta_k);
\end{aligned}
\tag{48}
$$

and the corresponding components of the torque are

$$
\begin{aligned}
M_{1k} &= 0\\
M_{2k} &= \frac{6GMmR^2}{5r^3}\, \mathcal{C}_k \cos\sigma_k \sin\theta \sin(2\varphi - \beta_k)\\
M_{3k} &= \frac{3GMmR^2}{5r^3}\Big(\mathcal{C}_k \cos\sigma_k \sin 2\theta \cos(2\varphi - \beta_k) - \ \mathcal{C}''_k \cos\sigma''_k \sin 2\theta \cos\beta''_k\Big).
\end{aligned}
\tag{49}
$$

6.2.1 Forces and Torques Acting on an Equatorial Companion

The variables θ and φ are the co-latitude and longitude of M. Since M is assumed to lie in the equatorial plane of m, $\theta = \pi/2$ and $\varphi = v + \omega$. Hence

$$
\begin{aligned}
F_{1k} &= -\frac{3GMmR^2}{5r^4}\Big(3\mathcal{C}_k \cos\sigma_k \cos(2v - (2-k)\ell - \sigma_k) - \ \mathcal{C}''_k \cos\sigma''_k \cos(k\ell - \sigma''_k)\Big)\\
F_{2k} &= 0\\
F_{3k} &= -\frac{6GMmR^2}{5r^4}\, \mathcal{C}_k \cos\sigma_k \sin(2v - (2-k)\ell - \sigma_k),
\end{aligned}
\tag{50}
$$

and

$$
\begin{aligned}
M_{1k} &= 0\\
M_{2k} &= \frac{6GMmR^2}{5r^3}\, \mathcal{C}_k \cos\sigma_k \sin(2v - (2-k)\ell - \sigma_k)\\
M_{3k} &= 0.
\end{aligned}
\tag{51}
$$

7 Tidal Evolution: The Primary's Rotation

To study the rotation of the primary, we use the equation $C\dot{\Omega} = M_2$ where C is the moment of inertia with respect to the rotation axis.[9] This equation deserves some comments. First, we note that it neglects the variation of the moment of inertia C. Second, it is obtained after two sign inversions (which cancel themselves). The first of these sign inversions is done because of the adopted frame of reference. The component M_2 is directed downwards (the co-latitude is the polar angle) and so the z-component of the torque acting on the companion is $-M_2$. The second inversion is done because the M_{2k} given above are the components of the moment of the forces acting on the companion and what we need in the equation for $\dot{\Omega}$ is the reaction of the primary to the torque acting on the companion. Hence

$$\dot{\Omega} = -\frac{3GM\overline{\epsilon}_\rho}{2a^3} \sum_{k\in\mathbb{Z}} E_{2,k} \cos\sigma_k \sum_{j+k\in\mathbb{Z}} E_{2,k+j} \sin(j\ell + \sigma_k), \tag{52}$$

where we have simplified the coefficient by using the homogeneous body value $C = \frac{2}{5}mR^2$ and introduced the actual values of $\mathcal{C}_k$. The summations are done over all terms of order less than or equal to a chosen N. (Remember that $E_{2,k} = \mathcal{O}(e^k)$.)

One important characteristic of this equation, due to the invariance of the torque to rotations of the reference system, is that the right-hand side is independent of the attitude of the primary. The arguments of the periodic terms do not include the azimuthal angle fixing the position of the rotating body. Therefore, this is a true first-order differential equation and there are no free oscillations. The corresponding physical librations are forced oscillations. This is totally different from the classical spin-orbit dynamics of rigid bodies where a permanent azimuthal asymmetry in the mass distribution of the body (potential terms with coefficients J_{22} or J_{31}) gives rise to terms including the azimuthal angle in the arguments and the equation to be considered is a second-order differential equation.

The average of Eq. (52) with respect to ℓ is

$$<\dot{\Omega}> = -\frac{3GM\overline{\epsilon}_\rho}{4a^3} \sum_{k\in\mathbb{Z}} E_{2,k}^2 \sin 2\sigma_k. \tag{53}$$

We remind that (see Eq. (39))

$$\sin 2\sigma_k = \frac{2\gamma(\nu + kn)}{\gamma^2 + (\nu + kn)^2}, \tag{54}$$

where

$$\nu = 2\Omega - 2n \tag{55}$$

[9] Remind that we are only considering the orthogonal case in which the rotation axis of the primary is perpendicular to the orbital plane.

is the semi-diurnal frequency. (N.B. In the Keplerian approximation, the sidereal and the anomalistic mean-motions are equal.)

Far from the equilibrium solutions, the more important term in the series is $k = 0$. When the approximation $k = 0$ is adopted, the average is reduced to

$$\langle \dot{\Omega} \rangle = -\frac{3GM\overline{\epsilon}_\rho}{4a^3} E^2_{2,0} \frac{2\gamma\nu}{\gamma^2 + \nu^2}. \tag{56}$$

This equation has a classical interpretation. The sign of $\langle \dot{\Omega} \rangle$ is opposite to the sign of ν. That is, $\langle \dot{\Omega} \rangle$ is negative (resp. positive) when $\Omega > n$ (resp. $\Omega < n$). Therefore, far from the equilibrium, the variation of the rotation of the primary is always oriented towards the equilibrium.

7.1 Synchronization

The study of the full set of solutions of this system is difficult because of the small values of the torques. In order to get a picture of the solutions space, we construct a map that associates to each value Ω its increment in one orbital period. Formally, these maps are $\Omega(\ell) \rightarrow \Omega(\ell+2\pi)-\Omega(\ell)$. The increments are calculated into a grid of values of ν/n (horizontal axis) and the intersections with zero are the stationary solutions of the system (see Fig. 6)

The use of a first-order integrator is enough. $\dot{\Omega}$ is too small and we are allowed to assume Ω constant (that is, σ_k constant), in the r.h.s., and just integrate the periodic terms over one cycle.

Because of the small values of the variation of ν/n, the results in Fig. 7 appear multiplied by 10^6.

For large values of γ (that is, for $\gamma \geq n$), the curves intersect the axis $\Delta\Omega = 0$ just once (Fig. 7 top). There exists one and only one attractor (or stable stationary solution). We notice that for $e \neq 0$, this attractor is super-synchronous, The intersection is situated at $\nu/n \simeq 12e^2$, the same value found in all Darwin-type theories (see Sect. 11.1) and in the creep tide theory when $\gamma \gg n$ (see [19, Eqn. 35]).

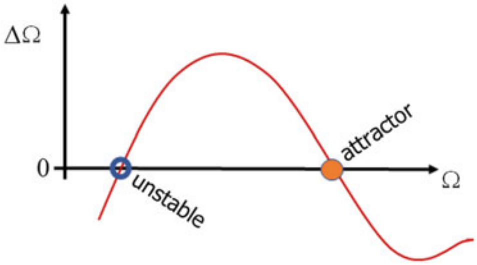


Fig. 6 Map showing the variation of $\Delta\Omega$ per period as function of Ω. The stationary solutions, stable and unstable, appear as intersections of the map with the axis $\Delta\Omega = 0$

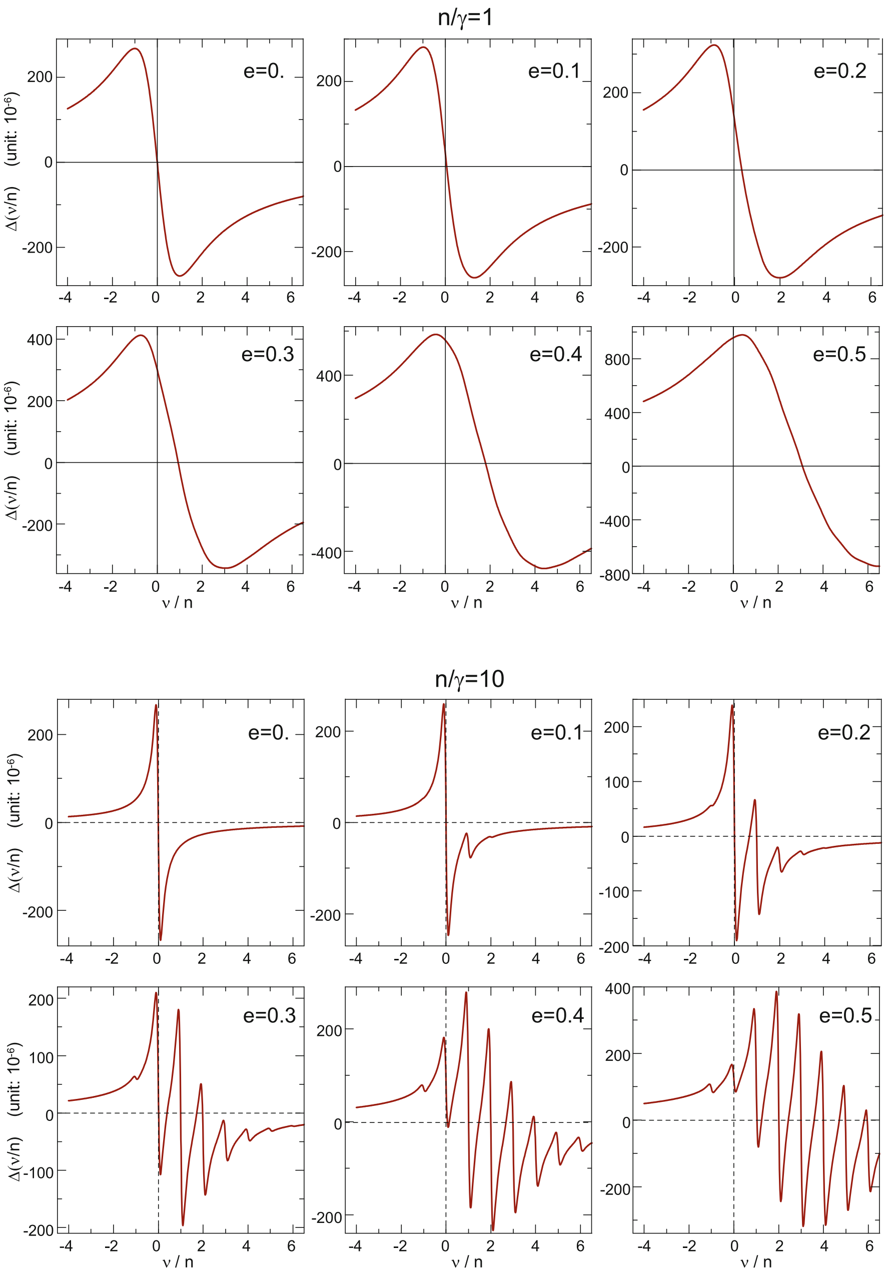


Fig. 7 Maps showing the variation of ν/n per period for $n = \gamma$ and $n/\gamma = 10$. Remind that for stars and giant gaseous planets, $\gamma/n \gg 1$, while for planetary satellites and terrestrial planets, $\gamma/n \ll 1$ (see Table 4). Reprinted from [20] with permission

For smaller values of γ, the curves may intersect the axis $\Delta\Omega = 0$ many times and there are many attractors (Fig. 7 bottom). These attractors are located at $\nu = -n, 0, n, 2n, \cdots$. Figure 7 bottom also shows how the existence of the attractors depends on the orbital eccentricity. If $e = 0$, the only attractor is the synchronous solution. When the eccentricity increases, the other attractors at $\nu/n \simeq -1, 1, 2, 3, \ldots$ gradually appear. It is worth noting that the attractor $\nu = 0$ does not show, in this case, the offset seen in the cases shown in Fig. 7 top. The averages are now very close to the actual synchronization.

The maps show the same features shown in the plots of the average torque vs. rotation frequency of [8]. The similarity is a consequence of the virtual equivalence of the creep tide theory and the Maxwell model (see Sect. 17).

7.2 The 3/2 Spin-Orbit Resonance: Mercury

The values of n/γ adopted to construct Fig. 7 bottom correspond to stiff bodies in large orbits as Mercury and some distant satellites. The evolution scenario, according to classical theories, is the following: The body had primordially a rotation much faster than the current one which slowly evolved, due to tidal dissipation, up to reach one attractor where it remained *trapped*. The trapping depends on the orbital eccentricity. The spin-orbit synchronous attractor $\nu/n = 0$ may only be reached if the eccentricity is small. If the eccentricity is high, attractors corresponding to higher values of ν/n will be reached before the synchronous attractor, and the rotation of the body will remain trapped there without reaching the synchronous condition. For instance, in the case of Mercury, the planet is trapped in the attractor $\nu/n = 1$, i.e. $\Omega = \frac{3}{2}n$. The rotation period is 2/3 of the orbital period. Since the orbital eccentricity is variable in the long term, we cannot discard the possibility that the rotation was for a while, in the past, trapped in the attractor $\nu/n = 2$ (i.e. $\Omega = 2n$), but escaped that resonance in one event in which the eccentricity plunged to its smallest values, close to 0.1 (see [35]), and the attractor $\nu/n = 2$ temporarily disappeared. When the eccentricity grew again and the attractor was restored, the rotation had already evolved to slower states and Mercury's rotation was driven to its present situation. Additionally, the fact that Mercury remained trapped in the 3/2 resonance shows that never in the past, the orbital eccentricity has been much below 0.1 since, for such small eccentricities, the 3/2 attractor disappears, and the rotation would no longer stay trapped there.

The facts that the rotation of Mercury was able to cross the 2/1 resonance without staying trapped there forever and eventually became trapped into the 3/2 solution and that no significant drift from the 3/2 commensurability could be measured [52], allow one to estimate that the relaxation factor of Mercury lies in the interval $4 < \gamma < 30 \times 10^{-9}\ \mathrm{s}^{-1}$ [20].

In the existing theories of rotation capture in spin-orbit resonance, probabilities of capture are calculated. In fact, the differential equations of models exploring the damped rotation of one rigid body are pendulum-like second-order equations with a separatrix between the two regimes of motion: resonant and non-resonant. The ability of the solution to go across one resonance or to be captured into the resonance depends on the phase of the corresponding angle when the separatrix is reached. This does not happen in the theory presented here. Here, the differential equation for Ω is of first-order and no pendulum-like separatrices exist. Capture follows necessarily when the solution reaches the basin of attraction of one resonance.

8 Tidal Evolution: Orbital Elements

The main tools to study the variation of the elements of the perturbed Keplerian motion of one body are the Lagrange variational equations, or, equivalently, if the disturbing forces are known instead of the disturbing potential, the Gauss variational equations. As already discussed in Sect. 4, in the study of tidal evolution, the perturbation acting on the primary is *internal* to the system of bodies and the potential of the disturbing forces acting on the primary must be multiplied by $(1+\frac{M}{m})$ to take into account the reaction on the companion of its tidal action on the primary. We thus consider the force per unit mass acting on the primary minus its reaction on the companion (see discussion in [23, Section 18.1]).

8.1 Semi-major Axis

The variation of the osculating semi-major axis due to the tides raised on the primary may be obtained using the Lagrange variational equation [4, Chap. XI]:

$$\dot{a} = \frac{2}{na}\frac{\partial \mathcal{R}}{\partial \ell}, \tag{57}$$

where the disturbing function is $\mathcal{R} = -(1+M/m)\delta U$ and δU is the potential of the tidal forces acting on the primary:

$$\delta U = -\frac{GmR^2}{5a^3}\sum_{k\in\mathbb{Z}}\sum_{j+k\in\mathbb{Z}}\Big(3\mathcal{C}_k\cos\sigma_k E_{2,j+k}\cos\big((2-k)\ell^*+2\omega^*+\sigma_k+(j+k-2)\ell-2\omega\big)$$

$$-\,\mathcal{C}''_k\cos\sigma''_k E_{0,j+k}\cos\big(-k\ell^*+\sigma''_k+(j+k)\ell\big)\Big). \tag{58}$$

In this expression, we have indicated with an asterisk the mean anomaly and the argument of the pericenter of Diana (ℓ^*, ω^*) (see Sect. 6.1). This is now necessary because the derivatives appearing in the Lagrange variational equations refer only to the elements of the body on which the force derived from the potential δU is acting, which are introduced into δU when the coordinates r, φ are substituted by the Keplerian variables of the motion of the companion around the primary. Once the derivatives are calculated, we may introduce the identities $\ell^* = \ell$, $\omega^* = \omega$, and obtain

$$\dot{a} = -\frac{2nR^2}{5a} \sum_{k\in\mathbb{Z}} \sum_{j+k\in\mathbb{Z}} \Big(3(j+k-2)\mathcal{C}_k \cos\sigma_k E_{2,j+k} \sin\big(j\ell + \sigma_k\big) \\ - (j+k)\mathcal{C}''_k \cos\sigma''_k E_{0,j+k} \sin\big(j\ell + \sigma''_k\big)\Big). \tag{59}$$

We may compare the Lagrange variational equation to the rate of variation of the work done by the force derived from the potential δU:

$$\dot{W} = \delta\mathbf{f} \cdot \mathbf{V} = -M.\mathrm{grad}_{\mathbf{r}}\delta U \cdot \mathbf{V} = -Mn\frac{\partial \delta U}{\partial \ell}. \tag{60}$$

If the third Kepler law ($n^2a^3 = G(M+m)$) is used, the comparison to the Lagrange equation gives

$$\dot{a} = \frac{2a^2}{GmM}\dot{W}. \tag{61}$$

This equation is sometimes used [19, 23] as an alternative to the Lagrange equation for the variation of the semi-major axis.

8.2 Eccentricity

The variation of the eccentricity is given by the corresponding Lagrange variational equation [4]:

$$\dot{e} = \frac{1-e^2}{na^2e}\frac{\partial\mathcal{R}}{\partial\ell} - \frac{\sqrt{1-e^2}}{na^2e}\frac{\partial\mathcal{R}}{\partial\omega}. \tag{62}$$

An alternative is to use the equivalent equation

$$\dot{e} = \frac{1-e^2}{e}\left(\frac{\dot{a}}{2a} - \frac{\dot{\mathcal{L}}}{\mathcal{L}}\right), \tag{63}$$

where $\mathcal{L} = \frac{GMm}{na}\sqrt{1-e^2}$ is the orbital angular momentum, derived from the equation for the variation of $\mathcal{L}$.

After some manipulation, we obtain:

$$\dot{e} = -\frac{3nR^2}{5a^2e}\sum_{k\in\mathbb{Z}}\mathcal{C}_k\cos\sigma_k\sum_{j+k\in\mathbb{Z}}\Big(2\sqrt{1-e^2}+(j+k-2)(1-e^2)\Big)E_{2,j+k}\sin(j\ell+\sigma_k)$$

$$+\frac{nR^2}{5a^2e}\sum_{k\in\mathbb{Z}}\mathcal{C}''_k\cos\sigma''_k\sum_{j+k\in\mathbb{Z}}(j+k)(1-e^2)E_{0,j+k}\sin(j\ell+\sigma''_k) \qquad (64)$$

9 Energy Variation and Dissipation

The bulk dissipation can be predicted by a mere application of the energy conservation principle. If the companion is considered as a mass point, the energy tidally dissipated in the primary body may only take origin in its rotation and in the relative orbital motion of the two bodies. The secular variations of the semi-major axis and of the rotation of the body are the two gauges allowing us to evaluate the energy lost by the system. No other mechanical process exists able to continuously supply energy to be dissipated in the system. We thus consider the energy exchanged with the orbit due to the direct attraction of the two bodies and the rotational energy stored in the primary. Other energy storing mechanisms are negligible [28]. The physical processes responsible for the dissipation inside the primary (see [16, 37, 43]) are not considered here, as well as the case of differentiated bodies, in which some parts may be much more efficient to dissipate energy than others [48].

The time rate of the work done by the tidal forces acting on the primary is obtained directly from the equations of Sect. 8.1.[10] It is

$$\dot{W} = -\frac{GMmnR^2}{10a^3}\sum_{k\in\mathbb{Z}}\sum_{j+k\in\mathbb{Z}}\Big(3(j+k-2)\overline{\epsilon}_\rho E_{2,k}E_{2,j+k}\cos\sigma_k\sin\big(j\ell+\sigma_k\big)$$

$$+(j+k)(E_{0,k}\overline{\epsilon}_\rho+2\delta_{0,k}\overline{\epsilon}_z)E_{0,j+k}\cos\sigma''_k\sin\big(j\ell+\sigma''_k\big)\Big), \qquad (65)$$

the average of which over ℓ is

$$<\dot{W}>= -\frac{GMmnR^2\overline{\epsilon}_\rho}{20a^3}\sum_{k\in\mathbb{Z}}\Big(3(k-2)E^2_{2,k}\sin 2\sigma_k + kE^2_{0,k}\sin 2\sigma''_k\Big). \qquad (66)$$

[10] The reader may pay attention to the opposite signs appearing in the definitions of $\mathcal{C}_k$ and $\mathcal{C}''_k$, which is often a source of mistakes in the transformation of the equations.

Additionally, the time rate of the energy variation associated with the rotation of the primary is $\dot{W}_{\rm rot} = C\Omega\dot{\Omega}$, that is,

$$\dot{W}_{\rm rot} = -\frac{3GMm\Omega R^2\overline{\epsilon}_\rho}{5a^3}\sum_{k\in\mathbb{Z}}\sum_{j+k\in\mathbb{Z}} E_{2,k}E_{2,k+j}\cos\sigma_k \sin(j\ell+\sigma_k) \tag{67}$$

(see Eq. (52)), or, in the average,

$$< \dot{W}_{\rm rot} >= -\frac{3GMm\Omega R^2\overline{\epsilon}_\rho}{10a^3}\sum_{k\in\mathbb{Z}} E^2_{2,k}\sin 2\sigma_k. \tag{68}$$

Some approximations were done in the above calculations which deserve mention: (1) It is assumed the homogeneous sphere value $C = \frac{2}{5}mR^2$ and variations of C due to the shape of the body are neglected. (2) In the averaging process, ν is assumed to be constant. However, Ω has small forced librations which become important near $\nu = 0$ Therefore, the averages given above are only valid far from $\nu = 0$ and the non-averaged equations giving $\dot{W}$ and $\dot{W}_{\rm rot}$ must be used if the motion is close to a commensurability where the variation of Ω may affect the result.

If the two averages are added, there results:

$$< \dot{W}_{\rm total} >= -\frac{GMmR^2\overline{\epsilon}_\rho}{20a^3}\sum_{k\in\mathbb{Z}}\Big(3(\nu+kn)E^2_{2,k}\sin 2\sigma_k + knE^2_{0,k}\sin 2\sigma''_k\Big). \tag{69}$$

If we consider that

$$\sin 2\sigma_k = \frac{2\gamma(\nu+kn)}{\gamma^2+(\nu+kn)^2}, \qquad \sin 2\sigma''_k = \frac{2\gamma kn}{\gamma^2+k^2n^2}, \tag{70}$$

we may see that the result is always negative (there is a loss of the total mechanical energy); the signs of $\sin 2\sigma_k$ and $\sin 2\sigma''_k$ are compensated by the signs of the factors $(\nu+kn)$ and kn so that the sum of the terms inside the brackets is a sum of squares.

The modulus of $< \dot{W}_{\rm total} >$ is the total energy dissipated inside the primary.

Figure 8 shows the dissipation in two cases in which $|\nu/n| = 2.5$. In the faster case, $(\nu > 0)$ the body rotation is much faster than the orbital motion; in the other, $(\nu < 0)$ the rotation is slightly retrograde. The values were chosen so as to avoid being close to the stationary solutions.

Figure 8 shows that the dissipation is dominated by two power laws. When $\gamma \gg n$ (gaseous bodies), the dissipation is inversely proportional to γ/n (In the right-hand side part of the log-log plot the curve is a straight line with inclination equal to -1); This is the same regime adopted by Darwin [11] in his theory in which the dissipation is proportional to the frequency of the main harmonic. On the contrary, in the left-hand side part of the plot, where $\gamma \ll n$ (stiff bodies), the dissipation is proportional to γ/n. One regime of this kind but with a less steep power law has

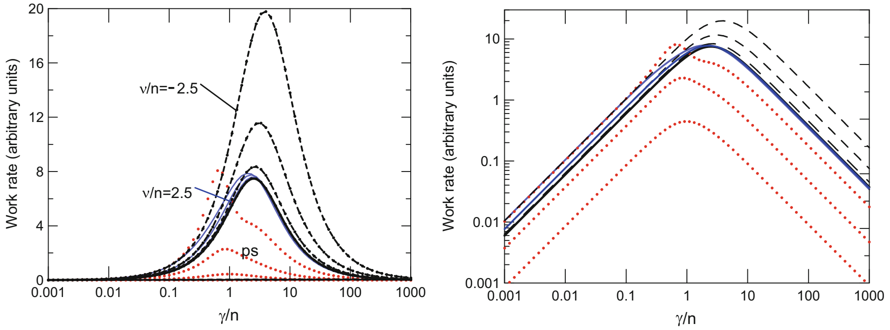


Fig. 8 *Left:* Time rate of the energy dissipated in free rotating bodies in two cases: $\nu = -2.5n$ (dashed/black) and $\nu = 2.5n$ (solid/blue) for eccentricities between 0.0 (thick line) and 0.3. In the two cases, the results coincide for $e = 0$ and the dissipation increases with the eccentricity. *Right:* Same in logarithmic scale to show the power laws ruling the dissipation in the two regimes: $\gamma \ll n$ (Efroimsky-Lainey) and $\gamma \gg n$ (Darwin). The pseudo-synchronous solution is included in the figure (dots/red) for comparison. Reprinted from [20] with permission

been first adopted by Efroimsky and Lainey [15] to represent the tidal lag of stiff bodies.

10 Darwin's Theory

The next three sections are devoted to Darwin's [11] theory. This theory was proposed 130 years ago, time enough to have it revisited by many authors ([1, 13, 14, 17, 23, 30, 32–34, 41], etc.) and to the proposition of many modifications. In this text we follow mainly [33] and [23].

The starting point in Darwin's theory is the static tide (see Eq. (17)):

$$\delta\zeta = R\sum_{k\in\mathbb{Z}}\Big(\mathcal{C}_k\sin^2\widehat{\theta}\cos(2\widehat{\varphi}+(k-2)\ell-2\omega)+\mathcal{C}''_k(\cos^2\widehat{\theta}-\frac{1}{3})\cos k\ell\Big). \tag{71}$$

The dynamic tide is assumed to be formed by the same harmonics as the static tide, each delayed by a given phase delay, resp. ε_k and ε''_k. Besides, each term is assumed to have an amplitude attenuation equal to the cosine of the corresponding phase delay. Then

$$\delta\zeta = R\sum_{k\in\mathbb{Z}}\Big(\mathcal{C}_k\sin^2\widehat{\theta}\cos\varepsilon_k\cos(2\widehat{\varphi}+(k-2)\ell-2\omega-\varepsilon_k)$$
$$+\,\mathcal{C}''_k(\cos^2\widehat{\theta}-\frac{1}{3})\cos\varepsilon''_k\cos(k\ell-\varepsilon''_k)\Big). \tag{72}$$

The comparison to the creep tide theory shows that the expression of the dynamic tide in Darwin's theory is the same as that of the creep tide theory, with the only difference that now we have the given quantities ε_k and ε''_k instead of the parameters σ_k and σ''_k introduced by the integration of the creep equation. This is not just a coincidence. Before proposing his theory for the orbital evolution of a companion revolving about a tidally distorted primary [11], Darwin briefly considered a model based on the Newtonian creep and obtained equations like those of the creep tide theory. He abandoned that model, but adopted in his new model, the same formal characteristics of his earlier studies.

This approximation, known as *weak friction approximation* [1], is often used. There, not only the ad hoc lags ε_k and ε''_k are assumed small, as their cosines appearing as attenuation of the harmonics amplitudes, are considered as equal to 1.

10.1 The Anelastic Tide

In the weak friction approximation, the arguments $\cos(2\varphi + \cdots - \varepsilon_k)$ may be Taylor expanded about $\varepsilon_k = 0$ giving $\cos(2\varphi + \cdots) + \varepsilon_k \sin(2\varphi + \cdots) + \mathcal{O}(\varepsilon_k^2)$. The first term in this expansion is the corresponding term in the static (a.k.a. elastic) tide, thoroughly studied in Sect. 4. The second (or linear) term is sometimes called anelastic (or visco-elastic) tide. The deformation of the primary corresponding to it is

$$\delta\zeta_{\rm anel} = R \sum_{k\in\mathbb{Z}} \Big(\varepsilon_k \mathcal{C}_k \sin^2\widehat{\theta} \sin(2\widehat{\varphi} + (k-2)\ell - 2\omega) + \varepsilon''_k \mathcal{C}''_k (\cos^2\widehat{\theta} - \frac{1}{3}) \sin k\ell\Big). \tag{73}$$

It is usual to restrict the study of the tidal evolution to the anelastic tide (as in [23]). As already discussed, the static tide does not affect the rotation of the primary and the averaged variations of the semi-major axis and eccentricity of the system. However, it needs to be considered when one is interested in knowing the perturbations of the main angles of the system (e.g. the argument of pericenter—see [23, Appendix B]).

One important point to emphasize with respect to the anelastic tide is that the maximum of its leading term ($k = 0$) is not at $\varphi = \ell + \omega (\mathrm{mod}\pi)$ as the leading term of the static tide, but at $\varphi = \ell + \omega - 45^\circ(\mathrm{mod}\pi)$. In other words: The vertex of the anelastic tide is trailing the vertex of the static tide by $\sim$45°.

10.2 Forces and Torques

The development of Darwin's theory is like that presented above for the creep tide theory and we do not need to remake every calculation. Because of the similarity of the expression of $\delta\zeta$ in both theories, we may use the results of the earlier sections and just replace the σ's by the corresponding ε's, and, when convenient use the weak friction approximation.

For instance, the forces and torques (see (50)–(51)) are:

$$\begin{aligned} F_{1k} &= -\frac{3GMmR^2}{5r^4}\Big(3\mathcal{C}_k \cos(2v-(2-k)\ell-\varepsilon_k) - \mathcal{C}''_k \cos(k\ell-\varepsilon''_k)\Big) \\ F_{2k} &= 0 \\ F_{3k} &= -\frac{6GMmR^2}{5r^4}\,\mathcal{C}_k \sin(2v-(2-k)\ell-\varepsilon_k), \end{aligned} \tag{74}$$

$$\begin{aligned} M_{1k} &= 0 \\ M_{2k} &= \frac{6GMmR^2}{5r^3}\,\mathcal{C}_k \sin(2v-(2-k)\ell-\varepsilon_k) \\ M_{3k} &= 0. \end{aligned} \tag{75}$$

Low-order Darwin's theory is friendly and the above equations can easily be adapted to different models (ex: core/mantle bodies [26, 27], effects due to response attenuation [23], etc.). High-order Darwin theories are easy to construct. However, given our ignorance of the actual rheology of celestial bodies, the accuracy of expansions to higher-orders may be illusory.

10.3 Ad Hoc Rheologies

Existing versions of Darwin's theory may differ by the law they adopt for the body response to tidal stresses. Some versions (e.g. [23, 33, 34]) do not define any a priori rheology. The only followed rule is that the responses are the same when the frequencies are equal. However, the most common case is to have one law relating the lags and the frequencies fixed a priori. Some usual choices are

- *The lags are proportional to the frequencies* (Darwin)

  In these theories the lags are proportional to the frequencies of the corresponding radial oscillations of one point on the surface of the body, as given by the static tide. The more important frequencies are given in Table 3 (N.B. $\varphi = \varphi_0 + \Omega t$). They can be fixed a priori [11, 41] or have the linearity as result of the choice of the dissipation laws [17, 32]. They are known as linear or CTL (constant time lag) theories. They were thoroughly discussed in [23]. We notice that this is the same law of the creep tide theory (see Eq. (39)) when we assume that the lags are small. The proportionality coefficient is $1/\gamma$. The difference between the two theories comes from the weak friction approximation that imposes that the lags are small and thus limits the validity of the theory to large γ's (i.e. to gaseous bodies. See Table 4).

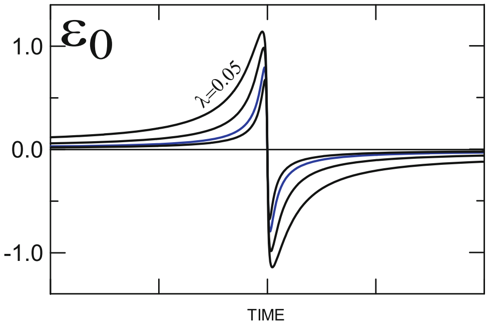


Fig. 9 Time evolution of the geodetic tide lag when the frequency of the semi-diurnal tide crosses 0 and the tidal bulge changes of side with respect to the sub-companion point (see Sect. 16). Reprinted from [19] with permission

- *The lags do not depend on the frequencies*

In these theories all harmonics have equal lags and these lags do not change during the evolution of the system. They are known as CPL (constant phase lag) theories. They were used in the study of the evolution of Solar System bodies. Because of the fast rotation of the planets, the frequencies $\nu - kn$ of the main harmonics are almost the same. All terms are "semi-diurnal" and the results are very similar to those obtained with CTL theories. They were also used in the study of the Moon and other planetary satellites. In these cases, the eccentricities are small and only the harmonics with $|k| \le 1$ matter. They have equal frequencies (they are "monthly". See Table 3).

The extension of CPL theories to exoplanets, where generally $\Omega \ll n$, is however not acceptable. Because of the slow rotation of the stars, the main harmonics have, in this case, completely different periods.

- *The lags follow an inverse power law* (Efroimsky-Lainey)

In the case of stiff bodies, to be in agreement with the observed lags of the seismic waves in the inner Earth, it has been suggested [15] the use of an inverse power law $\varepsilon_k \sim \text{cte}[\text{frequency}]_k^{-\alpha}$ with α chosen in the range 0.2–0.4. This law, however, cannot be uniformly used since a pure inverse power law leads to infinite values of ε_0 when $\nu \to 0$. It has then to be combined with some modifications when the semi-diurnal frequency ν approaches zero and, in the actual applications [37], the time behavior of the lag ε_0, when ν crosses the zero, is the same shown in Fig. 9, with a fast, but continuous, sign reversion near $\nu = 0$.

- *Constant geometric lag* (MacDonald)

In this very popular model [36], the whole equilibrium ellipsoid is delayed of a constant geometric lag. This hypothesis greatly simplifies the algebra involved in a theory, but is generally considered as unphysical [18, 55] because it does not define a rheology. The Fourier decomposition of the equations shows that there is no law relating lags and the frequencies of the harmonics. In other words,

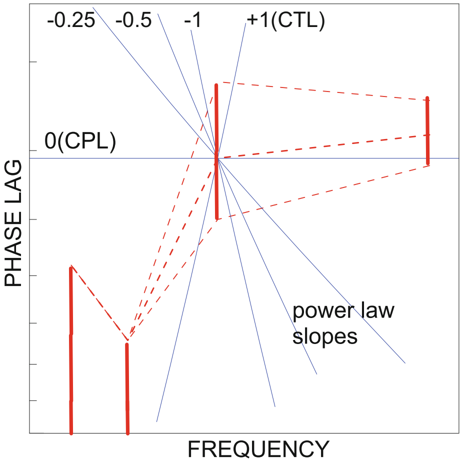


Fig. 10 Log-log plot of ε_k vs. frequency as determined from the laser ranging observations of the Moon. The blue straight lines show several different power laws passing by the mid-point of the value of ε_k corresponding to 1 year

no simple rheology is intrinsically fixed by MacDonald's equations. The same is true of the modification proposed in [50]. Another difficulty intrinsic to the choice done by MacDonald is the value of the lag when the primary is oscillating around the spin-orbit synchronism. In that case, it is necessary to switch the lag sign by changing it at every crossing of $\nu = 0$. The above-mentioned problems were fixed by Williams and Efroimsky through a proper modification of the original equations that transformed the original model into a CTL model [55].

One difficulty in assessing the right rheology is the absence of direct observations. The only body for which direct observations exist is the Moon. The laser ranging observations of the Moon show that the quality factor Q ($\simeq 1/\varepsilon_k$) is 38±4 at 1 month, 41 ±9 at 1 year, ≥74 at 3 years and ≥58 at 6 years [54], The corresponding values of ε_k are plotted against the value of the frequencies in a log-log plot in Fig. 10. The straight lines show several different power laws passing by the mid-point of the value of ε_k corresponding to 1 year. The disagreement between them and the observations is noteworthy. The best representation of the two better determined values is obtained with the CPL model.

11 Darwin's Theory: Tidal Evolution

The equations of tidal evolution in the frame of Darwin's theory may be obtained from those already discussed in the creep tide theory just replacing the σ's by the corresponding ε's, and using the weak friction approximation.

11.1 Synchronization

The differential equation for the rotation of the primary, considering the anelastic tide, is:

$$\dot{\Omega} = -\frac{3GM\overline{\epsilon}_\rho}{2a^3}\sum_{k\in\mathbb{Z}} E_{2,k} \sum_{j+k\in\mathbb{Z}} \varepsilon_k E_{2,k+j}\cos j\ell, \tag{76}$$

the average of which with respect to ℓ is

$$<\dot{\Omega}>= -\frac{3GM\overline{\epsilon}_\rho}{2a^3}\sum_{k\in\mathbb{Z}} \varepsilon_k E_{2,k}^2, \tag{77}$$

or, substituting the actual values of the Cayley functions (see Table 2):

$$<\dot{\Omega}>= -\frac{3GM\overline{\epsilon}_\rho}{8a^3}\Big(4\varepsilon_0 + e^2(-20\varepsilon_0 + \varepsilon_1 + 49\varepsilon_{-1}) + \cdots\Big). \tag{78}$$

The first remark concerning this equation is that the result depends on several distinct lags. So, the continuation depends on the adopted rheology.

Let us assume that, in the neighborhood of the synchronisation, $\varepsilon_0 \ll |\varepsilon_1|$ ($\varepsilon_0 \sim 0$) and $\varepsilon_{-1} = -\varepsilon_1$ (terms with equal but opposite frequencies have opposite lags). Hence $<\dot{\Omega}> \propto (\varepsilon_0 - 12e^2\varepsilon_1) \neq 0$, That is, the rotation cannot be synchronous, unless the motion is circular. If we solve the equation $<\dot{\Omega}>= 0$, we obtain

$$\varepsilon_0 \simeq 12e^2\varepsilon_1. \tag{79}$$

If Darwin's CTL rheology is adopted, $\varepsilon_0 \propto 2\Omega - 2n$, $\varepsilon_1 \propto 2\Omega - n$, and the latest equation becomes

$$\Omega_{\rm stat} \simeq n(1 + 6e^2). \tag{80}$$

In Darwin's theory, the only stationary solution is super-synchronous. This is the same result shown in the top panels of Fig. 7 where the only stationary solution appears at the right of the origin.

In Darwin's theory, a synchronous solution cannot exist if $e \neq 0$. So, in this case, any stationary solution is either a super-synchronous solution or the tidal torque is not the only one acting on the primary. It is often assumed that the primary has a permanent axial asymmetry, and, at the equilibrium, the torque created by this asymmetry compensates the tidal torque (see [23, sec 5.3]). Darwin's theory is also not able to produce the possibility of different stationary solutions as shown in the bottom panel of Fig. 7 without assuming the action of additional torques.

11.2 Dissipation

To obtain the energy tidally dissipated in the primary, we consider the contributions of the orbit of the system and the rotation of the primary. They are obtained from the corresponding equations in the creep tide theory. After substitution, adoption of the weak friction approximation and averaging over ℓ, we obtain for the secular variation of the orbital energy

$$\langle \dot{W}_{\text{orb}} \rangle = -\frac{GmMnR^2\overline{\epsilon}_\rho}{10a^3} \sum_{k\in\mathbb{Z}} \Big(3(k-2)E^2_{2,k}\varepsilon_k + kE^2_{0,k}\varepsilon''_k \Big). \tag{81}$$

and for the secular variation of the rotational energy of the primary,

$$\langle \dot{W}_{\text{rot}} \rangle = -\frac{3GmM\Omega R^2\overline{\epsilon}_\rho}{5a^3} \sum_{k\in\mathbb{Z}} E^2_{2,k}\varepsilon_k. \tag{82}$$

The sum of the two components gives

$$\langle \dot{W}_{\text{tot}} \rangle = -\frac{GmMR^2\overline{\epsilon}_\rho}{10a^3} \sum_{k\in\mathbb{Z}} \Big(6(\Omega-n)E^2_{2,k}\varepsilon_k + 3knE^2_{2,k}\varepsilon_k + knE^2_{0,k}\varepsilon''_k \Big). \tag{83}$$

In the CTL theories, ε_k and ε''_k are proportional to the frequencies $2\Omega+(k-2)n$, and kn, respectively, and the parenthesis of the above equation may be reduced to a sum of squares. As expected, the final result for $\langle \dot{W}_{\text{tot}} \rangle$ is negative (energy is lost). This is the energy that the system is dissipating inside the primary.

Two approximations are important in the applications. One is the case of one system in free rotation (i.e. far from the equilibrium) and small eccentricity. In this case, the dissipation is dominated by the term corresponding to $k=0$, that is,

$$\langle \dot{W}_{\text{tot}} \rangle_{k=0} = -\frac{3GmMR^2\overline{\epsilon}_\rho}{5a^3}(\Omega-n)E^2_{2,0}\varepsilon_0. \tag{84}$$

The other is the case close to the stationary rotation. In this case $\varepsilon_0 \sim 0$ and the terms with $|k|=1$ have also to be considered in the approximation. Since we have, in the stationary solution, $\Omega - n = 6ne^2$, the dissipation is given by

$$\langle \dot{W}_{\text{tot}} \rangle_{stat} = -\frac{21GmMR^2\overline{\epsilon}_\rho ne^2\varepsilon_1}{5a^3}. \tag{85}$$

(We remind that ε''_1 and ε''_{-1} correspond to opposite frequencies and so $\varepsilon''_{-1} = -\varepsilon''_1$; besides, when $\Omega \sim n$, we similarly have $\varepsilon_{-1} = -\varepsilon_1$ and also $\varepsilon''_1 = \varepsilon_1$. See Table 3.)

These results show some classical properties: The dissipation in a free rotating primary is controlled by ε_0, while in a body whose rotation is trapped in the stationary solution, it is controlled by ε_1. The quality factors found in the applications are thus not the same in both cases. We have $Q \sim 1/\varepsilon_0$ in the first case, and $Q \sim 1/\varepsilon_1$ in the second case. The other important property is that the dissipation is proportional to the lags; If we assume that the lags are proportional to the frequencies and plot the dissipation in a log-log plot, as it was done in the creep tide theory (Fig. 8 right), we get just the right-hand side straight line going downwards; the left-hand branch characteristic of the Maxwell bodies with an inverted behavior does not appear. We remind that because of the weak friction approximation, Darwin's theory is only valid for frequencies much smaller than γ.

11.3 Orbital Evolution

We present in this section the average variation of the semi-major axis and eccentricity as given by Darwin's theory. The equations for the variation of the angular elements (longitude at epoch, argument of pericenter, and, in non-orthogonal theories, also the longitude of the node) are not given as these quantities are never considered. Their calculation follows from the variational equations of Lagrange in the same way as the others. The only difference to keep in mind is that the contribution of the static tide to these variations is important and even more important than the contribution of the anelastic tide.

After introduction of the ε's, adoption of the weak friction approximation and averaging over ℓ, we obtain

$$\langle \dot{a} \rangle = -\frac{nR^2\overline{\epsilon}_\rho}{5a} \sum_{k \in \mathbb{Z}} \Big(3(k-2) E^2_{2,k} \varepsilon_k + k E^2_{0,k} \varepsilon''_k \Big). \tag{86}$$

If we keep only the leading terms ($|k| \leq 1$) and make $\varepsilon''_{-1} = -\varepsilon''_1$:

$$\langle \dot{a} \rangle = \frac{nR^2\overline{\epsilon}_\rho}{5a} \Big(6E^2_{2,0}\varepsilon_0 + 3E^2_{2,1}\varepsilon_1 + 9E^2_{2,-1}\varepsilon_{-1} - 2E^2_{0,1}\varepsilon''_1 \Big). \tag{87}$$

or, considering the actual expression of the Cayley functions (see Table 2),

$$\langle \dot{a} \rangle = \frac{3nR^2\overline{\epsilon}_\rho}{10a} \Big(4\varepsilon_0 - e^2 (20\varepsilon_0 - \frac{1}{2}\varepsilon_1 - \frac{147}{2}\varepsilon_{-1} + 3\varepsilon''_1) \Big). \tag{88}$$

The averaged variation of the eccentricity, in Darwin's theory, is obtained in a similar way:

$$\langle \dot{e} \rangle = -\frac{3nR^2\overline{\epsilon}_\rho}{10a^2 e} \sum_{k \in \mathbb{Z}} \Big(2\sqrt{1-e^2} + (k-2)(1-e^2) \Big) E^2_{2,k} \varepsilon_k$$

$$-\frac{nR^2\overline{\epsilon}_\rho}{10a^2e}\sum_{k\in\mathbb{Z}}k(1-e^2)E^2_{0,k}\varepsilon''_k. \tag{89}$$

The term depending on $\overline{\epsilon}_z$ does not appear since $k\delta_{0,k} = 0$. If we note that $\Big(2\sqrt{1-e^2}+(k-2)(1-e^2)\Big) = k+(1-k)e^2+\mathcal{O}(e^4)$, we obtain for the leading terms

$$\langle\dot{e}\rangle = -\frac{nR^2\overline{\epsilon}_\rho}{10a^2e}\sum_{k\in\mathbb{Z}}\Big(3\big[k+(1-k)e^2\big]E^2_{2,k}\varepsilon_k + k(1-e^2)E^2_{0,k}\varepsilon''_k\Big) \tag{90}$$

or, keeping only terms with $|k|\leq 1$,

$$\langle\dot{e}\rangle = -\frac{nR^2\overline{\epsilon}_\rho}{10a^2e}\Big(3e^2E^2_{2,0}\varepsilon_0+3E^2_{2,1}\varepsilon_1-3E^2_{2,-1}\varepsilon_{-1}+2E^2_{0,1}\varepsilon''_1\Big). \tag{91}$$

Finally, considering the actual expression of the Cayley functions (see Table 2), we obtain [30]

$$\langle\dot{e}\rangle = -\frac{3nR^2\overline{\epsilon}_\rho}{20a^2}e\Big(2\varepsilon_0+\frac{1}{2}\varepsilon_1-\frac{49}{2}\varepsilon_{-1}+3\varepsilon''_1\Big). \tag{92}$$

The continuation of these derivations depends on the chosen rheology.

12 Evolution Equations in the CTL Model

We give below the orbital evolution results when the lags are assumed to be proportional to the frequencies of the respective tide harmonics (see Table 3) with τ as coefficient of proportionality (or *time lag*).

$$\langle\dot{a}\rangle = \frac{12nR^2\overline{\epsilon}_\rho}{5a}\Big(\Omega(1+\frac{27}{2}e^2)-n(1+23e^2)\Big)\tau. \tag{93}$$

$$\langle\dot{e}\rangle = \frac{3nR^2\overline{\epsilon}_\rho}{5a^2}e(11\Omega-18n)\tau. \tag{94}$$

These equations are the same found in several papers on tidal friction using Hut's formulas (e.g. [12, 38]). It is worth mentioning that Hut's results [32] are not given by expansions but by closed expressions. By construction, they must also be equivalent to the results of the creep tide theory for $\gamma \gg n$ with $\tau = 1/\gamma$. The basic formulation of the two theories is the same when $\gamma \gg n$. Indeed, in this case we may apply the weak friction approximation for σ_k, σ''_k like it was done for ε_k, ε''_k and the equations of the two theories become the same differing just by the used notation for the lag and the Taylor expansion about $\sigma_k = \sigma''_k = 0$.

One last remark concerns theories in which the radial terms arising from the tidal contraction of the polar axis are neglected. In these cases, the terms with lags ε''_k are absent and the eccentricity dependent coefficients of n in the above equations are 181/8 (instead of 23) and 69/4 (instead of 18).

12.1 Fast-Rotating Planets

If the primaries are fast-rotating planets, as Jupiter, for example, and the companion one satellite, we have $n \ll \Omega$ and we may neglect the contribution of the term proportional to n in the above equations. There results:

$$\langle \dot{a} \rangle = \frac{12nR^2\Omega\overline{\epsilon}_\rho}{5a}(1 + \frac{27}{2}e^2)\tau, \tag{95}$$

$$\langle \dot{e} \rangle = \frac{33nR^2\Omega\overline{\epsilon}_\rho}{5a^2}e\tau. \tag{96}$$

These equations are found in a great deal of applications. We find it even in the early applications of the tidal theory to exoplanetary systems, but this was not correct because the hypothesis $n \ll \Omega$ is not satisfied in systems where the rotation of the star is slow (see below).

12.2 Slow-Rotating Stars

If the primaries are slow-rotating stars hosting close-in exoplanets, we have $\Omega \ll n$ and we may neglect the contribution of the term proportional to Ω in the above equations.

$$\langle \dot{a} \rangle = -\frac{12n^2R^2\overline{\epsilon}_\rho}{5a}(1 + 23e^2)\tau, \tag{97}$$

$$\langle \dot{e} \rangle = -\frac{54n^2R^2\overline{\epsilon}_\rho}{5a^2}e\tau. \tag{98}$$

The important point to stress here, as compared to the case of Sect. 12.1, is the change of sign in both equations. When the central body is rotating fast, the tide in the primary result in the companion moving away from the primary in an orbit of increasing eccentricity. However, when the central body is rotating slowly, as evolved host stars, the tide in the primary circularizes the orbit of the companion and makes it to fall over the primary.

It is worth emphasizing that the rotation of the central star of a planetary system may suffer a great variation, being initially very fast and being later braked due to its activity [5, 25]. Besides, the discovered exoplanetary systems show stars with many different rotational states. Therefore, in general studies, the approximation given in this section may be insufficient and the general formulas of Sect. 12 must be used.

12.3 Hot Jupiters

Evolution studies show that close-in hot Jupiters tend to the stationary rotation in relatively short times (some Myr). In Darwin's theory, they are driven to a super-synchronous rotation with a rotation velocity $\Omega = n(1 + 6e^2)$. If this value is introduced in the above general equations, we obtain:

$$\langle \dot{a} \rangle = -\frac{42 n^2 R^2 \overline{\epsilon}_\rho}{5a} e^2 \tau. \tag{99}$$

$$\langle \dot{e} \rangle = -\frac{21 n^2 R^2 \overline{\epsilon}_\rho}{5a^2} e \tau. \tag{100}$$

The decay of the orbit due to the tide in the almost synchronous hot Jupiter is proportional to the square of the eccentricity, so its contribution should stop once the orbit is circularized. These formulas are sometimes used to study the tidal decay of super Earths and of planetary satellites. However, these bodies are stiff and are not expected to have a CTL rheology (see Sect. 10.3).

13 Evolution Equations in the CPL Model

A great deal of investigations of the tidal evolution of planetary satellites has been done using a couple of equations taken from the CPL (constant phase lag) model [45, 56]

In the CPL model, when the primary is a fast-rotating planet, all lags are taken with the same value as ε_0. In this case, all sectorial terms are semi-diurnal (see Table 3) and the result will not differ very much from the corresponding result in the CTL model when $\Omega \gg n$. However, the radial term (whose lag is ε''_1) is monthly. Since it is also taken equal to ε_0, we get a similar formula but with different $\mathcal{O}(e^2)$ contributions: 51/4 instead of 54/4 in the equation for $\langle \dot{a} \rangle$ and 19/4 instead of 22/4 in the equation for $\langle \dot{e} \rangle$ (In the comparison of the two models' formulas, remind that in the CTL model, in this case, $\varepsilon_0 \sim 2\Omega\tau$).

$$\langle \dot{a} \rangle = \frac{6nR^2\overline{\epsilon}_\rho}{5a}(1 + \frac{51}{4}e^2)\varepsilon_0. \tag{101}$$

$$\langle \dot{e} \rangle = \frac{3nR^2\overline{\epsilon}_\rho}{5a^2}(\frac{19}{4}e)\varepsilon_0. \tag{102}$$

In the CPL model, in the case of an almost synchronous companion, the lag ε_0 is taken at the super-synchronous stationary value defined by Eq. (79) and the others have same modulus (they are monthly), but the frequency of the term $k = -1$ is negative and so ε_{-1} may be taken as $-\varepsilon_1$ [30]. Hence,

$$\langle \dot{a} \rangle = -\frac{6nR^2\overline{\epsilon}_\rho}{5a}(7e^2)\varepsilon_1. \tag{103}$$

$$\langle \dot{e} \rangle = -\frac{3nR^2\overline{\epsilon}_\rho}{5a^2}(7e)\varepsilon_1. \tag{104}$$

13.1 Cumulative Orbital Variations Due to Tides in Both Bodies

In general, we must consider, simultaneously, the variations of the semi-major axis and eccentricity due to the tides raised in both the primary and the companion. For the sake of writing the two contributions in only one equation, we change to more universal notations in the following way. In the equations giving the variations due to the tides in the more massive central body, we make the substitutions $\varepsilon_j = \varepsilon_{jA}$, $m = m_A$, $M = m_B$, and $R = R_A$ In the equations giving the variations due to the tides in the almost synchronous smaller body, we make the substitutions $\varepsilon_j = \varepsilon_{jB}$, $m = m_B$, $M = m_A$ and $R = R_B$. We then introduce the factor [45, 56]

$$D = \left(\frac{m_A}{m_B}\right)^2 \left(\frac{R_B}{R_A}\right)^5 \left(\frac{\varepsilon_{1B}}{\varepsilon_{0A}}\right). \tag{105}$$

where we have preferred to use the ratio of lags $\varepsilon_{1B}/\varepsilon_{0A}$ instead of the equivalent ratio of quality factors Q_A/Q_B, as generally done.

From the above equations, we obtain:

$$\langle \dot{a} \rangle = \frac{9nm_B R_A^5 \varepsilon_{0A}}{2m_A a^4}\Big(1 + \frac{51}{4}e^2 - 7De^2\Big), \tag{106}$$

$$\langle \dot{e} \rangle = \frac{9nm_B R_A^5 \varepsilon_{0A}}{4m_A a^5}\big(\frac{19}{4}e - 7De\big). \tag{107}$$

If the same kind of combination is done with the corresponding equations in the CTL model, the results are very similar to those given above, We just have 54/4 instead of 51/4 in the equation for $\langle \dot{a} \rangle$ and 22/4 instead of 19/4 in the equation for $\langle \dot{e} \rangle$. This similarity validates the use of the CTL model even on problems where the lags are better represented as frequency independent (see Fig. 10), at least for times short enough to allow us to consider that the frequencies do not change. It also explains the non-existence of big differences between results of the CTL and CPL models for limited times [2].

When the full synchronization ($\Omega = n$) is assumed, the term $7De^2$ in the CPL-model equation for $\langle \dot{a} \rangle$ appears sometimes replaced by $19De^2$. However, in the frame of Darwin's theory synchronous solutions may only exist when $e \sim 0$.

14 Mignard's Theory

Mignard's theory [41] is an alternative to Darwin's theory constructed in terms of closed expressions with no Fourier expansions at all. It considers the fictitious three-body arrangement primary-companion-Diana, with Diana being responsible for the deformation of the primary and the deformed primary is interacting gravitationally with the companion (see Fig. 5). Let $\mathbf{r}$, $\mathbf{r}^*$ be the position vectors of the companion and of Diana, respectively, referred to the center of the primary. The other notations are the same as before.

The static tide of the primary is a Jeans prolate ellipsoid (the contribution of the rotation to the polar oblateness is not considered in [41]), The disturbing gravitational potential of the primary due to the tidal deformation, in a generic point $\mathsf{P}(\mathbf{r})$ can be obtained from Eq. (22). Taking into account the definitions of the moments of inertia A, B ($C = B > A$) and the expressions of a, b ($c = b < a$) in terms of ϵ_J, we obtain

$$U_2(\mathbf{r}) = -\frac{GmR^2}{5r^3}\epsilon_J(3\cos^2\Psi - 1), \tag{108}$$

where Ψ is the angle between the directions of $\mathbf{r}$ and $\mathbf{r}^*$ (see Fig. 5), or

$$U_2(\mathbf{r}) = -\frac{3GMR^5}{4r^5 r^{*5}}(3(\mathbf{r}\cdot\mathbf{r}^*)^2 - r^2 r^{*2}). \tag{109}$$

Note that the Jeans prolateness ϵ_J introduces in this expression, because of its definition (Eq. (10)), the position vector of Diana, $\mathbf{r}^*$, and its modulus r^*.

To consider the non-instantaneous response of the primary to the tidal potential raised by Diana, Mignard substitutes, in the above equation, Diana's radius vector $\mathbf{r}^*$ by the *delayed* vector

$$\mathbf{r}_1^* = \mathbf{r}^* - \mathbf{v}^* \Delta t + \mathbf{\Omega} \Delta t \times \mathbf{r}^*$$

where $\mathbf{\Omega}$ is the rotational velocity vector of the primary, $\mathbf{v}^*$ is Diana's velocity, and Δt is a time delay. In the applications to the Earth tide, Mignard adopted $\Delta t = 10$ min, so that his results could reproduce the current variation of the Earth's length-of-the-day. However, the resulting phase lag is much larger than Earth's tidal lag as given by modern observations (see Sect. 16).

After the substitution $\mathbf{r}^* \to \mathbf{r}_1^*$, U_2 is transformed into $U_2 + \delta U$, where

$$\delta U(\mathbf{r}) = \frac{9GMR^5}{2r^5 r^{*5}} \Delta t \Big((\mathbf{r} \cdot \mathbf{r}^*) \Big(\mathbf{r}^* \cdot (\mathbf{\Omega} \times \mathbf{r}) + \mathbf{r} \cdot \mathbf{v}^* \Big) - \frac{(\mathbf{r}^* \cdot \mathbf{v}^*)}{2r^{*2}} \Big(5(\mathbf{r} \cdot \mathbf{r}^*)^2 - r^2 r^{*2} \Big) \Big). \tag{110}$$

We may add here the same comments done when using potentials in the creep tide theory: (a) Eq. (110) is enough to study the rotational and orbital evolution of the system since the contribution of the static tide is ineffective in this respect (see Sect. 4); (b) $\delta U(\mathbf{r})$ is a time dependent potential, the dependence on t being introduced in the potential through the radius vector and velocity of Diana; (c) The disturbing force acting on the companion is $\delta F = M \nabla_{\mathbf{r}} \delta U$ ($\mathbf{r}^*, \mathbf{v}^*$ are treated as constants in the calculation of the gradient); (d) Once the gradient is calculated, we identify Diana with the companion making $\mathbf{r}^* = \mathbf{r}$ and $\mathbf{v}^* = \mathbf{v}$. We thus obtain

$$\delta \mathbf{F} = -\frac{9GM^2R^5}{2r^{10}} \Delta t \Big(2(\mathbf{r} \cdot \mathbf{v})\mathbf{r} + r^2 (\mathbf{r} \times \mathbf{\Omega} + \mathbf{v}) \Big) \tag{111}$$

and the torque

$$\mathcal{M} = \mathbf{r} \times \delta \mathbf{F} = -\frac{9GM^2R^5}{2r^8} \Delta t \Big((\mathbf{r} \cdot \mathbf{\Omega})\mathbf{r} - r^2 \mathbf{\Omega} + \mathbf{r} \times \mathbf{v} \Big). \tag{112}$$

It is worth mentioning that after [41], other theories were proposed [17, 32] leading to the same closed expression for the disturbing force, despite their very different formulations. We also mention that an expansion of this force to third order in eccentricity and inclination leads to the same series expansions used in Darwin's CTL model (when the rotational contribution to ϵ_z is neglected).

The closed expression for $\delta \mathbf{F}$ may be simply used as an additional disturbing force in N-body models allowing us to study systems more complex than the two-body model usual in tidal evolution studies (see next section). One difficulty in this case is the smallness of the tidal forces which makes the numerical integration of the equations for the variation of the coordinates too slow. One alternative in these applications is to artificially increase the time delay Δt [22, 49]. This *scaling* is however a non-rigorous procedure and the results must be carefully checked against some exact simulations to guarantee that the errors thus introduced remain at an acceptable level.

15 Three-Body Models: Transfer of Angular Momentum

Mignard's force has been included in an N-body code used to study the secular behavior of one system of two close-in planets in orbit around a Sun-like star. The planets are a hot mini-Neptune and a more massive outer planet. The parameters of the CoRoT 7 system were used. The mini-Neptune is very close to the star (as the super-Earth CoRoT 7b). The outer Jupiter (as CoRoT 7c), is also close to the planet, but not so close, so that its tidal interactions with the star may be neglected [49]. A scaling factor 100 was used to allow the study of a great number of cases.

Figure 11 shows that the orbit of the inner planet is circularized due to the tides raised by the star on the planet (the tides raised on the star are negligible) and fall toward the star. The fall however stops when the orbit becomes circular (as expected from Darwin's theory. See Eq. (99)). The outer planet, while far enough and not in tidal interaction with the star, is also affected by the tidal interaction of the inner planet and the star, and its eccentricity decreases. The decrease almost stops when the orbit of the inner planet becomes circular. Figure 12 helps to understand the role of the outer planet in the evolution of the innermost one. It shows the time evolution of the inner planet semi-major axis (solid curve in the left panel) and

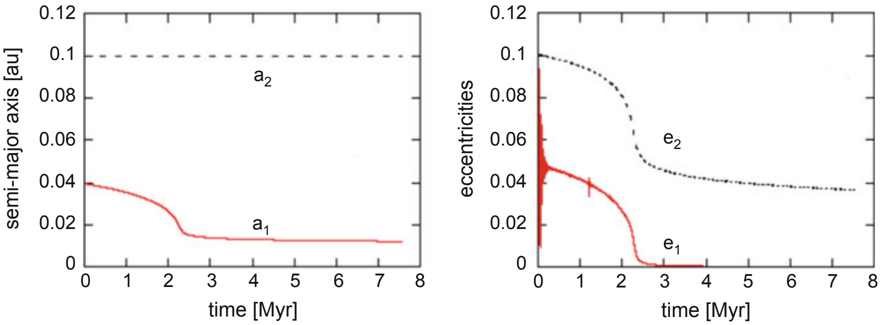


Fig. 11 Long-term evolution of semi-major axes and eccentricities in a Sun—mini Neptune—hot Jupiter system. Time scaling = 100

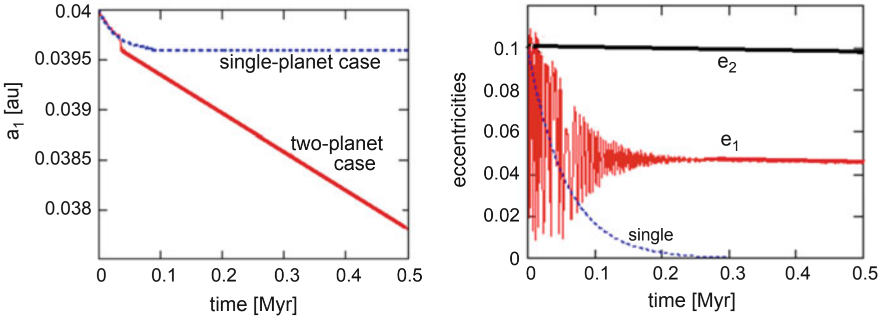


Fig. 12 Short-term evolution of semi-major axes and eccentricities. The evolution in the case in which the inner planet is the only planet is also shown. Adapted from [49] with permission

the two eccentricities (right panel) over the first 50 Myr. Due to the large mass of the outer planet, e_2 is almost constant over the time interval shown in this figure. On the contrary, e_1 initially suffers large oscillations, but they are quickly damped to the so-called first *equilibrium eccentricity* [39]. The trapping in this eccentric stationary solution delays the orbital circularization and accelerates the fall toward the star [49]. The same graph shows the variation of e_1 in the case of a single-planet system, when the interactions with one outer planet do not exist. The presence of the outer planet increases significantly the time needed to have the inner planet orbit circularized by the tidal effects. Moreover, in the single-planet case, the semi-major axis a_1 just decreases slightly, but stops decreasing after circularization.

16 The Figure of the Tidally Deformed Primary

The deformed shape of the primary due to the dynamic tide is given by Eq. (37). After the transient phase, that is, for $\gamma t \gg 1$, only the forced terms matter and the shape is dominated by the semi-diurnal component

$$\frac{1}{2} R\overline{\epsilon}_\rho E_{2,k} \sin^2 \widehat{\theta} \cos\sigma_0 \cos(2\widehat{\varphi} - 2\ell - 2\omega - \sigma_0) \tag{113}$$

whose maximum is reached when the argument of the trigonometric term is 0, that is, when the angle between the vertex of the primary's figure to the sub-companion point $(\widehat{\varphi} - \ell - \omega)$ is $\frac{1}{2}\sigma_0$. We remind that the angle σ_0, univocally determined by the integration of the creep equation, is not necessarily small. It is

$$\sigma_0 = \arctan \frac{\nu}{\gamma}. \tag{114}$$

In the case of a perfect fluid (static tide), the tide highest point stays aligned with the mean direction of the tide raising body (the companion). However, in the real case of a rocky planet, $\gamma \ll \nu$ and σ_0 will approach 90°. This result is in contradiction with the very small measured geodetic lag of the Earth's bodily semi-diurnal tide (0.20 ± 0.05°) [47]. In order to conciliate the theory and the Earth's measured tidal lag, we have to assume that the actual tide is not restricted to the dynamic component arising from the creep equation, but has also an elastic component [19]. This additional elastic component is defined in each point by its height over the sphere and is given by

$$\delta\zeta_{\rm el}(\widehat{\varphi}, \widehat{\theta}) = \lambda \delta\rho(\widehat{\varphi}, \widehat{\theta}), \tag{115}$$

where ρ is the radius vector of the static tide (equilibrium surface) and λ is a quantity related to the maximum height of the tide (see Sect. 10.2). For the Earth, for instance, $\lambda \sim 0.2$, which is the ratio of the observed maximum height of the lunar tide (26 cm cf. [40]) and the maximum height of the static tide (1.34 m cf. Table 1)

The sum of the (local) heights of the added elastic tide and of the main term of the creep tide is (in the circular approximation):

$$H = \frac{1}{2} R\overline{\epsilon}_\rho \Big(\lambda \cos(2\widehat{\varphi} - 2\ell - 2\omega) + \cos\sigma_0 \cos(2\widehat{\varphi} - 2\ell - 2\omega - \sigma_0) \Big), \tag{116}$$

where, for the sake of simplicity, we have set $E_{2,0}(e) = 1$, $\sin\widehat{\theta} = 1$ (equator), and restricted the result to the dominant semi-diurnal component.

If we introduce, in the creep differential equation (26), a new variable describing the above composition of the dynamic creep tide and the added elastic tide,

$$Z = \zeta + \lambda\delta\rho, \tag{117}$$

that equation becomes

$$\dot{Z} + \gamma Z = (1 + \lambda)\gamma\rho + \lambda\dot{\rho} - \gamma\lambda R. \tag{118}$$

which is an equation with the same characteristics of a Maxwell model and which is reduced to the creep model when $\lambda = 0$.

We remind that the elastic tide is torque free and conservative (see Sect. 4).[11] Therefore, the addition to the dynamic tide derived from the creep equation of one component $\lambda\rho$ proportional to the elastic tide does not affect the orbital and

[11] The results of Sect. 4 are consistent with those obtained with the creep tide theory (or with Darwin's theory) when all lags are made equal to zero. Indeed, the expressions for $\dot{a}$, $\dot{e}$, $\dot{W}$ of Sects. 8 and 9 are trigonometric series in the arguments $\sin(j\ell + \sigma_k)$ and $\sin(j\ell + \sigma''_k)$, which average to zero when the lags vanish. The vanishing of the torque when the σ_k vanish is less obvious. However, the auxiliary expansions given in the Online Supplement to [20], allow one to see that $\sum_{k\in\mathbb{Z}} E_{2,k} \sin\big(2v - (2-k)\ell\big) = 0$, and so that $M_2 = 0$ when the lags vanish.

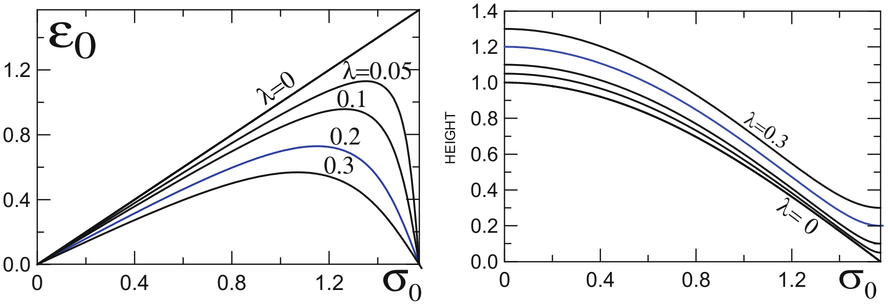


Fig. 13 *Left:* Geodetic lag ε_0 of the semi-diurnal tide as a function of σ_0. *Right:* Maximum height of the composite tide in units $\frac{1}{2}R\overline{\epsilon}_\rho$. The blue lines correspond to $\lambda = 0.2$ (Earth). Reprinted from [19] with permission

rotational evolution of the system. The orbital elements (a, e) will indeed have an additional variation, but with zero averages and not affecting the evolution.

16.1 The Geodetic Lag

The maximum tide height (the maximum of H) is, now, no longer reached at $\widehat{\varphi} - \ell - \omega = \frac{1}{2}\sigma_0$, as for the creep tide, but at $\widehat{\varphi} - \ell - \omega = \frac{1}{2}\varepsilon_0$, where

$$\varepsilon_0 = \arctan \frac{\sin 2\sigma_0}{1 + 2\lambda + \cos 2\sigma_0}. \tag{119}$$

This function is shown in Fig. 13 (left). We see that, as far as $\lambda \neq 0$, ε_0 tends to 0 when σ_0 tends to $\frac{\pi}{2}$, that is, when $\gamma \ll n$.

16.2 The Maximum Height of the Tide

The maximum height of the composite tide (once the transient phase is over) is a function of the semi-diurnal frequency and the relaxation factor. It is the value of the function H at its maximum:

$$H_{\max} = \frac{1}{2}R\overline{\epsilon}_\rho\sqrt{\lambda^2 + (1 + 2\lambda)\cos^2\sigma_0}. \tag{120}$$

The relative value of the maximum height of the actual tide is shown in Fig. 13 (right). In that figure, the unit is the maximum height of the static tide ($\frac{1}{2}R\overline{\epsilon}_\rho$). One may note that when $\sigma_0 \to \pi/2$ (i.e. $\gamma \ll \nu$), the height of the creep tide tends to zero and the maximum height of the geodetic tide is the maximum height of the added elastic tide: ($\frac{1}{2}\lambda R\overline{\epsilon}_\rho$). We emphasize that the frequency-dependent height of the tide was not considered in the majority of modern tide theories ([33, 34, 36, 41], etc.). In those theories, the Love theorem was used to calculate the potential of the tidally deformed body, and no attention was paid to the shape of the primary.

17 Dynamical Tide: The Maxwell Body Model

A theory of the dynamic tide virtually equivalent to the creep tide theory may be built using, instead of the creep differential equation (26), the constitutive equation of one Maxwell body [8] (Fig. 14):

$$\dot{Z} + \gamma Z = \gamma\rho + \lambda\dot{\rho}. \tag{121}$$

This equation defines the relationship between the radii vectors $Z = Z(\widehat{\theta}, \widehat{\varphi}, t)$ of the points in the actual surface of the body and the radii vectors $\rho = \rho(\widehat{\theta}, \widehat{\varphi}, t)$, of the corresponding points in the surface of the instantaneous static tide. The angles $\widehat{\theta}, \widehat{\varphi}$ are the co-latitude and longitude of the surface points. If we introduce the variable ζ using the same transformation used in the introduction of the geodetic lag (Eq. (117)), i.e. $\zeta = Z - \lambda\delta\rho$, the above equation becomes:

$$\dot{\zeta} + \gamma\zeta = (1-\lambda)\gamma\rho + \lambda\gamma R. \tag{122}$$

This equation shows the virtual equivalence of the creep tide and the Maxwell viscoelastic model [21]. The two theories are however different because of the factor $(1-\lambda)$ that multiplies $\gamma\rho$, and the constant $\lambda\gamma R$. If the tidal evolution theory is constructed using this equation, the results are the same that have been obtained in the Sects. 5–9, with just the factor $(1-\lambda)$ multiplying all results. The additional constant $\lambda\gamma R$ does not influence the solution because of the normalization to the mean radius of the body implicit in the construction of the potential. The transformation of the results thus obtained to those given by Correia et al. [8] is done making

$$\gamma = \frac{1}{\tau} \qquad \text{and} \qquad \lambda = \frac{\tau_e}{\tau},$$

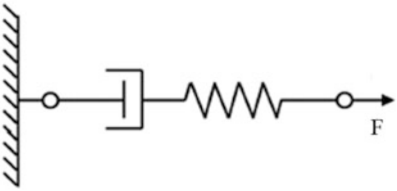


Fig. 14 String-dashpot model of a Maxwell body

where τ is the main relaxation time (inverse of the relaxation factor) and τ_e is the elastic relaxation time. It is worth noting that $\lambda < 1$. If $\lambda = 1$, the radial velocities of the two surfaces are equal when they coincide, and we have just a continuous damping to the equilibrium. It would be necessary to adopt a nonlinear approximation to obtain the tidal harmonics.

If we change the boundary conditions used in the study of the Navier-Stokes equation (Sect. 5.1) using the same transformation as above, $Z = \zeta + \lambda\delta\rho$, the first boundary condition used there becomes $V_r(\rho) = \lambda\dot{\rho}$ (instead of $V_r(\rho) = 0$). Consequently, the constant C_1 becomes $C_1 = \frac{w\rho}{6\eta} + \lambda\frac{\dot{\rho}}{\rho}$, (instead of $C_1 = \frac{w\rho}{6\eta}$) and the linear approximation of the Navier-Stokes equation becomes $V_r(Z) = \gamma(\rho - Z) + \lambda\dot{\rho}$, which is nothing but Eq. (121). The Maxwell constitutive equation is thus, it also, an approximated solution of the Navier-Stokes equation of a laminar flow (low Reynolds number) transversal to the surface. But with a set of boundary conditions forcing the surface of the dynamical tide to have a radial velocity dependent on that of the static tide, when both surfaces coincide.

18 Summary

This series of lectures was devoted to a synthetic presentation of planetary tide theories in the simple case of a homogeneous primary rotating around an axis orthogonal to the orbital plane of the companion. These restrictions are important, but planetary tide theories are complex, and to present them in an approximated, but simple setting, may be more adequate for a proper discussion of the main concepts without the complex algebraic developments of the full theory.

The theories considered here belong to a group of theories in which the starting point of the study is the figure of the tidally deformed primary and the orbital and rotational evolution are derived using standard physical laws. The central theory in this presentation is the creep tide theory, a first-principles theory where the tidal deformation of the body is calculated using the classical Navier-Stokes equation with boundary conditions such that the radial velocity of the dynamical tide vanishes when the deformations due to the dynamical tide and to the static tide coincide and the equation is simplified by the assumption that this motion is a low-Reynolds-number flow (laminar flows with no turbulence). The dynamic figure of the tidally deformed body is an ellipsoid delayed with respect to the motion of the companion. The delay, however, is not arbitrary, but determined by that equation. The pure hydrodynamical nature of the creep equation allows us to obtain a complete description of the orbital and rotational evolution and of the energy dissipation in the primary, but the results derived using it for the shape of free rotating stiff bodies, does not seem correct. In the only case for which a measurement is available, it gives a very different result (see Sect. 16). In order to obtain both the tidal evolution and the dynamic figure of the Earth, it is necessary to add an elastic component that converts the creep tide into a Maxwell model. This addition can be done either by adding a correction to the solution of the creep tide model or by using a Maxwell-like model ab initio. In both cases the tidal evolution is the same, the only difference being a factor $(1 - \lambda)$ that appears multiplying the results in the viscoelastic Maxwell model.

The Darwin's theory was also considered at length in these lectures. The difference between it and the hydrodynamical theories lies on the nature of the lags. While in the hydrodynamical theories, the lags are univocally determined by the solution of one first-order differential equation, in Darwin theories they are considered as arbitrary quantities introduced ad-hoc in the arguments of the equations of the static tide. From the formal point of view, the tidal evolution results obtained with the creep tide theory may be used to write the results of the Darwin theories, without having to derive every step of the theory again. It is enough to substitute the hydrodynamical lags σ_k by the arbitrary quantities ε_k and adopt the weak friction approximation hypothesis after which the ε_k are small quantities. The various rheologies used in Darwin theories are discussed, with an emphasis on the CTL (*constant time lag*) theories, in which the arbitrary lags are assumed proportional to the frequency of the considered tidal harmonic, and the CPL (*constant phase lag*) theories, where the lags are assumed to be frequency-independent. The CTL theories are equivalent to the creep tide theory when $1/\gamma \to 0$ (or $\gamma \gg n$). A special section is also devoted to Mignard's formulation of the CTL theories in closed form and its applications.

Acknowledgements This text follows closely the course given at the University of São Paulo in 2017 and it is a pleasure to thank Profs. J.A.S. de Lima and T.A.Michtchenko and Drs. H.A.Folonier and E.Andrade-Ines, for their many comments. Thanks are also due to G.O.Gomes for reading the entire manuscript and for the suggestions improving the clarity. The investigation on the creep tide theory is funded by the National Research Council, CNPq, grant 302742/2015-8, by FAPESP, grants 2014/13407-4, 2016/13750-6 and 2017/10072-0 and by the INCT Inespaço procs. FAPESP 2008/57866-1 and CNPq 574004/2008-4.

References

1. Alexander, M.E.: The weak friction approximation and tidal evolution in close binary systems. Astrophys. Space Sci. **23**, 459–510 (1973)
2. Barnes, R.: Tidal locking of habitable exoplanets. Celest. Mech. Dyn. Astron. **129**(4), 509–536 (2017)
3. Beutler, G.: Methods of Celestial Mechanics. Springer, Berlin (2005)
4. Brouwer, D., Clemence, M.: Methods of Celestial Mechanics. Academic Press, New York (1961)
5. Carone, L.: Tidal interactions of short-period extrasolar transit planets with their host stars: constraining the elusive stellar tidal dissipation factor. Dissertation, Universität zu Köln (2012)
6. Cayley, A.: Tables of developments of functions in the theory of elliptic motion. Mem. R. Astron. Soc. **29**, 191–306 (1861)
7. Chandrasekhar, S.: Ellipsoidal Figures of Equilibrium. Yale University Press, New Haven (1969). Chap. VIII
8. Correia, A.C.M., Boué, G., Laskar, J., Rodríguez, A.: Deformation and tidal evolution of close-in planets and satellites using a Maxwell viscoelastic rheology. Astron. Astrophys. **571**, A50 (2014)

9. Darwin, G.H.: On the influence of geological changes on the Earth's axis of rotation. Philos. Trans. **167**, 271–312 (1877)
10. Darwin, G.H.: On the bodily tides of viscous and semi-elastic spheroids and on the ocean tides upon a yielding nucleus. Philos. Trans. **170**, 1–35 (1879). Repr. Scientific Papers Vol. II, Cambridge, 1908
11. Darwin, G.H.: On the secular change in the elements of the orbit of a satellite revolving about a tidally distorted planet. Philos. Trans. **171**, 713–891 (1880). Repr. Scientific Papers Vol. II, Cambridge, 1908
12. Dobbs-Dixon, I., Lin, D.N.C., Mardling, R.A.: Spin-orbit evolution of short-period Planets. Astrophys. J. **610**, 464–476 (2004)
13. Efroimsky, M.: Tidal dissipation compared to seismic dissipation: in small bodies, Earths, and super-Earths. Astrophys. J. **746**, 150 (2012)
14. Efroimsky, M.: Bodily tides near spin-orbit resonances. Celest. Mech. Dyn. Astron. **112**, 283–330 (2012)
15. Efroimsky, M., Lainey, V.: Physics of bodily tides in terrestrial planets and the appropriate scales of dynamical evolution. J. Geophys. Res. **112**, E12003 (2007)
16. Efroimsky, M., Makarov, V.V.: Tidal dissipation in a homogeneous spherical body. I. Methods. Astrophys. J. **795**, 6 (2014)
17. Eggleton, P.P., Kiseleva, L.G., Hut, P.: The equilibrium tide model for tidal friction. Astrophys. J. **499**, 853–870 (1998)
18. Ferraz-Mello, S.: Earth tides in MacDonald's model (2013). arXiv: 1301.5617 astro-ph.EP
19. Ferraz-Mello, S.: Tidal synchronization of close-in satellites and exoplanets. A rheophysical approach. Celest. Mech. Dyn. Astron. **116**, 109–140 (2013). arXiv: 1204.3957
20. Ferraz-Mello, S.: Tidal synchronization of close-in satellites and exoplanets: II. Spin dynamics and extension to Mercury and exoplantes host stars. Celest. Mech. Dyn. Astr. **122**, 359–389 (2015). Errata: Celest. Mech. Dyn. Astr. **130**, 78, 20–21 (2018). arXiv: 1505.05384
21. Ferraz-Mello, S.: On large and small tidal lags. The virtual identity of two rheophysical theories. Astron. Astrophys. **579**, A97 (2015). arXiv.org/abs/1504.04609
22. Ferraz-Mello, S., Beaugé, C., Michtchenko, T.A.: Evolution of migrating planet pairs in resonance. Celest. Mech. Dyn. Astron. **87**, 99–112 (2003)
23. Ferraz-Mello, S., Rodríguez, A., Hussmann, H.: Tidal friction in close-in satellites and exoplanets. The Darwin theory re-visited. Celest. Mech. Dyn. Astron. **101**, 171-201 (2008). Errata: Celest. Mech. Dyn. Astron. **104**, 319–320 (2009). arXiv: 0712.1156
24. Ferraz-Mello, S., Grotta-Ragazzo, C., Ruiz, L.S.: Dissipative Forces on Celestial Mechanics, Chap. 3. Soc. Bras. Matem., Rio de Janeiro (2015)
25. Ferraz-Mello, S., Folonier, H., Tadeu dos Santos, M., Csizmadia, Sz., do Nascimento, J.D., Pätzold, M.: Interplay of tidal evolution and stellar wind braking in the rotation of stars hosting massive close-in planets. Astrophys. J. **807**, 78 (2015). arXiv: 1503.04369
26. Folonier, H.A.: Tide on differentiated planetary satellites. Application to Titan. Dr.Thesis, IAG/Univ. São Paulo (2016)
27. Folonier, H.A., Ferraz-Mello, S.: Tidal synchronization of an anelastic multi-layered satellite. Titan's synchronous rotation. Celest. Mech. Dyn. Astron. **129**, 359–396 (2017). arXiv: 1706.08603
28. Folonier, H.A., Ferraz-Mello, S., Andrade-Ines, E.: Tidal synchronization of close-in satellites and exoplanets: III. Tidal dissipation revisited and application to Enceladus. Celest. Mech. Dyn. Astron. **130**, 78 (2018). arXiv: 1707.09229v2
29. Folonier, H., Ferraz-Mello, S., Kholshevnikov, K.V.: The flattenings of the layers of rotating planets and satellites deformed by a tidal potential. Celest. Mech. Dyn. Astron. **122**, 183–198 (2015, online supplement). arXiv: 1503.08051
30. Goldreich, P.: On the eccentricity of satellite orbits in the Solar System. Mon. Not. R. Astron. Soc **126**, 257–268 (1963)
31. Happel, J., Brenner, H.: Low Reynolds Number Hydrodynamics. Kluwer, Dordrecht (1973)
32. Hut, P.: Tidal evolution in close binary systems. Astron. Astrophys. **99**, 126–140 (1981)

33. Jeffreys, H.: The effect of tidal friction on eccentricity and inclination. Mon. Not. R. Astron. Soc. **122**, 339–343 (1961)
34. Kaula, W.M.: Tidal dissipation by solid friction and the resulting orbital evolution. Rev. Geophys. **3**, 661–685 (1964)
35. Laskar, J.: Large scale chaos and marginal stability in the Solar System. Celest. Mech. Dyn. Astron. **64**, 115–162 (1996)
36. MacDonald, G.F.: Tidal friction. Rev. Geophys. **2**, 467–541 (1964)
37. Makarov, V.V., Efroimsky, M.: Tidal dissipation in a homogeneous spherical body. II. Three examples: Mercury, Io, and Kepler-10 b. Astrophys. J. **795**, 7 (2014)
38. Mardling, R.A., Lin, D.N.C.: On the survival of short-period terrestrial planets. Astrophys. J. **614**, 955–959 (2004)
39. Mardling, R.: Long-term tidal evolution of short-period planets with companions. Mon. Not. R. Astron. Soc. **382**, 1768–1790 (2007)
40. Melchior, P.: The Tides of the Planet Earth. Pergamon Press, Oxford (1983)
41. Mignard, F.: The evolution of the lunar orbit revisited - I. Moon and Planets **20**, 301–315 (1979)
42. Moulton, F.R.: An Introduction to Celestial Mechanics. Macmillan, New York (1914)
43. Ogilvie, G.I., Lin, D.N.C.: Tidal dissipation in rotating giant planets. Astrophys. J. **610**, 477–509 (2004)
44. Oswald, P.: Rheophysics: The Deformation and Flow of Matter. Cambridge University Press, Cambridge (2009)
45. Peale, S.J.: Origin and evolution of the natural satellites. Annu. Rev. Astron. Astrophys. **37**(1), 533–602 (1999)
46. Ragazzo, C., Ruiz, L.S.: Viscoelastic tides: models for use in Celestial Mechanics. Celest. Mech. Dyn. Astron. **128**, 19–59 (2017)
47. Ray, R.D., Eanes, R.J., Lemoine, F.G.: Constraints on energy dissipation in the Earth's body tide from satellite tracking and altimetry. Geophys. J. Int. **144**, 471–480 (2001)
48. Remus, F., Mathis, S., Zahn, J.P., Lainey, V.: The surface signature of the tidal dissipation of the core in a two-layer planet. Astron. Astrophys. **573**, A23 (2015)
49. Rodríguez, A., Ferraz-Mello, S., Michtchenko, T.A., Beaugé, C., Miloni, O.: Tidal decay and orbital circularization in close-in two-planet systems. Mon. Not. R. Astron. Soc. **415**, 2349–2358 (2011)
50. Singer, S.F.: The origin of the Moon and geophysical consequences. Geophys. J.R. Astron. Soc. **15**, 205–22 (1968)
51. Sommerfeld, A.: Lectures on Theoretical Physics, vol. 2. Mechanics of Deformable Bodies. Academic Press, New York (1950)
52. Taylor, P.A., Margot, J.-L.: Tidal evolution of close binary asteroid systems. Celest. Mech. Dyn. Astron. **108**, 315–338 (2010)
53. Tisserand, F.: Traité de Mécanique Céleste, tome II. Gauthier-Villars, Paris (1891)
54. Williams, J.G., Boggs, D.: Tides on the Moon: Theory and determination of dissipation. J. Geophys. Res. Planets **120**, 689–724 (2015)
55. Williams, J.G., Efroimsky, M.: Bodily tides near the 1:1 spin-orbit resonance. Correction to Goldreich's dynamical model. Celest. Mech. Dyn. Astron. **114**, 387–414 (2012)
56. Yoder, C.F., Peale, S.J.: The tides of Io. Icarus **47**, 1–35 (1981)

Perturbation Methods in Celestial Mechanics

Antonio Giorgilli

The real trouble with this world of ours is not that it is an unreasonable world, nor even that it is a reasonable one. The commonest kind of trouble is that it is nearly reasonable, but not quite. Life is not an illogicality; yet it is a trap for logicians. It looks just a little more mathematical and regular than it is; its exactitude is obvious, but its inexactitude is hidden; its wildness lies in wait. (G. K. Chesterton)

Abstract A concise, not too technical account of the main results of perturbation theory is presented, paying particular attention to the mathematical development of the last 60 years, with the work of Kolmogorov on one hand and of Nekhoroshev on the other hand. The main theorems are recalled with the aim of providing some insight on the guiding ideas, but omitting most details of the proofs that can be found in the existing literature.

Keywords Perturbation theory · KAM theory · Nekhoroshev theory · Exponential stability · Superexponential stability

1 Ouverture

The present lectures are concerned with some fundamental results in the framework of perturbation theory, with particular emphasis on the long-standing problem of stability of the Solar System.

The ancient astronomy, starting (according to our current knowledge) with the tables collected by Caldean and Egyptian astronomers, has been based on the periodic character of the planetary motions. The same concept lies at the very basis of Greek astronomy and of the clever schemes of eccentrics, epicycles and equants

A. Giorgilli (✉)
Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy
e-mail: antonio.giorgilli@unimi.it

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_2

that we know mainly through the work of Ptolemy. In modern terms, the guiding idea of Greek astronomy is: *the motion of the planets is a composition of periods, which can be empirically determined through observations*.

The development of astronomy after Newton makes wide use of the concept of quasiperiodic motions; that is, epicycles represented in the modern form of Fourier series as introduced by Lagrange in the eighteenth century. The actual difference is that *the periods can be calculated on the basis of the theory of gravitation*.

Newton himself pointed out that the theory of gravitation raises the fundamental problem of *stability of the Solar System*: *the mutual attraction acting on a long time might modify the orbits of the planets until the whole system needs a restoration* [57]. On the other hand, after Poincaré's work we know that the dynamics of the planetary system is actually an elaborated and intriguing combination of order and chaos. The problem of stability of our Solar System, as well as of the extrasolar system that we are discovering, remains open.

1.1 Apology

My plan is to give a concise, not too technical account of the main results of perturbation theory, paying particular attention to the mathematical development of the last 60 years, with the work of Kolmogorov on one hand and of Nekhoroshev on the other hand. I should stress that the literature on the arguments discussed here is now so wide that a detailed account is actually unpractical, and a complete list of references would exhaust the available space. Therefore I will put severe restrictions on my approach. First, I will present a personal view of the mathematical development of some crucial results. It is unavoidably incomplete, since it reflects my personal limited experience. I present some theorems, but I will avoid most technical elements of the proofs, trying rather to put the accent on the guiding ideas—selected, as I have already said, on the basis of my experience. Detailed proofs may be found in the references, or elsewhere. Second, I will pay particular attention to explicit algorithms that may be actually worked out, possibly using appropriate tools of algebraic manipulation. This is a strong limit due to my personal belief that if one wants to exploit a mathematical model of our physical world then he is bound to make his calculation feasible and to extend his work until an applicable result is found. In this spirit, an existence theorem is a beautiful and often priceless step, but one should not stop there.

1.2 The Dawn of Perturbation Theory

It is known that Kepler discovered the elliptic form of planetary orbits while working out the calculation of the *Tabulæ Rudolphinæ*. But he also wanted to compare the results of his calculations with observations performed in the past, and available to him. He discovered deviations from the elliptic motions which were particularly evident for Jupiter and Saturn (see [44]). In the introduction to the *Tabulæ* he

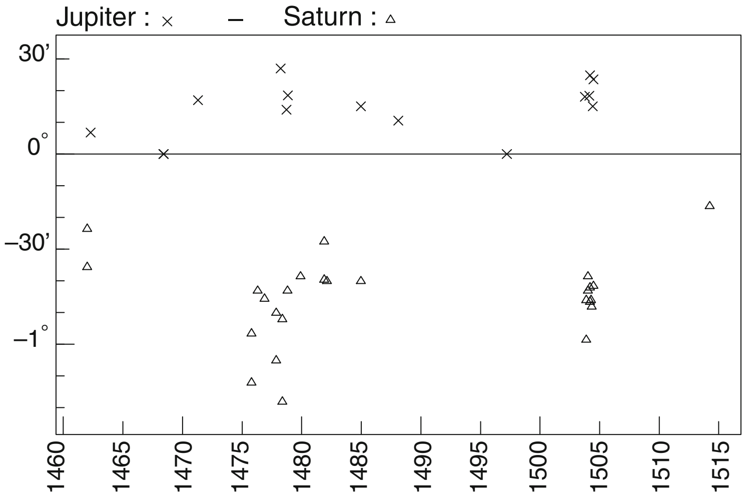


Fig. 1 Comparison between the calculation via the *Rudolphine tables* and the observations by Regiomontanus (Johannes Müller der Könisberg) and Bernhard Walther, between 1461 and 1514. (figure from [22])

promised to explain in a booklet how the elliptic orbits should be corrected by introducing *secular equations*, namely periodic changes in the elements of the orbits to be determined through observations *over many centuries*. He also wrote a preliminary draft of the booklet, that was never published until 1860, when it was included in the collection of Kepler's works [36].

A synthesis of Kepler's calculations concerning Jupiter and Saturn is reported in Fig. 1. The difference between the observed longitude and the calculated one is represented. The data exhibit a wide dispersion, due to errors in visual observations, but it is evident that Jupiter seems to accelerate, while Saturn seems to slow down. Later this phenomenon has been named *the great inequality*. Kepler could not figure out the secular equations to be introduced. He only mentioned in letters to some friends that he had discovered the deviation: the contents of his note remained unknown for a couple of centuries. Later, the increase both in number and in precision of astronomical observations confirmed the phenomenon, thus opening a challenging question.

A first attempt to identify a secular correction was made by Halley. Pragmatically, he introduced in his new tables a correction of the semimajor axes of Jupiter and Saturn which was *linear* in time, thus claiming that he had been able to make precise predictions over an interval of 6000 years before and after 1700. After Halley, the adjective "secular" was intended to mean "linear in time".

A few decades later the question was raised whether the theory of gravitation of Newton could explain the observed deviation. After some attempts due to Euler, who had the great merit of creating the bases of perturbation theory, Lagrange succeeded in developing his theory of secular motions for the nodes of the planetary orbits [38, 40, 41], soon after extended also to eccentricities by Laplace [16]. Meanwhile, Lagrange had announced his proof of stability of the planetary system [39]: *the time evolution of the semimajor axes does not contain secular terms in the approximation of the first order in the masses.* Thus, secular terms in Halley's sense could be excluded for the evolution of the semimajor axes. In 1786 Laplace succeeded in answering the question concerning the great inequality of Jupiter and Saturn, showing that it is due to the closeness of the periods to the 5:2 resonance [17]. The skeleton of perturbation theory was well established.

Most of the work of astronomers during the nineteenth century has been devoted to proving the complete validity of the stability result of Lagrange. It is convenient to refer to the classification of terms in perturbation expansions due to Poincaré [59]. One may find three different kinds of terms: pure trigonometric terms of the form $\exp(i\omega t)$, so that the time appears as argument of a trigonometric function; pure secular terms, with powers of time such as t^s; mixed secular terms $t^s \exp(i\omega t)$, namely trigonometric terms with powers of time as coefficients. The question may be formulated as: *prove that there are neither pure nor mixed secular terms in the evolution of the semimajor axes, at every order in the masses.*

The dream was soon dissolved, for in 1809 Poisson found that at second order in the masses there are mixed secular terms, but *no pure secular terms* [60]. One was left with the question whether pure secular terms could appear at higher order. A few such terms were found at third order by Spiru Haretu [32, 33]. A few years later methods that could produce pure trigonometric expansion were developed by Lindstedt [45] and Gyldén [31].

1.3 The Hurricane

In 1885 the Swedish Academy announced a prize for the 60-th birthday of King Oscar. One of the questions proposed can be stated in short as: *Write the solutions of the planetary problem as series uniformly convergent for all times (possibly pure trigonometric series).* The prize was awarded to Poincaré who actually found (among a lot of new results) that there are plenty of solutions which are not quasiperiodic, and so can not be written as trigonometric expansion. In particular he pointed out the existence of *asymptotic solutions*, and in a corrected version of his memoir he discovered the existence of *homoclinic orbits*, which generate a chaotic behaviour.

A few years later Poincaré published his treatise *Les méthodes nouvelles de le Mécanique Céleste* [58]. In sect. 13 he formulates the *general problem of dynamics*:

Investigate the dynamics of a canonical system with Hamiltonian

$$H(p,q) = H_0(p) + \varepsilon H_1(p,q) + \varepsilon^2 H_2(p,q) + \dots\ , \qquad p \in \mathcal{G} \subset \mathbb{R}^n,\ q \in \mathbb{T}^n$$

> *where $p \in \mathcal{G} \subset \mathbb{R}^n$, an open subset, and $q \in \mathbb{T}^n$ are action–angle variables. The Hamiltonian is assumed to be holomorphic in p, q and expanded in convergent power series for small ε.*

This is the problem that I will discuss in the present lectures.

2 Integrability and Non Integrability

The discussion may start with the classical concept of *integrability by quadratures*: the solution of a system of differential equations should be written in terms of algebraic operations, including inversion of functions, and of integrals of known functions.

Nowadays it is more common to restrict attention to systems that can be written in action-angle variables, say I, φ, as in the general problem of dynamics, and with a Hamiltonian $H(I)$ independent of the angles.

2.1 *The Theorems of Liouville and of Arnold–Jost*

A general framework for integrability has been provided by Liouville [46] and elaborated in a more geometric form by Arnold [4] and Jost [35]. One needs to introduce the concept of *complete involution system*. Two functions $f(p,q)$ and $g(p,q)$ are said to be in involution in case their Poisson bracket satisfies $\{f,g\} = 0$. A complete involution system on a $2n$-dimensional phase space (n degrees of freedom) is a set $\Phi_1(p,q),\ldots,\Phi_n(p,q)$ of n functions which are independent and in involution.

Theorem 1 *Let the Hamiltonian system $H(q,p)$ possess a complete involution system of first integrals $\Phi_1(q,p)\ldots,\Phi_n(q,p)$. Then the following statements hold true.*

(i) The system is integrable by quadratures (Liouville [46]).
(ii) Let the invariant manifold defined by $\Phi_1(p,q) = c_1,\ldots,\Phi_n(q,p) = c_n$ possess a compact and connected component Σ_c. Then in a neighbourhood of Σ_c there are action-angle variables I,φ such that the Hamiltonian depends only on the actions I, i.e., $H = H(I)$ (Arnold [4] and Jost [35]).

The dynamics of an integrable system $H = H(I)$ is described as follows: *the phase space is foliated into invariant tori, carrying a Kronecker flow with frequencies $\omega(I) = \frac{\partial H}{\partial I}$*. If the Hamiltonian $H(I)$ is non degenerate, i.e. if

$$\det\left(\frac{\partial^2 H}{\partial I_j \partial I_k}\right) \neq 0\ ,$$

then there are no independent first integrals depending on the angles φ.

Among the classical examples of integrable systems one finds: the Kepler problem, for which a convenient set of action-angle variables has been introduced by Delaunay [14]; the free rigid body and the Lagrange top, with action-angle variables introduced by Andoyer [1].

2.2 The Non Integrability Theorem of Poincaré

The first negative result of Poincaré states that the general problem of dynamics is generically non integrable. Restoring the notation $p \in \mathcal{G} \subset \mathbb{R}^n$ and $q \in \mathbb{T}^n$ for action-angle variables, the perturbation $H_1(p, q)$ may be expanded in Fourier series as

$$H_1(p, q) = \sum_{k \in \mathbb{Z}^n} h_k(p) e^{i\langle k, q\rangle} .$$

Theorem 2 *Let the Hamiltonian $H(p, q) = H_0(p) + \varepsilon H_1(p, q)$ satisfy the following hypotheses:*

(i) nondegeneracy, i.e.,

$$\det\left(\frac{\partial^2 H_0}{\partial p_j \partial p_k}\right) \neq 0 ;$$

(ii) genericity: no coefficient $h_k(p)$ of the Fourier expansion of $H_1(p, q)$ is identically zero on the manifold $\langle k, \omega(p)\rangle = 0$.

Then there is no analytic first integral independent of H.

It is worth giving a short outline of the proof, because it helps to understand how the problem of resonances shows up, taking the form of small divisors. The reader is referred to ch. VIII of [58] for a detailed exposition. The attempt is to construct a first integral expanded as $\Phi(p, q) = \Phi_0(p) + \varepsilon\Phi_1(p, q) + \ldots$ by looking for a solution of the equation $\{H, \Phi\} = 0$. Replacing the expansions in ε we obtain the recurrent system

$$\{H_0, \Phi_0\} = 0 , \quad \{H_0, \Phi_1\} = -\{H_1, \Phi_0\} , \quad \{H_0, \Phi_2\} = -\{H_1, \Phi_1\} , \ \ldots$$

The first equation is solved by any $\Phi(p)$ independent of the angles q, in view of non degeneration. The proof then proceeds in two steps. The first step consists in proving that if Φ is independent of H then one can find Φ_0 independent of H_0. This part requires a clever, delicate argument. The second step consists in proving that $\Phi_0(p)$ can not be independent of $H_0(p)$, as a consequence of the genericity hypothesis. In rough, short terms the argument proceeds as follows. Expand the second equation as

$$i\sum_k \langle k, \omega(p)\rangle \varphi_k(p) e^{i\langle k,q\rangle} = i\sum_k \left\langle k, \frac{\partial \Phi_0}{\partial p}\right\rangle h_k(p) e^{i\langle k,q\rangle} , \quad \omega(p) = \frac{\partial H_0}{\partial p} ,$$

with coefficients $\varphi_k(p)$ to be found. Therefore we must solve the infinite set of equations

$$\langle k, \omega(p)\rangle \varphi_k(p) = \left\langle k, \frac{\partial \Phi_0}{\partial p} \right\rangle h_k(p) \ , \quad 0 \neq k \in \mathbb{Z}^n \ .$$

On the resonant manifold $\langle k, \omega(p)\rangle = 0$ either case must occur:

$$\left\langle k, \frac{\partial \Phi_0}{\partial p} \right\rangle = 0 \quad \text{or} \quad h_k(p) = 0 \ .$$

Now, in view of the genericity condition we have $h_k(p) \neq 0$; therefore we must have $\left\langle k, \frac{\partial \Phi_0}{\partial p} \right\rangle = 0$. The conclusion follows by exploiting the fact that resonances are dense in $\mathcal{G}$, which forces the gradients of $\Phi_0(p)$ and $H_0(p)$ to be parallel in a dense set of points. Thus, one must conclude that Φ_0 can not be independent of H_0 .

The immediate consequence of Poincaré's theorem is that, generically, *the geometric structure of the invariant unperturbed tori does not persist under perturbation*, for the theorem of Liouville does not apply.

2.3 A Puzzling Example

The condition of genericity appears to be a too strong one. Poincaré was well aware of this fact, and he did discuss how to relax it, still keeping the validity of the result. Here I want to illustrate a puzzling example with the aim of showing how the condition of genericity, in same sense, is eventually recovered. At the same time the example suggests a way out of the difficulties raised by Poincaré.

Consider the Hamiltonian of a system of two coupled rotators $H = H_0 + \varepsilon H_1$ with

$$H_0 = \frac{1}{2}(p_1^2 + p_2^2) \ , \quad H_1 = \cos q_1 + \cos(q_1 - q_2) + \cos(q_1 + q_2) + \cos q_2 \ . \qquad (1)$$

It satisfies the non degeneracy condition, but it is definitely not generic, because it contains only a finite number of Fourier modes. Let us work out the construction of a first integral by choosing e.g., $\Phi_0(p) = p_1$, clearly independent of H_0. Using the complex representation of trigonometric functions calculate

$$\{H_1, p_1\} = \frac{i}{2} \left[\left(e^{iq_1} + e^{i(q_1 - q_2)} + e^{i(q_1 + q_2)} \right) - \text{c.c.} \right] \ ,$$

where c.c. stands for the complex conjugate. Therefore a solution for Φ_1 is found to be

$$\Phi_1 = -\frac{1}{2}\left[\left(\frac{e^{iq_1}}{p_1} + \frac{e^{i(q_1-q_2)}}{p_1-p_2} + \frac{e^{i(q_1+q_2)}}{p_1+p_2}\right) + \text{c.c.}\right] .$$

We remark immediately that the solution is well defined if we remove from the plane p_1, p_2 the resonant manifolds (actually straight lines) $p_1 = 0$, $p_1 - p_2 = 0$, $p_1 + p_2 = 0$. The remark applies, of course, for any perturbation which is a trigonometric polynomial of finite degree, provided we remove a finite number of resonant manifolds.

We now go to the next step. We must consider the equation at order ε^2, namely $\{H_0, \Phi_2\} = \{\Phi_1, H_1\}$. Without performing a complete calculation, let us focus our attention on the Fourier modes that are generated. The process is illustrated in Fig. 2, where the Fourier modes that appear in the functions Φ_s are represented for orders $s = 1, 2, 3$. The Poisson braket makes the exponentials to be multiplied; hence $\{H_1, \Phi_1\}$ contains new Fourier modes, including in particular

$$e^{i(2q_1-q_2)} \,,\; e^{i(2q_1+q_2)} \,,\; e^{i(q_1-2q_2)} \,,\; e^{i(q_1+2q_2)} \,,$$

which are not multiples of the previous ones. The generated modes are represented by grey squares in the figure for order 2. Therefore Φ_2 contains the new divisors $2p_1 - p_2$, $2p_1 + p_2$, $p_1 - 2p_2$, $p_1 + 2p_2$ and we must remove the additional resonant manifolds

$$2p_1 - p_2 = 0 \,,\; 2p_1 + p_2 = 0 \,,\; p_1 - 2p_2 = 0 \,,\; p_1 + 2p_2 = 0 \,.$$

At order ε^3 we get the new modes represented in the figure as open squares. With a moment's thought we realize that at order ε^s the right member $\{H_1, \Phi_{s-1}\}$ contains the Fourier modes $e^{i(k_1q_1+k_2q_2)}$ with $|k_1| + |k_2| \le 2s$; a finite number, but increasing with s. Therefore we must remove more and more resonant straight lines $k_1p_1 + k_2p_2 = 0$ with $|k_1| + |k_2| \le 2s$. We conclude that *for $s \to \infty$ we must remove a dense set of resonances, so that a first integral independent of H can not be constructed on an open domain, even formally.*

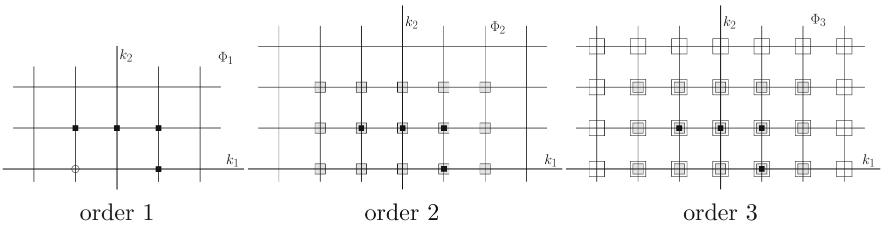


Fig. 2 Illustrating the propagation of Fourier modes through the process of construction of a first integral for the Hamiltonian (1). Only modes (k_1, k_2) with $k_2 \ge 0$ are included, because (k_1, k_2) and $-(k_1, k_2)$ actually represent the same mode

2.4 How to Proceed?

It is a fact that the non integrability theorem of Poincaré did not trouble the astronomers too much: the perturbation expansions had proven to be very useful during a couple of centuries. But a few mathematicians continued their fight with small divisors.

A first attempt, taken since the beginning of the twentieth century, has been to remove the hypothesis of non degeneration of the unperturbed Hamiltonian H_0, considering for instance the dynamics in a neighbourhood of an elliptic equilibrium. This problem will be discussed later, so let me put it aside for a while. I will rather focus the attention on two different approaches that have been developed in the second half of the twentieth century.

The first way has been opened by Kolmogorov in 1954 [37]: *Look only for a set of invariant tori which are characterized by a condition of strong non resonance, thus putting restrictions on the initial values of the actions.* The work of Kolmogorov has marked the beginning of what we call now *KAM theory*, the acronym standing for Kolmogorov–Arnold–Moser.

The second way has been proposed by Moser [53] and Littlewood [47, 48], and has been formulated in a general form by Nekhoroshev [55, 56]: *put restrictions on time, but consider initial data in open sets. Look for results valid over a finite but very long time.* These two approaches will be the backbone of the rest of the present lectures.

3 The Renaissance of Epicycles

This section is devoted to presenting the celebrated theorem of Kolmogorov on the persistence of quasi-periodic motions in nearly integrable systems. The theorem has been announced at the International Congress of Mathematicians at Amsterdam, in 1954. A sketch of the proof has been published in [37]. Kolmogorov gave a complete proof in a series of lectures, but it seems that the text has not been published (not in western countries, at least). The first published proofs are due to Moser [54] and Arnold [2].

3.1 The Normal Form of Kolmogorov

Consider the Hamiltonian in action-angle variables p, q

$$H(p, q) = \langle \omega, p \rangle + F(p, q) \ , \quad \omega \in \mathbb{R}^n \ .$$

Let us say that the Hamiltonian is in *normal form of Kolmogorov* in case $F(p,q)$ is at least quadratic in the actions p, i.e., $F(p,q) = \mathcal{O}(p^2)$. Write the canonical equations

$$\dot{q} = \omega + \mathcal{O}(p) \ , \quad \dot{p} = \mathcal{O}(p^2) \ .$$

and select initial data with $p(0) = 0$ and $q(0) = q_0$ arbitrary. Then we get the solution

$$q(t) = \omega t + q_0 \ , \quad p(t) = 0 \ ,$$

i.e., *the torus $p = 0$ is invariant and carries a Kronecker flow with frequencies ω.*

The idea of Kolmogorov is to cast a Hamiltonian into the normal form above in a neighbourhood of a non resonant unperturbed torus. Here I shall sketch the main ideas by closely following the short note of Kolmogorov [37]. Later I will do some recasting in order to illustrate a constructive method of proof.

Let us consider a Hamiltonian

$$H(p,q) = h(p) + f(p,q)$$

with $h(p)$ non degenerate and $f(p,q)$ small. The reader may want to add a parameter ε in front of $f(p,q)$ if this helps in tracking the smallness of some terms. However he or she will realize at some point that in Kolmogorov's scheme one must get rid of a perturbation parameter. We also assume that $h(p)$ is quadratic and $f(p,q)$ may be expanded as a Fourier series in the angles with coefficients at most quadratic in p. This simplifies the argument while keeping the crucial difficulties of the problem.

The dynamics of the unperturbed Hamiltonian $h(p)$ is quasi-periodic with frequencies $\omega(p) = \frac{\partial h}{\partial p}$. Let us select an initial condition p^* such that the corresponding frequencies $\omega(p^*)$ are non resonant, i.e., $\langle k, \omega\rangle \neq 0$ for $k \neq 0$. With a translation $p' = p - p^*$ the Hamiltonian takes the form, omitting primes and writing $\omega \in \mathbb{R}^n$ in place of $\omega(p^*)$,

$$H(p,q) = \eta + \langle \omega, p\rangle + \frac{1}{2}\big\langle \mathsf{C}(q)p, p\big\rangle + A(q) + \big\langle B(q), p\big\rangle \tag{2}$$

where terms of different degrees in p have been separated introducing the function $A(q)$, the vector function $B(q)$ and the symmetric matrix $\mathsf{C}(q)$, that can be calculated as

$$A(q) = f(p^*, q) \ , \quad B_j(q) = \frac{\partial f}{\partial p_j}(p^*, q) \ , \quad \mathsf{C}_{jk}(q) = \frac{\partial^2 h}{\partial p_j \partial p_k}(p^*) + \frac{\partial^2 f}{\partial p_j \partial p_k}(p^*, q) \ .$$

Here $A(q)$ and $B(q)$ are of the same order as f, while $\mathsf{C}(q)$ includes a small correction to the quadratic part of $h(p)$. The constant η may be ignored.

In a more general setting the Hamiltonian will contain polynomials of higher degrees in the actions, or even will be a power series in p. However, as I have already said, this is just a technical complication which does not add essential difficulties.

I come now to reformulating the method of Kolmogorov using the algorithm of Lie series in order to perform near the identity canonical transformations. A short reminder of the algorithm is given in Appendix. The suggestion of Kolmogorov is to kill the unwanted parts $A(q)$ and $\langle B(q), p\rangle$ by using a near the identity generating function

$$\chi(p,q) = \langle \xi, q\rangle + X(q) + \langle Y(q), p\rangle\ , \quad (\xi \in \mathbb{R}^n)\ .$$

The transformation of coordinates is written as (with a minor abuse of notation)

$$\begin{aligned} q' &= \exp\bigl(L_\chi\bigr)q = q + Y(q) + \frac{1}{2}\left\langle \frac{\partial Y}{\partial q}, Y(q)\right\rangle + \dots\ , \\ p' &= \exp\bigl(L_\chi\bigr)p = p + \xi + \frac{\partial X}{\partial q} + \left\langle \frac{\partial Y}{\partial q}, p\right\rangle + \dots\ , \end{aligned}$$

the dots denoting higher order terms. Thus we have a small translation and deformation of the actions combined with a deformation of the angles.

Let us transform the Hamiltonian as $H' = \exp\bigl(L_\chi\bigr)H$. Here the Poisson bracket with $\langle\omega, p\rangle$ plays a special role, so let us introduce the notation $\partial_\omega = L_{\langle\omega,p\rangle}$. We get

$$\begin{aligned} H' =& \langle\omega, p\rangle + \frac{1}{2}\langle \mathsf{C}(q)p, p\rangle \\ &+ A(q) - \partial_\omega X + \langle\omega, \xi\rangle \\ &+ \bigl\langle B(q), p\bigr\rangle + \left\langle \mathsf{C}(q)p, \xi + \frac{\partial X}{\partial q}\right\rangle - \partial_\omega\langle Y(q), p\rangle + \dots \end{aligned}$$

Here the first line contains the part already in normal form, and the dots stand for smaller terms that are left unhandled, and must be removed later. The second and third lines contain the parts that should be cleared. To this end, ignoring the constant $\langle\omega, \xi\rangle$ in the Hamiltonian, we write the equations

$$\begin{aligned} A(q) - \partial_\omega X &= 0\ , \\ \overline{\mathsf{C}}\xi + \overline{B + \frac{\partial X}{\partial q}} &= 0\ , \\ B(q) + \mathsf{C}(q)\left(\xi + \frac{\partial X}{\partial q}\right) - \partial_\omega Y &= 0\ . \end{aligned} \tag{3}$$

Here the overline denotes the average with respect to the angles q, i.e., the term independent of q in the Fourier expansion of a function. The first and third equation (usually called homological equations) can be solved provided the average of the known term is zero (see next section). In the first equation the average is a constant that can be neglected. The second equation aims at determining the real vector ξ precisely in order to clear the average of the known term in the third equation. It may be solved provided the constant matrix $\overline{\mathsf{C}}$ is not degenerate, which is initially assured by non degeneracy of $h(p)$. The translation vector ξ keeps the frequency fixed.

Having determined the generating function we may perform the transformation and then rearrange the Hamiltonian in the same form as (2), namely

$$H' = \langle \omega, p\rangle + \frac{1}{2}\bigl\langle \mathsf{C}'(q)p, p\bigr\rangle + A'(q) + \bigl\langle B'(q), p\bigr\rangle$$

with $A'(q)$ and $B'(q)$ hopefully smaller than $A(q)$ and $B(q)$ and with a new symmetric matrix $\mathsf{C}'(q)$ which is a small correction of the previous one. Roughly, if we assume that $A(q)$ and $B(q)$ were of order ε, then we may expect $A'(q)$, $B'(q)$ and $\mathsf{C}'(q) - \mathsf{C}(q)$ to be of order ε^2.

Thus, the consistency of the procedure depends on the existence of the solution of Eq. (3). If so, then the procedure may be iterated in order to (hopefully) reduce the size of the unwanted terms to zero, thus giving the Hamiltonian the normal form of Kolmogorov.

3.2 Small Divisors and the Problem of Convergence

The problem of solving the homological equation can be stated in the general form: given a known function $\psi(p, q)$ with zero average, namely $\overline{\psi} = 0$, find χ such that $\partial_\omega \chi = \psi$. The actions p here are just parameters. The procedure is quite standard: we have already used it while discussing the non integrability result of Poincaré. Expand in Fourier series

$$\psi(p, q) = \sum_{0\neq k\in\mathbb{Z}^n} \psi_k(p) \exp\bigl(i\langle k, q\rangle\bigr)\ , \qquad \chi(p, q) = \sum_{k\in\mathbb{Z}^n} c_k(p) \exp\bigl(i\langle k, q\rangle\bigr)\ ,$$

with $\psi_k(p)$ known and $c_k(p)$ to be found. Calculate

$$\partial_\omega \chi = i\sum_k \langle k, \omega\rangle c_k(p) \exp\bigl(i\langle k, q\rangle\bigr)\ .$$

Therefore, assuming that the frequencies ω are non resonant, we get the formal solution with coefficients

$$c_k(p) = -i\frac{\psi_k(p)}{\langle k, \omega\rangle}\ .$$

Since the expressions $\langle k, \omega\rangle$ at the denominators may become arbitrarily small, we must introduce a suitable condition of *strong* non resonance. Kolmogorov actually used the *diophantine condition* already introduced by Siegel, namely

$$|\langle k, \omega\rangle| > \frac{\gamma}{|k|^\tau} \,, \qquad \gamma > 0 \,, \ \tau > n-1 \,.$$

It is known that such a condition is satisfied by a large set of frequencies, the complement having measure $\mathcal{O}(\gamma)$. The solution can be proved to be holomorphic on the basis of the following considerations, already made by Poincaré. If $\psi(p,q)$ is holomorphic, then the coefficients $\psi_k(p)$ decay exponentially, i.e., $\left|\psi_k(p)\right| \sim e^{-|k|\sigma}$ for some σ. Therefore one gets $\left|c_k(p)\right| \sim |k|^\tau e^{-|k|\sigma} \sim e^{-|k|\sigma'}$ with some $\sigma' < \sigma$. This shows that $\chi(p,q)$ is still holomorphic, thus making every single step of Kolmogorov to be formally consistent.

The problem now is that iterating the procedure we produce an *accumulation of small divisors*: at every step the coefficients gain a new small divisor, which makes convergence doubtful. Here comes the second idea of Kolmogorov. Do not use expansions in a parameter. Collect all contributions independent of and linear in p in a single pair of functions $A(q)$ and $\langle B(q), p\rangle$. In very rough heuristic terms this is what happens. Starting with functions of size ε and forgetting for a moment the contribution of small divisors the procedure reduces step by step the size of the unwanted terms to ε^2, ε^4, ε^8, $\ldots$; i.e., they decrease quadratically, as in Newton's method (as remarked by Kolmogorov himself). Such a strong decrease compensates the growth of the number of factors with small divisors, eventually assuring the convergence of the procedure. The latter heuristic argument was commonly used in the past, and often it has been synthetized in the words "quadratic method", "quadratic convergence", "Newton method", "superconvergence" and so on. A complete proof along the lines suggested by Kolmogorov may be found, e.g., in [5].

I will avoid here using the method of fast convergence for two reasons. The first one is that the procedure of Kolmogorov is not constructive, for dealing with infinite Fourier series is conceptually simple, but hardly practical: some form of truncation must be introduced. The second reason is that the fast convergence hides the actual process of accumulation of divisors: it just dominates it. My aim is instead to show that in some cases, including Kolmogorov's one, the divisors accumulate in a polite way.

3.3 The Formal Constructive Algorithm

I simplify again the discussion by considering a rather simple model: a system of coupled rotators as described by the Hamiltonian $H(p,q) = H_0(p) + \varepsilon H_1(p,q)$, where

$$H_0(p) = \frac{1}{2}\sum_{j=1}^{n} p_j^2 \,, \quad H_1(p,q) = \sum_{|k|\le K} c_k(p) e^{i\langle k,q\rangle} \,, \quad p \in \mathbb{R}^n \,, \ q \in \mathbb{T}^n \,, \tag{4}$$

with a fixed $K > 0$ and coefficients $c_k(p)$ that are polynomials of degree at most 2. The choice is made in order to reduce technicalities to a minimum (though there remain enough), but all the crucial difficulties of the problem are accounted for. The extension to the general case is matter of not being scared by long and boring calculations.

The aim is to construct an infinite sequence $H^{(0)}(p,q)$, $H^{(1)}(p,q)$, $H^{(2)}(p,q)$, … of Hamiltonians, with $H^{(0)}$ coinciding with H in (4), which after r steps of normalization turn out to be written in the general form

$$H^{(r)} = \omega\cdot p + \sum_{s=0}^{r} h_s(p,q) + \sum_{s>r} \varepsilon^s \left[A_s^{(r)}(q) + B_s^{(r)}(p,q) + C_s^{(r)}(p,q) \right] , \tag{5}$$

where $H^{(r)}(p,q)$ is in Kolmogorov's normal form up to order r. Here $h_1(p,q), \ldots, h_r(p,q)$ are quadratic in p, so that they are in normal form, and do not change after step r. Moreover: (i) $A_s^{(r)}(q)$ is independent of p; (ii) $B_s^{(r)}(p,q)$ is linear in p; (iii) $C_s^{(r)}(p,q)$ is a quadratic polynomial in p; (iv) $A_s^{(r)}(q)$, $B_s^{(r)}(p,q)$ and $C_s^{(r)}(p,q)$ are trigonometric polynomials of degree sK in q, where K is the degree of H_1 in the original Hamiltonian. The algorithm should preserve at every step the properties (i)–(iv) above.

Some remarks are mandatory here concerning the simplifications introduced in the model. Adding a factor ε^s to every function with lower label s the reader will immediately realize that we are actually working with an ε-expansion, as it was customary in the past in perturbation theory. However, a generic perturbation will not fulfill the requests of being at most quadratic and a trigonometric polynomial of finite degree K. With a sufficient amount of patience the reader may see that adding further powers of p or even an infinite series is essentially harmless. The expansion in infinite trigonometric series is definitely more puzzling, for we can not deal explicitly with an infinite number of terms. This problem has been also pointed out by Poincaré (see [58], Ch. XIII, § 147), who suggested the way out. We may exploit the fact that the size of the coefficients of the Fourier expansion of a holomorphic function decreases exponentially with the degree. Therefore we may choose a truncation parameter $K > 0$ and expand the Hamiltonian in the requested form (5) by splitting the series into trigonometric polynomials of the requested order. A naive argument would lead to setting $K \sim -\log\varepsilon$, the perturbation parameter. It is remarkable indeed that the best choice is to set K to a constant independent of ε, which does not need to be a large one; e.g., setting $K = 1/\sigma$ is often enough. This makes the expansion in a parameter unpractical, but everything works fine if one pays attention not to powers of ε, but to the size of the various terms, in some norm.

The normalization process is worked out with a minor recasting of the method of Kolmogorov. At every step r we apply a first canonical transformation with generating function $\chi_1^{(r)}(q) = X^{(r)}(q) + \langle\xi^{(r)}, q\rangle$, followed by a second transformation with generating function $\chi_2^{(r)}(p, q) = \langle Y^{(r)}(q), p\rangle$.

The explicit constructive algorithm for a single step is presented in Table 1. Assuming that $r - 1$ steps have been performed, so that the Hamiltonian $H^{(r-1)}(p, q)$ has the wanted form (5) with $r - 1$ in place of r, we construct in sequence the new Hamiltonians

$$\hat{H}^{(r)} = \exp\bigl(L_{\chi_1^{(r)}}\bigr)H^{(r-1)} , \quad H^{(r)} = \exp\bigl(L_{\chi_2^{(r)}}\bigr)\hat{H}^{(r)} ,$$

the first one being an intermediate Hamiltonian, and the second one being in normal form up to order r. The generating functions are determined by solving a pair of homological equations, which is possible in view of the non resonance condition. All functions entering the transformed Hamiltonians are explicitly expressed in terms of Lie derivatives, as the reader may easily check. The Hamiltonian remains quadratic because the action of the Lie derivative $L_{\chi_1^{(r)}}$ decreases the degree in p by one, while $L_{\chi_2^{(r)}}$ leaves it unchanged. As to the trigonometric degree, the rules stated for the Hamiltonian are respected because the homological equation does not change it, so that $\chi_1^{(r)}$ and $\chi_2^{(r)}$ have degree rK, and if, say, f_s has degree sK then $L_{\chi_j^{(r)}} f_s$ clearly has degree $(s + r)K$. With these remarks, the reader should be able to check that the algorithm is actually applicable, so that the construction of the normal form is formally consistent. The challenge now is: *Prove that the sequence of Hamiltonians $H^{(r)}$ in normal form up to order r converges to a holomorphic Hamiltonian, $H^{(\infty)}$ say, in Kolmogorov's normal form.*

3.4 Analytical Estimates

I come now to the crucial problem that has challenged mathematicians for a couple of centuries: the accumulation of small divisors. The argument makes essential use of a real, non increasing sequence $\{\alpha_r\}_{r\geq 0}$ defined as

$$\alpha_0 = 1 , \quad \alpha_r = \min\Bigl(1, \min_{0<|k|\leq rK} \bigl|\langle k, \omega\rangle\bigr|\Bigr) . \tag{6}$$

That is, α_r is the smallest divisor that may appear in the solution of the homological equation for the generating functions $\chi_1^{(r)}$ and $\chi_2^{(r)}$ at step r of the normalization process. If the frequencies are non resonant then the sequence has zero limit for $r \to \infty$.

Let us introduce a convenient norm adapted to our case as follows. For a homogeneous polynomial of degree s in the actions p write (in multi-index notation) $g = \sum_{|j|=s} g_j p^j$, and define its norm as $\|g\| = \sum_{|j|=s} |g_j|$, namely the sum of

Table 1 The formal constructive algorithm for Kolmogorov's normal form

- Equations for the generating functions $\chi_1^{(r)} = X^{(r)} + \langle \xi^{(r)}, q\rangle$ and $\chi_2^{(r)} = \langle Y^{(r)}(q), p\rangle$:
  $\partial_\omega X^{(r)} - A_r^{(r-1)} = 0\,, \quad \langle \xi^{(r)}, p\rangle = \overline{B_r^{(r-1)}} = 0\,,$

$$\partial_\omega \chi_2^{(r)} - \hat{B}_r^{(r)} = 0\,, \quad \hat{B}_r^{(r)} = \left\langle \frac{\partial X^{(r)}}{\partial q}, p \right\rangle + B_r^{(r-1)} - \overline{B_r^{(r-1)}}\,.$$

- Intermediate Hamiltonian $\hat{H}^{(r)} = \exp\bigl(L_{\chi_1}\bigr) H^{(r-1)}$:
  $\hat{A}_r^{(r)} = 0$

$$\hat{A}_s^{(r)} = \begin{cases} A_s^{(r-1)}\,, & r < s < 2r\,; \\ \frac{1}{2} L^2_{\chi_1^{(r)}} h_{s-2r} + L_{\chi_1^{(r)}} B_{s-r}^{(r-1)} + A_s^{(r-1)}\,, & 2r \le s < 3r\,; \\ \frac{1}{2} L^2_{\chi_1^{(r)}} C_{s-2r}^{(r-1)} + L_{\chi_1^{(r)}} B_{s-r}^{(r-1)} + A_s^{(r-1)}\,, & s \ge 3r\,. \end{cases}$$

$$\hat{B}_s^{(r)} = \begin{cases} L_{\chi_1^{(r)}} h_{s-r} + B_s^{(r-1)}\,, & r \le s < 2r\,; \\ L_{\chi_1^{(r)}} C_{s-r}^{(r-1)} + B_s^{(r-1)}\,, & s \ge 2r\,. \end{cases}$$

- Transformed Hamiltonian $H^{(r)} = \exp\bigl(L_{\chi_2^{(r)}} \hat{H}^{(r)}\bigr)$ (set $s = kr + m$ with $0 \le m < k$):
  $h_r = L_{\chi_2^{(r)}} h_0 + C_r^{(r)}\,.$

$$A_s^{(r)} = \sum_{j=0}^{k-1} \frac{1}{j!} L^j_{\chi_2^{(r)}} \hat{A}_{s-jr}^{(r)}\,, \quad s > r\,.$$

$$B_s^{(r)} = \begin{cases} \frac{k-1}{k!} L^{k-1}_{\chi_2^{(r)}} \hat{B}_r^{(r)} + \sum_{j=0}^{k-2} \frac{1}{j!} L^j_{\chi_2^{(r)}} \hat{B}_{s-jr}^{(r)}\,, & k \ge 2\,,\ m = 0\,; \\ \sum_{j=0}^{k-1} \frac{1}{j!} L^j_{\chi_2^{(r)}} \hat{B}_{s-jr}^{(r)}\,, & k \ge 1\,,\ m \ne 0\,. \end{cases}$$

$$C_s^{(r)} = \frac{1}{k!} L^k_{\chi_2^{(r)}} h_m + \sum_{j=0}^{k-1} \frac{1}{j!} L^j_{\chi_2^{(r)}} C_{s-jr}^{(r)}, \qquad s > r\,.$$

the absolute values of the coefficients. For a trigonometric polynomial $f(p,q) = \prod_k f_k(p) e^{i\langle k,q\rangle}$ with coefficients $f_k(p)$ that are homogeneous polynomials we define a norm parameterized by $\sigma > 0$ as

$$\|f\|_\sigma = \sum_k \|f_k\|\, e^{|k|\sigma}\,.$$

The choice of the parameter σ is rather arbitrary for trigonometric polynomials, as considered here. For a real analytic function σ it is related to the width of a complex strip $\mathbb{T}^n_\sigma$ (as defined by (22), see Appendix) where the function is holomorphic and bounded.

Recalling that the algorithm of Kolmogorov normal form uses Lie derivatives and homological equations we need to know how these operations affect the norms. We have the following estimates.

Table 2 Quantitative estimates for the algorithm for Kolmogorov's normal form

- Generating functions $\chi_1^{(r)} = X^{(r)} + \langle \xi^{(r)}, q \rangle$ and $\chi_2^{(r)} = \langle Y^{(r)}(q), p \rangle$:
$$\|X_r\|_{(1-d_{r-1})\sigma} \le \frac{1}{\alpha_r}\|A_r^{(r-1)}\|_{(1-d_{r-1})\sigma}\ , \quad \left|\xi_{r,j}\right| \le \left\|\overline{B_r^{(r-1)}}\right\|_{(1-d_{r-1})\sigma}$$
$$\|\chi_2^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma} \le \frac{1}{\alpha_r}\|\hat{B}_r^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma}$$
- Intermediate Hamiltonian $\hat{H}^{(r)} = \exp\left(L_{\chi_1}\right)H^{(r-1)}$. Set
$$G_{r,1} = \frac{2e}{\sigma}\left(\|A_r^{(r-1)}\|_{(1-d_{r-1})\sigma} + \alpha_r\delta_r\sigma\left\|\overline{B_r^{(r-1)}}\right\|_{1-d_{r-1}}\right) .$$
For $r < s < 2r$, $2r \le s < 3r$ and $s \ge 3r$, respectively, get
$$\|A_s^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma} \le$$
$$\begin{cases} \|A_s^{(r)}\|_{(1-d_{r-1})\sigma}\ ; \\ \left(\frac{G_{r,1}}{\delta_r\alpha_r}\right)^2 \|h_{s-2r}\|_{(1-d_{s-2r})\sigma} + \frac{G_{r,1}}{\delta_r\alpha_r}\|B_{s-r}^{(r-1)}\|_{(1-d_{r-1})\sigma} + \|A_s^{(r)}\|_{(1-d_{r-1})\sigma}\ ; \\ \left(\frac{G_{r,1}}{\delta_r\alpha_r}\right)^2 \|C_{s-2r}^{(r-1)}\|_{(1-d_{s-2r})\sigma} + \frac{G_{r,1}}{\delta_r\alpha_r}\|B_{s-r}^{(r-1)}\|_{(1-d_{r-1})\sigma} + \|A_s^{(r)}\|_{(1-d_{r-1})\sigma}\ ; \end{cases}$$
For $r \le s < 2r$ and $s \ge 2r$, respectively, get
$$\|\hat{B}_s^{(r)}\|_{(1-d_{r-1}-\delta_r)} \le$$
$$\begin{cases} \frac{G_{r,1}}{\delta_r\alpha_r}\|h_{s-r}\|_{(1-d_{s-r})\sigma} + \|B_s^{(r-1)}\|_{(1-d_{r-1})\sigma}\ ; \\ \|\hat{B}_s^{(r)}\|_{(1-d_{r-1}-\delta_r)} \le \frac{G_{r,1}}{\delta_r\alpha_r}\|C_{s-r}^{(r-1)}\|_{(1-d_{s-r})\sigma} + \|B_s^{(r-1)}\|_{(1-d_{r-1})\sigma}\ . \end{cases}$$
- Transformed Hamiltonian $H^{(r)} = \exp\left(L_{\chi_2^{(r)}}\hat{H}^{(r)}\right)$. Set $G_{r,2} = \frac{3}{\sigma}\|\hat{B}_r^{(r)}\|_{(1-d_{r-1}-\delta_r)}$.
For $s \ge r$ get
$$\|h_r\|_{(1-d_r)\varrho,\sigma} \le \frac{G_{r,2}}{\delta_r\alpha_r}\|h_0\|_\sigma + \|C_r^{(r)}\|_{(1-d_{r-1})\sigma}\ .$$
$$\|A_s^{(r)}\|_{(1-d_r)\varrho,\sigma} \le \sum_{j=0}^{k-1}\left(\frac{G_{r,2}}{\delta_r\alpha_r}\right)^j \|\hat{A}_{s-jr}^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma}\ ;$$
$$\|B_s^{(r)}\|_{(1-d_r)\varrho,\sigma} \le \sum_{j=0}^{k-1}\left(\frac{G_{r,2}}{\delta_r\alpha_r}\right)^j \|\hat{B}_{s-jr}^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma}\ ;$$
$$\|C_s^{(r)}\|_{(1-d_r)\varrho,\sigma} \le \frac{G_{r,2}}{\delta_r\alpha_r}\|h_m\|_{(1-d_{r-m})\sigma} + \sum_{j=0}^{k-1}\left(\frac{G_{r,2}}{\delta_r\alpha_r}\right)^j \|\hat{C}_{s-jr}^{(r)}\|_{(1-d_{r-1}-\delta_r)\sigma};$$

(i) *Let*

$$\psi^{(r)} = \sum_{0<|k|\le rK} \psi_k(p)e^{i\langle k,q\rangle}$$

be a trigonometric polynomial of degree rK. *Then the zero-averaged solution of the homological equation* $\partial_\omega\chi^{(r)} = \psi^{(r)}$ *satisfies*

$$\left\|\chi^{(r)}\right\|_\sigma \le \frac{1}{\alpha_r}\left\|\psi^{(r)}\right\|_\sigma\ .$$

(ii) *The action of Lie derivatives is estimated by the inequalities*

$$\left\| \frac{1}{s!} L^s_{\chi_1^{(r)}} f \right\|_{(1-d)\sigma} \le \left(\frac{2e\|\chi_1^{(r)}\|_\sigma}{d\sigma} \right)^s \|f\|_\sigma \ ,$$

$$\left\| \frac{1}{s!} L^s_{\chi_2^{(r)}} f \right\|_{(1-d)\sigma} \le \left(\frac{3e\|\chi_2^{(r)}\|_\sigma}{d\sigma} \right)^s \|f\|_\sigma \ .$$

where $0 < d < 1$.

The reader will remark that the estimate requires a reduction of the value of σ similar to the restriction of domains illustrated in Appendix (it is the same thing, indeed).

Applying the estimates above to the algorithm for the Kolmogorov normal form is a boring but straightforward matter: replace every operation in the recurrent formulæ of Table 1 with the corresponding estimate for the norm. The result is summarized in Table 2. There is a point to be carefully taken into account. At every step we need a restriction of σ parameterized by an increasing positive sequence d_r, depending on the step r. But the sequence must have a finite limit $d < 1$. The sequence is arbitrary, so let us set $d_0 = 0$ and $d_r = 2(\delta_1 + \ldots + \delta_r)$ with

$$\delta_r = \frac{1}{\pi^2} \cdot \frac{1}{r^2} \ , \quad \sum_{r \ge 1} \delta_r = \frac{1}{6} \ .$$

A remark is mandatory. Lie derivatives introduce divisors δ_r which are small, and may affect convergence as well as the small divisors α_r , but we shall see that they always appear as products $\beta_r = \delta_r \alpha_r$, thus we shall control all divisors with the same method.

Without entering a very technical discussion, let me point out the common structure of all estimates. It is immediately seen that the norm of every function is bounded by a sum of different terms. Moreover:

(i) Every term comes either from a function of previous order or from a (possibly multiple) Lie derivative of a previous function.

(ii) A constant factor $G_{r,1}$ or $G_{r,2}$ coming from the r-independent constants in the estimates is associated to every Lie derivative $L_{\chi_2^{(r)}}$ or $L_{\chi_2^{(r)}}$, respectively.

(iii) Every factor $G_{r,1}$ or $G_{r,2}$ comes paired with a divisor β_r : the small divisor α_r and the restriction δ_r, that always appear in pairs.

The suggestion is to look for an uniform estimate for every term in the sums on the right hand sides, letting aside for the moment the problem of estimating the sums. It is expected that every term at order r is bounded by an expression such as $b\eta^{r-1}\bigl(\prod_\ell \beta_\ell\bigr)^{-1}$, with positive constants b, η, the product running over a set of indices to be determined. We are thus led to focus attention to the divisors in the product $\prod_\ell \beta_\ell$. Better, the suggestion is: *forget the actual values of the divisors; pay attention only to the indices*. Indeed the indices control the process of *accumulation of small divisors.*

3.5 The Game of Small Divisors

Let me first get rid of a naive argument that leads to a pessimistic conclusion. It seems that at every step a new divisor β_r is added:

$$\begin{array}{llll}
\chi_1^{(1)} \ \text{has denominator} & \beta_1 : & \chi_2^{(1)} \ \text{has denominator} & \beta_1^2 \ ; \\
\chi_1^{(2)} \ \text{has denominator} & \beta_1^2\beta_2 : & \chi_2^{(2)} \ \text{has denominator} & \beta_1^2\beta_2^2 \ ; \\
\chi_1^{(3)} \ \text{has denominator} & \beta_1^2\beta_2^2\beta_3 : & \chi_2^{(3)} \ \text{has denominator} & \beta_1^2\beta_2^2\beta_3^2 \ ; \\
\quad \cdots\cdots\cdots & & & \\
\chi_1^{(r)} \ \text{has denominator} & \beta_1^2\ldots\beta_{r-1}^2\beta_r : & \chi_2^{(r)} \ \text{has denominator} & \beta_1^2\ldots\beta_r^2 \ ;
\end{array}$$

Now, if we assume the diophantine inequality $\alpha_r \sim 1/r^{\tau}$ and set $\delta_r \sim 1/r^2$ (our choice), then we get

$$\|\chi_1^{(r)}\| \sim (r!)^{2\tau+3} \ , \quad \|\chi_2^{(r)}\| \sim (r!)^{2\tau+4} \ .$$

The immediate conclusion is that the naive argument can not be used for proving convergence; the accumulation of small divisors seems to be explosive.

We are thus confronted with the question: *can we control the explosive behaviour of the divisors?* As I have already pointed out, the great revolutionary idea introduced by Kolmogorov is that an efficient control is provided by the quadratic convergence due to a method "similar to that of Newton" (in his own words). This allowed him to open a breach in a two centuries old problem. Without diminishing the enormous importance of the work of Kolmogorov, we are now able to understand that the accumulation of divisors is not so bad as instinctively expected.

3.6 The Kindness of Small Divisors

Looking at Tables 1 and 2 we should pay attention to the actual mechanism of accumulation of divisors. As I have already pointed out, it is better to concentrate on the indices of the divisors. A convenient method is to organize the indices in a list (repetitions are allowed). Here I shall illustrate the argument in a synthetic but hopefully complete way.

Let me first point out how divisors show up.

(i) The generating function $\chi_1^{(r)}$ adds a new divisor α_r to the existing ones in $A_r^{(r-1)}$.

(ii) The first term affected in $\hat{H}^{(r)}$ is $\hat{B}_r^{(r)}$, by addition of $L_{\chi_1^{(r)}}h_0$ with a divisor β_r .

(iii) The latter term enters $\chi_2^{(r)}$, thus adding a further divisor α_r , to be promoted to β_r by Lie derivatives.

Now I come to the process of accumulation. Heuristically, let us try to reduce the complicated scheme of estimates to the following elementary operation. The notation here is reduced to a minimum. Let ψ_r be a trigonometric polynomial of degree r which owns a list $\mathcal{I}_r$ of divisors. Solving the homological equation $\partial_\omega \chi_r = \psi_r$ makes χ_r to own a list of divisors $\{r\} \cup \mathcal{I}_r$, the union meaning concatenation of lists. Let now f_s be a trigonometric polynomial of degree s that owns a list of indices $\mathcal{I}_s$. Then $L_{\chi_r} f_s$ owns the list $\{r\} \cup \mathcal{I}_r \cup \mathcal{I}_s$. With some patience the reader will realize that this is precisely the mechanism of generation of the lists associated to every term in the expressions that appear in the algorithm.

Let me now propose a little detour, exploiting the mechanism above independently of its application to the algorithm of Kolmogorov. Let $\mathcal{I}_s = \{j_1, \ldots, j_{s-1}\}$ be a list of $s - 1$ non negative indices, that we may collect in non decreasing order. A partial ordering on lists of indices is introduced as follows. For two given lists $\mathcal{I}_s = \{j_1, \ldots, j_{s-1}\}$ and $\mathcal{I}'_s = \{j'_1, \ldots, j'_{s-1}\}$ in non decreasing internal order we say that $\mathcal{I}$ precedes $\mathcal{I}'$ in case $j_1 \le j'_1, \ldots, j_{s-1} \le j'_{s-1}$; we write $\mathcal{I} \lhd \mathcal{I}'$. Let us also introduce the special lists

$$\mathcal{I}^*_s = \left\{ \left\lfloor \frac{s}{s} \right\rfloor, \left\lfloor \frac{s}{s-1} \right\rfloor, \ldots, \left\lfloor \frac{s}{2} \right\rfloor \right\} . \tag{7}$$

Lemma 3 *For the list of indices $\mathcal{I}^*_s$ the following statements hold true:*

(i) for $0 < r \le s$ we have

$$\left(\{r\} \cup \mathcal{I}^*_r \cup \mathcal{I}^*_s \right) \lhd \mathcal{I}^*_{r+s} .$$

(ii) for every $k \in \{1, \ldots, j_{\max}\}$ the index k appears exactly $\lfloor \frac{s}{k} \rfloor - \lfloor \frac{s}{k+1} \rfloor$ times;

The first claim concerns precisely the mechanism of accumulation of small divisors in the normalization algorithm for the Kolmogorov's normal form. Thus *$\mathcal{I}^*_s$ represents the worst list of indices that can be generated.*

The second claim contains the control of the action of small divisors. Let any sequence $1 = \alpha_0 \ge \alpha_1 \ge \alpha_2 \ge \ldots$ be given; we should estimate the product

$$\prod_{j \in \mathcal{I}^*_s} \frac{1}{\alpha_j} = \left(\alpha_1^{q_1} \cdot \ldots \cdot \alpha_{\lfloor s/2 \rfloor}^{q_{\lfloor s/2 \rfloor}} \alpha_s^{q_s} \right)^{-1} ,$$

where $q_k = \lfloor \frac{s}{k} \rfloor - \lfloor \frac{s}{k+1} \rfloor$ is the number of indices in $\mathcal{I}^*_s$ which are equal to k, and $q_s = 1$. We have

$$\ln \prod_{j \in \mathcal{I}^*_s} \frac{1}{\alpha_j} \le - \sum_{k=1}^{s} \left(\left\lfloor \frac{s}{k} \right\rfloor - \left\lfloor \frac{s}{k+1} \right\rfloor \right) \ln \alpha_k \le -s \sum_{k \ge 1} \frac{\ln \alpha_k}{k(k+1)} .$$

We are thus led to introduce

Condition $\boldsymbol{\tau}$ *The sequence* $\{\alpha_r\}_{r\geq 0}$ *satisfies*

$$-\sum_{r\geq 1}\frac{\ln\alpha_r}{r(r+1)}=\Gamma<\infty\,. \tag{8}$$

Letting α_r be the sequence defined by (6) we have thus found a condition of strong non resonance for the frequencies ω of the invariant torus.

A few remarks allow us to compare condition $\boldsymbol{\tau}$ with other commonly used conditions.

(i) The diophantine condition introduced by Siegel says $\alpha_r = r^{-k}$ with $k > 1$ (an innocuous multiplicative constant is omitted). This gives

$$-\sum_{r\geq 1}\frac{\ln\alpha_r}{r(r+1)}=k\sum_{r\geq 1}\frac{\ln r}{r(r+1)}<\infty\,.$$

By the way, this also shows that if the sequence α_r satisfies condition $\boldsymbol{\tau}$ then so does the sequence $\beta_r = \delta_r\alpha_r$ with $\delta_r \sim r^{-2}$ that appears in our estimates for the case of Kolmogorov.

(ii) Condition $\boldsymbol{\tau}$ is weaker than the diophantine one. E.g., if $\alpha_r = e^{-r/\ln^2 r}$ then

$$-\sum_{r\geq 1}\frac{\ln\alpha_r}{r(r+1)}=\sum_{r\geq 1}\frac{1}{(r+1)\ln^2 r}<\infty\,.$$

(iii) There are ω's that violate condition $\boldsymbol{\tau}$. For instance, if $\alpha_r = e^{-r}$ then

$$-\sum_{r\geq 1}\frac{\ln\alpha_r}{r(r+1)}=\sum_{r\geq 1}\frac{1}{(r+1)}=\infty.$$

(iv) The widely used condition of Bruno writes

$$-\sum_{r\geq 1}\frac{\ln\alpha_{2^{r-1}}}{2^r}=\text{Б}<\infty\,.$$

It is equivalent to condition $\boldsymbol{\tau}$, for one gets $\Gamma < \text{Б} < 2\Gamma$. However, condition $\boldsymbol{\tau}$ may present some advantages since sometimes it helps in finding better convergence estimates; e.g., see the application to the Poincaré–Siegel problem in [23].

In the case of interest here, namely the algorithm for the normal form of Kolmogorov, we must go back to Table 2, considering only the contribution of the divisors β_r. The problem is to identify the worst possible product of divisors in every coefficient of every function. To this end, to every coefficient we may associate two informations, namely: (i) the *number* of divisors β_j, and (ii) a *selection*

Table 3 The number of divisors and the selection rule for the functions h_r , $A_s^{(r)}$, $B_s^{(r)}$ and $C_s^{(r)}$ for $1 \le r < s$

Function	Number of divisors	Selection rule
h_r	$2r$	$\mathcal{I}_r^* \cup \mathcal{I}_r^* \cup \{r\} \cup \{r\}$
$A_s^{(r)}, \hat{A}_s^{(r)}$	$2s-2$	$\mathcal{I}_s^* \cup \mathcal{I}_s^*$
$B_s^{(r)}, \hat{B}_s^{(r)}$	$2s-1$	$\mathcal{I}_s^* \cup \mathcal{I}_s^* \cup \{r\}$
$C_s^{(r)}$	$2s$	$\mathcal{I}_s^* \cup \mathcal{I}_s^* \cup \{r\} \cup \{r\}$

rule, i.e., the maximal list of coefficients according to our partial ordering. With some patience, and using recursion, one finds the rules summarized in Table 3. In view of condition $\boldsymbol{\tau}$ we conclude that the products of divisors grow not faster than geometrically with s.

3.7 Sketch of the Proof of the Theorem of Kolmogorov

The estimate concerning the divisors is the most challenging part of the proof. The question now is to find upper bounds for the generating functions and prove that they satisfy the condition of Proposition 9 in appendix for convergence of the sequence of canonical transformations. This part requires a couple of tons of patience, but no really new ideas. On the other hand, a detailed exposition would exceed the limits of the present note. Therefore I give here only a hint on how to proceed.

Looking at Table 2, one sees that the norm of every function $A_s^{(r)}$, $B_s^{(r)}$, $C_s^{(r)}$, $\chi_1^{(r)}$ and $\chi_2^{(r)}$ will likely be estimated up to a multiplicative factor by a quantity $\nu_{r,s} T_{r,s} C^{s-1}$ where:

- the power C^{s-1} is due to products of the quantities $G_{r,1}$ and $G_{r,2}$ that estimate the norms of the generating functions;
- $T_{r,s} = \prod_j \beta_j^{-1}$ is the product of divisors with indexes j obeying the selection rules of the previous table;
- $\nu_{r,s}$ a numerical factor that takes into account the number of terms produced by Lie derivatives.

All these quantities are actually bounded geometrically, and the constant C is proportional to the size ε of the perturbation. On the other hand, the norm used here provides an upper bound for the supremum norm of a function. Hence, for ε small enough, the norms of the generating functions satisfy the condition of Proposition 9, thus assuring convergence of the normal form of Kolmogorov. Adding a further couple of tons of patience, the reader may check that the argument applies to any Hamiltonian of the general problem of dynamics. Thus I conclude with a (not too) formal statement.

Theorem 4 *Consider the Hamiltonian*

$$H(p,q) = H_0(p) + \varepsilon H_1(p,q) \ , \quad p \in \mathcal{G} \subset \mathbb{R}^n \ , \quad q \in \mathbb{T}^n \ .$$

Assume:

(i) *$H_0(p)$ is non degenerate, i.e.*

$$\det\left(\frac{\partial^2 H_0}{\partial p_j \partial p_k}\right) \neq 0 \ ;$$

(ii) *$H_0(p)$ possesses an invariant torus p^* with frequencies ω satisfying condition $\boldsymbol{\tau}$.*

Then there exists a positive ε^* such that the following holds true: for $\varepsilon < \varepsilon^*$ there exists a perturbed invariant torus carrying quasiperiodic motions with frequencies ω, which is close to the unperturbed one.

The relevance of the theorem of Kolmogorov for the planetary system has been often emphasized as being the proof that the dynamics of the planetary system is quasiperiodic, i.e., it may be described by the classical method of epicycles. But the crucial question is: *how small should be ε?*

3.8 Application to the Sun–Jupiter–Saturn System

The question concerning the actual applicability of the theorem of Kolmogorov to our Solar System has received some attention in the last, say, 30 years (see, e.g., [7, 8] and [49]). It should be noted that the Hamiltonian of the Solar System is degenerate. However, it has been shown by Arnold that degeneration can be removed, at least for the problem of three bodies, by following the lines of Lagrange's theory for secular motions. Here I give a very brief report on the paper [50] where the actual applicability to the problem of three bodies in the Sun–Jupiter–Saturn case has been argued (if not proven in strict mathematical sense).

The paper exploits the ideas of Arnold and the constructive character of our algorithm. The main steps are the following.

(i) Write the Hamiltonian of the problem of three bodies in Delaunay variables, in a heliocentric coordinate system.
(ii) Choose the value of the semimajor axes corresponding to the actual frequencies of Jupiter and Saturn, and expand around it up to the second order in the masses.
(iii) Average over the fast angles (mean anomalies), introduce the variables of Poincaré and expand the Hamiltonian in power series in the neighbourhood of the orbit with zero eccentricity and inclination.
(iv) Construct a Birkhoff normal form up to degree 6, thus removing the degeneration of the Hamiltonian, and find the torus with the actual frequencies of the system.
(v) Expand the Hamiltonian around that unperturbed torus in the form required by the algorithm for the normal form of Kolmogorov.

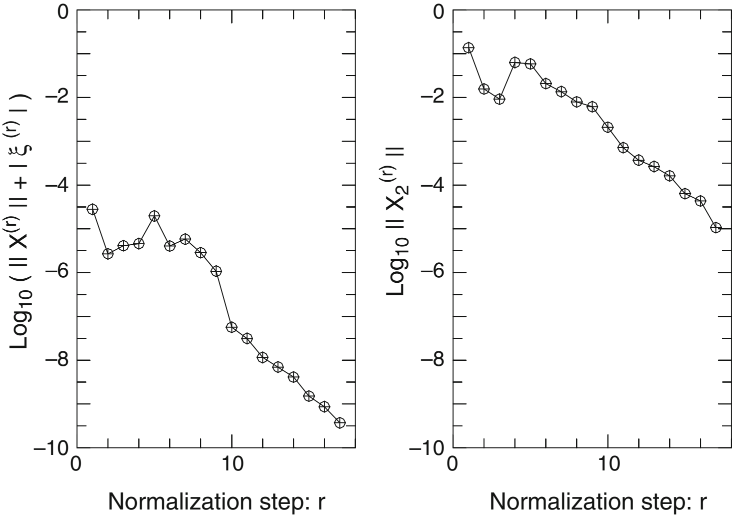


Fig. 3 The norms of the generating functions for the case Sun–Jupiter–Saturn, with *actual* parameters for the orbits taken from JPL database. (figure from [50])

(vi) Calculate the normal form up to some reachable order with the algorithm of Table 1.

(vii) From the explicit expansion calculate the norms of the generating functions up to that order.

The values of the norms are plotted in Fig. 3. It is clearly seen that after a few steps the norms begin to decrease geometrically. This provides a strong support to the thesis that the normal form of Kolmogorov is convergent, so that the orbit of Jupiter and Saturn (if we neglect the action of the other planets) is close to an invariant torus.

4 The Normal Form of Poincaré and Birkhoff

The aim of this section is to investigate the dynamics in a neighbourhood of either an elliptic equilibrium or an invariant torus of Kolmogorov exploiting the method of normal form of Poincaré and Birkhoff. I shall put emphasis on the problem of stability.

In a neighbourhood of an elliptic equilibrium the Hamitonian may generally be written as a power series

$$H(x,y) = H_0(x,y)+H_1(x,y)+H_2(x,y)+\dots\,, \qquad H_0(x,y) = \frac{1}{2}\sum_{l=1}^{n}\omega_l(x_l^2+y_l^2)\,, \tag{9}$$

where $\omega = (\omega_1, \dots, \omega_n) \in \mathbb{R}^n$ are the frequencies in the linear approximation, and $H_s(x, y)$ is a homogeneous polynomial of degree $s + 2$ in the canonical variables $(x, y) \in \mathbb{R}^{2n}$. The series is assumed to be convergent in some neighbourhood of the origin.

In a neighbourhood of an invariant torus of Kolmogorov the Hamiltonian may be expanded in power series of the actions as

$$H(p, q) = H_0(p) + H_1(p, q) + H_2(p, q) + \dots \,, \qquad H_0(p) = \langle \omega, p \rangle \,, \qquad (p, q) \in \mathcal{G} \times \mathbb{T}^n \tag{10}$$

with coefficients expanded as Fourier series in the angles q. As a consequence of the theorem of Kolmogorov we may always assume that the series is uniformly convergent in a neighbourhood $\mathcal{G}$ of the origin.

Dealing with an infinite Fourier series is clearly unsuitable for a practical calculation. However, exploiting again a suggestion of Poincaré we may split the Hamiltonian so that $H_s(p, q)$ is *at least quadratic in* p and is a trigonometric polynomial of degree sK with some positive integer K.

I should stress that although the two problems seem to be different they can be treated with the same approach. Let me also stress that the frequencies ω are not assumed to be non resonant.

4.1 Formal Normalization

I shall discuss the construction of a normal form using the method of Lie transform, recalled in Appendix section "An Algorithm for Lie Transform". The composition of Lie series may be used as well: I leave this as an exercise for an interested reader. I stress, however, that here I discuss only *formal* methods and results: accepting the common attitude of astronomers, all series and trigonometric expansions are performed without taking care of convergence problems. The problem of (non)-convergence will be discussed later, starting with Sect. 4.6.

Let us say that the Hamiltonian (9) or (10) is in Poincaré–Birkhoff normal form in case $\partial_\omega H = 0 \,, \quad \partial_\omega \cdot = \{\cdot, H_0\}$. I shall use the notation

$$Z = H_0 + Z_1 + Z_2 + \dots \,, \qquad \partial_\omega Z_s = 0 \,, \qquad s \geq 1$$

thus stressing with the symbol Z that the Hamiltonian is in normal form. The problem is: *find the generating sequence of a near the identity transformation that gives the Hamiltonian* (9) *or* (10) *a normal form.*

The question is formally answered by solving the equation $T_\chi Z = H$, the unknowns being the normal form Z itself and the generating sequence $\chi =$

$\{\chi_1, \chi_2, \ldots\}$, with χ_s a generating function of order s. The formal algorithm is found by recalling the definition (19) of the operator T_χ , namely

$$T_\chi = \sum_{s\geq 0} E_s \ , \quad E_0 = \mathbf{1} \ , \quad E_s = \sum_{j=1}^{s} \frac{j}{s} L_{\chi_j} E_{s-j} \ .$$

The following algorithm is found: *for $s \geq 1$ find χ_s and Z_s by recursively solving the homological equation*

$$Z_s - \partial_\omega \chi_s = \Psi_s \ , \quad s = 1, \ldots, r \ , \tag{11}$$

where

$$\begin{aligned} \Psi_1 &= H_1 \ , \\ \Psi_s &= H_s - \sum_{j=1}^{s-1} \frac{j}{s} \bigl(L_{\chi_j} H_{s-j} + E_{s-j} Z_j \bigr) \quad \text{for } 2 \leq s \leq r \ . \end{aligned} \tag{12}$$

In order to obtain the formulæ above it is convenient to recall the triangle for Lie transform

$$\begin{tikzcd}
H_0 & H_0 \arrow[d] & & & & \\
H_1 & E_1 H_0 \arrow[d] & Z_1 \arrow[d] & & & \\
H_2 & E_2 H_0 \arrow[d] & E_1 Z_1 \arrow[d] & Z_2 \arrow[d] & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & \\
H_r & E_r H_0 & E_{r-1} Z_1 & E_{r-2} Z_2 & \cdots & Z_r
\end{tikzcd}$$

Using the explicit expression

$$E_s H_0 = L_{\chi_s} H_0 + \sum_{j=1}^{s-1} \frac{j}{s} L_{\chi_j} E_{s-j} H_0$$

one finds an expression for Ψ_s similar to that in (11). A little play with algebra is needed in order to remove from Ψ_s all terms depending on H_0, using the homological equation for the previous orders.

4.2 Solving the Homological Equation

We need a definition. To the frequency vector ω we associate the *resonance module* (a subgroup of $\mathbb{Z}^n$)

$$\mathcal{M}_\omega = \bigl\{ k \in \mathbb{Z}^n \ : \ \langle k, \omega \rangle = 0 \bigr\} \ .$$

The dimension $\dim \mathcal{M}_\omega$ is often called the *multiplicity of the resonance.*

The interesting point is that the linear operator $\partial_\omega \cdot = \{\cdot, H_0\}$ may be diagonalized. In the case of an elliptic equilibrium we should perform a canonical transformation to complex variables ξ, η by setting

$$x_l = \frac{1}{\sqrt{2}}(\xi_l + i\eta_l) \ , \quad y_l = \frac{i}{\sqrt{2}}(\xi_l - i\eta_l)$$

for $l = 1, \ldots, n$. The unperturbed Hamiltonian H_0 takes the form

$$H_0 = \sum_{l=1}^{n} \omega_l I_l \ , \quad I_l = i\xi_l \eta_l \ ,$$

while the polynomials $H_s(\xi, \eta)$ are still homogeneous. The operator ∂_ω turns out to be diagonal over the basis $\xi^j \eta^k$ of monomials (in multiindex notation), for

$$\partial_\omega \xi^j \eta^k = i \langle j - k, \omega \rangle \, \xi^j \eta^k \ .$$

In the case of a torus the linear operator ∂_ω is already diagonal on the Fourier basis, for

$$\partial_\omega f_k(p) e^{i\langle k, q \rangle} = i \langle k, \omega \rangle \, f_k(p) e^{i\langle k, q \rangle} \ .$$

The kernel $\mathcal{N}_\omega$ and the range $\mathcal{R}_\omega$ of ∂_ω are defined as usual. Denoting by $\mathcal{P}$ the linear space under consideration (either homogeneous polynomials of Fourier series) we define

$$\mathcal{N}_\omega = \partial_\omega^{-1}\bigl(\{0\}\bigr) \ , \quad \mathcal{R}_\omega = \partial_\omega\bigl(\mathcal{P}\bigr) \ .$$

Since the linear operator ∂_ω maps $\mathcal{P}$ into itself the kernel and the range are subspaces of the same space $\mathcal{P}$. Moreover, since ∂_ω is diagonalizable we have

$$\mathcal{N}_\omega \cap \mathcal{R}_\omega = \{0\} \ , \quad \mathcal{N}_\omega \cup \mathcal{R}_\omega = \mathcal{P} \ .$$

Therefore the operator ∂_ω restricted to $\mathcal{R}_\omega$ is uniquely inverted.

4.3 The Solution of Poincaré and Birkhoff

The straightforward solution, namely the one proposed by Poincaré and Birkhoff, is the following. Project the right hand side of (11) on $\mathcal{N}_\omega$ and $\mathcal{R}_\omega$, i.e., (with obvious meaning of the superscripts)

$$\Psi_s = \Psi_s^{(\mathcal{N})} + \Psi_s^{(\mathcal{R})} , \quad \Psi_s^{(\mathcal{N})} \in \mathcal{N}_\omega , \quad \Psi_s^{(\mathcal{R})} \in \mathcal{R}_\omega .$$

Then set

$$Z_s = \Psi_s^{(\mathcal{N})} , \quad \chi_s = \partial_\omega^{-1} \Psi_s^{(\mathcal{R})} ,$$

so that $\chi_s \in \mathcal{R}_\omega$ is uniquely defined. An arbitrary term $\tilde{\chi}_s \in \mathcal{N}_\omega$ to χ_s may be added, but usually it is not necessary.

Concluding, we may state

Proposition 5 *The Hamiltonian $H = H_0 + H_1 + \ldots$ with $H_0 = \sum_l \omega_l I_l$ linear in the actions may be cast formally in normal form of Poincaré and Birkhoff*

$$Z = H_0 + Z_1 + Z_2 + \ldots , \quad \partial_\omega Z = 0 .$$

In complex variables for the elliptic equilibrium we have

$$Z_s(\xi, \eta) = \sum_{j-k \in \mathcal{M}_\omega} c_{j,k} \xi^j \eta^k .$$

In action-angle variables for an invariant torus we have

$$Z_s(p, q) = \sum_{k \in \mathcal{M}_\omega} c_k(p) \exp\bigl(i \langle k, q \rangle\bigr) .$$

4.4 Action-Angle Variables for the Elliptic Equilibrium

The dynamics is better described in action-angle variables $p = (p_1, \ldots, p_n) \in \mathbb{R}_+^n$ and $q = (q_1, \ldots, q_n) \in \mathbb{T}^n$. Thus, let us rewrite the Hamiltonian for the elliptic equilibrium in action-angle variables, by transforming

$$x_l = \sqrt{2p_l} \cos q_l , \quad y_l = \sqrt{2p_l} \sin q_l , \quad l = 1, \ldots, n .$$

From complex variables we easily write the Hamiltonian by using the exponential form of trigonometric functions, namely

$$\xi_l = \sqrt{p_l}\, e^{iq} , \quad \eta_l = -i\sqrt{p_l}\, e^{-iq} , \quad l = 1, \ldots, n .$$

The unperturbed Hamiltonian becomes linear in the actions, since

$$H_0 = \langle \omega, p \rangle \ , \quad \partial_\omega = \left\langle \omega, \frac{\partial}{\partial q} \right\rangle .$$

A homogeneous polynomial $f(\xi,\eta) = \sum_{|j+k|=s} f_{j,k}\xi^j\eta^k$ is changed into a trigonometric polynomial of the same degree that we may write as

$$f(q,p) = \sum_{|k|\leq s} c_k(p)\exp(i\langle k,q\rangle)$$

with coefficients $c_k(p)$ that are homogeneous polynomials in $p^{1/2}$. The square root is a little unpleasant, because it introduces a singularity. This often makes cartesian coordinates more useful. The normal form is expanded as a series of trigonometric polynomials that contain only Fourier harmonics $\langle k,q\rangle$ with $k \in \mathcal{M}_\omega$, i.e.,

$$Z_s(q,p) = \sum_{k\in\mathcal{M}_\omega\,,\,|k|=\leq s} c_k(p)\exp\big(i\langle k,q\rangle\big)$$

where, again, $c_k(p)$ are homogeneous polynomials of degree s in $p^{1/2}$.

4.5 First Integrals and Action-Angle Variables

The Lie transform formalism allows us to considerably simplify the search for first integrals.

Proposition 6 *The Hamiltonian $H = H_0 + H_1 + \ldots$ with H_0 linear in the actions possesses $n - \dim(\mathcal{M}_\omega)$ formal first integrals of the form*

$$\Phi(p,q) = \Phi_0(p) + \Phi_1(p,q) + \ldots \ , \quad \Phi_0(p) = \langle \mu, p \rangle \ , \quad 0 \neq \mu \perp \mathcal{M}_\omega \, .$$

which are independent and in involution.

The construction goes as follows: find the first integrals of Z; then prove that every first integral for Z generates a first integral for H that inherits the properties of independence and involution.

Let $\Phi_0 = \langle \mu, p\rangle$ with $0 \neq \mu \in \mathbb{R}^n$, that we want to satisfy $\{\Phi_0, Z\} = 0$. Writing $Z = \sum_{k\in\mathcal{M}_\omega} c_k(p)\exp\big(i\langle k,q\rangle\big)$ calculate

$$\{\Phi_0, Z\} = -i \sum_{k\in\mathcal{M}_\omega} \langle k,\mu\rangle\, c_k(p)\exp\big(i\langle k,q\rangle\big) \ ,$$

which is zero if $\mu \perp \mathcal{M}_\omega$. Therefore there are $n - \dim(\mathcal{M}_\omega)$ such functions which are independent. Moreover Z itself is a first integral which is independent of the ones so found in the resonant case, since Z will typically depend also on the angles. In general there are no further independent first integrals unless Z has a very special form.

We may now see that $\Phi = T_\chi \Phi_0 = \Phi_0 + \Phi_1 + \ldots$ is a first integral for H. For by the properties of the Lie transform operator T_χ we have

$$\{\Phi, H\} = \{T_\chi \Phi_0, T_\chi Z\} = T_\chi \{\Phi_0, Z\} = 0 \,,$$

i.e., $\Phi = T_\chi \Phi_0$ is a first integral of the wanted form. The $n - \dim(\mathcal{M}_\omega)$ first integrals so found are obviously independent. They are also in involution. For, let $\Phi_0 = \langle \mu, I\rangle$ and $\Phi'_0 = \langle \mu', I\rangle$ be independent. Then $\{\Phi, \Phi'\} = \{T_\chi \Phi_0, T_\chi \Phi'_0\} = T_\chi\{\Phi_0, \Phi'_0\} = 0$.

In the non resonant case, $\dim(\mathcal{M}_\omega) = 0$, the condition $\partial_\omega Z = 0$ of normal form implies that $\frac{\partial Z}{\partial q} = 0$, i.e., we have $Z = Z(p_1, \ldots, p_n)$. Then the system is formally integrable: $p_1, \ldots, p_n$ are first integrals, and are the action variables. Moreover, the normal form is formally expanded in power series of p (powers of $p^{1/2}$ entail a dependence on the angles). Thus, $Z_1 = Z_3 = \ldots = 0$. The usual description of the dynamics applies in this case. The phase space is foliated into invariant tori parameterized by $p_1, \ldots, p_n$, carrying quasi periodic motions with frequencies

$$\Omega(p) = \omega + \frac{\partial Z_2}{\partial p}(p) + \frac{\partial Z_4}{\partial p}(p) + \ldots$$

Thus, generically, the dynamics of the normal form is not isochronous.

Let us now come to the resonant case, $0 < \dim(\mathcal{M}_\omega) = r < n$, which is more intriguing. As we have seen, the normal form depends on the actions p and on the combinations of the angles $\langle k, q\rangle$ with $k \in \mathcal{M}$, i.e

$$Z(p, q) = \langle \omega, p\rangle + Z_1\bigl(p_1, \ldots, p_n, \langle k^{(1)}, q\rangle, \ldots, \langle k^{(r)}, q\rangle\bigr) + \ldots$$

For $r = 1$ the Hamiltonian Z possesses $n - 1$ independent first integrals which are linear combinations of the actions p. Moreover, the Hamiltonian itself is a first integral, and if $Z(q, p)$ does depend on q then it is independent of the previous first integrals. Therefore the system is still Liouville-integrable.

A resonant system with $\dim(\mathcal{M}_\omega) = r > 1$ is not expected to be integrable, except for very particular cases. However the first integrals may be used in order to reduce the number of degrees of freedom by r. A general procedure is the following. First, find a basis $k^{(1)}, \ldots, k^{(r)}$ for $\mathcal{M}_\omega$. That is, we should choose r integer vectors in $\mathcal{M}_\omega$ which are independent, and satisfy the further property that $\operatorname{span}\bigl(k^{(1)}, \ldots, k^{(r)}\bigr) \cap \mathbb{Z}^n = \mathcal{M}_\omega$. Denote by

$$\begin{pmatrix} k_{1,1} & k_{1,2} & \ldots & k_{1,n} \\ \vdots & \vdots & \ldots & \vdots \\ k_{r,1} & k_{n-r,2} & \ldots & k_{r,n} \end{pmatrix}$$

the matrix whose lines are the vectors of the basis. Then the matrix can be completed with integer entries in the form

$$\mathsf{M} = \begin{pmatrix} k_{1,1} & k_{1,2} & \ldots & k_{1,n} \\ \vdots & \vdots & \ldots & \vdots \\ k_{r,1} & k_{n-r,2} & \ldots & k_{r,n} \\ m_{1,1} & m_{1,2} & \ldots & m_{1,n} \\ \vdots & \vdots & \ldots & \vdots \\ m_{n-r,1} & m_{n-r,2} & \ldots & m_{n-r,n} \end{pmatrix}, \quad \det \mathsf{M} = \pm 1$$

Such a matrix is said to be unimodular. The interesting fact is that it provides a linear transformation on a torus that preserves all periods.

Apply the canonical transformation with generating function $S(I, q) = \langle I, \mathsf{M}q \rangle$, i.e.,

$$\varphi = \mathsf{M}q \ , \quad p = \mathsf{M}^\top I \ .$$

Then the Hamiltonian is transformed as

$$H_0(I) = \langle \omega', I \rangle \ , \quad Z_s(I, \varphi) = Z_s(I_1, \ldots, I_n, \varphi_1, \ldots, \varphi_r) \ ,$$
$$\omega' = \mathsf{M}\omega = (0, \ldots, 0, \omega'_{n-r+1}, \ldots, \omega'_n) \ .$$

The Hamiltonian turns out to depend only on the *resonant angles* $\varphi_1, \ldots, \varphi_r$. Hence the actions $I_{r+1}, \ldots, I_n$ are first integrals that may be considered as parameters, and we may forget $H_0(I_{n-r+1}, \ldots, I_n)$, which is constant. We conclude that the dynamics is determined by the family of reduced systems of $r < n$ degrees of freedom with Hamiltonian

$$Z(I_1, \ldots, I_n, \varphi_1, \ldots, \varphi_r) = Z_1(I_1, \ldots, I_n, \varphi_1, \ldots, \varphi_r) + Z_2(I_1, \ldots, I_n, \varphi_1, \ldots, \varphi_r) + \ldots$$

parameterized by the initial values of the constants $I_{r+1}, \ldots, I_n$. However, the latter Hamiltonian, in general, is not a perturbation of an integrable system: its dynamics may well be chaotic, typically over a slow time scale.

A last *caveat:* we should never forget that *all claims made in this section are just formal:* we still lack a discussion of the (non) convergence of the normal form.

4.6 The Dark Side of Small Divisors

Let us now come to the problem of convergence of the normal form of Poincaré–Birkhoff. To this end, let us associate again to the frequency vector ω the non increasing sequence $\{\alpha_r\}_{r>0}$ defined as

$$\alpha_r = \min_{0<|k|\le rK} (|\langle k, \omega \rangle|) \ . \tag{13}$$

The trouble here is that the naive argument concerning the accumulation of divisors illustrated at the beginning of Sect. 3.5 fully applies to the present case. For, looking at the recurrent formulæ (11) and (12) we see that the following happens: the generating function χ_r appears to have the divisors $\alpha_1 \cdots \alpha_r$, which cause the norm to grow as a factorial. One may hope that a kind mechanism similar to that of the algorithm of Kolmogorov applies, thus leading again to a kind accumulation of divisors, but no such mechanism has been discovered.

Thus, there is a strong suggestion that the normal form of Poincaré–Birkhoff does not converge. Indeed, Siegel in 1941 proved that the normal form is generically non convergent [63]: in a suitable topology on the space of Hamiltonians in the neighbourhood of an equilibrium divergence occurs in the majority of cases, and the divergent Hamiltonians are dense. Things are actually quite complicated, since in a different topology the set of Hamiltonians with a convergent normal form is also dense.

The mechanism of divergence has been investigated by Contopoulos, Efthymiopoulos and the author [13, 19] on the basis of some considerations on maps in [62]. The conclusion was that the estimates of accumulation via diophantine inequality are close to optimal.

In view of divergence one may be tempted to reject all methods based on the normal form of Poincaré–Birkhoff—a conclusion in sharp contrast with the old standing tradition of Celestial Mechanics, quite successful in describing many phenomena. But there is a much better attempt, already suggested by Poincaré: *exploit the asymptotic character of perturbation series.*

4.7 Old Fashioned Numerical Exploration

At the dawn of numerical simulations of dynamics, between 1955 and 1960, the method of Poincaré section has been used in order to visualize the dynamics of systems of two harmonic oscillators with a cubic nonlinearity, a simple model that may describe the dynamics of stars in a Galaxy. Many such studies have been performed by Contopoulos, who also had the idea of calculating the so called *third integral* (to be added to the energy and the angular momentum) by a series expansion similar to that obtained via the Poincaré–Birkhoff normal form, and to compare the results with the Poincaré section [11]. The series had to be truncated at low order, of course, due to the limited power of computers available at that time. The aim of this section is to perform a similar comparison paying attention to aspects related to the convergence of the series.

The starting point is nothing but the traditional one: truncate the expansions at a finite order, $r \geq 1$ say (degree $r + 2$). That is, get a truncated normal form

$$H^{(r)} = \langle \omega, p \rangle + Z_1(p) + \ldots + Z_r(p) + \mathcal{F}^{(r)}(p, q) \ ,$$

with $\mathcal{F}^{(r)}$ a non normalized remainder of order at least r. This is done by constructing the Lie triangle up to the r-th line, and so a truncated generating sequence $\{\chi_1, \ldots, \chi_r\}$. The problems connected with resonances appear here in a weak form: only resonances $\langle k, \omega \rangle = 0$ with $|k| \leq r + 2$ should be avoided, or taken into account by constructing a resonant normal form. Having determined the generating sequence, we are able to construct a truncated first integral, e.g.,

$$\Phi^{(r)} = I_l + \Phi_1(p, q) + \ldots + \Phi_r(p, q)$$

taking as I_l one of the actions (or a combination adapted to the resonance).

As a model problem, let us consider the Hamiltonian

$$H = \frac{\omega_1}{2}(x_1^2+y_1^2)-\frac{\omega_2}{2}(x_2^2+y_2^2)+x_1^2x_2-\frac{1}{3}x_2^3\,, \quad \omega_1 = 1,\ \omega_2 = \frac{\sqrt{5}-1}{2}\,. \tag{14}$$

Note that the frequencies here have different signs, which makes the model reminiscent of the case of triangular equilibria of the planar restricted problem of three bodies: the energy integral can not be used as a Lyapounov function in order to assure stability.

The Poincaré section on the energy surface $E = 0.0025$ is calculated by setting $x_1 = 0$, and the result is reported in Fig. 4. One will remark that a stable region exists close to the origin, that can be investigated by constructing a suitable first integral. The region is bounded by the separatrices of an unstable periodic orbit.

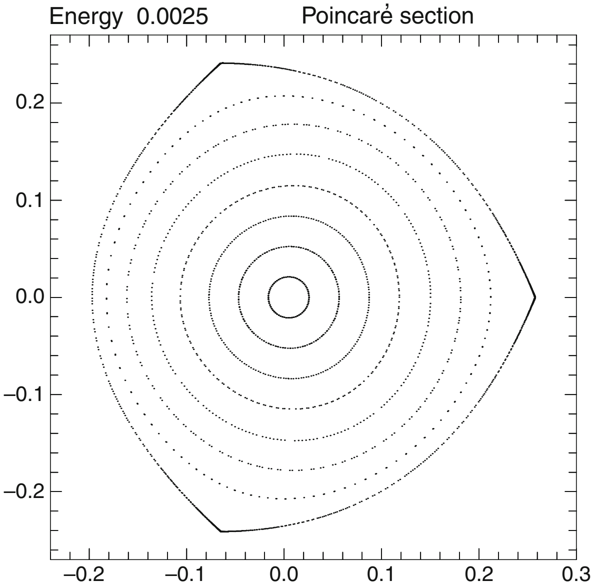


Fig. 4 The Poincaré section for the Hamiltonian (14) on the energy surface $E = 0.0025$

There are also three separated islands outside the region represented here, which are not described by our normal form.

The comparison with the first integral is not difficult. Choosing a point (x_2, y_2) on the section plane, with $x_1 = 0$, the value y_1 is found by solving the equation $H(0, y_1, x_2, y_2) = E$, thus giving $y_1(x_2, y_2)$. Then one replaces these values in Φ, thus getting a function $\Phi(0, y_1(x_2, y_2), x_2, y_2)$ of two variables. If Φ is a true first integral (e.g., in case of convergence) then the level lines of the functions should describe the Poincaré section of the orbit.

The comparison between the Poincaré section and the level lines of the first integral is represented in Figs. 5 and 6, for truncation orders 5, 9, 12, 24, 38, 45, 60, 70. It goes without saying that an expansion up to order 70 has been made possible thanks to the increased power of computers, compared to the ones available in the sixties: the calculation has been performed on a desktop

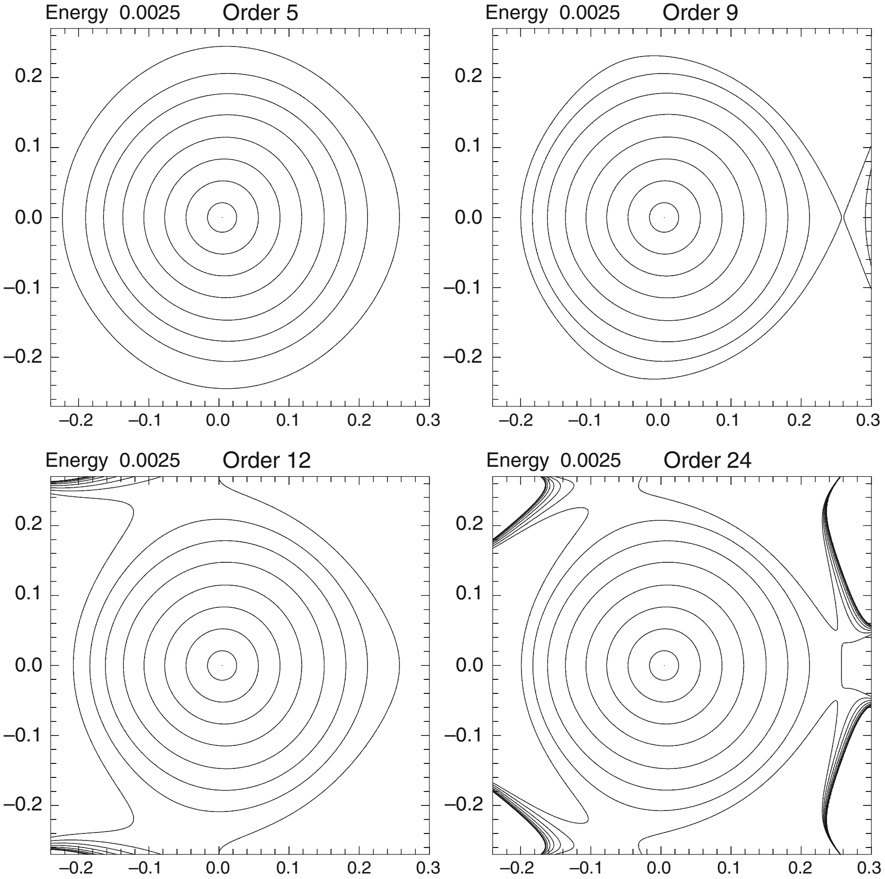


Fig. 5 Comparison of the Poincaré section of the Hamiltonian (14) with the level lines of the first integral, truncated at different orders

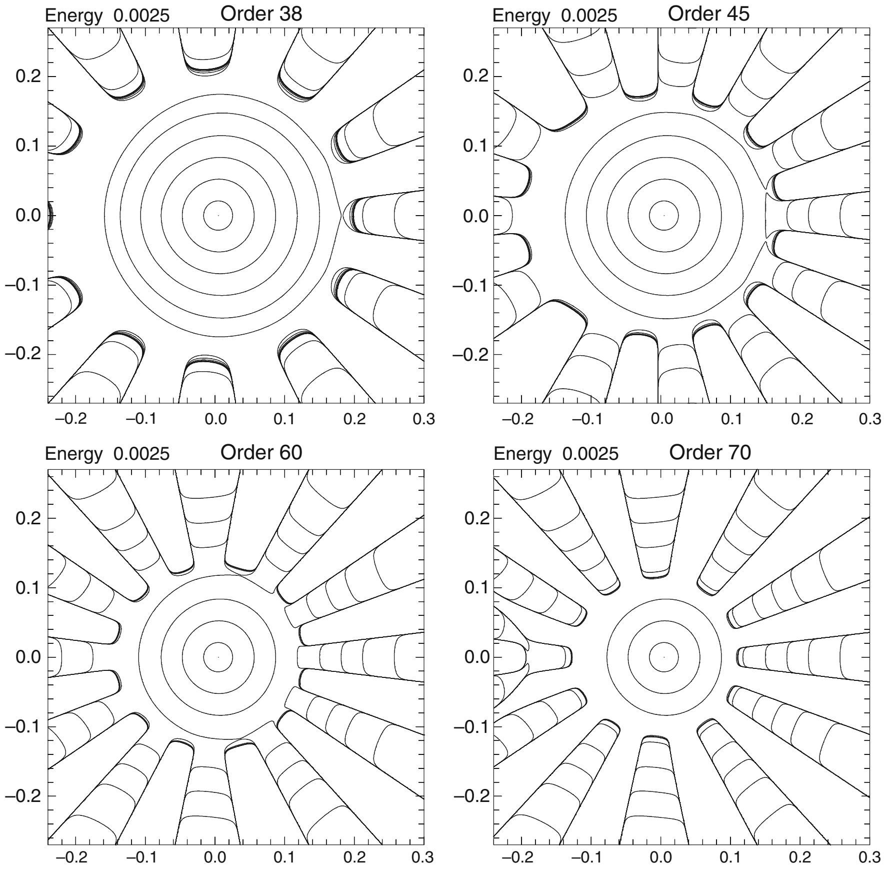


Fig. 6 Continuation of Fig. 5

computer around 1997. One sees that there is a good (visual) coincidence of the curves close to the origin, with some global improvement up to a truncation order 24 where singularities non dramatically far from the unstable orbit seem to show up. But a further increase of the truncation makes the region of correspondence to gradually shrink, suggesting that its size will reduce to zero for $r \to \infty$. This is precisely the expected behaviour of an asymptotic series.

4.8 Qualitative Description of Dynamics

A description of the dynamics in terms of truncated first integrals is based on the following considerations. Pick $r \geq 1$, and consider the truncated first integral

$$\Phi^{(r)} = I_l + \Phi_1 + \ldots + \Phi_r \ . \tag{15}$$

A combination of the actions may be taken in the resonant case. It is easily checked that by construction we have

$$\dot{\Phi}^{(r)} = -\{H_1, \Phi_r\} ,$$

which is a polynomial of degree $r + 3$. For a Hamiltonian with a full power series expansion we get a series starting with terms of degree $r + 3$.

Consider now a domain of initial data which is a polydisk of radius ϱ centered at the origin, namely

$$\Delta_\varrho = \left\{(x, y) \in \mathbb{R}^n \ : \ x^2 + y^2 \le \varrho^2\right\} .$$

Suppose for a moment that (by some unexpected miracle) $\Phi^{(r)}$ is an exact first integral, so that it keeps its initial value during the evolution. Recall, however, that we are usually able to observe only the action I_l : determining hopefully better quantities that may be calculated by perturbation methods is a more difficult task. But from (15) we know that in the domain Δ_ϱ we have

$$\left|\Phi^{(r)}(x, y) - I_l\right| < D_r \varrho^3 ,$$

with some constant D_r depending on the truncation order. In other words, the torus $\Phi^{(r)}(x, y) = \text{constant}$ appears in the original variables x, y as a deformed torus carrying a quasiperiodic motion. Therefore the value of $I_l(t)$ will actually oscillate in a strip of width $D_r\varrho^3$, and its evolution will be quasiperiodic, as illustrated in the upper part (a) of Fig. 7.

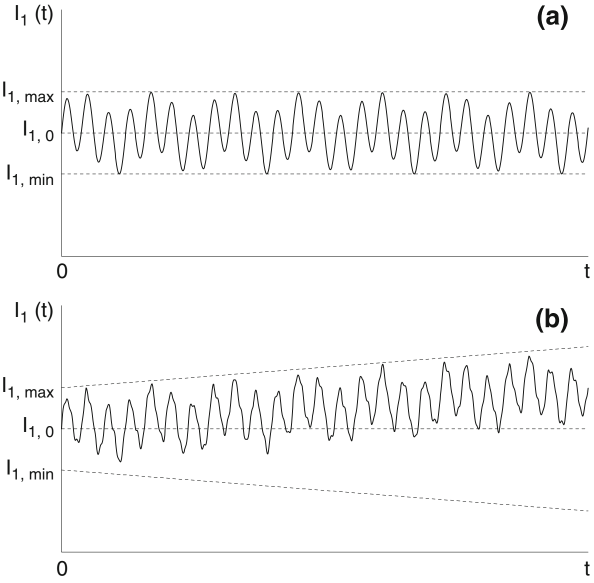


Fig. 7 Illustrating the effect of the deformation and of the noise

But $\Phi^{(r)}(x, y)$ is not exactly constant. Therefore we can say only that (by the triangle inequality)

$$\left|I_l(t) - I_l(0)\right| < \left|I_l(t) - \Phi^{(r)}(t)\right| + \left|\Phi^{(r)}(t) - \Phi^{(r)}(0)\right| + \left|\Phi^{(r)}(0) - I_l(0)\right|$$

with

$$\left|\Phi^{(r)}(t) - \Phi^{(r)}(0)\right| \le |t| \cdot \left|\dot{\Phi}^{(r)}\right| < t \cdot C_r \varrho^{r+3} \ ,$$

with some constant C_r depending on the order of truncation. We conclude that $I(t)$ oscillates in a strip whose width increases very slowly, with a slope $O(\varrho^{-r})$, being also subjected to a *noise* that is superimposed to the quasiperiodic motion, as illustrated in the lower part (b) of Fig. 7. The effect of the noise can not be predicted, and induces small, possibly chaotic deviations with respect to the quasiperiodic evolution.

These considerations force us to accept the idea that stability may occur only for a long time, but not forever. This will be discussed in detail in the next section.

5 Long Time Stability

When discussing stability the most common reference is the theory of Lyapounov, paying particular attention to stability of an equilibrium. But for a physical system that evolves equilibrium is an exceptional state. We need a more refined approach that takes into account at the same time the existence of action variables for an integrable system and the lack of integrability introduced by a small perturbation.

We may reformulate the problem of stability as follows:

Prove that the actions I satisfy an inequality such as

$$\left|I(t) - I(0)\right| < \varepsilon^b \quad \text{for } |t| \le T(\varepsilon) \tag{16}$$

with $T(\varepsilon)$ large, in a sense to be made precise, and with some positive $b < 1$.

In the case of an elliptic equilibrium the role of the perturbation parameter ε is played by the size ϱ of the neighbourhood of the equilibrium.

The latter formulation is particularly suited to the case of the Solar System. The actions are (in rough but essentially correct terms) the semimajor axes, the eccentricities and the inclinations of the orbits. The angles are the mean anomaly (related to an area according to the second Kepler's law), the argument of the perihelion and the argument of the node. For the development of life it is essential that the semimajor axes, the eccentricities and the inclinations do not change too much for a substantial fraction of the life of the Solar System itself. The question is: *Does our model of the Solar System account for such a long stability time?*

5.1 A Note on the Concept of Stability

Different approaches to the problem of stability of perturbed, near to integrable systems have been developed. The basic question is expressed by (16). The dependence of $T(\varepsilon)$ on the size of the perturbation makes the difference.

(i) $T(\varepsilon) \simeq 1/\varepsilon$: *adiabatic invariants*. This is essentially the theory of Lagrange for the Solar System, related to the method of averaging. It could be noted that this concept has played a major role in the development of Quantum Mechanics.

(ii) $T(\varepsilon) \simeq 1/\varepsilon^r$ with $r > 1$: *complete stability*. It was introduced by Birkhoff [6] for the dynamics around an equilibrium, as we have said in Sect. 4.8. It is based on a bound of type $C_r \varrho^{r+3}$ on the *noise*, with some undetermined constant C_r depending on r.

(iii) $T(\varepsilon) \simeq \exp(1/\varepsilon^a)$ with $0 < a \le 1$: *exponential stability*. It has been proposed by Moser [53] and Littlewood [47, 48] for an elliptic equilibrium. Its general form has been developed by Nekhoroshev [55, 56].

(iv) $T(\varepsilon) \simeq \exp\bigl(\exp(1/\varepsilon^a)\bigr)$ with $0 < a \le 1$: the *superexponential stability*. It has been investigated by Morbidelli and the author [24, 52].

(v) $T(\varepsilon) = \infty$: *perpetual stability*. It has been the dream of many mathematicians and astronomers of the nineteenth century: to prove that the Newtonian model of the Solar System is integrable. It is also the guiding idea of the theory of Lyapounov [51].

5.2 Adiabatic Theory and Complete Stability

Let us recall that having fixed $r \ge 1$ we may construct the normal form of the Hamiltonian truncated at order r, namely.

$$H^{(r)} = \langle\omega, p\rangle + Z_1(p) + \ldots + Z_r(p) + \mathcal{F}^{(r)}(p, q) \,, \quad \mathcal{F}^{(r)} = \mathcal{O}(\varepsilon^{r+1}) \,.$$

We may also construct truncated first integrals, e.g.,

$$\Phi^{(r)} = p_l + \Phi_1(p, q) + \ldots + \Phi_r \,, \quad \dot{\Phi}^{(r)} = \mathcal{O}(\varepsilon^{r+1}) \,.$$

Taking into account both the deformation and the noise we conclude

$$\bigl|p(t) - p(0)\bigr| = \mathcal{O}(\varepsilon) \quad \text{for} \quad |t| \sim \frac{1}{\varepsilon^r} \,.$$

For $r = 1$ this corresponds to applying the averaging method. The result is the typical estimate of adiabatic theory: the actions remain almost constant for a time of order $1/\varepsilon$.

The concept of complete stability corresponds essentially to performing a higher order averaging. The result is qualitative. For instance, in the case of a neighbourhood of size ϱ of an elliptic equilibrium it may be reformulated as

$$\bigl|p(t) - p(0)\bigr| = \mathcal{O}(\varrho^3) \quad \text{for} \quad |t| \sim \frac{1}{\varrho^r} \,,$$

adding the usual claim that this is true for ϱ small enough. In slightly more precise terms, the size of the remainder should be estimated as

$$\bigl|\mathcal{F}^{(r)}\bigr| < C_r \varrho^{r+3} \,,$$

with a constant C_r strongly affected by accumulation of small divisors and growing very fast with r. Birkhoff did not try to estimate the dependence of C_r on r. The natural question is: *Can we make more precise and possibly improve the complete stability of Birkhoff?*

5.3 Exponential Stability

In order to be definite, let us assume a condition of non resonance. The resonant case may be treated in a similar manner.

An estimate of the constant C_r may be found by implementing a scheme of analytic estimates, as we did for the theorem of Kolmogorov. However, let us avoid technical and boring calculations. The interested reader may find a detailed exposition in [20] or [27]. In the case of Kolmogorov we have seen that the crucial problem is the accumulation of small divisors coming both from the solution of the homological equation and from Cauchy's estimates for Lie derivatives. As we have already remarked in Sect. 4.6 the function χ_r of the generating sequence is expected to have a product $\alpha_1 \cdots \alpha_r$ of divisors, with the sequence of α's defined by (13). Let us assume a diophantine condition on the frequencies, i.e., $\alpha_r \sim r^{-\tau}$ with $\tau > n - 1$. Hence we may guess that the constant C_r may be replaced by $C^r (r!)^a$ with a constant C not depending on r and with $a > \tau$, in order to account for the estimate of Lie derivatives. So it is, indeed, and one may also find $a = \tau + 1$.

Accepting the argument above, the remainder of the normal form (forgetting unessential constants) is estimated as

$$\mathcal{F}^{(r)} \sim (r!)^a \varepsilon^{r+1} \,.$$

The estimate depends on two quantities: (i) ε, given by Nature, and (ii) r, which is our choice. It would be desirable to remove our arbitrary choice of r. To this end, for a given ε let us look for the best choice of r, in the sense that the remainder $\mathcal{F}^{(r)}$ is reduced to a minimum. Write the right hand side as $(r!)^a\varepsilon^r = r^a\varepsilon\cdot\bigl((r-1)!\bigr)^a\varepsilon^{r-1}$, and remark that it clearly takes a minimum for

$$r = r_{\rm opt} = (1/\varepsilon)^{1/a}\ .$$

Using Stirling's formula calculate

$$(r_{\rm opt}!)^a\varepsilon^{r_{\rm opt}} \sim \left(\frac{r_{\rm opt}}{e}\right)^{ar_{\rm opt}}\varepsilon^{r_{\rm opt}} \sim \exp\left[-a\left(\frac{1}{\varepsilon}\right)^{1/a}\right]\ .$$

The latter estimate depends only on ε, as wanted. Hence for a given ε we get the estimate

$$\bigl|p(t)-p(0)\bigr| = \mathcal{O}(\varepsilon)\quad\text{for}\quad |t|\sim\exp\left[a\left(\frac{1}{\varepsilon}\right)^{1/a}\right]\ ,$$

a time exponentially long with a power of $1/\varepsilon$. This is the *exponential stability*. Littlewood commented: *"If not eternity, this is a considerable slice of it."*

5.4 Using Computer Algebra

The argument of the previous section may be implemented numerically using a suitable package of algebraic manipulation. The aim is to obtain good stability estimates by using the truncated first integrals explicitly constructed. I will illustrate the procedure for the case of an elliptic equilibrium, also giving an explicit example.

Suppose that we have constructed the normal form according to the algorithm of formulæ (11) and (12), up to some order r. Then we may also construct truncated first integrals $\Phi^{(r)} = I_l + \Phi_1 + \ldots + \Phi_r$, with $I_l = (x_l^2 + y_l^2)/2$, which are polynomials of degree $r+2$. All this may be done, e.g., with the same program used for Figs. 5 and 6.

The aim is to perform a numerical optimization of the order r. Consider a domain

$$\Delta_\varrho = \left\{(x,y)\in\mathbb{R}^{2n}\ :\ I_l(x,y)\le\frac{\varrho^2}{2}\,,\ l=1,\ldots,n\right\}\ ,$$

namely a polydisk of radius ϱ centered at the origin (the equilibrium). A preliminary problem is to evaluate the supremum norm of polynomials in such a domain. This can be done in different ways: the simplest one is to add up the absolute values of the coefficients and multiply by a power ϱ^s if the polynomial is homogeneous of degree s. A better method will produce better estimates of the time of stability, of course.

Take an initial point $(x_0, y_0) \in \Delta_{\varrho_0}$. The orbit is subject to the combined effect of deformation and noise, and can possibly escape from a larger disk Δ_ϱ after some time. We want to establish an estimate such as $\varphi^t(x_0, y_0) \in \Delta_\varrho$ for $|t| \le \tau(\varrho_0, \varrho)$ for $\varrho_0 < \varrho$. To this end let us recall the inequality

$$\bigl|I(t) - I(0)\bigr| \le \underbrace{\bigl|I(t) - \Phi^{(r)}(t)\bigr|}_{\delta_r(\varrho)} + \bigl|\Phi^{(r)}(t) - \Phi^{(r)}(0)\bigr| + \underbrace{\bigl|I(0) - \Phi^{(r)}(0)\bigr|}_{\delta_r(\varrho_0)} ,$$

$$\delta_r(\varrho) = \sup_{\Delta_\varrho} \bigl|\Phi^{(r)}(x, y) - I_l(x, y)\bigr| ,$$

which holds true provided $I(t) < \varrho^2/2$. The quantities $\delta_r(\varrho_0)$ and $\delta_r(\varrho)$ measure the deformation, of order ϱ^3, that can be determined using the estimate of the supremum norm, since we know the expansion. The quantity $\bigl|\Phi^{(r)}(t) - \Phi^{(r)}(0)\bigr|$ is the contribution of the noise, that is estimated as

$$\bigl|\Phi^{(r)}(t) - \Phi^{(r)}(0)\bigr| < |t| \sup_{(x,y)\in\Delta_\varrho} \bigl|\dot{\Phi}_l(x, y)\bigr| .$$

For given $\varrho_0,\ \varrho$ let

$$D_r(\varrho_0, \varrho) = \frac{\varrho^2 - \varrho_0^2}{2} - \delta_r(\varrho) - \delta_r(\varrho_0) .$$

This is the quantity left for diffusion. In view of $\delta_r(\varrho) \sim \varrho^3$, the qualitative behaviour of the function $D_r(\varrho_0, \varrho)$ is as represented in the upper left panel of Fig. 8. The allowed interval for ϱ is determined by the compatibility condition $D_r(\varrho_0, \varrho) \ge 0$.

Using all the first integrals available, the escape time from Δ_ϱ is then estimated to be not less than

$$\tau_r(\varrho_0, \varrho) = \min_{l=1,\ldots,n} \frac{D_r(\varrho_0, \varrho)}{\sup_{(x,y)\in\Delta_\varrho} \bigl|\dot{\Phi}^{(r)}(x, y)\bigr|} \tag{17}$$

The qualitative graph of the denominator of the latter expression is represented in the right upper panel of Fig. 8: it grows as fast as $r!\varrho^{r+3}$, thus strongly depending on r.

The estimate (17) depends on r, ϱ_0 and ϱ, and we want to optimize it against r and ϱ. That is: *to look for an optimal estimate*

$$T^*(\varrho_0) = \max_r \sup_\varrho \tau_r(\varrho_0, \varrho) .$$

depending only on the initial radius ϱ_0.

The qualitative behaviour of the function $\tau_r(\varrho_0, \varrho)$ for a fixed r is represented in the left lower panel of Fig. 8. It has a maximum in the allowed interval of ϱ at

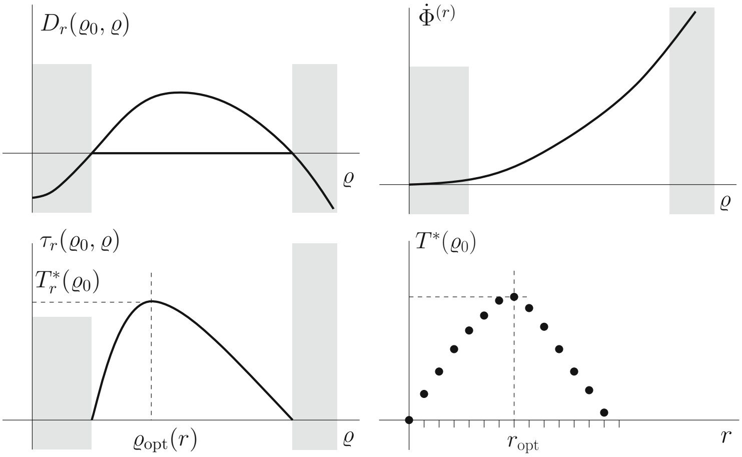


Fig. 8 Illustrating the scheme of calculation of the optimal stability time

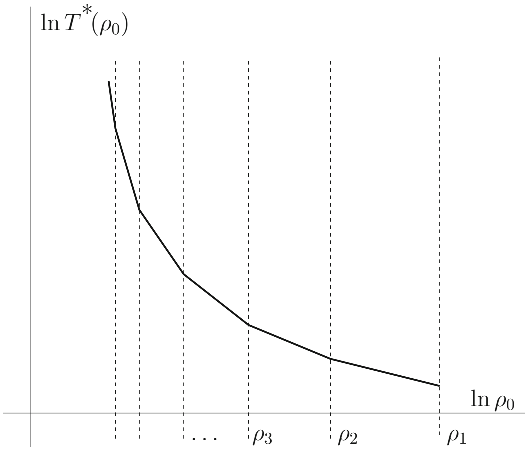


Fig. 9 The expected behaviour of the optimal stability time as a function of the initial radius ϱ_0, in log-log scale

a point ϱ_{opt} with a corresponding value $T_r^*(\varrho_0)$ which represents the best estimate of the time for a given r. The last step is the optimization against r. In view of the asymptotic character of the series the values $T_r^*(\varrho_0)$ are expected to distribute as in the right lower panel of Fig. 8, thus allowing us to select an optimal value r_{opt} for r corresponding to the maximum and depending only on ϱ_0, as requested. The wanted value $T^*(\varrho_0)$ is the maximum so found, to which an optimal value ϱ_{opt} is associated.

The qualitative behaviour of $T^*(\varrho_0)$ is represented in Fig. 9, in log-log scale. According to the behaviour of a function such as $r!\varrho^r$ it is expected that there is a decreasing sequence $\varrho_1, \varrho_2, \varrho_3, \ldots$ of values of ϱ which mark an increase of the optimal order r_{opt}. Here, ϱ_1 plays the role of a threshold above which nothing useful is provided by perturbation methods: the perturbation is too big. Conversely, in every interval $[\varrho_r, \varrho_{r-1}]$ the estimated stability time grovs as ϱ^{-r}, as represented in Fig. 9. The resulting graph is a sequence of segments with increasing slope for $\varrho \to 0$. The exponential behaviour of the estimated stability time is actually a lower bound to the sequence of segments.

The method illustrated here can be adapted to the study of stability in a neighbourhood of an invariant torus: it is just matter of using a scheme of algebraic manipulation adapted to that case. Once the generating sequence of the normal form is constructed, the procedure is the same. As a last remark, the numerical estimates of the stability time may be increased if one accepts to work in the coordinates of the normal form. This removes the need of taking into account the deformation of coordinates. The domain of stability turns out to be a deformed disk, but the procedure is correct.

5.5 An Application to the Sun–Jupiter–Saturn–Uranus system

Some applications of the method above are available in the literature. E.g., applications to the case of the triangular Lagrangian equilibria in the case of the Sun–Jupiter system have been worked out in [9, 25] and [18]; while applications to the Sun–Jupiter–Saturn system can be found in [28, 50] and in [61], where also Uranus is considered. Here I report the results for the case of the Sun–Jupiter–Saturn–Uranus system [29], investigating the long time stability in a neighborhood of an invariant KAM torus which approximates very well the secular orbits. Specifically, we consider a planar secular model that can be regarded as a major refinement of the Lagrange-Laplace theory.

Hereafter, I only sketch the procedure, referring to [29] for a detailed exposition. First the Hamiltonian is expanded in Poincaré variables. In our calculations we truncate the expansion as follows. The Keplerian part is expanded up to the quadratic terms in the fast actions, while the perturbation, due to the mutual interactions between the planets, include: (i) the linear terms in the so-called fast actions, (ii) all terms up to degree 18 in the secular variables, (iii) all terms up to the trigonometric degree 16 with respect to the fast angles. Our choice of the limits allows to include the effects of near mean-motion resonances (5:2 Jup.–Sat., 7:1 Jup.–Uran., 7:5:3 Jup.–Sat.–Uran.).

Then, following the approach described in [50], we perform two “Kolmogorov-like” normalization steps so as to remove the main perturbation terms depending on the fast angles. This allows us to improve the classical circular approximation, by replacing it with a solution that is invariant up to order two in the masses.

The secular Hamiltonian is then obtained just by averaging over the fast angles. After the diagonalization of the quadratic part, the Hamiltonian essentially describes a system of three perturbed harmonic oscillators. Thus, we can construct a secular

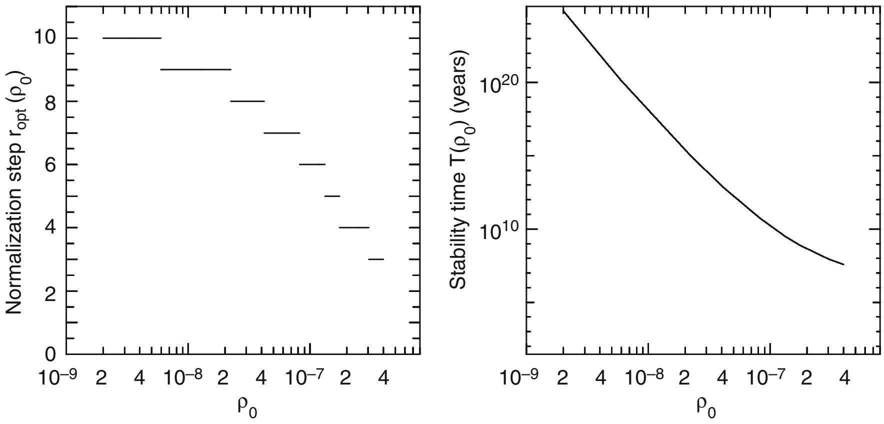


Fig. 10 The optimal normalization order (left panel) and the estimated stability time (right panel) for the planar Sun–Jupiter–Saturn–Uranus system (figure from [29])

invariant KAM torus near an elliptic equilibrium and compute the estimated stability time in a neighborhood of that torus. The actual implementation consists in the explicit computation of the Kolmogorov normal form (up to order 5), followed by a high-order Birkhoff normalization via an iterative scheme of estimates. Finally the stability time is numerically estimated.

The result is reported in Fig. 10. In the left panel the value of $r_{\rm opt}$ is reported as a function of ϱ_0. In the right panel we report the estimated stability time as a function of the initial distance ϱ_0 from the invariant KAM torus.

The ideal goal is to show that there is a neighborhood of that torus for which the estimated stability time is larger than the lifetime of the Solar System. In our result, the actual size of such a neighborhood, compared with the uncertainties of the astronomical observations, is about ten times smaller.

6 A Considerable Slice of Eternity

The last section is devoted to an informal exposition of the theorem of Nekhoroshev on exponential stability and of its extension named superexponential stability.

The theorem of Nekhoroshev may be seen as the global version of the exponential stability discussed in Sects. 4 and 6.5. The results in the previous sections are *local*, being concerned with a neighbourhood of either an elliptic equilibrium or an invariant torus. The theory of Nekhoroshev investigates the stability of dynamics in a possibly large open set of the phase space; more precisely in an open set of the actions domain.

The theory of superexponential stability aims at showing that the stability time may be much longer than exponential.

6.1 Back to the General Problem of Dynamics

Let us consider a Hamiltonian

$$H(p,q) = h(p) + H_1(p,q) + H_2(p,q) + \ldots \ , \quad (p,q) \in \mathcal{G} \times \mathbb{T}^n \ ,$$

where $H_s = \mathcal{O}(\varepsilon^s)$ is a trigonometric polynomial of degree sK for some $K > 0$. As already observed, every holomorphic perturbation may be cast in this form.

We have learned that for such a system first integrals do not exist (the theorem of Poincaré), because the divisors $\langle k, \omega(p)\rangle$ are not constant, and resonances are dense. However, the example of Sect. 2.3 shows that truncated first integrals may be constructed in suitable domains. On the other hand, the local theory of normal form applies also in case of resonance. These ideas are exploited in the theory of Nekhoroshev.

We proceed in two steps, named analytic part and geometric part. In the analytic part a *local* result in a region around given resonances is found; it sheds some light on the local behaviour of orbits, but is unable to provide a global description. The geometric part makes the picture *global* by introducing a clever geography of resonances.

6.2 Local Analytic Results

We need a definition. A *non-resonance domain* is an open subset $\mathcal{V}$ of the action space $\mathcal{G}$ characterized by:

(i) a given resonance module $\mathcal{M} \subset \mathbb{Z}^n$, with $0 \le \dim \mathcal{M} < n$;
(ii) a non resonance condition on $\mathcal{V}$: for some $r \ge 1$ we want

$$|\langle k, \omega(p)\rangle| > \alpha \quad \text{for all} \quad p \in \mathcal{V} \ , \quad k \in \mathbb{Z}^n \setminus \mathcal{M} \quad \text{and} \quad |k| \le rK \ .$$

The non resonance condition assures that a normal form up to order r can be constructed in the domain $\mathcal{V}$. The normal form will be either the non resonant or the resonant one, depending on the resonance module $\mathcal{M}$. Therefore we may construct a Hamiltonian in normal form

$$H^{(r)}(p,q) = h(p) + Z_1(p,q) + \ldots + Z_r(p,q) + \mathcal{F}^{(r)}(p,q) \ ,$$
$$Z_s(p,q) = \sum_{k \in \mathcal{M}\,,\ |k| \le sK} z_k(p) e^{i\langle k,q\rangle} \ , \quad \mathcal{F}^{(r)} = \mathcal{O}(\varepsilon^r) \ .$$

The construction is the same as for a neighbourhood of an invariant torus, the only difference being that the frequencies $\omega(p) = \frac{\partial H_0}{\partial p}$ do depend on the actions. However, the presence of zero divisors is excluded in the non resonance domain $\mathcal{V}$.

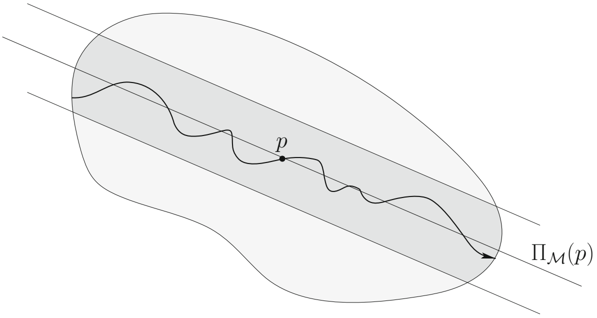


Fig. 11 The dynamics inside a non resonance domain

The Hamiltonian $H^{(r)} - \mathcal{F}^{(r)}$ possesses $n - \dim \mathcal{M}$ independent first integrals

$$\Phi(p) = \langle \lambda, p \rangle \quad \text{with} \quad \lambda \perp \mathcal{M} \ .$$

The intersection $\Pi_{\mathcal{M}}$ of the planes $\langle \lambda, p \rangle = c$, that we name *resonant plane*, is invariant for the flow of $H^{(r)} - \mathcal{F}^{(r)}$. However we must take into account the deformation induced by the transformation to normal form, and also the (very slow) noise induced by the perturbation. Thus we formulate a *local stability lemma: the orbit lies in a cylinder of radius $\delta(\varepsilon)$ around the resonant plane for a time $\mathcal{O}(1/\varepsilon^r)$ unless it leaves the domain $\mathcal{V}$ through a base of the cylinder (the intersection with the border of $\mathcal{V}$).*

The meaning is illustrated in Fig. 11. The initial point p determines the resonant plane $\Pi_{\mathcal{M}}(p)$ that is invariant in the coordinates of the normal form. Due to deformation, in the original coordinates the orbit is confined in a small neighbourhood of the invariant plane; call it a *cylinder*. The flow due to $H^{(r)} - \mathcal{F}^{(r)}$ causes a *fast drift* along the plane; the noise induces a slow drift, possibly transversal to the plane, that makes the size of the cylinder to grow during time: see the description of dynamics in Sect. 4.8.

An escape due to noise may occur only after a long time. However, an escape due to fast drift may well occur in a time $1/\varepsilon$. The question is: *what happens if the orbit leaves $\mathcal{V}$*? An escape could cause a *diffusion* of the orbit inside the action domain $\mathcal{G}$.

Diffusion may actually be generated by two different mechanisms. The first one is due to a *channel of diffusion*, which may appear if a resonant plane coincides or is too close to the manifold $\langle k, \omega(p) \rangle = 0$. The second mechanism is the so called *overlapping of resonances*, illustrated in Fig. 12. The fast drift along the resonant plane may drive the orbit inside a different non resonance domain, where it can follow a different resonant plane, and so on. The latter mechanism has been identified a long time ago as the responsible of chaos [10, 12]. Nowadays it can be observed in many nice figures representing some parameter that characterizes chaos or diffusion. A stability result should avoid such situations, at least for resonances of not too high order.

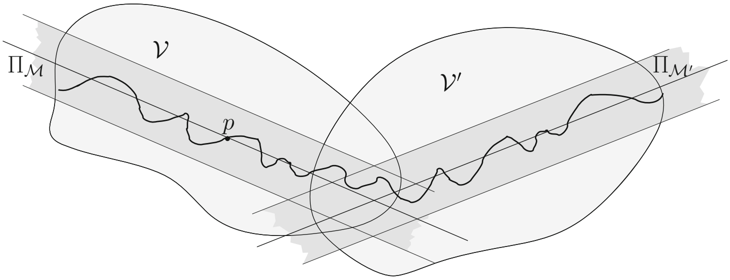


Fig. 12 Illustrating the mechanism of overlapping of resonances that may drive an orbith through different non resonance domains, thus causing diffusion

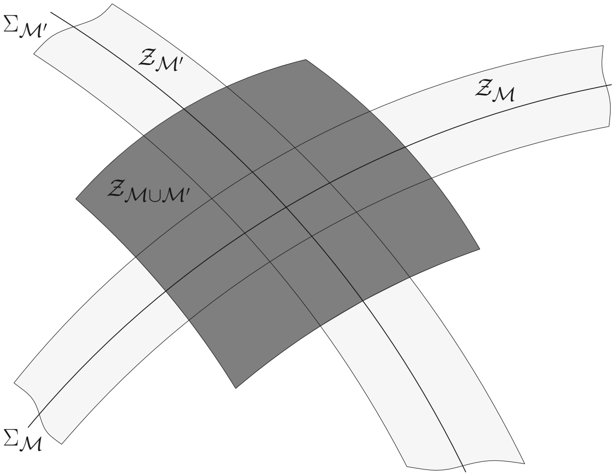

$\Sigma_{\mathcal{M}'}$
$\mathcal{Z}_{\mathcal{M}'}$
$\mathcal{Z}_{\mathcal{M}}$
$\mathcal{Z}_{\mathcal{M}\cup\mathcal{M}'}$
$\Sigma_{\mathcal{M}}$

Fig. 13 Resonant manifolds and resonant zones. For two resonance modules $\mathcal{M}$ and $\mathcal{M}'$ of dimension 1 the resonant manifolds are actually two curves. The intersection between the two curves (a point in the figure) is the resonant manifold associated to the resonance module $\mathcal{M}\cup\mathcal{M}'$. The resonant zones of multiplicity 0 is the whole domain $\mathcal{G}$. The resonant zones of multiplicity 1 associated to $\mathcal{M}$ and $\mathcal{M}'$ are the light grey strips around the corresponding resonant manifolds. The resonant zone of multiplicity 2 associated to $\mathcal{M}\cup\mathcal{M}'$ is the dark grey square

6.3 Geography of Resonances

My aim now is to illustrate in a not too technical form the geometric part of the theorem of Nekhoroshev. The definitions here are of a general character. What is omitted is the part concerning the choice of parameters that determine the size of different parts. A reader who is looking for a detailed proof in a reference that uses the same scheme and language of the present notes may see, e.g., [26] or [21].

The backbone of the geography of resonance is provided by the *resonant manifolds* (see Fig. 13) Pick a positive integer N, and select all resonance modules

$\mathcal{M}$, of any dimension $s = \dim \mathcal{M}$, that contain s independent integer vectors with $|k| \le N$. To every resonance module associate a resonant manifold

$$\Sigma_{\mathcal{M}} = \bigl\{ p \ : \ \langle k, \omega(p) \rangle = 0 \quad \text{for} \quad k \in \mathcal{M} \bigr\} \, .$$

The number of resonance modules satisfying the condition above is clearly finite; therefore the structure of resonant manifolds is not dense. Around a resonant manifold we select a strip of width β_s increasing with $s = \dim \mathcal{M}$. The increase of β_s with the multiplicity allows us to avoid overlapping of resonances, as will be explained later. We define a *resonant zone* associated to $\mathcal{M}$ as

$$\mathcal{Z}_{\mathcal{M}} = \bigl\{ p \ : \ |\langle k, \omega(p) \rangle| \le \beta_s \text{ for } k \in \mathcal{M} \, , \ |k| \le N \bigr\} \, .$$

Let us say that the resonant zone has multiplicity $s = \dim \mathcal{M}$.

A *resonant block* of multiplicity s is constructed by taking out from a zone everything that belongs to a zone of multiplicity $s + 1$. I.e., if $s = \dim \mathcal{M}$ then

$$\mathcal{B}_{\mathcal{M}} = \mathcal{Z}_{\mathcal{M}} \setminus \mathcal{Z}^*_{s+1} \, , \quad \mathcal{Z}^*_{s+1} = \bigcup_{\dim \mathcal{M}' = s+1} \mathcal{Z}_{\mathcal{M}'} \, .$$

Remark that a block of multiplicity s has empty intersection with every block of multiplicity $s + 1$, but its intersection with blocks of multiplicity larger than $s + 1$ may well be non empty (see Fig. 14).

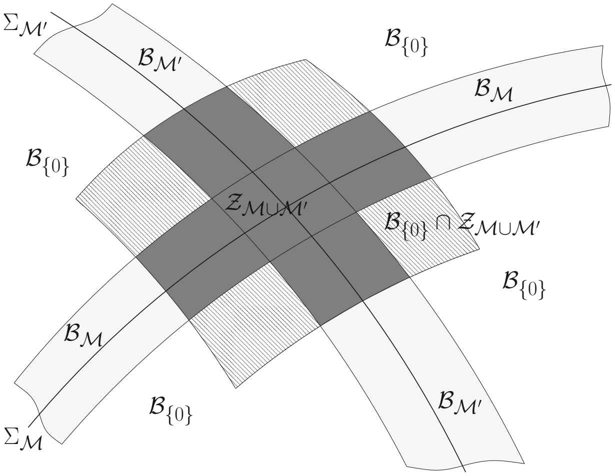

$\Sigma_{\mathcal{M}'}$
$\mathcal{B}_{\mathcal{M}'}$
$\mathcal{B}_{\{0\}}$
$\mathcal{B}_{\mathcal{M}}$
$\mathcal{B}_{\{0\}}$
$\mathcal{Z}_{\mathcal{M}\cup\mathcal{M}'}$
$\mathcal{B}_{\{0\}} \cap \mathcal{Z}_{\mathcal{M}\cup\mathcal{M}'}$
$\mathcal{B}_{\{0\}}$
$\mathcal{B}_{\mathcal{M}}$
$\mathcal{B}_{\mathcal{M}'}$
$\mathcal{B}_{\{0\}}$
$\Sigma_{\mathcal{M}}$

Fig. 14 Resonant blocks. From every zone of multiplicity s subtract the part belonging to a zone of multiplicity $s + 1$. E.g., $\mathcal{B}_{\{0\}}$ is the whole domain minus the light grey cross of the zones of multiplicity 1 in Fig. 13; $\mathcal{B}_{\mathcal{M}}$ is the zone $\mathcal{Z}_{\mathcal{M}}$ minus the dark grey square of multiplicity 2; $\mathcal{B}_{\mathcal{M}\cup\mathcal{M}'}$ is the dark grey square. The dashed region belongs to both $\mathcal{B}_{\{0\}}$ and $\mathcal{B}_{\mathcal{M}\cup\mathcal{M}'}$, but not to a zone of multiplicity 1

The structure of the resonant block is a good basis for constructing the non resonance domains requested by the analytic part of the theory. For, we know exactly which resonances are inside the block, by construction of the zones, and resonances not belonging to the module $\mathcal{M}$ associated to the block have been excluded. However, the clean structure of the blocks conflicts with the dynamics (and with the restrictions of domains requested by analytic estimates, that I do not take care of here). Hence we should continue our construction making the structure of blocks somehow fuzzy.

Let $p \in \mathcal{B}_{\mathcal{M}}$ be a point of an orbit. The analytic lemma tells us that we should take into account the fast drift along the plane, with a possible creation of diffusion channels. This is what can not be excluded and is likely to happen if the resonant plane and the resonant manifold are too close. In order to avoid such a situation the unperturbed Hamiltonian H_0 must satisfy a suitable condition that guarantees a separation between the two manifolds (see Fig. 15). In the original paper of Nekhoroshev such a condition was identified with *steepness* (roughly: a tangency of finite order between $\Sigma_{\mathcal{M}}$ and $\Pi_{\mathcal{M}}(p)$). Other proofs require the more manageable condition of convexity, that implies transversality, or at least convexity on the energy surface.

This is not enough, however. The analytic lemma also claims that in the original coordinates (that we are using in our construction) we must take into account the

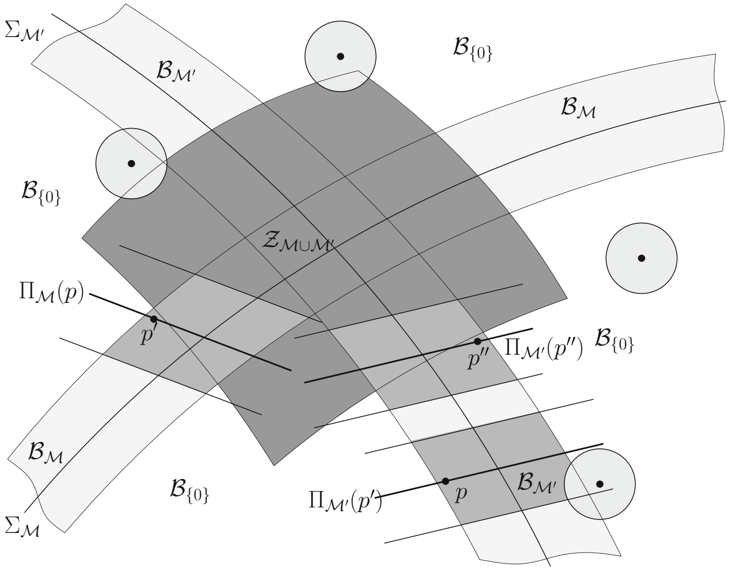


Fig. 15 Resonant planes, cylinders and extended blocks. Through every point of the domain we draw the resonant plane parallel to the resonance module $\mathcal{M}$. Next we add a strip of width δ_s around the resonant plane and intersect it with the corresponding zone. This makes a cylinder. E.g., the cylinder around a point inside the non resonant block $\mathcal{B}_{\{0\}}$ is actually disk. The union of cylinders constructed around all points of a block makes an extended block

deformation and the noise. Therefore we enlarge the plane with a strip of width δ_s, depending on the multiplicity of the resonance, and associate to every point $p \in \mathcal{B}_{\mathcal{M}}$ of a given block a *cylinder of width* δ_s defined as

$$\mathcal{C}_{\mathcal{M},\delta_s}(p) = \Pi_{\mathcal{M},\delta_s}(p) \cap \mathcal{Z}_{\mathcal{M}} \ , \quad \Pi_{\mathcal{M},\delta_s}(p) = \bigl\{p' \in \mathbb{R}^n \ : \ \mathrm{dist}\bigl(p', \Pi_{\mathcal{M}}(p)\bigr) \le \delta_s\bigr\} \ .$$

Remark that the cylinder lies inside a single resonance zone, due to the intersection: the bases coincide with the border of the zone. However, it may overlap a zone of higher multiplicity (see Fig. 15). The latter case is harmless provided the condition of *non overlapping of resonances* holds true: cylinders with different resonant modules and same multiplicity should have empty intersection. This is assured by the property of convexity (or steepness) together with a suitable choice of the parameters β_s and δ_s, that should be increasing fast enough with s. I will not enter the details of such a choice, that may be found in published papers.

Finally, all cylinders around every point of a given block are collected together in an extended block. Formally, we define

$$\mathcal{B}_{\mathcal{M},\delta_s} = \bigcup_{p \in \mathcal{B}_{\mathcal{M}}} \mathcal{C}_{\mathcal{M},\delta_s}(p) \ .$$

The extended blocks are the domains of non resonance required by the analytic part of the theory: inside an extended block there are known resonances, and the intersection of blocks of the same multiplicity is empty. The fact that blocks of different multiplicity may have a non empty intersection makes the painting definitely fuzzy, but this is unavoidable due to the dynamics.

Thus, we should exploit the property of convexity (or steepness) making a choice of the parameters β_s (the width of the resonant zones) and δ_s (the width of the cylinders) so as to satisfy the request of non overlapping of resonance. All parameters must be defined in terms of the perturbation ε, and decrease to zero with it.

6.4 Confinement of Orbits and Exponential Stability

Let us now come to the final argument of the proof, that I illustrate making reference to Fig. 16. As stated by the analytic lemma, an orbit with initial point on a block moves inside the corresponding cylinder, thus wandering inside the extended block. Recall that the fast drift has velocity $\mathcal{O}(\varepsilon)$. Conversely, the motion due to the noise is slow, with velocity $\mathcal{O}(\varepsilon^r)$. Therefore the orbit may leave a cylinder in a time less that $\mathcal{O}(1/\varepsilon^r)$ only through a base, but in that case *it must enter a zone of lower multiplicity*.

The latter remark has far reaching consequences: *an orbit remains confined in a cylinder for an exponentially long time.* Let me illustrate the trace of the argument.

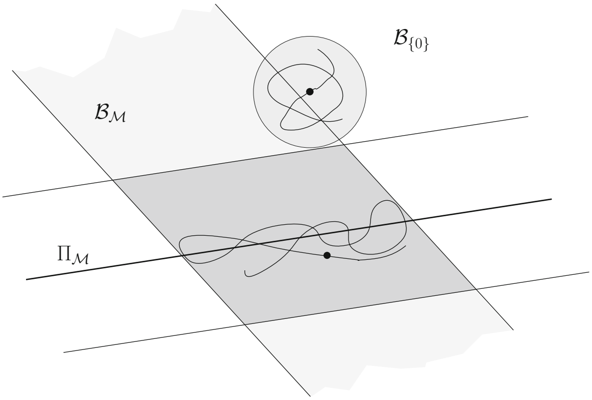


Fig. 16 An orbit may visit zones of different multiplicity, but it remains confined inside a cylinder of minimal multiplicity

An orbit may visit different zones, but there is a point p' belonging to a zone of minimal multiplicity. Now: *the orbit with initial point p' remains in the corresponding cylinder up to a time $\sim 1/\varepsilon^r$*. The proof relies on a simple argument by contradiction: assume that the orbit leaves the cylinder; then it must exit through a base of the cylinder, thus entering a zone with lower multiplicity. This contradicts the hypothesis that p' was in the zone with minimal multiplicity. Thus we come to the conclusion:

For every orbit we have

$$\bigl|p(t) - p(0)\bigr| < \delta \quad \text{up to} \quad |t| \sim \frac{1}{C_r \varepsilon^r}$$

where δ is the diameter of the largest cylinder.

The claim follows from analytical estimates according to which the size of the remainder is $\sim C_r\varepsilon^r$ with C_r growing as a (power of a) factorial with r, as in the case of an equilibrium.

It remains to make an optimal choice of a common normalization order r for all extended blocks. Here we exploit the asymptotic character of the series, as in the case of an elliptic equilibrium. Recalling that every extended block has his own parameters β_s and δ_s and a corresponding value of r, take the worst values among them. Then the process of optimization is similar to that we have made for an elliptic equilibrium.

Theorem 7 *Let $H = h(p) + \varepsilon f(p, q, \varepsilon)$ be analytic in all variables in a domain $\mathcal{G} \times \mathbb{T}^n$, with $\mathcal{G} \subset \mathbb{R}^n$ open, and let $h(p)$ be a convex function. Then there exist positive constants μ_* and T^* such that the following statement holds true: if*

$$\mu_* \varepsilon < 1$$

then for every orbit $p(t), q(t)$ *satisfying* $p(0) \in \mathcal{G}$ *one has the estimate*

$$\mathrm{dist}\bigl(p(t) - p(0)\bigr) \le (\mu_* \varepsilon)^{1/4}$$

for all times t *satisfying*

$$|t| \le \frac{T^*}{\varepsilon} \exp\left[\left(\frac{1}{\mu_* \varepsilon} \right)^{1/a} \right] ,$$

for some positive a *depending on* n*, e.g.,* $a \sim 2n^2$.

6.5 Superexponential Stability

The stability of an invariant torus of Kolmogorov has been discussed in Sect. 4, together with the stability of an elliptic equilibrium, using the normal for of Poincaré-Birkhoff. The conclusion there was that one can prove exponential stability in Nekhoroshev's sense. My aim now is to show that the stability of an invariant torus of Kolmogorov (or a strongly non resonant elliptic equilibrium) may be definitely stronger than exponential. For definiteness, I will focus on the case of a torus.

First, we should remind the theorem of Kolmogorov. Assuming that the frequencies are diophantine, namely

$$\bigl|\langle k, \omega \rangle\bigr| > \gamma |k|^{-\tau} , \quad 0 \neq k \in \mathbb{Z}^n , \ \gamma > 0 , \ \tau \ge n-1 ,$$

the Hamiltonian is given a holomorphic normal form in the neighbourhood of the torus

$$H(p,q) = \langle \omega, p \rangle + \mathcal{R}(p,q) , \quad \mathcal{R}(p,q) = \mathcal{O}(p^2) .$$

The second step is the construction of the normal form of Poincaré-Birkhoff in the neighbourhood of the torus. We should make a suitable expansion of the Hamiltonian

$$H(p,q) = \langle \omega, p \rangle + H_1(p,q) + H_2(p,q) + \ldots$$

where $H_s(p,q)$ is small of order ε^s in some norm. Moreover $H_s(p,q)$ is at least quadratic in p and a trigonometric polynomial of degree sK in the angles q. As we have already seen, this can be done. The Poincaré-Birkhoff normal form up to a finite order r is written as

$$H^{(r)}(p,q) = \langle \omega, p \rangle + Z^{(r)}(p) + \mathcal{F}^{(r)}(p,q) .$$

With analytic estimates and an optimal choice of r one finds that if $\bigl|\mathcal{R}(p,q)\bigr| \sim \varepsilon$ then the perturbation becomes exponentially small, i.e.,

$$\bigl|\mathcal{F}^{(r)}(p,q)\bigr| \sim \exp\bigl(-1/\varepsilon^{1/n}\bigr) .$$

The third step is the application of the theorem of Nekhoroshev in its general form. Rewrite the Hamiltonian as

$$H(p,q) = h(p) + \mathcal{F}(p,q) , \quad h(p) = \langle\omega, p\rangle + Z^{(r)}(p) .$$

Here we must assume that $h(p)$ satisfies the convexity condition. If so, then the Hamiltonian has the form required by the theorem of Nekhoroshev, but with an exponentially small perturbation. Thus, from the theorem of Nekhoroshev we conclude

$$\bigl|p(t) - p(0)\bigr| \sim \exp\bigl(-1/\varepsilon^{1/n}\bigr) \quad \text{for} \quad |t| \sim \exp\bigl(\exp\bigl(1/\varepsilon^{1/n}\bigr)\bigr) .$$

This is a local result of superexponential stability, first stated in [52].

A stronger version of the theorem may be found in [24]. The proof in this case is definitely longer, and is inspired by the proof of the theorem of Kolmogorov due to Arnold [3]. In [3] it is proved that in the phase space there exists a set of invariant tori which are deformations of strongly non resonant unperturbed tori. The difference with respect to the proof illustrated in the present lectures is that the result is global, since it applies to an open set of the action space, and the existence of a set of invariant tori with large measure is proved at once.

Superexponential stability follows by replacing the quadratic scheme used by Arnold with the analytic part of the theorem of Nekhoroshev for a non resonant domain. The starting point is, again, the general problem of dynamics, namely a Hamiltonian system $H(p,q) = h(p) + \varepsilon F(p,q)$.

(i) Excluding from the action space all resonant zones of a finite order, less than some $N > 0$, one is left with an open non resonant domain where the Hamiltonian can be given a normal form similar to the one above, but with a perturbation of size $\varepsilon' \sim \exp(-1/\varepsilon^a)$ with some positive $a < 1$. Thus, in the open non resonant domain one has stability for a time $\sim \exp(1/\varepsilon^a)$.

(ii) The step (i) is iterated infinitely many times by suitably increasing N, so that at every step the perturbation is exponentially small with respect to the previous step. Accordingly, the domain is reduced at every step by subtracting the resonant zones created by the new resonances. Thus one finds a sequence of boxed domains where one has stability for a time which is successively estimated as

$$\exp\bigl(\sim \exp(1/\varepsilon^a)\bigr) , \ \exp\Bigl(\exp\bigl(\exp(1/\varepsilon^a)\bigr)\Bigr) , \ \ldots$$

adding an exponential at every step.

(iii) In the limit of infinitely many steps one finds a set of invariant tori (similar to the one found by Arnold) wich has an open dense complement.

Thus, with respect to the result of Arnold, one adds the remarkable information that the dynamics around the tori is frozen much more than exponentially.

7 A Final Question

In these notes I have presented a series of results from perturbation theory, organized in a personal picture that starts from the discoveries of Kepler and ends with very recent and current research. I included in particular the theorem of Kolmogorov on persistence of invariant tori and the theories of exponential and superexponential stability.

There is a big question that remains unanswered: *does all this matter have some significance for the stability of the Solar System or, more generally, of a planetary system?*

The theorem of Kolmogorov is hardly applicable to the whole Solar System, even including only the giant planets and the internal planets: numerical simulations have shown that a chaotic behaviour actually occurs [42, 43]. However, it constitutes the skeleton that lies behind the flesh of theories about long time stability.

The theorem of Nekhoroshev is much more robust. First, the theorem does not exclude a chatic behaviour: it says that chaos is confined for a long time in a small region. Second, one can prove a version where a small time dependent perturbation is allowed, not even periodic or quasi periodic [26]. Therefore, one may attempt to take into account the action of small bodies as a generic and small time dependence.

Thus, a possible question is: *can we prove that the theory of Nekhoroshev—or some variant of it—is meaningful for a planetary system, possibly with suitable limitations on its configuration?* We have some suggestions in a few particular cases. If we ask more then the answer, I think, is only: *there is still a lot of work to be done*.

> *Oritur sol, et occidit, et ad locum suum revertitur, ibique renascens gyrat per meridiem, et flectitur ad aquilonem . . .*
>
> *. . . et proposui in animo meo quaerere et investigare sapienter de omnibus quae fiunt sub sole. Hanc occupationem pessimam dedit Deus filiis hominum, ut occuparentur in ea.*
>
> (*Qohelet*)

Acknowledgements First of all, I should express my gratitude to the organizers of the School “SDSM 2017” in S. Martino al Cimino. They made a major effort in making the school interesting and fruitful, and so it was, indeed. I thank them for their kind invitation, and I should also thank all the participants for their remarkable patience in enduring my lectures. Last but not least, Marco Sansottera and Giovanni Gronchi made an accurate reading of the manuscript. I greatly appreciated their help and remarks.

Appendix: A Short Overview on Lie Series Methods

Here I recall a few notions concerning Lie series and Lie transforms that are used in the text. Throughout the appendix all functions will be assumed to be holomorphic.

Lie Series

For a given *generating function* $\chi(p, q)$ the Lie series operator is defined as the exponential of the Lie derivative $L_\chi \cdot = \{\cdot, \chi\}$, namely

$$\exp(\varepsilon L_\chi) = \sum_{s\geq 0} \frac{\varepsilon^s}{s!} L_\chi^s \ . \tag{18}$$

This is actually the *autonomous* flow of the canonical vector field generated by $\chi(p, q)$. The flow at time ε is used in order to produce a one-parameter family of canonical transformations that is written as

$$\begin{aligned} p &= \exp(\varepsilon L_\chi)p' = p' - \varepsilon \left.\frac{\partial \chi}{\partial q}\right|_{p',q'} + \frac{\varepsilon^2}{2} L_\chi \left.\frac{\partial \chi}{\partial q}\right|_{p',q'} - \ldots \\ q &= \exp(\varepsilon L_\chi)q' = q' + \varepsilon \left.\frac{\partial \chi}{\partial p}\right|_{p',q'} + \frac{\varepsilon^2}{2} L_\chi \left.\frac{\partial \chi}{\partial p}\right|_{p',q'} + \ldots \ ; \end{aligned}$$

As an operator acting on holomorphic functions the exponential operator is linear and invertible, and has the remarkable properties of distributing over the products and the Poisson brackets of functions, i.e., $\exp(L_\chi)(f \cdot g) = \big(\exp(L_\chi)f\big) \cdot \big(\exp(L_\chi)g\big)$ and $\exp(L_\chi)\{f, g\} = \{\exp(L_\chi)f, \exp(L_\chi)g\}$. The inverse of $\exp\big(\varepsilon L_\chi\big)$ is $\exp\big(\varepsilon L_{-\chi}\big)$, for the flow is autonomous.

The most useful property of the exponential operator has been named *exchange theorem* by Gröbner [30]. It is stated (in a somehow puzzling form) as

$$f(p, q)\Big|_{p=\exp(\varepsilon L_\chi)p',q=\exp(\varepsilon L_\chi)q'} = \exp(\varepsilon L_\chi) f\Big|_{p',q'} \ .$$

The meaning is that an operation of substitution of a near the identity transformation followed by an expansion on the parameter (left side) is replaced by a direct application of the exponential operator to the function (right side): substitutions are avoided.

The application of the operator to a function $f(p, q) = f_0(p, q) + \varepsilon f_1(p, q) + \varepsilon^2 f_2(p, q) + \ldots$ expanded in power series of the parameter ε is nicely represented by the triangular diagram for Lie series of Table 4. Terms of the same order in ε are aligned on the same row. Remark that the triangle is generated *by columns*: every

Table 4 The triangular diagram for Lie series

g_0	f_0					
	↓					
g_1	$L_{\chi_1} f_0$	f_1				
	↓	↓				
g_2	$\frac{1}{2} L_{\chi_1}^2 f_0$	$L_{\chi_1} f_1$	f_2			
	↓	↓	↓			
g_3	$\frac{1}{3!} L_{\chi_1}^3 f_0$	$\frac{1}{2} L_{\chi_1}^2 f_1$	$L_{\chi_1} f_2$	f_3		
	↓	↓	↓	↓		
g_4	$\frac{1}{4!} L_{\chi_1}^4 f_0$	$\frac{1}{3!} L_{\chi_1}^3 f_1$	$\frac{1}{2} L_{\chi_1}^2 f_2$	$L_{\chi_1} f_3$	f_4	
	↓	↓	↓	↓	↓	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 5 The triangular diagram for a generating function of order ε^2

g_0	f_0					
	↓					
g_1	·	f_1				
	↓	↓				
g_2	$L_{\chi_2} f_0$	·	f_2			
	↓	↓	↓			
g_3	·	$\frac{1}{2} L_{\chi_2} f_1$	·	f_3		
	↓	↓	↓	↓		
g_4	$\frac{1}{2} L_{\chi_2}^2 f_0$	·	$L_{\chi_2} f_2$	·	f_4	
	↓	↓	↓	↓	↓	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

column may be calculated separately once the upper term is known. If the function f is known, then the coefficients of the ε expansion of the transformed function $g = \exp(L_\chi) f$ are calculated by adding up all terms on the same line. The result may be expressed by the formula

$$g_0 = f_0 \ , \quad g_s = \sum_{j=0}^{s} \frac{1}{j!} L_{\chi_1}^j f_{s-j} \ , \quad s \geq 1 \ .$$

A generating function χ_2 of order ε^2 generates a similar triangle, which, however, will contain many empty cells, as represented in Table 5.

A general formula for the transformation of a function with a generating function of order ε^r is

$$\begin{aligned} &g_0 = f_0 \ , \ldots, \ g_{r-1} = f_{r-1} \ , \\ &g_s = \sum_{j=0}^{k} \frac{1}{j!} L_{\chi_r}^j f_{s-jr} \ , \quad k = \left\lfloor \frac{s}{r} \right\rfloor \ , \quad s \geq r \end{aligned}$$

Remark that the first change occurs at order $r+1$.

Lie series operators of increasing order may be formally composed as follows. Let $\chi = \{\chi_1(p,q),\ \chi_2(p,q),\ \ldots\}$ be a sequence of generating functions of increasing orders $\varepsilon,\ \varepsilon^2,\ \ldots$; the composition is formally defined as

$$S_\chi = \ldots \circ \exp\left(L_{\chi_3}\right) \circ \exp\left(L_{\chi_2}\right) \circ \exp\left(L_{\chi_1}\right)$$

We may also use the recursive definition of a sequence of operator $S_1,\ S_2,\ S_3,\ \ldots$

$$S_1 = \exp\left(L_{\chi_1}\right)\ , \quad S_r = \exp\left(L_{\chi_r}\right) \circ S_{r-1}\ ,$$

considering S_χ as the limit (in formal sense) of the latter sequence for $r \to \infty$.

Compositions of Lie series are unavoidable in view of the following property: *every near the identity canonical transformation of coordinates*

$$p = p' + \varphi_1(p',q') + \varphi_2(p',q') + \ldots\ , \quad q = q' + \psi_1(p',q') + \psi_2(p',q') + \ldots$$

may be represented by a composition of Lie series. In general this is untrue for a single Lie series. For this reason the composition of Lie series is often replaced by the algorithm of *Lie transform*, introduced independently by Hori [34] and Deprit [15]. The two methods are formally equivalent. However, the composition of Lie series is in definitely better position as regards the *convergence* question (for instance in the case of Kolmogorov's theorem). If the reader tries to reformulate the control of small divisors in the present notes using the Lie transform he or she will likely fail.

An Algorithm for Lie Transform

Contrary to Lie series, Lie transform can be constructed in a number of different ways. Here I present one of the formulations. Given a sequence $\{\chi_1,\ \chi_2,\ \ldots\}$ of generating functions define the *Lie transform operator* as

$$T_\chi = \sum_{s\geq 0} E_s \tag{19}$$

with the sequence E_s of linear operators recursively defined as

$$E_0 = 1\ , \quad E_s = \sum_{j=1}^{s} \frac{j}{s} L_{\chi_j} E_{s-j}\ . \tag{20}$$

The operator may be seen as a generalization of the exponential operator of Lie series. A straightforward remark is that if we choose the generating sequence

$\chi = \{\chi_1, 0, 0, \ldots\}$ then $T_\chi = \exp(L_{\chi_1})$. Moreover T_χ has the same properties of the exponential operator of Lie series: it is linear and invertible, and distributes over products and Poisson brackets, i.e., $T_\chi(f \cdot g) = T_\chi f \cdot T_\chi g$ and $T_\chi \{f, g\} = \{T_\chi f, T_\chi g\}$ for any pair $f,\ g$ of functions. The inverse requires some care: it has an elaborate expression which requires a second sequence of operators:

$$T_\chi^{-1} = \sum_{s\geq 0} G_j\ , \quad G_0 = \mathsf{1}\ , \quad G_s = -\sum_{j=1}^{s} \frac{j}{s} G_{s-j} L_{\chi_j}\ . \tag{21}$$

However, using the latter formula for an actual calculation is not recommended: we shall see in a short a more effective method. The formula is useful for analytical convergence estimates. It should be remarked that the inverse is not elementary because the Lie transform may be interpreted as generated by the flow of a *non autonomous* vector field, which can not be inverted by a mere change of sign of the vector field (as it happens for Lie series). Precisely the latter idea is developed in the paper of Deprit [15].

Finally, T_χ possesses the property expressed by the exchange theorem, namely

$$f(p,q)\Big|_{p=T_\chi p', q=T_\chi q'} = T_\chi f\Big|_{p',q'}\ .$$

The scheme of application of T_χ may also be represented by a triangular diagram similar to that of Lie series, as represented in Table 6. Here too the triangle is filled in by columns, and a function $g = T_\chi f$ is found by adding up all contributions on the same line. The diagram also provides a straightforward method for calculating the inverse $f = T_\chi^{-1} g$. Just proceed as follows: from the first line get $f_0 = g_0$, and fill the column for f_0 ; from the second line get $f_1 = g_1 - E_1 f_0$, and fill the column for f_1 ; from the third line get $f_2 = g_2 - E_1 f_1 - E_2 f_0$, and fill the column for f_2 , and so on.

Table 6 The triangular diagram for Lie transform

g_0	f_0					
	↓					
g_1	$E_1 f_0$	f_1				
	↓	↓				
g_2	$E_2 f_0$	$E_1 f_1$	f_2			
	↓	↓	↓			
g_3	$E_3 f_0$	$E_2 f_1$	$E_1 f_2$	f_3		
	↓	↓	↓	↓		
g_4	$E_4 f_0$	$E_3 f_1$	$E_2 f_2$	$E_1 f_3$	f_4	
	↓	↓	↓	↓	↓	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Analytical Tools

Here I introduce some basic tools that allow us to discuss the convergence of Lie series and of composition of Lie series. I shall restrict my attention to the case of a phase space endowed with action-angle variables $p \in \mathcal{G} \subset \mathbb{R}^n$ and $q \in \mathbb{T}^n$, as considered in the present notes. However, the whole argument is based on the theory of holomorphic functions.

The first step requires introducing a family of complex domains

$$\mathcal{D}_{(1-d)(\varrho,\sigma)} = \Delta_{(1-d)\varrho} \times \mathbb{T}^n_{(1-d)\sigma}$$

with fixed $\varrho, \sigma > 0$ and $0 \le d < 1$; here

$$\Delta_\varrho = \left\{ p \in \mathbb{C}^n \; : \; |p| \le \varrho \right\} , \quad \mathbb{T}^n_\sigma = \left\{ q \in \mathbb{C}^n \; : \; |\operatorname{Im} q| \le \sigma \right\} . \tag{22}$$

In the case of one degree of freedom the domain is represented in Fig. 17. The action domain here is a polydisk Δ_ϱ centered at the origin of $\mathbb{C}^n$, which is enough for the proof of the theorem of Kolmogorov. However, the whole argument may be extended to the case of a complex domain $\mathcal{G}_\varrho = \bigcup_{p\in\mathcal{G}} \Delta_\varrho(p)$ constructed by making the union of all complex disks of radius ϱ centered at every point of the real domain $\mathcal{G}$ of the actions.

The second step is concerned with the extension of Cauchy's estimates for the derivatives of holomorphic functions to the case of Lie derivatives. For a function $f(p,q)$ which is holomorphic in $\mathcal{D}_{(\varrho,\sigma)}$ we shall use the supremum norm

$$\left| f \right|_{(\varrho,\sigma)} = \sup_{(p,q)\in\mathcal{D}_{(\varrho,\sigma)}} |f(p,q)| . \tag{23}$$

We assume that $\left| f \right|_{(\varrho,\sigma)}$ is finite. Following Cauchy, the derivatives of the function $f(p,q)$ are estimated as

$$\left| \frac{\partial f}{\partial p} \right|_{(1-d)(\varrho,\sigma)} \le \frac{1}{d\varrho} \left| f \right|_{(\varrho,\sigma)} , \quad \left| \frac{\partial f}{\partial q} \right|_{(1-d)(\varrho,\sigma)} \le \frac{1}{d\sigma} \left| f \right|_{(\varrho,\sigma)} .$$

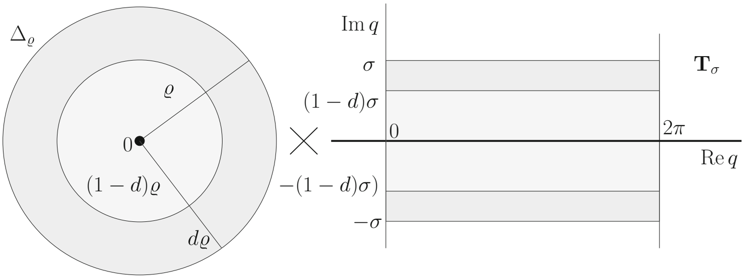


Fig. 17 Construction of the family of complex domains

Higher order derivatives can be estimated, too. However, for our purposes, it is better to obtain estimates for Lie derivatives. An appropriate approach is the following. Assume that we know the norm $|\chi|_{\varrho,\sigma}$ of a generating function χ on the whole domain and the norm $\|f\|_{(1-d')(\varrho,\sigma)}$ in a possibly smaller domain, with $0 \le d < 1$. Then for $d' < d < 1$ one gets generally an estimate such as

$$\left|L_\chi f\right|_{(1-d)(\varrho,\sigma)} \le \frac{C}{d(d-d')\varrho\sigma}|\chi|_{\varrho,\sigma}|f|_{(1-d')(\varrho,\sigma)}\ . \tag{24}$$

with some constant $C \ge 1$ depending on the choice of the norm (and on the method of estimate). In the present case of the supremum norm a straightforward calculation gives $C = 2n$, because the Poisson bracket is expressed by the sum of $2n$ products of derivatives. However, a more careful estimate, using the fact that we are performing a derivative in a given direction, provides the better value $C = 1$.

The estimate of multiple Lie derivatives is more delicate. Suppose we know $|\chi|_{\varrho,\sigma}$ and $|f|_{\varrho,\sigma}$ on the common domain $\mathcal{D}_{\varrho,\sigma}$. If we want the evaluate $\left|L^s_\chi f\right|_{(1-d)(\varrho,\sigma)}$ in a restricted domain we can define $\delta = d/s$ and estimate, in sequence,

$$\left|L_\chi f\right|_{(1-\delta)(\varrho,\sigma)}\ ,\ \left|L^2_\chi f\right|_{(1-2\delta)(\varrho,\sigma)}\ ,\ \ldots\ ,\ ,\ \left|L^s_\chi f\right|_{(1-s\delta)(\varrho,\sigma)}\ .$$

To this end we apply by recurrence the estimate (24) for a single derivative, setting step by step $d' = 0, \delta, \ldots, (s-1)\delta$. With some calculations we end up with the estimate (setting $C = 1$)

$$\frac{1}{s!}\left|L^s_\chi f\right|_{(1-d)(\varrho,\sigma)} \le \frac{1}{e}\left(\frac{e}{d^2\varrho\sigma}\right)^s |\chi|^s_{\varrho,\sigma}|f|_{(1-d')(\varrho,\sigma)}\ .$$

Convergence of Lie Series and of the Composition of Lie Series

Substituting the latter estimate in the expression of Lie series we prove

Lemma 8 *Let $\chi(p,q)$ be holomorphic and bounded in $\mathcal{D}_{(\varrho,\sigma)}$. If the convergence condition*

$$|\chi|_{(\varrho,\sigma)} < \frac{d^2\varrho\sigma}{2e}\ ,\quad d < 1/2$$

is satisfied, then the near the identity canonical transformation

$$p' = \exp\left(L_\chi\right)p\ ,\quad q' = \exp\left(L_\chi\right)q$$

is holomorphic in $\mathcal{D}_{(1-d)(\varrho,\sigma)}$, and

$$|p - p'| < d\varrho\ ,\quad |q - q'| < d\sigma\ .$$

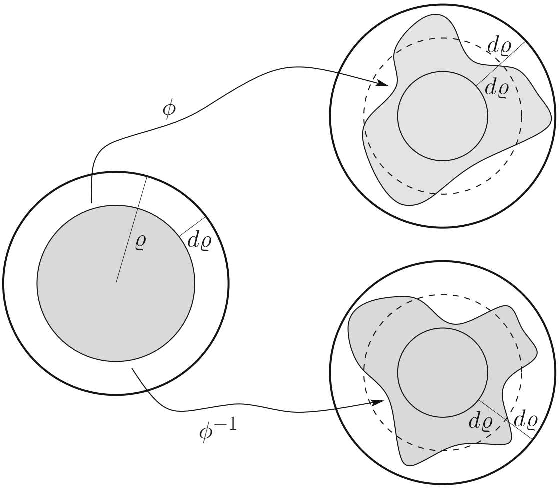


Fig. 18 The deformation induced by the near the identity transformation of Lemma 8. The flow is denoted by ϕ, with inverse ϕ^{-1}

By the way, the lemma is actually a reformulation of Cauchy's theorem on existence and uniqueness of a local flow in the analytic case. The implications of the lemma can be understood looking at Fig. 18 and recalling that the transformation is defined through the flow generated by $\chi(p, q)$. The transformation is essentially a deformation of coordinates. Therefore if we consider it as defined on a domain $\mathcal{D}_{(1-d)(\varrho,\sigma)}$, with $d < 1/2$ then the relation

$$\mathcal{D}_{(1-2d)(\varrho,\sigma)} \subset \mathcal{D}_{(1-d)(\varrho,\sigma)} \subset \mathcal{D}_{\varrho,\sigma} \tag{25}$$

holds true, so that there is a domain where the transformation is well defined.

Coming to the composition of Lie series, we may intepret it as a composition of flows. Therefore we should check that the relations (25) are still satisfied. The final result is expressed by

Proposition 9 *Let the sequence of generating functions $\chi = \{\chi_1, \chi_2, \ldots\}$ be holomorphic and bounded in $\mathcal{D}_{(\varrho,\sigma)}$. If the convergence condition*

$$\sum_{j\geq 1} \left|\chi_j\right|_{(\varrho,\sigma)} < \frac{d^2\varrho\sigma}{4e} , \quad d < 1/2$$

is satisfied, then the near the identity canonical transformation

$$p' = S_\chi p , \quad q' = S_\chi q$$

generated by the infinite composition of Lie series

$$S_\chi = \ldots \circ \exp\left(L_{\chi_3}\right) \circ \exp\left(L_{\chi_2}\right) \circ \exp\left(L_{\chi_1}\right)$$

is holomorphic in $\mathcal{D}_{(1-d)(\varrho,\sigma)}$, *and*

$$|p - p'| < d\varrho \ , \quad |q - q'| < d\sigma \ .$$

Similar results may be obtained also for the algorithm of Lie transform. However, they are not needed for the purposes of the present notes, thus I omit them.

References

1. Andoyer, H.: Cours De mécanique Céleste. Atlantis Press, Paris (1923)
2. Arnold, V.I.: Proof of a theorem of A.N. Kolmogorov on the invariance of quasi–periodic motions under small perturbations of the Hamiltonian. Usp. Mat. Nauk. **18**, 13 (1963); Russ. Math. Surv. **18**, 9 (1963)
3. Arnold, V.I.: Small denominators and problems of stability of motion in classical and celestial mechanics. Usp. Math. Nauk. **18**(6), 91 (1963); Russ. Math. Surv. **18**(6), 85 (1963)
4. Arnold, V.I.: A theorem of Liouville concerning integrable problems of dynamics. Sibirsk. Math. Zh. **4**, 471–474 (1963)
5. Benettin, G., Galgani, L., Giorgilli, A., Strelcyn, J.M.: A proof of Kolmogorov's theorem on invariant tori using canonical transformations defined by the Lie method. Il Nuovo Cimento **79**(B), 201–223 (1984)
6. Birkhoff, G.D.: Dynamical Systems. American Mathematical Society, New York (1927)
7. Celletti, A., Chierchia, L.: On the stability of realistic three body problems. Commun. Math. Phys. **186**, 413–449 (1997)
8. Celletti, A., Falcolini, C.: Construction of invariant tori for the spin-orbit problem in the Mercury-Sun system. Celest. Mech. Dyn. Astron. **53**, 113–127 (1992)
9. 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)
10. Chirikov, B.V.: A universal instability of many dimensional oscillator system. Phys. Rep. **52**, 263–379 (1979)
11. Contopoulos, G.: A third integral of motion in a Galaxy. Z. Astrophys. **49**, 273–291 (1960)
12. Contopoulos, G.: In: Nahon, F., Hénon, M. (eds.) Les nouvelles Méthodes de la Dynamique Stellaire. CNRS, Paris (1966); see also Bull. Astron. Ser. 3, **2**, Fasc. 1, 233 (1967)
13. Contopoulos, G., Efthymiopoulos, C., Giorgilli, A.: Non-convergence of formal integrals of motion. J. Phys. A Math. Gen. **36**, 8639–8660 (2003)
14. Delaunay, C.: Théorie du mouvement de la lune. Memoir **28** Academy of Sciences France, Paris (1860)
15. Deprit, A.: Canonical transformations depending on a small parameter. Celest. Mech. **1**, 12–30 (1969)
16. de Laplace, P.-S.: Mémoire sur le principe de la gravitation universelle et sur les inégalités séculaires des planètes qui en dependent. Mémoires de l'Académie Royale des Sciences de Paris (1773). Reprinted in Oeuvres complètes de Laplace. Gauthier–Villars, Paris (1891), tome VIII, pp. 201–275
17. de Laplace, P–S.: Théorie de Jupiter et Saturne. Mémoires de l'Académie Royale des Sciences de Paris, année 1785, (1788). Reprinted in Oeuvres complètes de Laplace. Gauthier–Villars, Paris (1891), tome XI, p. 95
18. Efthymiopoulos, C., Sándor, Z.: Optimized Nekhoroshev stability estimates for the Trojan asteroids with a symplectic mapping model of co-orbital motion. Mon. Not. R. Astron. Soc. **364**(1), 253–271 (2005)

19. Efthymiopoulos, C., Contopoulos, G., Giorgilli, A.: Non-convergence of formal integrals of motion II: improved estimates for the optimal order of truncation. J. Phys. A Math. Gen. **37**, 10831–10858 (2004)
20. Giorgilli, A.: Rigorous results on the power expansions for the integrals of a Hamiltonian system near an elliptic equilibrium point. Ann. de l'I.H.P. Prog. Theor. Phys. **48**, 423–439 (1988)
21. Giorgilli, A.: Notes on exponential stability of Hamiltonian systems. In: Dynamical Systems, Part I: Hamiltonian Systems and Celestial Mechanics, 87–198. Pubblicazioni del Centro di Ricerca Matematica Ennio De Giorgi, Pisa (2003)
22. Giorgilli, A.: A Kepler's note on secular inequalities. Rendiconti dell'Istituto Lombardo Accademia di Scienze e Lettere, Classe di Scienze Matematiche e Naturali, **145**, 97–119 (2011)
23. Giorgilli, A., Marmi, S.: Convergence radius in the Poincaré-Siegel problem. DCDS Ser. S **3**, 601–621 (2010)
24. Giorgilli, A., Morbidelli, A.: Invariant KAM tori and global stability for Hamiltonian systems. ZAMP **48**, 102–134 (1997)
25. Giorgilli, A., Skokos, C.: On the stability of the Trojan asteroids. Astron. Astrophys. **317**, 254–261 (1997)
26. Giorgilli, A., Zehnder, E.: Exponential stability for time dependent potentials. ZAMP **5**, 827–855 (1992)
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. Differ. Equ. **20**, (1989)
28. Giorgilli, A., Locatelli, U., Sansottera, M.: Kolmogorov and Nekhoroshev theory for the problem of three bodies. Celest. Mech. Dyn. Astron. **104**, 159–175 (2009)
29. Giorgilli, A., Locatelli, U., Sansottera, M.: Secular dynamics of a planar model of the Sun–Jupiter–Saturn–Uranus system; effective stability into the light of Kolmogorov and Nekhoroshev theories. Regul. Chaotic Dyn. **22** 54–77 (2017)
30. Gröbner, W.: Die Lie–Reihen und Ihre Anwendungen. VEB Deutscher Verlag der Wissenschaften, Berlin (1967)
31. Gyldén, H.: Untersuchungen über die convergenz der reigen, welche zur darstellung der coordinaten der planeten angewendet werden. Acta **9**, 185–294 (1887)
32. Haretu, S.C.: Thèses Presentès a la Faculté des Sciences de Paris. Gauthier-Villars, Paris (1878)
33. Haretu, S.C.: Sur l'invariabilité des grands axes des orbites planétaires. Ann. Obs. Paris Mém. **18**, 1–39 (1885)
34. Hori, G.: Theory of general perturbations with unspecified canonical variables. Publ. Astron. Soc. Jpn. **18**, 287–296 (1966)
35. Jost, R.: Winkel–und Wirkungsvariable für allgemeine mechanische Systeme. Helv. Phys. Acta **41**, 965–968 (1968)
36. Kepler, J.: Consideratio observationum Regiomontani et Waltheri, published in: Johannis Kepleri astronomi opera omnia, MDCCCLX, Vol. VI, pp. 725–774
37. Kolmogorov, A.N.: Preservation of conditionally periodic movements with small change in the Hamilton function. Dokl. Akad. Nauk SSSR **98**, 527 (1954). English translation in: Los Alamos Scientific Laboratory translation LA-TR-71-67; reprinted in Lecture Notes in Physics **93**
38. Lagrange, J.L.: Recherche sur les équations séculaires des mouvements des noeuds et des inclinaisons des orbites des planètes. Mémoires de l'Académie Royale des Sciences de Paris (1774). Reprinted in Oeuvres de Lagrange, Gauthier–Villars, Paris (1870), tome VI, pp. 635–709
39. Lagrange, J.L.: Sur l'altération des moyens mouvements des planètes. Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Berlin (1776). Reprinted in Oeuvres de Lagrange. Gauthier–Villars, Paris (1867), tome IV, pp. 255–271

40. Lagrange, J.L.: Théorie des variations séculaires des éléments des planètes. Première partie contenant les principes et les formules générales pour déterminer ces variations. Nouveaux mémoires de l'Académie des Sciences et Belles–Lettres de Berlin (1781). Reprinted in Oeuvres de Lagrange, Gauthier–Villars, Paris (1870), tome V, pp.125–207
41. Lagrange, J.L.: Théorie des variations séculaires des éléments des planètes. Seconde partie contenant la détermination de ces variations pour chacune des planètes pricipales. Nouveaux mémoires de l'Académie des Sciences et Belles–Lettres de Berlin (1782). Reprinted in Oeuvres de Lagrange, Gauthier–Villars, Paris (1870), tome V, pp.211–489
42. Laskar, J.: The chaotic motion of the solar system: a numerical estimate of the size of the chaotic zones. Icarus **88**, 266 (1990)
43. Laskar, J.: Large scale chaos in the solar system. Astron. Astrophys. **287** (1994)
44. Laskar, J.: Lagrange et la Stabilité du Sytème Solaire. In: Sacchi Landriani, G., Giorgilli, A. (eds.) Sfogliando la Méchanique Analitique. LED edizioni, Milano (2008)
45. Lindstedt, A.: Beitrag zur integration der differentialgleichungen der differentialgleichungen der störungstheorie. Mém. Acad. Imp. des sciences St. Pétersbourg. **XXXI**, 4 (1883)
46. Liouville, M.J.: Sur l'intégrations des équations différentielles de la dynamique. J. de Mathématiques pures et appliquées tome XX, 137–138 (1855)
47. Littlewood, J.E.: On the equilateral configuration in the restricted problem of three bodies. Proc. Lond. Math. Soc. **9**(3), 343–372 (1959)
48. Littlewood, J.E.: The Lagrange configuration in celestial mechanics. Proc. Lond. Math. Soc. **9**(3), 525–543 (1959)
49. Locatelli, U., Giorgilli, A.: Invariant tori in the secular motions of the three-body planetary systems. Celest. Mech. Dyn. Astron. **78**, 47–74 (2000)
50. Locatelli, U., Giorgilli, A.: Invariant tori in the Sun–Jupiter–Saturn system. DCDS-B **7**, 377–398 (2007)
51. Lyapunov, A.M.: The General Problem of the Stability of Motion (in Russian). Doctoral dissertation, University of Kharkov, Kharkov (1892). French translation in: *Problème général de la stabilité du mouvement*, Annales de la Faculté des Sciences de Toulouse, deuxième série, Tome IX, 203–474 (1907). Reprinted in Ann. Math. Study, Princeton University Press, n. 17, (1949)
52. Morbidelli, A., Giorgilli, A.: Superexponential stability of KAM tori. J. Stat. Phys. **78**, 1607–1617 (1995)
53. Moser, J.: Stabilitätsverhalten kanonisher differentialgleichungssysteme. Nachr. Akad. Wiss. Göttingen, Math. Phys. K1 IIa. **6**, 87–120 (1955)
54. Moser, J.: On invariant curves of area-preserving mappings of an annulus. Nachr. Akad. Wiss. Gött., II Math. Phys. Kl. **1962**, 1–20 (1962)
55. Nekhoroshev, N.N.: Exponential estimates of the stability time of near-integrable Hamiltonian systems. Russ. Math. Surv. **32**, 1 (1977)
56. Nekhoroshev, N.N.: Exponential estimates of the stability time of near-integrable Hamiltonian systems, 2. Trudy Sem. Petrovs. **5**, 5 (1979)
57. Newton, I.: Opticks: or, A Treatise of the Reflections. Refractions, Inflections and Coulors of Light, London (1704)
58. Poincaré, H.: Les méthodes Nouvelles de la Mécanique Céleste. Gauthier–Villars, Paris (1892)
59. Poincaré, H.: Le\ccedillaons de Mécanique Céleste. Professées a la Sorbonne, Tome I, Théorie générale des Perturbations Planetaires. Gautier–Villars, Paris (1905)
60. Poisson, S.: Mémoire sur la Variation des Constantes Arbitraires Dans les Questions de Mécanique. J. de l'École Polythecnique quinzième cahier, tome VIII (1809)
61. Sansottera, M., Locatelli, U., Giorgilli, A.: On the stability of the secular evolution of the planar Sun–Jupiter–Saturn–Uranus system. Math. Comput. Simul. **88**, 1–14 (2013)
62. Servizi, G., Turchetti, G., Benettin, G., Giorgilli, A.: Resonances and asymptotic behaviour of Birkhoff series. Phys. Lett. A **95**, 11–14 (1983)
63. Siegel, C.L.: On the integrals of canonical systems. Ann. Math. **42**, 806–822 (1941)

Space Debris: From LEO to GEO

Anne Lemaître

Abstract The paper focuses on the dynamics of space debris in the Earth environment, with a celestial mechanics and theoretical point of view, and not with an operational perspective. The introduction describes the Earth space junk, with the description and the evolution of the debris population, and lists the main forces acting on them, their relative importance and the main regions of interest (Low, Medium and Geostationary Orbits, later called LEO, MEO and GEO). The resonances are present at several levels: gravitational resonances, for MEO and GEO, but also lunar-solar resonances, and secondary resonances involving the Sun. A classical Hamiltonian approach is proposed for GEO or MEO regions, with different associated toys models. The numerical integrations, their limits, their characteristics, symplectic or not, for short or long time integrations are presented, commented and compared, with the connected chaotic indicators (MEGNO in particular) which allow to put the stability of some regions into perspective. The solar radiation pressure is investigated with more details, without or with shadowing effects especially in the GEO region. For the LEO, the atmospheric drag plays an important role on the dynamics, dependent on the ballistic coefficient. Some comparisons are presented, concerning the solar activity and the consequences on the reentry times. A few words about the rotation of the debris, the explosions and collisions mechanisms, and the possibility to simulate those events in a synthetic population conclude the paper.

Keywords Space debris · Solar radiation pressure · Symplectic integration · Drag · Synthetic population

A. Lemaître (✉)
naXys, University of Namur, Namur, Belgium
e-mail: anne.lemaitre@unamur.be

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_3

1 Introduction

The term *space debris* is used to designate all the objects (fragments of satellites, rocket parts, remains of explosions or collisions), of all sizes and all chemical compositions, which orbit around the Earth at different altitudes.

The number of space debris has dramatically increased in the last decades; since Sputnik in October 1957, more than 6600 satellites have been launched and more than 200 exploded in space, for accidental or political reasons. Indeed because of the cold war or for similar political reasons, the nations did not hesitate to provoke the explosion of a satellite to keep their innovative technologies. Moreover some space missions did not care about the situation of a satellite after its lifetime and some of them are orbiting the Earth for years after their official end. Only 6 % of the objets in orbits are active satellites, etc.

It was commonly thought that the drag would rapidly bring back to the Earth most of the objects; it is true for the low orbits (LEO) with small altitudes, between 500 and 900 km, but certainly not for the highest ones, in particular for the geostationary orbits (GEO) situated at 36,000 km of altitude and characterized by periods of exactly 24 h. The lifetime of a classical satellite is estimated to a month, for an altitude of 300 km, a year for 400 km, 10 years for 500 km, decades for 700 km, centuries for 900 km and millennia for 1200 km.

Even if the drag is able to clean progressively the LEO region, the presence of a huge number of debris is responsible for collisions and consequently, generates a continuous re-population of the region.

A catalogue of about 20,000 debris is maintained and completed by NASA at each registered collision. It contains the objects larger than 10 cm for LEOs (and larger than 1 m for GEOs) which could really damage an active satellite and compromise its mission. They are stocked in a file called *the two line elements* or *TLE* as reference to their format. More precisely, a two-line element set is a data format used to convey sets of orbital elements that describe the orbits of Earth-orbiting objects. A computer program called a *model* can use the TLE to compute the position of any satellite or debris at any particular time. Usually the model is a powerful numerical integrator, but valid only for short time evolutions.

The risk of collision is real and avoidance manoeuvers are performed regularly, by the active satellites, the space shuttles or the ISS, increasing the cost of the missions by consuming fuel.

However the TLEs only refer to the huge objects and represent the tip of the iceberg. The smallest debris are much more numerous, with rough estimations of about 200,000 objects between 1 and 10 cm and more than 35 millions of objects between 0.1 and 1 cm. These debris are neither catalogued nor individually identified. Knowledge of space debris environment at sub-catalogue sizes is normally acquired in a statistical manner through experimental sensors with higher sensitivities.

While telescopes are mainly suited for GEO and high-altitude debris observations, radars are advantageous in the low-Earth orbit (LEO) regime, below 2000 km. Ground-based telescopes can detect GEO debris down to 10 cm in size, ground-based radars can detect LEO debris down to a few mm in size, and in situ impact detectors can sense objects down to a few micrometers in size.

We also can gain information on the small-size, sub-millimeter environment through the analysis of retrieved space hardware, such as the EURECA satellite, and the three solar arrays retrieved from the Hubble Space Telescope through the Space Shuttle.

Even if we stopped all launches, the number of debris would still increase for several years just by collisions and fragmentations of the present objects in orbits. Let imagine the situation with more than 100 launches per year!

Special equipments and armor plating protections are now systematically scheduled for the spacecrafts, increasing their cost and requiring always more powerful rockets (because of their weight). Even if they are expensive, these protections are efficient for the small debris below the centimeter size, but are often inefficient for larger ones. This is why, presently, the most dangerous population is the intermediary one, between 1 and 10 cm, where the objects are too small to be followed individually but too big to be considered only as a dusty environment altering the surfaces.

Our lifestyle is really dependent on the presence of spacecrafts: telecommunications, GPS or cellular phones, TV, Internet, climate watches, ecological studies, catastrophe prevention, military surveys, etc. Despite the technological progress, the costs and the risks due to the space debris are increasing and can really stop or drastically reduce the systematic replacement or extension of the present satellite constellations.

Since 1978, NASA has developed guidelines to keep down the amount of debris generated by space launches and to minimize the possibility of later fragmentations. Other countries soon followed suit, and in 2007, after 10 years of intense debate and negotiation, the United Nations General Assembly approved a set of guidelines for orbital space debris mitigation.

These can be summed up under seven points:

1. Limit debris released during normal operations
2. Minimize the potential for break-ups during operational phases
3. Limit the probability of accidental collision in orbit
4. Avoid intentional destruction and other harmful activities
5. Minimize potential for post-mission breakups resulting from stored energy
6. Limit the long-term presence of spacecraft and launch vehicle orbital stages in the low Earth orbit region after the end of their mission
7. Limit the long-term interference of spacecraft and launch vehicle orbital stages with the geosynchronous region after the end of their mission

However it will not be sufficient and the experts are quite pessimistic:

The buildup of space debris orbiting the Earth, which poses a threat to spacecraft and the environment, has reached a critical point. The space junk trend no longer can be reversed by full compliance with mitigation measures now in place; it will get worse without more-aggressive action such as active debris removal (ADR).

And that just might pose the biggest engineering challenge of the 21st century. As the international community gradually reaches a consensus on the need for ADR, the focus will shift from environment modeling to completely different challenges: technology development, systems engineering, and operations, J.-C. Liou, Orbital Debris Program Office at NASA's Johnson Space Center.

The European Space Agency (ESA) has developed the Space Situational Awareness (SSA) program, formally launched 1 January 2009. The objective of the SSA program is to support Europe's independent use of, and access to, space through the provision of timely and accurate information, data and services regarding the space environment, and particularly regarding hazards to infrastructure in orbit and on the ground. In general, these hazards stem from possible collisions between objects in orbit, harmful space weather and potential strikes by natural objects that cross Earth's orbit. The SSA program will, ultimately, enable Europe to autonomously detect, predict and assess the risk to life and property due to remnant man-made space objects, re-entries, in-orbit explosions and release events, in-orbit collisions, disruption of missions and satellite-based service capabilities, potential impacts of Near Earth Objects, and the effects of space weather phenomena on space- and ground-based infrastructure.

The technological challenge for ADR has to be supported by a better understanding of the evolution of the present future debris population.

At the end of the nineties, the space debris population has interested the community of celestial mechanicians, traditionally involved in the dynamics of natural bodies. Before that epoch, on the one hand, the space agencies used very efficient numerical integrations, including a maximum of forces and contributions in the dynamics, and integrating the motions on very short periods of time, as for the probes or spacecraft, limited to their lifetime. On the other hand, the astronomers developed more and more sophisticated tools to integrate natural bodies on longer time periods (using symplectic integrators or very fast mappings) or to produce global maps of chaos and stability, privileging the global behaviors to the individual ones.

The space debris dynamics, with uncontrolled objects present for thousands of years offers to the celestial mechanicians the opportunity of adapting, testing and developing known techniques and methods, in a new environment.

Let us mention a few topics, linked to resonances, chaos and perturbations, in which the celestial mechanics approach has recently given new tools to the space debris dynamics.

2 The Classical Hamiltonian Formulation

The different forces (at least the conservative ones) are expressed through their potential, and added to the two body basic expression in the Hamiltonian formalism. The gravitational potential and the lunisolar perturbations are the classical perturbations of a Keplerian 2-body problem.

2.1 The Gravitational Potential

The Earth is not a perfect sphere, and the space debris, as the artificial satellites, are close enough to the surface to suffer substantial perturbations due to the non sphericity coefficients, the spherical harmonics. The potential $\mathscr{U}$ is general expressed in the following way:

$$\mathscr{U}(\mathbf{r}) = -\mu \int_V \frac{\rho(\mathbf{r_p})}{\|\mathbf{r} - \mathbf{r_p}\|} \, dV \, , \quad \mu = G\, m_E, \tag{1}$$

where $\mathbf{r}$ is the position of the piece of debris, expressed by its three coordinates, x, y, z in the geocentric equatorial reference frame, and r is its norm. $\mathbf{r_p}$ is the position of any point of the Earth. Let us consider that μ is simply GM_S, G is the gravitational constant and M_E the Earth's mass.

We can expressed $\mathbf{r}$ in spherical coordinates, r, λ being its longitude and Φ its latitude:

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

and the geopotential becomes:

$$\mathscr{U}(r, \lambda, \Phi) = -\frac{\mu}{r} \sum_{n=0}^{\infty} \sum_{m=0}^{n} \left(\frac{R_e}{r}\right)^n \mathscr{P}_{nm}(\sin\Phi)(C_{nm} \cos m\lambda + S_{nm} \sin m\lambda) \tag{2}$$

with R_e the equatorial Earth's radius and $\mathscr{P}_{nm}$ is the Legendre polynomial of degree n and order m.

The coefficients C_{nm} and S_{nm} are given by:

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

$$S_{nm} = \frac{2 - \delta_{0m}}{M_S} \frac{(n-m)!}{(n+m)!} \int_V \left(\frac{r_p}{R_e}\right)^n \mathscr{P}_n^m(\sin\Phi_p) \sin(m\lambda_p) \, \rho(\mathbf{r_p}) \, dV$$

where δ_{0m} is the Kronecker symbol, (x_p, y_p, z_p) are the coordinates of $\mathbf{r}_p$ and are expressed in spherical coordinates by:

$$\begin{aligned} x_p &= r_p \, \cos\Phi_p \, \cos\lambda_p \\ y_p &= r_p \, \cos\Phi_p \, \sin\lambda_p \\ z_p &= r_p \, \sin\Phi_p \end{aligned}$$

with r_p the norm, Φ_p the latitude and λ_p the longitude.

The two largest coefficients are C_{20} and C_{22} and are directly linked to the principal momenta of inertia, A, B and C, the Earth's mass, M_E and R_e the equatorial radius.

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

After some simplifications due to the choice of the center of mass as the origin or the reference frame, and use of the polar formulation:

$$\mathscr{U}(r, \lambda, \Phi) = -\frac{\mu}{r} + \frac{\mu}{r} \sum_{n=2}^{\infty} \sum_{m=0}^{n} \left(\frac{R_e}{r}\right)^n \mathscr{P}_n^m(\sin\Phi)\, J_{nm}\, \cos m(\lambda - \lambda_{nm}) \tag{3}$$

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

Usually the development is replaced by *Kaula's formulation* [24] introducing explicit functions of the elliptic elements, the eccentricity e, the inclination i, the argument of perigee, ω, the longitude of the node Ω, the mean anomaly M, of the piece of debris, related to the orbital motion in the Earth's equatorial frame. θ is the sidereal time (representing the rotation of the Earth).

$$\mathscr{U} = -\frac{\mu}{r} - \sum_{n=2}^{\infty} \sum_{m=0}^{n} \sum_{p=0}^{n} \sum_{q=-\infty}^{+\infty} \frac{\mu}{a} \left(\frac{R_e}{a}\right)^n F_{nmp}(i)\, G_{npq}(e)\, S_{nmpq}(\Omega, \omega, M, \theta) \tag{4}$$

$$S_{nmpq}(\Omega, \omega, M, \theta) = \begin{bmatrix} +C_{nm} \\ -S_{nm} \end{bmatrix}_{n-m\,\text{odd}}^{n-m\,\text{even}} \cos\,\Theta_{nmpq}(\Omega, \omega, M, \theta) \tag{5}$$

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

The angle Θ_{nmpq} is called the *Kaula gravitational argument* and is given by:

$$\Theta_{nmpq}(\Omega, \omega, M, \theta) = (n - 2p)\,\omega + (n - 2p + q)\,M + m(\Omega - \theta) \tag{6}$$

2.2 The Lunisolar Perturbations

The acceleration due to an external body, exerted on the space debris, writes:

$$\ddot{\mathbf{r}} = -\mu_i \left(\frac{\mathbf{r} - \mathbf{r_i}}{\|\mathbf{r} - \mathbf{r_i}\|^3} + \frac{\mathbf{r_i}}{\|\mathbf{r_i}\|^3}\right). \tag{7}$$

The convention is to refer to the Sun by $i = 1$, with mass $M_1 = M_S$, and to the Moon by $i = 2$, with mass $M_2 = M_M$.

The associated potential can be easily calculated:

$$\mathscr{R}_i = \mu_i \left(\frac{1}{\|\mathbf{r} - \mathbf{r_i}\|} - \frac{\langle \mathbf{r} \, . \, \mathbf{r_i} \rangle}{\|\mathbf{r_i}\|^3} \right) \tag{8}$$

with $\mu_i = GM_i$, $< ., . >$ designates the scalar product and $\|.\|$ the norm.

The classical development of the inverse of the distance can be applied:

$$\mathscr{R}_i = \frac{\mu_i}{r_i} \sum_{n \geq 2} \left(\frac{r}{r_i} \right)^n \mathscr{P}_n(\cos \phi_i), \tag{9}$$

with r_i the distance between the body i and the Earth's center, ϕ_i the angle between the third body i and the piece of debris, and $\mathscr{P}_n$ the Legendre polynomial of degree n.

To separate the contribution of the third body from that of the debris, we express again the three components (x, y, z) of the position vector $\mathbf{r}$ in Keplerian elements $(a, e, i, \Omega, \omega, f)$ with f the true anomaly, we define the Cartesian coordinates X_i, Y_i and Z_i of the unit vector pointing towards the third body, and we use the usual developments of f and $\frac{r}{a}$ in series of e, $\sin \frac{i}{2}$ and M.

We obtain the following development, where the third body motion is only present in the coefficients A:

$$\mathscr{R}_i = \frac{\mu_i}{r_i} \sum_{n=2}^{+\infty} \sum_{k,l,j_1,j_2,j_3} \left(\frac{a}{r_i} \right)^n A^{(n)}_{k,l,j_1,j_2,j_3}(X_i, Y_i, Z_i) \; e^{|k|+2j_2} \left(\sin \frac{i}{2} \right)^{|l|+2j_3} \cos \Phi$$

with the angles defined as:

$$\Phi = j_1 \lambda + j_2 \varpi + j_3 \Omega, \quad \lambda = M + \omega + \Omega, \quad \varpi = \omega + \Omega. \tag{10}$$

2.3 The Poincaré's Variables

To use Hamiltonian formalism, we define first the Delaunay's canonical momenta L, G, and H associated to λ, ϖ and Ω:

$$L = \sqrt{\mu \, a}, \qquad G = \sqrt{\mu \, a(1 - e^2)}, \qquad H = \sqrt{\mu \, a(1 - e^2)} \, \cos i \tag{11}$$

Second, to avoid singularities, we switch to non singular Delaunay's elements, P and Q, associated to p and q, keeping L and λ unchanged:

$$\begin{aligned} P &= L - G \qquad & p &= -\omega - \Omega \\ Q &= G - H \qquad & q &= -\Omega \end{aligned} \tag{12}$$

Third, we introduce canonical Cartesian coordinates, called Poincaré's variables:

$$\begin{array}{ll} x_1 = \sqrt{2P}\ \sin p & x_4 = \sqrt{2P}\ \cos p \\ x_2 = \sqrt{2Q}\ \sin q & x_5 = \sqrt{2Q}\ \cos q \\ x_3 = \lambda = M + \Omega + \omega & x_6 = L \end{array} \tag{13}$$

Fourth, we choose dimensionless non canonical variables ξ_1, ξ_2, η_1 and η_2 directly linked to Poincaré's ones:

$$\xi_1 = U\ \sin p, \qquad \eta_1 = U\ \cos p, \qquad \xi_2 = V\ \sin q, \qquad \eta_2 = V\ \cos q. \tag{14}$$

with

$$U = \sqrt{\frac{2P}{L}} \qquad V = \sqrt{\frac{2Q}{L}} \tag{15}$$

The momenta U and V are proportional to e and i, and their exact dependance is given explicitly by:

$$e = U\left(1 - \frac{U^2}{4}\right)^{\frac{1}{2}} = U - \frac{1}{8}U^3 - \frac{1}{128}U^5 + \mathscr{O}(U^7) \tag{16}$$

and

$$2\ \sin\frac{i}{2} = V\left[1 - \frac{U^2}{2}\right]^{-\frac{1}{2}} = V + \frac{1}{4}\,VU^2 + \frac{3}{32}\,VU^4 + \mathscr{O}(U^6) \tag{17}$$

The details are given in [43].

2.4 The Hamiltonian Formulation

The Hamiltonian, based on these two main perturbations, writes, in terms of the new variables, for any fixed values of n_{max} and N_n:

$$\mathscr{H} = -\frac{\mu^2}{2\,L^2} + \dot\theta\,\Lambda + \sum_{n=2}^{n_{max}} \frac{1}{L^{2n+2}} \sum_{j=1}^{N_n} \mathscr{A}_j^{(n)}(\xi_1, \eta_1, \xi_2, \eta_2)\, \mathscr{B}_j^{(n)}(\lambda, \theta) \tag{18}$$

$$+ \sum_{i=1}^{2} \sum_{n=2}^{n_{max}} \frac{L^{2n}}{r_i^{n+1}} \sum_{j=1}^{N_n} \mathscr{C}_j^{(n)}(\xi_1, \eta_1, \xi_2, \eta_2, X_i, Y_i, Z_i)\, \mathscr{D}_j^{(n)}(\lambda) \tag{19}$$

A fourth degree of freedom has been introduced, through the angle θ representing the daily Earth's rotation or the sidereal time, associated to a virtual momentum Λ.

The associated dynamical system is then:

$$\dot{\xi}_i = \frac{1}{L}\frac{\partial \mathscr{H}}{\partial \eta_i} \qquad \dot{\eta}_i = -\frac{1}{L}\frac{\partial \mathscr{H}}{\partial \xi_i} \qquad i = 1, 2$$

and

$$\dot{\lambda} = \frac{\partial \mathscr{H}}{\partial L} - \frac{1}{2L}\left[\sum_{i=1}^{2}\frac{\partial \mathscr{H}}{\partial \xi_i}\xi_i + \sum_{i=1}^{2}\frac{\partial \mathscr{H}}{\partial \eta_i}\eta_i\right] \qquad \dot{L} = -\frac{\partial \mathscr{H}}{\partial \lambda}$$

The last ones are trivial:

$$\dot{\theta} = \frac{\partial \mathscr{H}}{\partial \Lambda} \qquad \dot{\Lambda} = -\frac{\partial \mathscr{H}}{\partial \theta}.$$

3 The Semi-Analytical Methods

As in the case of natural bodies, a first way to understand the dynamics is to expand the different forces in Poisson's series expansions, and to integrate the differential system. For more details about this chapter, please refer to [44].

3.1 The Non Resonant Case

Outside of resonances, we expand the Hamiltonian as a power series in the different variables; for example using MSNAM [31], the series manipulator of the University of Namur, we obtain the development for any given order. An example is given below, in Table 1 where the multiples of the angles λ and θ appear, besides the polynomials exponents of the other Cartesian variables. These techniques have been usually reserved to natural bodies dynamics but are easily adapted to space debris. The unit of length is chosen as the geostationary semi-major axis (42,164 km), the Earth mass is the unit of weight, and the unit of time corresponds to $\mu = 1$.

Table 1 Sample of a few terms obtained by MSNAM

	λ	θ	ξ_1	η_1	ξ_2	η_2	L	X_M	Y_M	Z_M	r_M	X_S	Y_S	Z_S	r_S	Coefficient
cos	(0	0)	(0	0	0	0	−6	0	0	0	0	0	0	0	0)	0.12386619D-04
cos	(0	0)	(0	0	0	2	−6	0	0	0	0	0	0	0	0)	−0.18579928D-04
cos	(0	0)	(0	0	0	4	−6	0	0	0	0	0	0	0	0)	0.46449822D-05

Table 2 Number of terms obtained by the expansions: after averaging and before averaging (in parentheses)

Perturbation	Number of terms			
n-order expansion $\xi_1^{i_1}\eta_1^{i_2}\xi_2^{i_3}\eta_2^{i_4}$ with $i_1+i_2+i_3+i_4 \le n$	$n=2$	$n=4$	$n=6$	$n=8$
Geopotential limited to J_2	5 (33)	15 (145)	31 (410)	53 (895)
External Body: Sun and Moon				
Up to degree 2	27 (205)	86 (836)	197 (2374)	390 (5480)
Up to degree 3	73 (645)	250 (2642)	611 (7854)	1227 (18380)

We perform an averaging transformation, over the fastest variable λ, and we integrate (numerically) the associated averaged dynamical system, with a much larger stepsize (from 200 s to 1 day). This is what we call the *semi-analytical averaged solution* for the space debris non resonant dynamics. A similar approach for space debris has been used by F. Deleflie and collaborators, via the software STELA, based on [29].

Table 2 gives an idea of the number of terms, following the order, for the non-averaged (in the parentheses) and averaged potentials.

3.2 The Gravitational Resonances

Let us remind the geopotential Kaula's formulation (4):

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

with the very important gravitational argument (6):

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

What we call a *gravitational resonance* is a resonance between the orbital motion of the space debris and the rotation of the Earth, which is different from the spin-orbit resonances, where the rotational and orbital motions are related to the same body. This means that the two periods, P_S (1 day) and P_{obj} are very close to a commensurability:

$$\frac{P_S}{P_{obj}} \simeq \frac{q_1}{q_2}. \tag{20}$$

If the ratio is equal to 1, we are close to the geostationary orbit (GEO), if it is close to 2, we are typically in the MEO (Medium Earth's Orbit) region. These situations correspond to:

$$\dot{\Theta}_{nmpq}(\dot{\Omega}, \dot{\omega}, \dot{M}, \dot{\theta}) = (n-2p)\,\dot{\omega} + (n-2p+q)\,\dot{M} + m(\dot{\Omega} - \dot{\theta}) \simeq 0 \tag{21}$$

and for $q = 0$:

$$\frac{\dot{M}}{\dot{\theta}} \simeq \frac{\dot{\lambda}}{\dot{\theta}} \simeq \frac{q_1}{q_2}. \tag{22}$$

The most important role is played by the coefficient J_{22}, the largest one in the geopotential containing the Earth's rotation angle.

3.3 The Geostationary Resonant Case

Let us now concentrate our attention on the geostationary case, i.e. $q_1 = q_2 = 1$, also called the *synchronous case*. Calculating the semi-major axis corresponding to an orbital period of 1 day, we obtain the well known value $a = 42,164$ km.

We only keep the J_{22} terms in the geopotential, developed in the same way as in the non resonant case:

$$\mathscr{H} = \mathscr{H}_{J_{22}}(\xi_1, \eta_1, \xi_2, \eta_2, \Lambda, \lambda, L, \theta) + \dot{\theta}\,\Lambda. \tag{23}$$

and we introduce the resonant angle:

$$\sigma = \lambda - \theta. \tag{24}$$

To keep σ and θ as canonical variables instead of λ and θ, we have to correct the two momenta, that we call now L' and Λ':

$$L' = L, \qquad \theta' = \theta, \qquad \Lambda' = \Lambda + L, \tag{25}$$

and the resonant Hamiltonian is written as:

$$\mathscr{H} = \mathscr{H}_{J_{22}}\left(\xi_1, \eta_1, \xi_2, \eta_2, \sigma, L', \theta\right) + \dot{\theta}\,\left(\Lambda' - L'\right). \tag{26}$$

θ is now the fast angle, and we obtain an averaged model, still dependent on σ, the slow resonant angle.

Table 3 Number of terms obtained by the resonant expansions: after averaging and before averaging (in parentheses)

Perturbation	Number of terms			
n-order expansion $\xi_1^{i_1}\eta_1^{i_2}\xi_2^{i_3}\eta_2^{i_4}$ with $i_1+i_2+i_3+i_4 \leq n$	$n=2$	$n=4$	$n=6$	$n=8$
Resonant perturbation due to J_{22}	10 (94)	40 (468)	104 (1392)	206 (3178)

Table 4 Sample of a few resonant terms obtained by MSNAM

	σ	θ	ξ_1	η_1	ξ_2	η_2	L	X_M	Y_M	Z_M	r_M	X_S	Y_S	Z_S	r_S	Coefficient
cos	(2	0)	(0	0	0	0	−6	0	0	0	0	0	0	0	0)	0.1077767255D-06
cos	(2	0)	(0	0	0	0	−6	0	0	0	0	0	0	0	0)	0.1080907167D-06
sin	(2	0)	(0	0	0	0	−6	0	0	0	0	0	0	0	0)	−0.6204881922D-07

$$\begin{tikzcd}
\mathscr{H}_{J_{22}}\,(\xi_1,\eta_1,\xi_2,\eta_2,L,\Lambda,\theta,\lambda) \arrow[d] \\
\mathscr{H}_{J_{22}}\left(\xi_1,\eta_1,\xi_2,\eta_2,L',\Lambda',\theta',\sigma\right) \arrow[d] \\
\overline{\mathscr{H}}_{J_{22}}\left(\bar{\xi}_1,\bar{\eta}_1,\bar{\xi}_2,\bar{\eta}_2,\bar{L}',\bar{\Lambda}',-,\bar{\sigma}\right)
\end{tikzcd}$$

The resonant averaged Hamiltonian is obtained by the same process as the non-resonant one: series expansions, averaging, averaged equations and numerical integration. Table 3 gives an idea about the number of terms in the averaged and non-averaged Hamiltonian (in parentheses).

The series are now of the form given in Table 4.

3.4 A Simple Analytical Geostationary Resonant Model

Thanks to the resonant averaged semi-analytical method, we identify the main terms in the expansions, and we build a toy model, able to describe qualitatively and locally the resonant dynamics in the geostationary region.

The simplest resonant averaged model, based on the first terms of the development, can be written as:

$$\mathscr{H}(L,\sigma,\Lambda) = -\frac{\mu^2}{2L^2} + \dot{\theta}(\Lambda - L) - \frac{1}{L^6}\left[\alpha_1\,\cos 2\sigma + \alpha_2\,\sin 2\sigma\right] \tag{27}$$

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

The equilibria are easily determined by solving:

$$\frac{\partial \mathscr{H}}{\partial L} = 0 = \frac{\partial \mathscr{H}}{\partial \sigma}. \tag{28}$$

Two stable equilibria are given by $(\sigma_{11}^*, L_{11}^*)$, $(\sigma_{12}^*, L_{12}^*)$ and two unstable equilibria by $(\sigma_{21}^*, L_{21}^*)$, $(\sigma_{22}^*, L_{22}^*)$

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

with $L_{11}^* = L_{12}^* = 0.99999971$, $L_{21}^* = L_{22}^* = 1.00000029$, where $L = 1$ corresponds to 42 164 km, and $\lambda^* \simeq 75.07°$.

The phase diagram is given in Fig. 1 in Cartesian coordinates $\sqrt{2L}\cos\sigma$ and $\sqrt{2L}\sin\sigma$ where the coefficients α_1 and α_2 have been amplified artificially, to distinguish the resonant islands. The resonant motion, associated to σ, has a period of period of 818.7 days $\simeq$ 2.5 years and the width of the resonant zone is $\simeq$ 69 km.

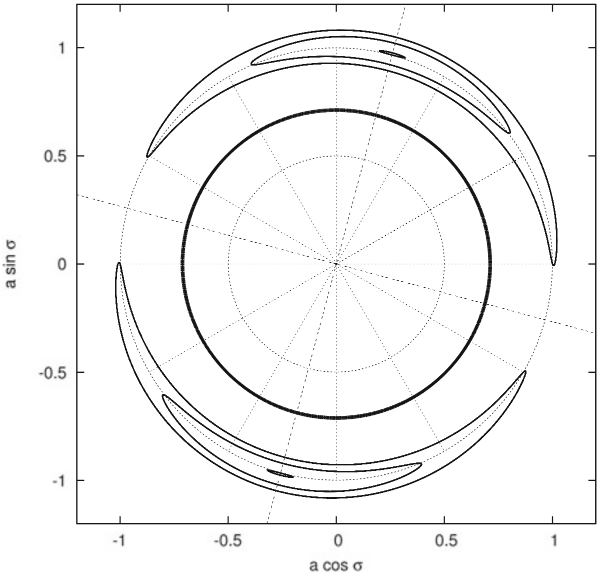


Fig. 1 Phase space in Cartesian coordinates for the geosynchronous averaged dynamics; to increase the visibility, the coefficients have been artificially amplified (source [44])

3.5 The Other Gravitational Resonances

Similar studies have been performed for the other gravitational resonances. In particular the 2:1 resonance case, corresponding to the MEO region, of great interest for the GPS and Galileo constellations, has been intensively studied. Let us mention the pioneer paper of Rossi [37] and the very complete analysis of Celletti and Galeş [7].

Other resonances, called minor, have also been investigated, with a description of their location, equilibria, width. Let us refer to two papers of the same team, inside the geostationary ring [8] and outside [9].

4 The Solar Radiation Pressure

The space debris present different shapes, forms and weights. It means that some of them are very perturbed by the solar radiation pressure, proportional to the coefficient A/m (Area over mass), and some others not at all. This perturbation could be the most important after the two body, and has to be considered, sometimes even before the gravitational potential and the lunisolar attractions. A complete analysis of this contribution can be found in the chapter 14 of Milani and Gronchi [30] or in [28].

Here we limit our study to the direct radiation pressure acceleration and we assume that the coefficient A/m is large. We are going to build an averaged model for the geostationary region.

Let us mention a very smart approach, using the eccentricity and the angular momentum as variables, also based on the averaged dynamics [35] and the complete perturbation theory with planetary motion developed by Gachet et al. [20] which generalizes and justifies the present basic analysis.

The acceleration due to the direct radiation pressure can be written in the form:

$$\mathbf{a_{rp}} = C_r\, P_r \left[\frac{a_S}{\|\mathbf{r}-\mathbf{r_S}\|}\right]^2 \frac{A}{m}\, \frac{\mathbf{r}-\mathbf{r_S}}{\|\mathbf{r}-\mathbf{r_S}\|} \simeq -C_r\, P_r\, \frac{A}{m}\, a_S^2 \sum_{n=1}^{n=N} \left(\frac{r}{a_S}\right)^n P_n(\cos\phi) \tag{29}$$

where C_r is the non-dimensional reflectivity coefficient ($0 < C_r < 2$), $P_r = 4.56 \cdot 10^{-6}$ N/m^2 is the radiation pressure per unit of mass for an object located at a distance of $a_S = 1$ AU, $\mathbf{r}$ is the geocentric position of the space debris; $\mathbf{r_S}$ is the geocentric position of the Sun, ϕ is the angle between $\mathbf{r}$ and $\mathbf{r}_S$, A is the exposed area to the Sun of the space debris, m is the mass of the space debris. Table 5 gives some examples of A/m coefficients, for natural and artificial bodies.

Some space debris, initially on circular orbits, show very large unexpected eccentricities [38]. It was explained by the detection of their A/m coefficients, particularly high (even larger than 50 for some objects) and the consequent high

Table 5 Examples of A/m coefficients

Object	A/m (m^2/kg)
Lageos 1 and 2	10^{-3}
Starlette	10^{-3}
GPS (Block II)	10^{-2}
Moon	10^{-10}
Space debris	??

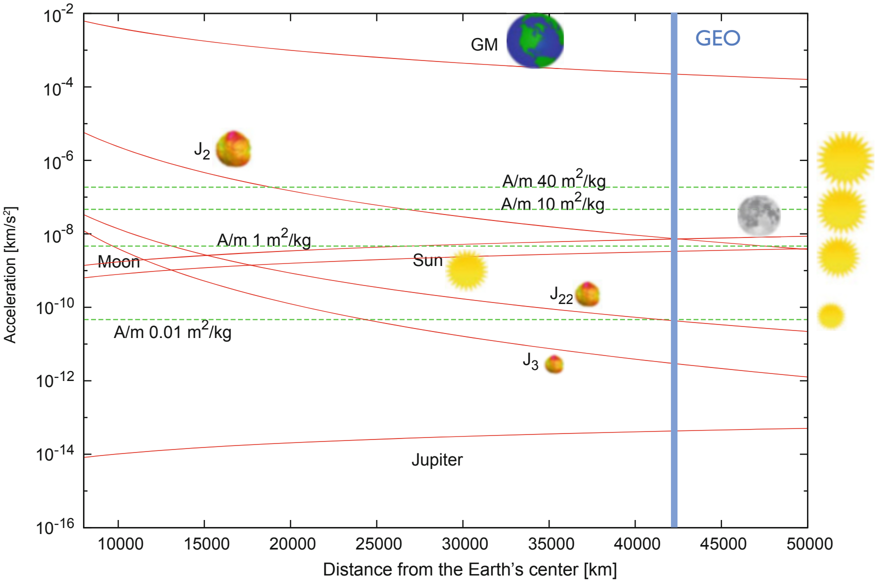


Fig. 2 The hierarchy of the perturbations for different values of A/m in the radial component of the acceleration, for the whole space debris region

perturbation caused by the solar radiation pressure in the dynamics of those bodies [13, 39, 40].

Figure 2 presents the order of magnitude of the different perturbations in the space debris environment, for different values of the coefficient A/m.

Assuming a high value of A/m, the solar radiation pressure becomes the main perturber of the Keplerian problem.

$$\mathscr{H}(\mathbf{v},\mathbf{r}) = \mathscr{H}_{kepl}(\mathbf{v},\mathbf{r}) + \mathscr{H}_{srp}(\mathbf{r}) \tag{30}$$

with $\mathbf{r}$ the geocentric position of the satellite, $\mathbf{v}$ its velocity, $\mathscr{H}_{kepl}(\mathbf{v},\mathbf{r})$ the attraction of the Earth, $\mathscr{H}_{srp}(\mathbf{r})$ the direct solar radiation pressure potential

$$\mathscr{H}_{kepl} = \frac{\|\mathbf{v}\|^2}{2} - \frac{\mu}{\|\mathbf{r}\|} \tag{31}$$

$$\mathscr{H}_{srp} = -C_r\,\frac{1}{\|\mathbf{r}-\mathbf{r}_S\|}\,P_r\,\frac{A}{m}\,a_S^2. \tag{32}$$

4.1 A First Toy Model for the Solar Radiation Pressure with Large A/m

Let us start with a very simple model, including the two-body and the direct solar radiation pressure. We use the development of the inverse of the distance in Legendre's polynomials in Eq. (31) and the Hamiltonian writes, after truncation at first order in the development:

$$\mathscr{H} = -\frac{\mu^2}{2L^2} + C_r\, P_r\, \frac{A}{m}\, r\, \overline{r}_S\, \cos\phi \tag{33}$$

with ϕ the angle between $\mathbf{r}$ and $\mathbf{r}_S$, $\overline{r}_S = \frac{r_S}{a_S}$.

$$\mathscr{H} = -\frac{\mu^2}{2L^2} + C_r\, P_r\, \frac{A}{m}\, a\, (u\,\xi + v\,\eta) \tag{34}$$

where the debris orbital motion is given by $u = \cos E - e$ and $v = \sin E\, \sqrt{1-e^2}$, E being the eccentric anomaly, and the Sun's influence is present through ξ and η, given by:

$$\xi = \xi_1\, \overline{r}_{S,1} + \xi_2\, \overline{r}_{S,2} + \xi_3\, \overline{r}_{S,3} \tag{35}$$

$$\eta = \eta_1\, \overline{r}_{S,1} + \eta_2\, \overline{r}_{S,2} + \eta_3\, \overline{r}_{S,3} \tag{36}$$

and in terms of the elliptic elements:

$$\begin{aligned}
\xi_1 &= \cos\Omega\ \cos\omega - \sin\Omega\ \cos i\ \sin\omega & \eta_1 &= -\cos\Omega\ \sin\omega - \sin\Omega\ \cos i\ \cos\omega \\
\xi_2 &= \sin\Omega\ \cos\omega + \cos\Omega\ \cos i\ \sin\omega & \eta_2 &= -\sin\Omega\ \sin\omega + \cos\Omega\ \cos i\ \cos\omega \\
\xi_3 &= \sin i\ \sin\omega & \eta_3 &= \ \sin i\ \cos\omega
\end{aligned}$$

Two periods are present in this formulation: the orbital period (through E) of 1 day, and the Sun orbital period (through $\overline{r}_{S,i}$) of 1 year.

The next step consists in averaging over the fast angle, M the mean anomaly, using $dM = (1 - e\,\cos E)\,dE$:

$$\begin{aligned}
\overline{\mathscr{H}} &= \frac{1}{2\pi}\int_0^{2\pi} \mathscr{H}\, dM \\
&= -\frac{\mu^2}{2\overline{L}^2} + \frac{1}{2\pi}\, C_r\ P_r\, \frac{A}{m}\, \overline{a}\, \int_0^{2\pi} (u\ \xi + v\ \eta)\, dM \\
&\simeq -\frac{\mu^2}{2\overline{L}^2} - \frac{3}{2}\, C_r\ P_r\, \frac{A}{m}\, \frac{\overline{L}^2}{\mu}\, \overline{e}\, \xi.
\end{aligned} \tag{37}$$

The bars designate the averaged variables, and will not be maintained in the further equations.

Using again Poincaré's variables:

$$\begin{array}{ll} p = -\omega - \Omega & P = L - G \\ q = -\Omega & Q = G - H \\ x_1 = \sqrt{2P}\ \sin p & y_1 = \sqrt{2P}\ \cos p \\ x_2 = \sqrt{2Q}\ \sin q & y_2 = \sqrt{2Q}\ \cos q \end{array}$$

using the approximations: $e \simeq \sqrt{\frac{2P}{L}}$, $\cos^2 \frac{i}{2} = 1 - \frac{Q}{2L}$, $\sin \frac{i}{2} \simeq \sqrt{\frac{Q}{2L}}$ and assuming a circular orbit for the Sun (with an obliquity ϵ):

$$\begin{aligned} \bar{r}_{S,1} &= \cos \lambda_S \\ \bar{r}_{S,2} &= \sin \lambda_S \cos \epsilon \\ \bar{r}_{S,3} &= \sin \lambda_S \sin \epsilon \end{aligned} \tag{38}$$

with $\lambda_S = n_S t + \lambda_{S,0}$, we can write:

$$\begin{aligned} \mathscr{H} &= \mathscr{H}(x_1, y_1, x_2, y_2, \lambda_S) \\ &\simeq -n_S\ \kappa\ \bar{r}_{S,1}\ (x_1\ R_2 + y_1\ R_1) \\ &+ n_S\ \kappa\ \bar{r}_{S,2}\ (x_1\ R_3 + y_1\ R_2) \\ &+ n_S\ \kappa\ \bar{r}_{S,3}\ (x_1\ R_5 - y_1\ R_4) \end{aligned} \tag{39}$$

with $\kappa = \frac{3}{2}\ C_r\ P_r\ \frac{A}{m}\ \frac{a}{\sqrt{L}}$ (directly proportional to A/m) and where $R_i(x_2, y_2)$ are second degree polynomials in x_2 and y_2.

The dynamical system associated is given by:

$$\begin{array}{ll} \dot{x}_1 = \frac{\partial \mathscr{H}}{\partial y_1} & \dot{y}_1 = -\frac{\partial \mathscr{H}}{\partial x_1} \\ \dot{x}_2 = \frac{\partial \mathscr{H}}{\partial y_2} & \dot{y}_2 = -\frac{\partial \mathscr{H}}{\partial x_2}. \end{array} \tag{40}$$

An analytical solution is calculated in three steps. First, assuming $x_2 = y_2 = 0$, we find the short periodic motion for x_1 and y_1:

$$\begin{array}{ll} x_1 = -\kappa\ \sin \lambda_S + C_x & = -\kappa\ (\sin \lambda_S - D_x) \\ y_1 = \kappa\ \cos \lambda_S\ \cos \epsilon + C_y & = \kappa\ (\cos \lambda_S\ \cos \epsilon + D_y). \end{array} \tag{41}$$

We conclude that e and ϖ follow a periodic motion (1 year), C_x and C_y or D_x and D_y being the initial conditions. If κ is larger, e_{max} (the maximal value of the eccentricity) increases. Figure 3 illustrates this annual motion.

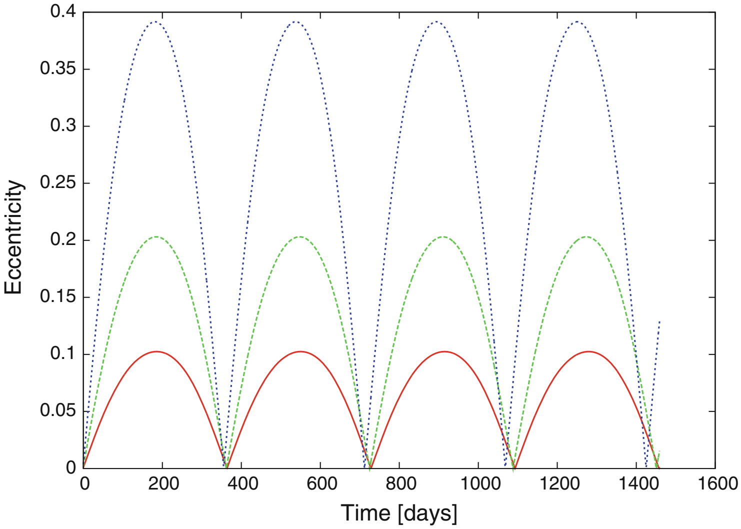


Fig. 3 The annual periodic motion of the eccentricity for $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$ (red), $A/m = 10\,\mathrm{m}^2/\mathrm{kg}$ (magenta) and $A/m = 20\,\mathrm{m}^2/\mathrm{kg}$ (green) (source [43])

Second, after averaging over the fast periods (1 year), we find an averaged motion for $\bar{x}_2$ and $\bar{y}_2$

$$\begin{cases} \bar{x}_2 = \mathscr{A} \sin \psi \\ \bar{y}_2 = \mathscr{A} \cos \psi - \frac{\rho}{\nu} = \mathscr{A} \cos \psi - \tan \epsilon \sqrt{L} \end{cases} \tag{42}$$

with $\psi = \nu t + \psi_0$, $\mathscr{A}$ and ψ_0 being the initial conditions.

We notice that the averaged values of the inclination and of the longitude of the ascending node, $\bar{i}$ and $\bar{\Omega}$, follow a long periodic motion (with a period of several dozens of years) with always the same maximal value of the inclination: $\bar{i}_{max} \simeq 2\epsilon$. If A/m increases, κ increases, then ν increases and the period of this motion decreases.

Third, we reinsert the short periodic terms (replacing x_1 and y_1 in terms of λ_S) into the Hamiltonian, so to obtain:

$$x_2 = \bar{x}_2 + \frac{\partial \mathscr{W}}{\partial y_2}(\lambda_S) \quad y_2 = \bar{y}_2 - \frac{\partial \mathscr{W}}{\partial x_2}(\lambda_S) \tag{43}$$

where

$$\mathscr{W} = -\kappa^2 \,(g_1 \ \sin \lambda_S - g_2 \ \cos \lambda_S + \frac{1}{2}\, g_3 \ \sin 2\lambda_S - \frac{1}{2}\, g_4 \ \cos 2\lambda_S). \tag{44}$$

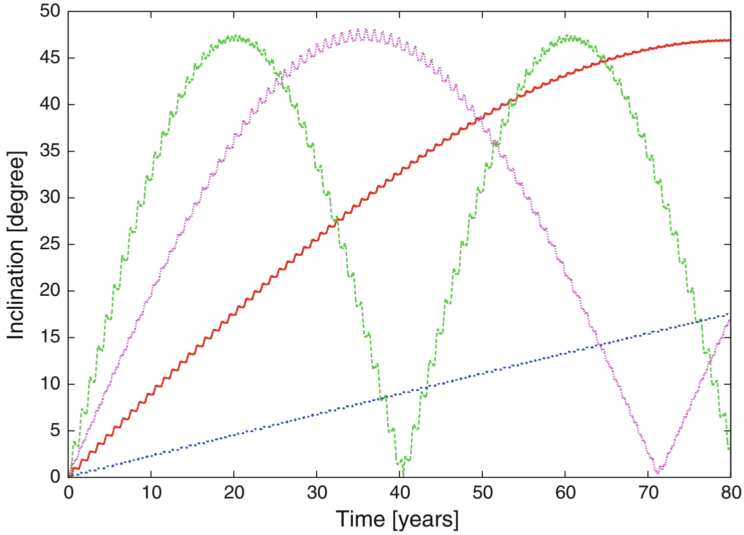


Fig. 4 The long periodic motion of the inclination for $A/m = 1\,\mathrm{m}^2/\mathrm{kg}$ (blue), $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$ (red), $A/m = 10\,\mathrm{m}^2/\mathrm{kg}$ (magenta) and $A/m = 20\,\mathrm{m}^2/\mathrm{kg}$ (green) (source [43])

The functions g_i, $i = 1, 2, 3, 4$ depend on x_2 and y_2 and on the initial conditions; their explicit expressions are given in [22].

Figure 4 describes that dynamics, for four different values of A/m; we distinguish the annual short periodic perturbations superposed on the long periodic motion.

5 The Earth's Umbra

The orbits of the space debris could cross the Earth's umbra, and, in that case, the solar radiation pressure stops affecting the dynamics, to reappear later. The geometrical cylindrical problem is described thanks to the shadow equation:

$$\begin{aligned} s_c(\mathbf{r}) = \frac{\mathbf{r} \cdot \mathbf{r}_S}{r_S} + \sqrt{r^2 - R_e^2} &< 0 \text{ inside Earth's shadows} \\ &> 0 \text{ outside Earth's shadows} \\ &= 0 \text{ entry and exit} \end{aligned} \tag{45}$$

The equation corresponds to a 4th degree polynomial in $\tan \frac{E}{2}$ solved by Cardan's formula. We denote by E_1 the entry eccentric anomaly $= E_1(a, e, i, \omega, \Omega, \overline{r}_S)$ and by E_2, the exit eccentric anomaly $= E_2(a, e, i, \omega, \Omega, \overline{r}_S)$. Figure 5 represents the cylindrical approach.

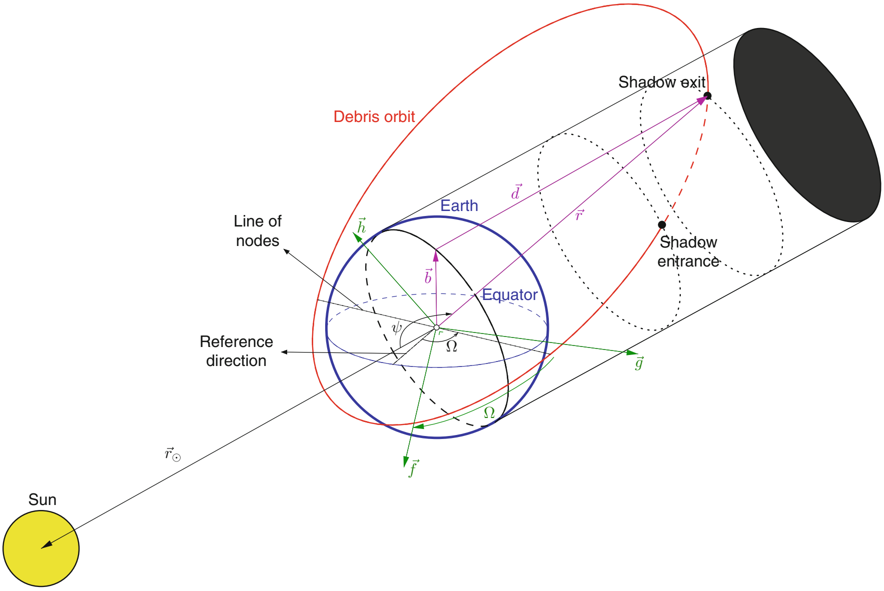


Fig. 5 The cylindrical model for the Earth's umbra. The Sun is assumed to be far enough (source [42])

We modify the toy model by inserting the Earth's shadows:

$$\mathscr{H} = -\frac{\mu^2}{2L^2} + \begin{cases} C_r\ P_r\ \frac{A}{m}\, r\, \overline{r}_S\ \cos(\phi) & \text{outside Earth's shadows} \\ 0 & \text{inside Earth's shadows} \end{cases} \tag{46}$$

5.1 *The Averaged Model*

We again average over the fast variable (M the mean anomaly) but we take into account the absence of the solar radiation pressure between M_1 and M_2:

$$\begin{aligned} \overline{\mathscr{H}} &= \frac{1}{2\pi} \int_0^{2\pi} \mathscr{H}\, dM \qquad (47) \\ &= -\frac{\mu^2}{2\overline{L}^2} + \frac{1}{2\pi}\ C_r\ P_r\ \frac{A}{m}\ \overline{a} \left[\int_0^{M_1} (u\ \xi + v\ \eta)\, dM + \int_{M_2}^{2\pi} (u\ \xi + v\ \eta)\, dM \right] \end{aligned}$$

following the pioneer works of Ferraz-Mello [18] or Aksnes [1].

The averaged Hamiltonian with shadowing effects writes now:

$$\overline{\mathscr{H}} = -\frac{\mu^2}{2\overline{L}^2} - \frac{3}{2}\, C_r\, P_r\, \frac{A}{m}\, \frac{\overline{L}^2}{\mu}\, \overline{e}\, \xi + \frac{1}{2\pi} C_r\, P_r\, \frac{A}{m}\, \frac{\overline{L}^2}{\mu}\, [\xi\, \mathscr{A} + \eta\, \mathscr{B}] \tag{48}$$

where

$$\begin{aligned}\mathscr{A} &= -2\,(1+\overline{e}^2)\, \cos\frac{S}{2}\, \sin\frac{D}{2} + \frac{3}{2}\,\overline{e}\, D + \frac{\overline{e}}{2}\, \cos S\, \sin D \\ \mathscr{B} &= \sqrt{1-\overline{e}^2}\, \left(-2\, \sin\frac{S}{2}\, \sin\frac{D}{2} + \frac{\overline{e}}{2}\, \sin S\, \sin D\right)\end{aligned} \tag{49}$$

and $S = E_1 + E_2$ and $D = E_2 - E_1$. The case $D = 0$ corresponds to the model without umbra.

The dynamical system is modified for $\bar{L}$ and then, for the semi-major axis $\bar{a}$, which is not constant anymore, but follows a long periodic motion:

$$\dot{\bar{a}} = \bar{a}^{\,3/2}\, \frac{2}{\pi\,\sqrt{\mu}}\, C_r\, P_r\, \frac{A}{m}\, \left[\overline{\xi}\sin\frac{S}{2} - \overline{\eta}\sqrt{1-\overline{e}^2}\, \cos\frac{S}{2}\right]\, \sin\frac{D}{2}. \tag{50}$$

To give an idea about the orders of magnitude, for $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$, the period $\simeq$ 13,000 years and for $A/m = 25\,\mathrm{m}^2/\mathrm{kg}$, the period $\simeq$ 1200 years. Figure 6 gives the evolution of the long periods as a function of the semi-major axis.

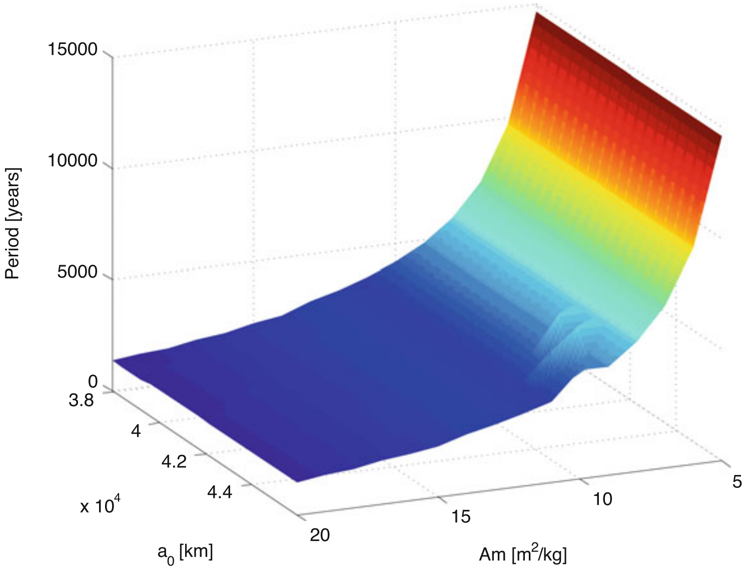


Fig. 6 The calculation of the very long period induced by the Earth's shadowing effects, as a function of the initial semi major axis, and of the coefficient A/m, in the geostationary region (source [22])

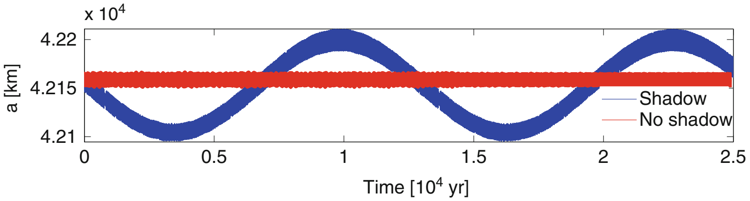


Fig. 7 Numerical integrations, without and with the shadowing effects, over 25,000 years, for $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$, for the Keplerian motion perturbed by the solar radiation pressure (source [22])

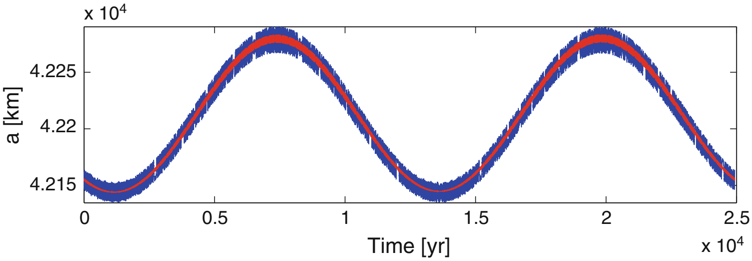


Fig. 8 Comparison between the analytical averaged model and the numerical integration, over 25,000 years, for $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$ (source [22])

The passage in the shadow is then responsible for a very long periodic motion for large values of A/m.

Numerical integrations (of the Keplerian problem, perturbed by the solar radiation pressure) show this very long periodic motion and the accuracy of our toy model. They have been obtained by the symplectic integrator SYMPLEC (see Sect. 7.2) with a simplified circular solar motion.

Figure 7 compares two numerical integrations, without and with the shadowing effects, over 25,000 years, for $A/m = 5\,\mathrm{m}^2/\mathrm{kg}$ and Fig. 8 shows the analytical solution versus the numerical integration, for the same case.

5.2 The Numerical Smoothing Function

The numerical integrations could be affected by the passage through the umbra, presented as a switch on-off. To avoid this non continuous situation, we replace locally the passage through the umbra by a smoothing function ν_C, depending on a parameter γ, and based on a hyperbolic tangent (see Fig. 9). The greater γ is, the shaper the function is

$$\nu_C = \frac{1}{2}(1 + \tanh(\gamma\ s_C(\mathbf{r}))) \simeq \begin{cases} 0 & \text{in cylindrical umbra} \\ 1 & \text{otherwise} \end{cases} \tag{51}$$

Thanks to this approach, we introduce the shadowing effects in the symplectic integrator SYMPLEC described in Sect. 7.2.

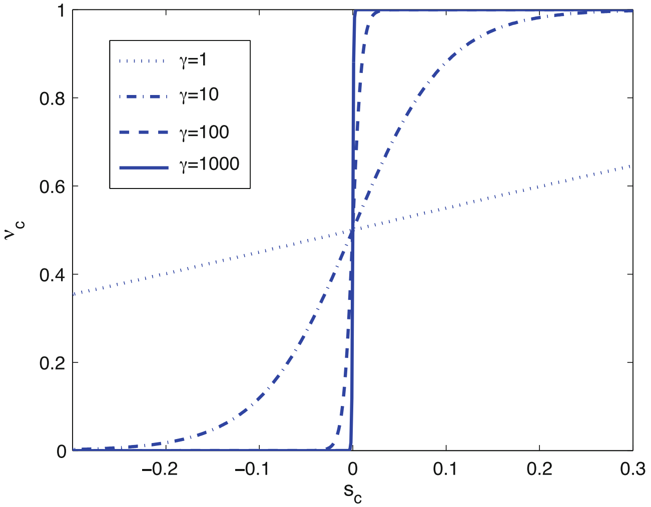


Fig. 9 The smoothing umbra ν_C described for different values of γ (source [21])

For the conical geometrical situation, a more complete analysis and model can be found in [22], describing and smoothing the passage through the umbra and the penumbra thanks to similar functions ν_u and ν_p depending on two parameters.

6 A More Complete Toy Model

If we observe Fig. 2, for huge values of A/m, we notice the importance of the solar radiation pressure on the dynamics. However the lunisolar perturbations and the J_2 flattening coefficient, can be considered of the same order of magnitude. Moreover up to now we have limited the solar radiation pressure to the first order (in Legendre's polynomials expansion) which can be improved.

6.1 The Toy Model, with Moon and Sun, Solar Radiation Pressure and J_2

We summarize here the approach developed in [6]. Let us start with J_2 perturbation:

$$\begin{aligned} H_{J_2}(\mathbf{r}) &= \frac{\mu}{r} J_2 \left(\frac{r_S}{r}\right)^2 P_2(\sin\lambda) \\ &= \frac{\mu}{r} J_2 \left(\frac{r_S}{r}\right)^2 \frac{1}{2}\left(3\left(\frac{z}{r}\right)^2 - 1\right) \end{aligned} \tag{52}$$

where λ represents the latitude of the satellite, and consequently $\sin\lambda = z/r$.

For the solar radiation pressure, we add the second order terms:

$$\begin{aligned}
H_{SRP}(\mathbf{r},\mathbf{r}_S) &= -C_r\,P_r\,\frac{A}{m}a_S^2\,\frac{1}{||\mathbf{r}-\mathbf{r}_S||}\\
&\simeq -C_r\,P_r\frac{A}{m}a_S^2\Big(\left(\frac{r}{a_S}\right)P_1(\cos\phi)+\left(\frac{r}{a_S}\right)^2\,P_2(\cos\phi)\Big)\\
&= H_{SRP_1}(\mathbf{r},\mathbf{r}_S)+H_{SRP_2}(\mathbf{r},\mathbf{r}_S) \qquad (53)
\end{aligned}$$

where ϕ is the angle between the satellite and the Sun.

For the third body Hamiltonian, we assume that the orbits of the Sun and of the Moon are circular. For the Sun:

$$\begin{aligned}
H_{3bS}(\mathbf{r},\mathbf{r}_S) &= -\mu_S\frac{1}{||\mathbf{r}-\mathbf{r}_S||}+\mu_S\frac{\mathbf{r}\cdot\mathbf{r}_S}{||\mathbf{r}_S||^3}\\
&\simeq -\frac{\mu_S}{a_S}\sum_{n\geq 0}\left(\frac{r}{a_S}\right)^n P_n(\cos\phi)+\mu_S\frac{r a_S\cos(\phi)}{a_S^3}\\
&\simeq -\frac{\mu_S}{a_S}(1+\left(\frac{r}{a_S}\right)^2 P_2(\cos\phi)), \qquad (54)
\end{aligned}$$

where $\mu_S = GM_S$, M_S is the mass of the Sun, and for the Moon:

$$H_{3bM}(\mathbf{r},\mathbf{r}_M)\simeq -\frac{\mu_M}{a_M}(1+\left(\frac{r}{a_M}\right)^2 P_2(\cos\phi_M)) \qquad (55)$$

where $\mu_M = GM_M$ with M_M the mass of the Moon, and ϕ_M is the angle between the satellite and the Moon.

Our complete toy model is now:

$$\begin{aligned}
&H_{SRP}(\mathbf{r},\mathbf{r}_S)+H_{3bS}(\mathbf{r},\mathbf{r}_S)+H_{3bM}(\mathbf{r},\mathbf{r}_M)\\
&\simeq H_{SRP_1}(\mathbf{r},\mathbf{r}_S)+H_{SRP_2}(\mathbf{r},\mathbf{r}_S)+H_{3bS}(\mathbf{r},\mathbf{r}_S)+H_{3bM}(\mathbf{r},\mathbf{r}_M)\\
&\simeq C_r\,P_r\,\frac{A}{m}\,a_S\,r\cos(\phi)-\frac{\mu_M}{a_M}\left(\frac{r}{a_M}\right)^2 P_2(\cos\phi_M)\\
&+\left[C_r\,P_r\frac{A}{m}a_S-\frac{\mu_S}{a_S}\right]\left(\frac{r}{a_S}\right)^2 P_2(\cos\phi) \qquad (56)
\end{aligned}$$

and, after average over the short periodic motion and some algebra, we obtain:

$$\begin{aligned}\overline{H}(x_1, y_1, x_2, y_2) &= \overline{H}_{kepler} + \overline{H}_{J_2}(x_1, y_1, x_2, y_2)\\ &+ \overline{H}_{SRP_1}(x_1, y_1, x_2, y_2, \mathbf{r}_S)\\ &+ \overline{H}_{SRP_2+3bS}(x_1, y_1, x_2, y_2, \mathbf{r}_S)\\ &+ \overline{H}_{3bM}(x_1, y_1, x_2, y_2, \mathbf{r}_M) \qquad (57)\end{aligned}$$

with

$$\overline{H}_{J_2} = C_p\ P + C_q\ Q = \frac{C_p}{2}(x_1^2 + y_1^2) + \frac{C_q}{2}(x_2^2 + y_2^2), \qquad (58)$$

$$\overline{H}_{SRP_1} = -\frac{3}{2}\,C_r\,P_r\,\frac{A}{m}\,a\,e\,\xi, \qquad (59)$$

$$\begin{aligned}\overline{H}_{SRP_2+3bS} &= -\left[C_r\,P_r\frac{A}{m}a_S - \frac{\mu_S}{a_S}\right]\,\frac{3a^2}{4a_S^2}w_S^2,\\ &= -\beta\,\frac{3a^2}{4a_S^2}w_S^2, \qquad (60)\end{aligned}$$

$$\overline{H}_{3bM} = \frac{\mu_M}{a_M}\,\frac{3a^2}{4a_M^2}w_M^2. \qquad (61)$$

and $\beta = \left[C_r\,P_r\frac{A}{m}a_S - \frac{\mu_S}{a_S}\right]\,\frac{3a^2}{4a_S^2}$. The coefficients w_S and w_M are given by:

$$w_S = -\sin q\ \ \sin i\ \mathbf{r}_{S,1} - \cos q\ \ \sin i\ \mathbf{r}_{S,2} + \cos i\ \mathbf{r}_{S,3} \qquad (62)$$

$$w_M = -\sin q\ \ \sin i\ \mathbf{r}_{M,1} - \cos q\ \ \sin i\ \mathbf{r}_{M,2} + \cos i\ \mathbf{r}_{M,3}, \qquad (63)$$

where q in defined in (12).

For the short (annual) periodic motion in eccentricity, we write, with n_S the mean motion of the Sun:

$$\dot{x_1}(t) = -C_2\ y_1 - n_S\ \kappa\ r_{S,1}, \qquad (64)$$

$$\dot{y_1}(t) = C_2\ x_1 - n_S\ \kappa\ r_{S,2}, \qquad (65)$$

with

$$C_2 = \frac{3}{2}\sqrt{\frac{\mu}{a^3}}J_2\frac{r_S^2}{a^2} \quad \text{and} \quad \kappa = \frac{3}{2}\ C_r\ P_r\ \frac{A}{m}\ \frac{a}{\sqrt{L}}.$$

The analytical solution is:

$$x_1(t) = C_x + \frac{k \sin(n_S t + \lambda_{S,0})}{1 - \eta^2} \, [\eta \, \cos\epsilon + 1] , \tag{66}$$

$$y_1(t) = C_y + \frac{k \cos(n_S t + \lambda_{S,0})}{1 - \eta^2} \, [\cos\epsilon + \eta] , \tag{67}$$

with C_x and C_y the initial conditions.

For the long periodic contributions (after averaging over the motion of the Sun and of the Moon), we define:

$$d_1 = n_S \, \frac{k^2}{4L} \, \cos\epsilon + \frac{C_q}{2} - \delta - \delta \, \cos^2\epsilon - \gamma - \gamma \, \cos^2\epsilon_M , \tag{68}$$

$$d_2 = n_S \, \frac{k^2}{4L} \, \cos\epsilon + \frac{C_q}{2} - 2\,\delta \, \cos^2\epsilon - 2\,\gamma \, \cos^2\epsilon_M , \tag{69}$$

$$d_3 = -n_S \, \frac{k^2}{2\sqrt{L}} \, \sin\epsilon + 2\,\delta \, \sqrt{L} \, \sin^2\epsilon + 2\,\gamma \, \sqrt{L} \, \sin^2\epsilon_M , \tag{70}$$

where $\delta = \beta \, \frac{3a^2}{16\, L\, a_S^2}$ and $\gamma = -\frac{\mu_M}{a_M} \, \frac{3a^2}{16\, L\, a_M^2}$.

Then we write the corresponding solution for $x_2(t)$ and $y_2(t)$:

$$x_2(t) = \mathscr{D} \, \sin(\sqrt{d_1 d_2}\, t - \psi) , \tag{71}$$

$$y_2(t) = \mathscr{D} \, \sqrt{\frac{d_2}{d_1}} \, \cos(\sqrt{d_1 d_2}\, t - \psi) - \frac{d_3}{d_1} , \tag{72}$$

$\mathscr{D}$ and ψ being the initial conditions.

We plot the motion of the inclination in Fig. 10, with or without the Moon, and for two values of A/m. The influence of the Moon decreases when A/m increases, it means when the solar radiation pressure is the main perturbation.

6.2 Quality of the Toy Model

We compare the four steps of the analytical model with a similar numerical integration (Fig. 11). The toy model describes quite well the qualitative behavior and gives a good first approximation of the periods and amplitudes. On a quantitative point of view, we can see differences in the maxima of the inclination of less than 10%.

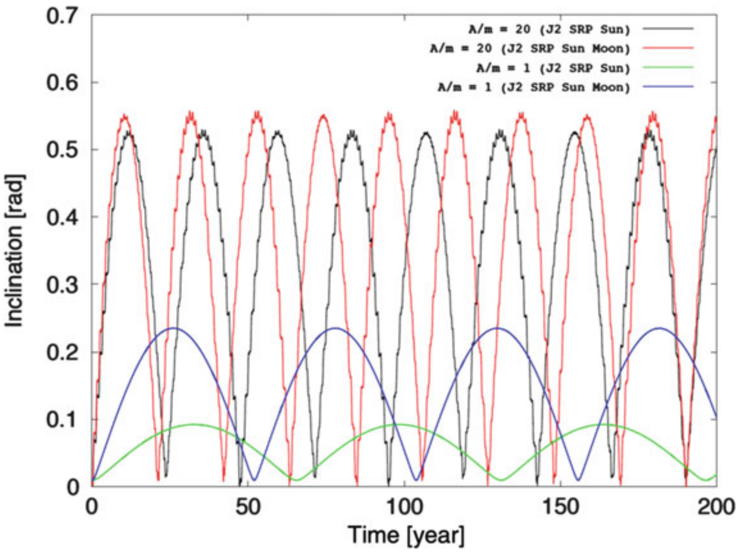


Fig. 10 The analytical motion of the inclination obtained for two values of A/m and for two models, with or without the Moon (source [6])

7 The Numerical Solutions

The toy models give analytical formulations of the dynamics, obtained by approximations and truncations. They show the main frequencies, they allow to approximate the periods, and to measure the minimal and maximal values of the elliptic elements. When we need more precision, we use semi-analytical solutions, we push further the expansions, and obtain huge dynamical systems, that we integrate numerically.

And to obtain quantitative precise results, especially in the artificial satellite world, the brute numerical integrations, including all the forces, in their full expressions, are still very performant, especially for short intervals of time.

However for space debris, mostly non operational and not anymore controlled, captured in stable regions, as for natural bodies, we are far away from a few years of life; the space debris could stay for hundreds, thousands of years in some zones of the space. A symplectic integrator, keeping a quasi constant energy, makes sense in that case.

7.1 The Classical Integrators

For space missions of a few years, classical numerical integrations are used. They can be refined by using several integrators at different orders (from a Runge-Kutta order 4 to an Adams–Bashforth–Moulton order 10 for example), by reducing the

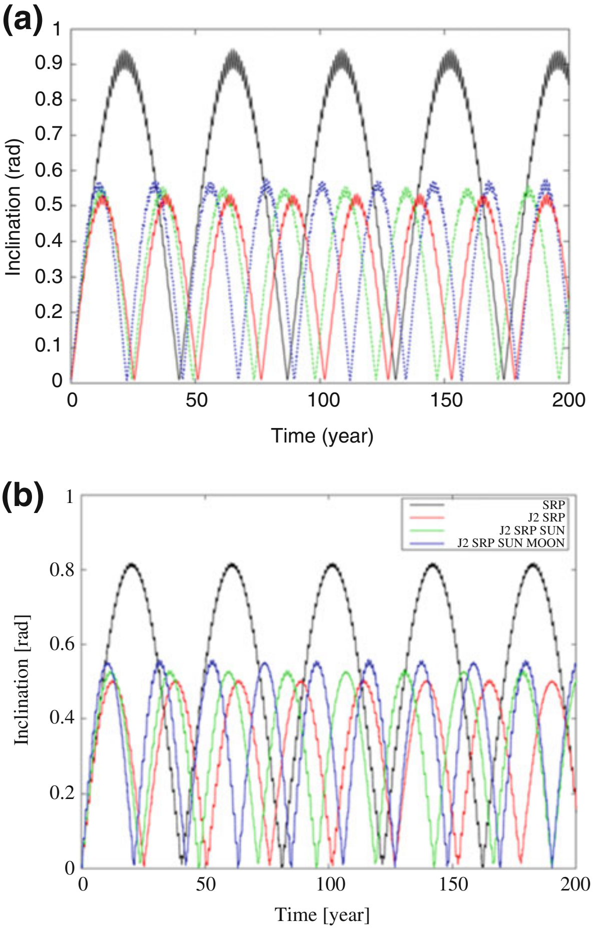


Fig. 11 The behavior of the inclination obtained for each of the four analytical models (**a**) and the results of the numerical integration for the same cases (**b**) (source [6])

integration step, or by using suitable coordinates. We use NIMASTEP developed by [17], as reference of any space debris orbit. The software has been intensively tested and compared, and is a robust tool. It is adaptable to any other telluric planet and has been used in particular for Mercury. Recently it has been completed by the different models of atmospheric drag (see Sect. 10) by Petit [33] so to be able to follow a motion from GEO to LEO and to predict reentry dates.

7.2 *The Symplectic Integrator*

The symplectic integrator basic idea is to divide the Hamiltonian into two separate parts, A and B, and to perform one half step with the first dynamical part, one step with the second one and again one half step with the first one (SABA). Other combinations are also possible, as SBAB for example. The principles and the details of the method are explained and developed in [26] and commented for different orders.

For the space debris, the implementation has been performed for 500 years by [21], the period of the chosen ephemeris of the Sun. However by replacing the Sun's ephemeris by another one or by an approximate analytical expression, it can be pushed well further (see Sect. 8.3). Let us remind that the motions of the Sun and Moon are introduced as given functions of time, and introduce periodic variations in the total hamiltonian energy.

The two parts A and B have been chosen as:

$$A(\mathbf{v}, \mathbf{r}, \Lambda) = H_K(\mathbf{v}, \mathbf{r}) + H_{Rot}(\Lambda) \tag{73}$$

$$B(\mathbf{r}, t) = H_{geo}(\mathbf{r}, t) + H_{3B}(\mathbf{r}, t) + H_{SRP}(\mathbf{r}, t) \tag{74}$$

with $H_{Rot}(\Lambda) = \dot{\theta}\ \Lambda$, $H_K(\mathbf{v}, \mathbf{r}) = \frac{v^2}{2} - \frac{\mu}{r}$, t the time, being present through θ the rotation of the Earth, in H_{geo}, and through the positions of the Sun or the Moon, in H_{SRP} and H_{3B}.

The efficiency and performance of SYMPLEC for different integrators and orders, are described in [21] and show how large could be the time step, keeping a quasi constant integral of motion. An example is reproduced in Fig. 12.

We present here a comparison between SYMPLEC version SABA order 4, with a stepsize of 4 h (about 14400 s) and NIMASTEP, with Adams–Bashforth–Moulton integrator of order 10, and for several stepsizes: 1152, 1004, 864 or 432 s. The differences are plotted in Fig. 13.

8 The Chaos

With so many perturbations, with the presence of resonances, it is obvious that the debris zone is chaotic. To measure at which degree, with which intensity, we can analyze the dynamics with chaos indicators. More precisely, we are going to use here the MEGNO (Mean Exponential Growth factor of Nearby Orbits) introduced by Cincotta and Simo [14] and developed by the same team [15]. The calculation has been inserted into NIMASTEP and into SYMPLEC, and for some specific zones, we refine the MEGNO analysis with the Frequency Map Analysis as introduced by Laskar [25].

8.1 The MEGNO Maps

In chaotic (irregular) regions of phase space, two initially nearby trajectories *separate roughly exponentially with time*; in quasi-periodic (regular) regions,

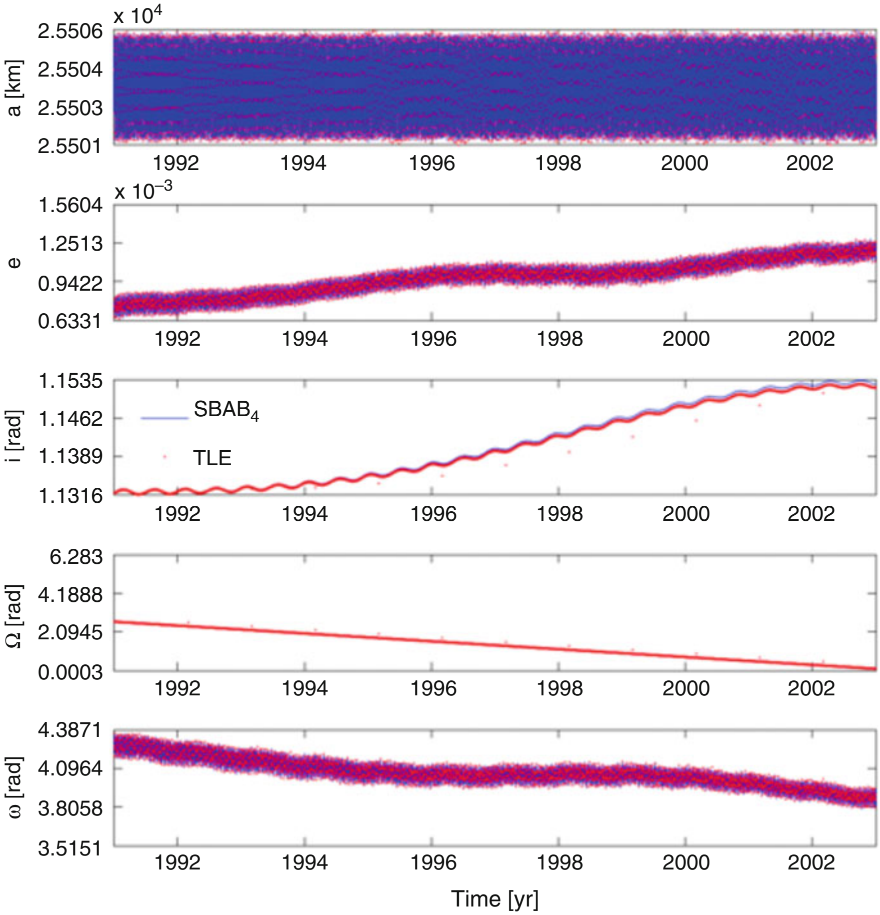


Fig. 12 Comparison between SYMPLEC performed with SABA order 4, and a TLE with a very precise orbit over 12 years ($A/m = 0.001\,\mathrm{m}^2/\mathrm{kg}$). The solar radiation pressure, with conical shadowing effects, the lunisolar perturbations and the geopotential up to the order 12, are included in the model (source [21])

neighboring trajectories *separate roughly linearly with time*. The chaotic indicator computes this rate of separation, the increasing divergence between two close orbits, and gives information about the sensitive dependence on initial conditions.

The Lyapunov coefficients γ and λ quantify this dependence, for a finite time or at infinity.

$$\gamma(t) = \frac{1}{t - t_0} \ln\left(\frac{d(t)}{d(t_0)}\right) \quad \text{and} \quad \lambda = \lim_{t\to\infty} \gamma(t)$$

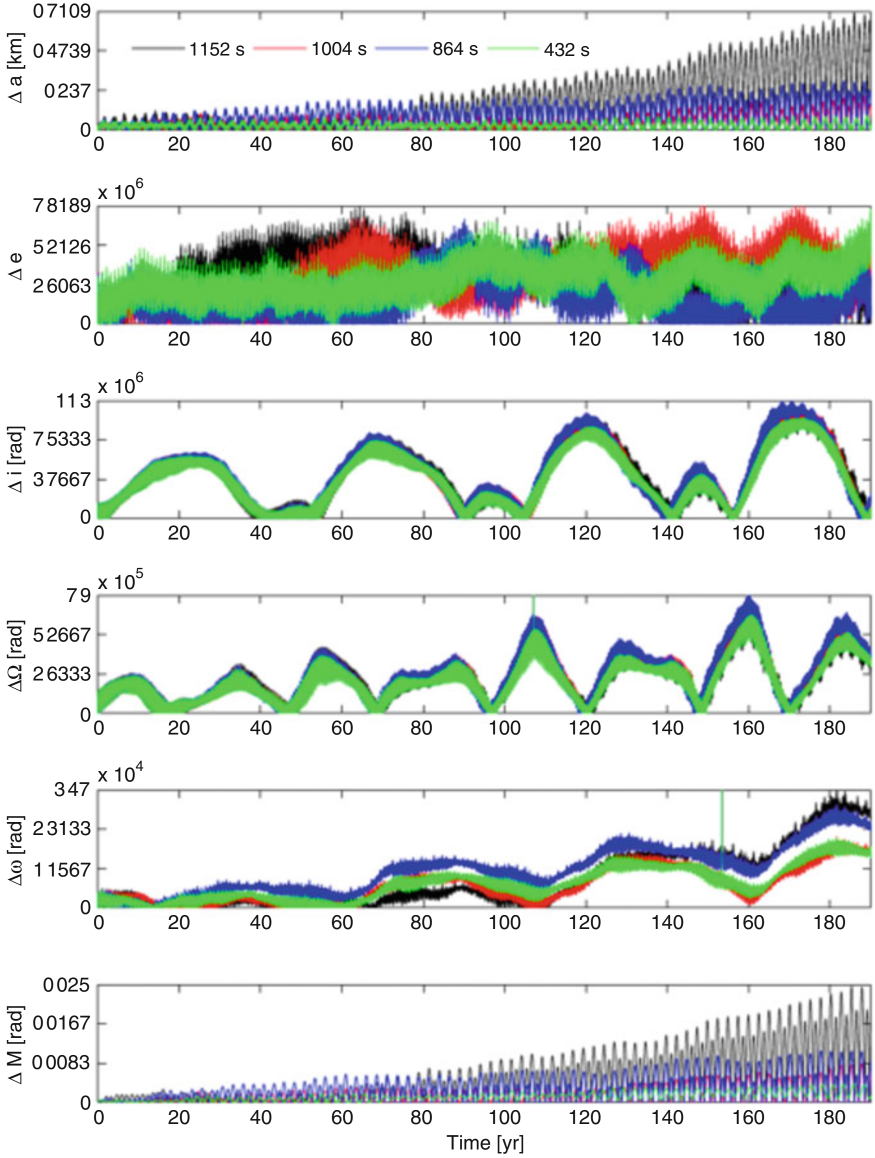


Fig. 13 Differences between SYMPLEC (stepsize of 4 h) and NIMASTEP, associated to stepsizes of 1152 s (black), 1004 s (blue), 864 s (red) or 432 s (green) (source [21])

where d is the Euclidian distance between two initially nearby trajectories. If the trajectories are chaotic (irregular), d grows exponentially (on the average), and γ approaches some positive constant; on the opposite, for quasi-periodic trajectories (regular), d grows linearly and γ approaches zero with a rate $ln(t)/t$.

The distance d is obtained via the solution of the variational dynamical system, δ associated to the main dynamics, $d = \|\delta\|$.

If the flow is given by:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\alpha}), \qquad x \in \mathbb{R}^{2n}, \tag{75}$$

where $\boldsymbol{\alpha}$ is a vector of parameters, the linear variational equations are:

$$\dot{\boldsymbol{\delta}} = \frac{d}{dt}\, \boldsymbol{\delta}(\boldsymbol{\phi}(t)) = J(\boldsymbol{\phi}(t))\, \boldsymbol{\delta}(\boldsymbol{\phi}(t)), \quad \text{with} \quad J(\boldsymbol{\phi}(t)) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\boldsymbol{\phi}(t)), \tag{76}$$

and $\boldsymbol{\phi}(t)$ a solution of the flow.

Concretely, the MEGNO (Mean Exponential Growth factor of Nearby Orbits) indicator, Y, as well as its mean value, $\bar{Y}$, are given by integrals, and their time derivatives are added as new differential equations in the dynamical system, main and variational.

$$Y(\boldsymbol{\phi}(t)) = \frac{2}{t} \int_0^t \frac{\dot{d}(\boldsymbol{\phi}(s))}{d(\phi(s))}\, s\, ds, \qquad \bar{Y}(\boldsymbol{\phi}(t)) = \frac{1}{t} \int_0^t Y(\boldsymbol{\phi}(s))\, ds \tag{77}$$

If the orbit is chaotic (irregular) $\bar{Y}(t) \simeq \lambda/2\ t$, if it is quasi-periodic (regular), $\bar{Y}(t) \to 2$ and for stable, isochronous periodic orbits, $\bar{Y}(t) \to 0$.

Breiter et al. [4] published the first paper applying the MEGNO to the space debris dynamics. We have also used MEGNO intensively in the geostationary region [45], to measure the stability of the different regions, for debris with different values of A/m. For example, for an integration of 30 years, Fig. 14 shows the spread up of the chaos zone, when $A/m = 1, 5, 10, 20$ m^2/kg. The pendulum-like space phase is more and more perturbed and the MEGNO values increase, far away from 2. In the first graph, after 30 years, except for initial conditions close to the separatrix, almost all trajectories are regular; for the last graph, only the central region of the pendulum is still stable.

The FLI (Fast Lyapounov Indicator) introduced by Froeschlé et al. [19] is the most popular chaos indicator and has been applied to the space debris population by several teams, in particular [12, 16]. Using the FLI or the MEGNO, the different position of the stability zones and resonant curves coincide quite well. However their position are very dependent on the force model used for their determination. To focus only on one angle of the geopotential makes the phase space very stable, which is not at all the case when we add successive harmonics.

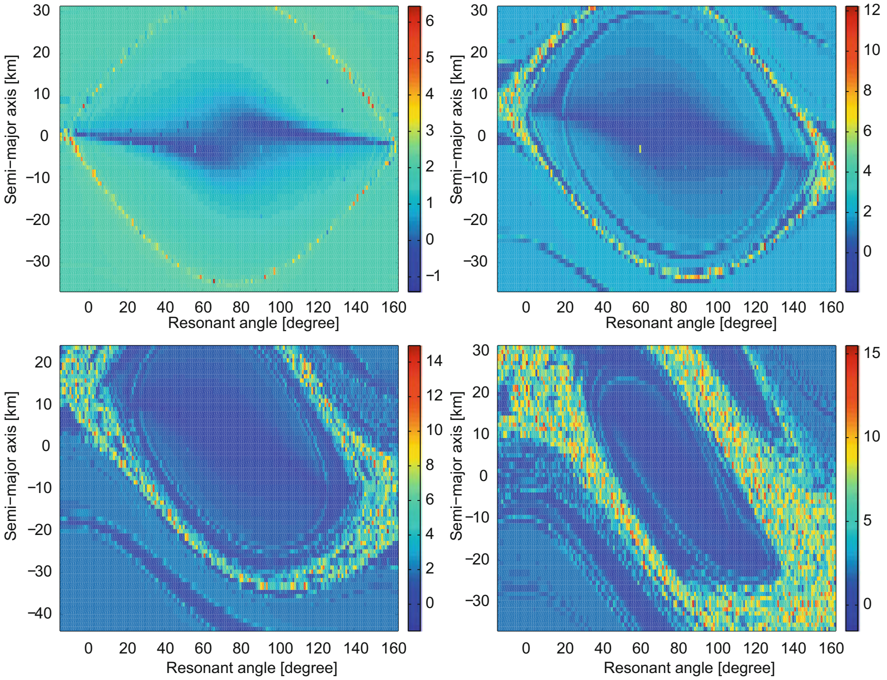


Fig. 14 The MEGNO maps for four cases of solar radiation pressure coefficients: $A/m = 1, 5, 10, 20\ \mathrm{m}^2/\mathrm{kg}$ (source [45])

8.2 The Frequency Map

Several improvements are performed to these figures; first of all, by an averaging process, the distorsion of the pendulum can be reduced as shown in [45]. Moreover, some specific regions are analyzed more precisely using the frequency map analysis introduced by Laskar [25] measuring the variations of the frequencies of a quasi-periodic approximation of the motion. Their variations are, in particular, measured by following their second derivatives. Figure 15 shows an example obtained for $A/m = 10\,\mathrm{m}^2/\mathrm{kg}$ and Fig. 16 gives the second derivatives calculated along a slice of the previous figure. The method allows to identify small islands of stability and curves of chaotic motions inside the stable zones.

More precisely, if we zoom on the resonant zone of the pendulum, we clearly see three stable islands near the separatrix (see Fig. 17). They correspond to secondary resonances and can be analyzed analytically and locally.

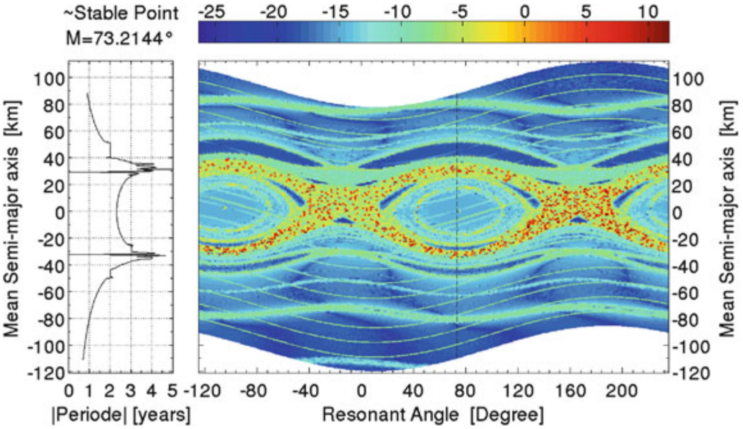


Fig. 15 The frequency map of the space phase (source [27])

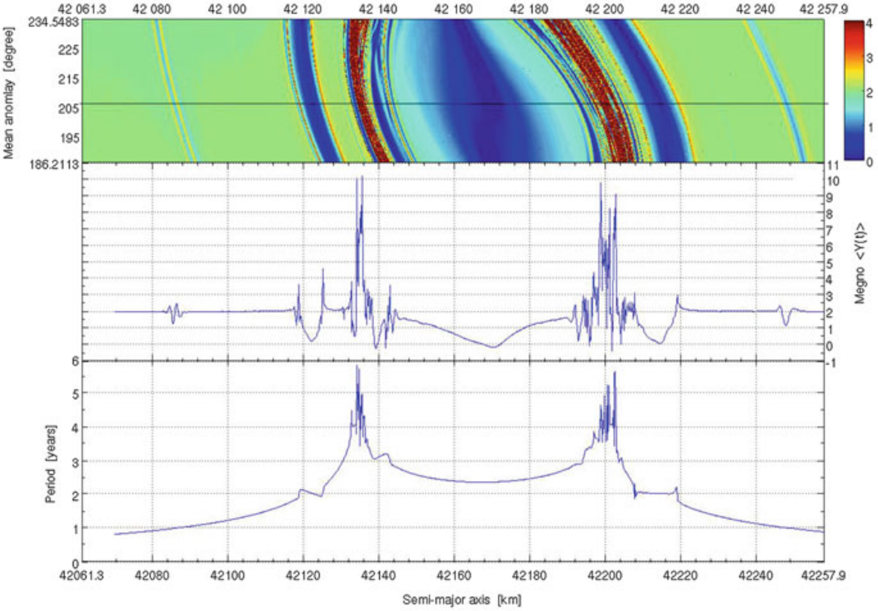


Fig. 16 The frequency map: details of a slice with the second derivative of the frequency as indicator (source [45])

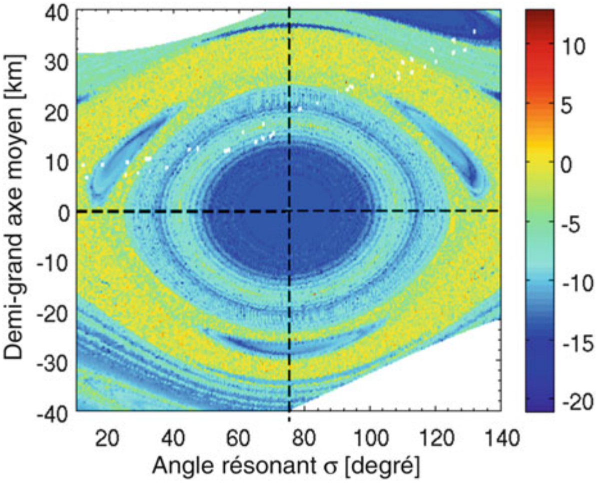


Fig. 17 The frequency map: a zoom on the resonant zone $A/m = 10\,\mathrm{m}^2/\mathrm{kg}$. The x-axis is the resonant angle, σ, given in degrees, and the y-axis is the semi-major axis, a, given in km

8.3 The Secondary Resonances

To explain the presence of the secondary resonances, in the geostationary region, for $A/m = 10\,\mathrm{m}^2/\mathrm{kg}$, let us start again with the Hamiltonian linked to the coefficient J_{22} (see [27] for the details of the calculations).

$$\mathscr{K} = -\frac{\mu^2}{2L^2} - \dot{\theta} L + \frac{3\mu^4}{L^6} R_e^2\, J_{22} \cos 2(\sigma - \sigma_0) - \frac{15\mu^4}{2L^6} R_e^2\, e^2\, J_{22} \cos 2(\sigma - \sigma_0) \tag{78}$$

where we keep the term in e^2 because we know that, in presence of high values of A/m, the variations of the eccentricity can be very large, with an annual motion.

We inject the solution obtained for the eccentricity (41), we make simplifications (null obliquity for example):

$$e^2 = \frac{\mathscr{Z}^2}{L^2 n_S^2} + \gamma^2 + \frac{2\mathscr{Z}}{L n_S}\, \gamma\, \cos(\lambda_S + \delta) \tag{79}$$

with $\mathscr{Z} = \kappa\sqrt{L}$, and the final (with all these successive approximations) Hamiltonian K is:

$$K(L, \sigma) = -\frac{\mu^2}{2L^2} - \dot{\theta} L + \cos(2\sigma - 2\sigma_0) \left[\frac{F}{L^6} - \frac{2G}{L^6} \cos(\lambda_S + \delta) \right], \tag{80}$$

with

$$F = 3\mu^4\, R_e^2\, J_{22} - \frac{15\mu^4}{2} R_e^2\, J_{22} \left(\frac{\mathscr{L}^2}{L^2 n_S^2} + \gamma^2 \right)$$

$$G = \frac{15\mu^4}{2} R_e^2\, J_{22}\, \frac{\mathscr{L}}{L n_S}\, \gamma, \tag{81}$$

in which we can rewrite :

$$2\cos(2\sigma - 2\sigma_0)\ \cos(\lambda_S + \delta) = \cos(2\sigma + \lambda_S - 2\sigma_0 + \delta) + \cos(2\sigma - \lambda_S - 2\sigma_0 - \delta).$$

The period of λ_S is clearly 1 year, the period of σ is about 2.5 years at the center of the libration zone and goes to infinity near the separatrix. Using a pendulum model, we can detect the regions where the two frequencies, $2\dot{\sigma}$ and $\dot{\lambda}_S$, are commensurable. All the calculations can be performed through the pendulum formulation, using elliptic integrals, and the positions of the main secondary resonances are then identified. We use, for each resonant case, a classical pendulum formulation with R the momentum, and r is the resonant angle, as:

$$h = \frac{R^2}{2} - b\ \cos r.$$

The initial conditions close to the separatrix can be chosen as $R = 0$ and $r = \pi - \epsilon$, and the energy level is given by

$$h_\epsilon = -b\cos(\pi - \epsilon) = b\cos\epsilon.$$

ϵ is then a parameter measured in radians from the separatrix, with $\epsilon = 0$ on the separatrix. In particular a (secondary) resonance 3:1 can be isolated, in the vicinity of the separatrix, at a distance $\epsilon = 0.9$ in the reduced variables. The calculated angular positions of the three islands are 60.26°, 180.26°, and 300.26° measured from the vertical positive axis, which is exactly what Fig. 17 shows.

Section 8.3 gives the positions of the different commensurabilities through the parameter ϵ and compares the values obtained by an analytical formula of the pendulum with a comparable numerical integration (Fig. 18).

A similar study has been performed in the circulation zone, with the detection of the main secondary resonances: 1:2, 1:3, etc.

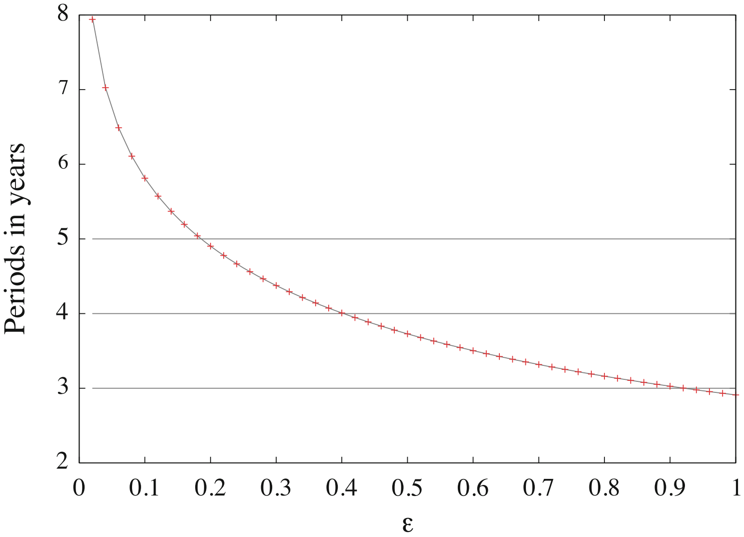


Fig. 18 Resonant case: the period of $2\sigma - 2\sigma_0$ calculated in years, through a numerical integration of the pendulum differential equations (dots) and through the analytical expressions (lines), as functions of ϵ measuring the distance from the separatrix ([27])

9 The Lunisolar Resonances

We have analyzed the gravitational resonances, characterized by a commensurability between the rotation of the Earth and the orbital period of space debris (GEO and MEO). We have mentioned the secondary resonances, which act inside a resonance, between the libration angle and another slow angle (between σ and λ_S in our case).

We introduce now the lunisolar resonances, which are secular resonances between ω and Ω of the space debris and the nodes and perigees of the Moon and the Sun, analyzed first by Breiter [3].

More recently several authors have systematically revisited the lunisolar resonances, in particular we can mention [16] or [36] but also [10–12]. Mixing the perturbation theories, the FLI to detect chaotic zones, the space phase is really sliced in different zones of stability and chaos, determined by the lunisolar resonances. Inside those resonances, structures are visible, probably secondary resonances inside the secular ones.

For the Moon, let us mention that the following combinations of angles are identified as potential secular resonances:

$$\dot{\Psi}_{2-2p,m,\pm s} = (2-2p)\,\dot{\omega} + m\,\dot{\Omega} \pm \dot{\Omega}_M \simeq 0, \qquad (82)$$

and for the Sun, the expressions are even simpler:

$$\dot{\Psi}_{2-2p,m} = (2-2p)\,\dot{\omega} + m\,\dot{\Omega} \simeq 0. \tag{83}$$

The challenge is now to identify the different layers obtained by these authors inside the secular ones. Analytical and numerical perspectives have to be mixed up, to understand the complexity of the dynamics in these regions.

10 The Atmospheric Drag

When space debris reach the LEO region, the conservative forces do not describe the complete dynamics and are not able to predict the reentry date. The atmospheric drag plays an important and efficient role, speeding up the loss of energy, and pushing down the debris to the Earth.

However this drag is not easy to model; different models and approaches exist and give different results (see [33]).

Let us mention the most popular density models:

- JB2006/JB2008: developed by [2], it is a semi-analytical model, based on the preliminary model Jacchia-71 (see [23]). It is still now the reference of the committee on Space Research (COSPAR).
- DTM2013: it is a drag temperature model (see [5]), also semi-analytical, which includes the data of the satellites Stella, Starlette, OGO-6, DE-2, AE-C, AE-E, CHAMP, GRACE and GOCE for altitudes between 200 and 900 km.
- TD88: it is an empirical model (see [41]), filled on the observation data, extended up to 1200 km.
- Many others could also be mentioned: other versions of Jacchia, MSIS, NRLMSISE00, GRAM, MET, GOST, TIEGCM, etc.

The density functions depend on many parameters, let us mention the most important ones: the solar flux, the geomagnetic activity, the local time, the length of the day, the latitude.

To compare the models and their prediction, two real orbits, corresponding to the satellites Stella and Starlette are given by their TLE. The three atmospheric models, JB2008, DTM2013 and TD88, are inserted into the software NIMASTEP associated to the Adam-Bashforth-Moulton integrator of order 10, and followed for more than 20 years (Fig. 19).

The qualitative behaviors of the three models is very similar, the differences appear each time an event occurs, decreasing sharply the altitude. The models are slightly different and could describe these events (due mainly to solar activity) in a slightly different way.

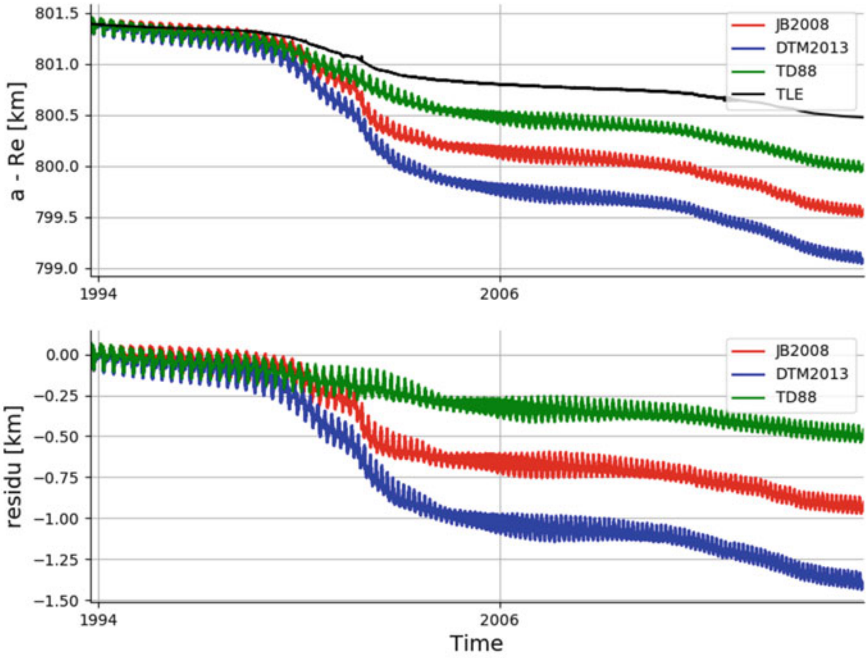


Fig. 19 Evolution of Starlette semi-major axis using JB2008, DTM2013, TD88, in comparison with the TLE (source [33])

Other comparisons obtained by different software are presented in [33], especially the follow up of the fragments of the explosion of the Chinese satellite Fengyun-1C (Fig. 20).

11 Yarkovsky-Schachs Effect

In the asteroidal context, the importance of Yarkovsky's thermal force is now obvious, on long periods of time, linked to the A/m coefficient and to the spin rate of the body. In the case of short space missions, the effect on the artificial satellites, over a few years, is certainly negligible; however on space debris, present for hundreds of years, and spinning in all the manners, the question is open. We have decided to test one of its aspects, the so-called Yarkovsky-Schachs effect, an orbital thermal contribution to the dynamics, proportional to the A/m coefficient.

We use the classical formulations reserved to asteroidal motions, and we adjust their expressions to the debris case. The main orbital effect is due to the differences of temperature of the piece of debris, due to the Sun's warmth. More precisely, the Yarkovsky-Schachs effect induces long-term semi-major axis variations, which

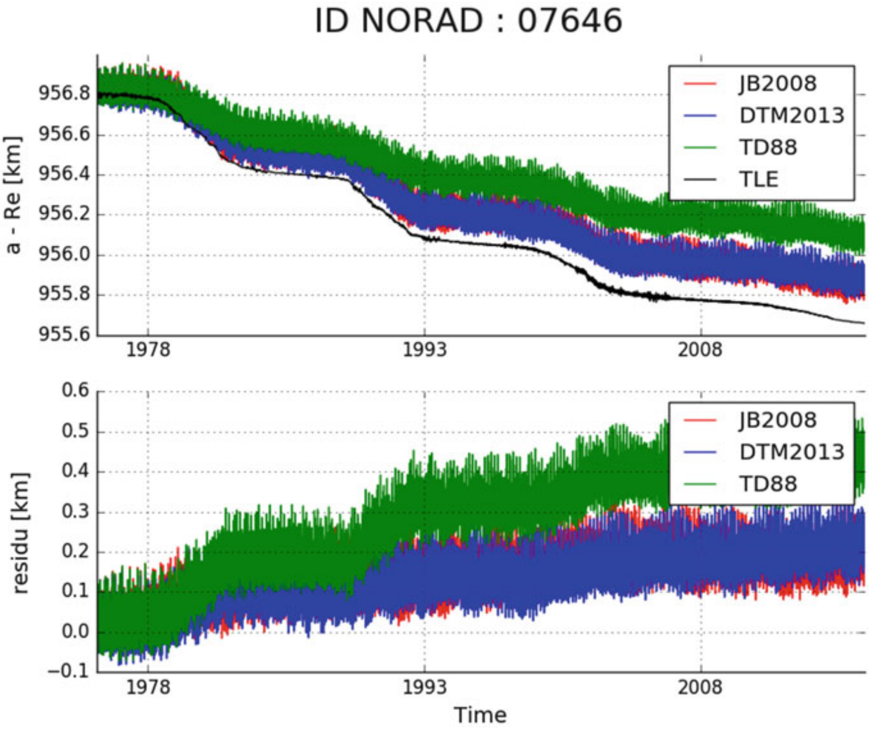


Fig. 20 Evolution of Stella semi-major axis using JB2008, DTM2013, TD88, in comparison with the TLE (source [33])

appear when the orbit crosses the Earth's shadow. The solar flux arriving at the satellite surface is then interrupted, the satellite surface cools down after entering the shadow, and heats up again after exiting from it. The recoil force does not average on one orbit, and the problem becomes therefore position-dependent.

The order of magnitude of this effect is very small, and in comparison with the other perturbations, it does not seem important to insert this force systematically in the dynamical models. For details see [32].

12 The Synthetic Populations of Space Debris

The building of virtual or synthetic populations is based on the work performed in naXys Institute (Namur Complex systems) by the research group in mobility and traffic: they are implicated in projects for more 40 years, to describe the traffic dynamics, from the classical Dijkstra algorithm of shortest paths in a graph, to psychological models about human behavior to choose a way. Facing the necessity of collecting data about families, ages, schools, supermarkets, rates of employment, they have realized the difficulties of getting a suitable set of data, usually very local and limited by the protection of private life.

During the last 10 years, they have succeeded in building a full synthetic population of Belgium, with more than ten millions of people, organized in families, with work, schools, habits, completely virtual but as close as possible to the reality (it means to the available local data). They have obtained an expertise in specific statistical methods adapted to this virtual population concept.

We develop a synthetic population of virtual space debris with similar characteristics to the real ones. Our known data are the 20,000 TLE corresponding to 10 cm or more objects. Our objective is to simulate a much larger population of objects, in particular objects of 1 cm, which are able, due to their high relative velocity, to create damages on the active satellites or probes, or even to the ISS.

The first tests are convincing, we simulate an event, like an explosion and a collision, creating new debris than the big ones observed and inserted in the TLE catalogue [34]. The method used is the Iterative Proportional Fitting (IPF), an iterative process for weighting data describing a population up to the convergence to a stable state. It is based on a matrix formulation, after discretization of the data $(a, e, i, \omega, \Omega, M, A/M)$. We follow the fragmentation of the satellite Ekran 2, and we compare the initial and the synthetic populations, and we obtain the convergence of the method. We have created artificially a new initial population, by modifying the ejection velocity by a factor 2, and we can measure the differences betweens the two observed clouds.

We hope to develop this tool as a real simulator of catastrophic events or predictor of developments of debris clouds.

Acknowledgements This paper is the summary of several papers concerning the space debris dynamics, published by some of my PhD students and post-docs. I would like to thank all of them for your collaboration and discussions: S. Valk, N. Delsate, Ch. Hubaux, A. Petit, M. Murawiecka, F. Deleflie and D. Casanova.

References

1. Aksnes, K.: Short-period and long-period perturbations of a spherical satellite due to direct solar radiation. Celest. Mech. **13**, 89–104 (1976)
2. Bowman, B.R., Kent Tobiska, W., Marcos, F.A., Valladares, C.: The JB2006 empirical thermospheric density model, J. Atmos. Sol. Terr. Phys. **70**, 774–793 (2008)
3. Breiter, S.: Lunisolar resonances revisited. Celest. Mech. Dyn. Astron. **81**, 81–91 (2001)
4. Breiter, S., Wytrzyszczak, I., Melendo, B.: Long-term predictability of orbits around the geosynchronous altitude. Adv. Space Res. **35**, 1313–1317 (2005)
5. Bruinsma, S.: The DTM-2013 thermosphere model. J. Space Weather Space Clim. **5**, A1 (2015)
6. Casanova, D., Petit, A., Lemaître, A.: Long-term evolution of space debris under the J_2 effect, the solar radiation pressure and the solar and lunar perturbations. Celest. Mech. Dyn. Astron. **123**, 223–238 (2015)
7. Celletti, A., Galeş, C.: On the dynamics of space debris: 1:1 and 2:1 resonances. J. Non Linear Sci. **24**, 1231–1262 (2014)

8. Celletti, A., Galeş, C.: Dynamical investigation of minor resonances for space debris. Celest. Mech. Dyn. Astron. **123**, 203–222 (2015)
9. Celletti, A., Galeş, C.: A study of the main resonances outside the geostationary ring. Adv. Space Res. **56**, 388–405 (2015)
10. Celletti, A., Galeş, C.: A study of the lunisolar secular resonance $2\dot{\omega} + \dot{\Omega} = 0$. Front. Astron. Space Sci. **3**, 11 (2016)
11. Celletti, A., Galeş, C., Pucacco, G., Rosengren, A.J: Analytical development of the lunisolar disturbing function and the critical inclination secular resonance. Celest. Mech. Dyn. Astron. **127**, 259–283 (2017)
12. Celletti, A., Efthymiopoulos, C., Gachet, F., Galeş, C., Pucacco, G.: Dynamical models and the onset of chaos in space debris. Int. J. Non Linear Mech. **90**, 147–163 (2017)
13. Chao, C.C.: Analytical investigation of GEO debris with high area-to-mass ratio, AIAA paper No. AIAA-2006-6514. In: Presented at the 2006 AIAA/AAS Astrodynamics Specialist Conference, Keystone, Colorado (2006)
14. Cincotta, P.M., Simó, C.: Simple tools to study global dynamics in non-axisymmetric galactic potentials. Astron. Astrophys. Suppl. **147**, 205–228 (2000)
15. 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. Phys. D **182**, 151–178 (2003)
16. Daquin, J., Rosengren, A.J., Alessi, E.M., Deleflie, F., Valsecchi, G.B., Rossi, A.: The dynamical structure of the MEO region: long-term stability, chaos, and transport. Celest. Mech. Dyn. Astron. **124**, 335–366 (2016)
17. Delsate, N., Compère, A.: NIMASTEP: a software to modelize, study and analyze the dynamics of various small objects orbiting specific bodies. Astron. Astrophys. **540**, A120 (2012)
18. Ferraz-Mello, S.: Analytical study of the Earth's shadowing effects on satellite orbits. Celest. Mech. **5**, 80–101 (1972)
19. Froeschlé, C., Lega, E., Gonczi, R.: Fast lyapunov indicators. Application to Asteroidal Motion. Celest. Mech. Dyn. Astron. **6**, 41–62 (1997)
20. Gachet, F., Celletti, A., Pucacco, G., Efthymiopoulos, C.: Geostationary secular dynamics revisited: application to high area-to-mass ratio objects. Celest. Mech. Dyn. Astron. **128**, 149–181 (2017)
21. Hubaux, Ch., Lemaître, A., Delsate, N., Carletti, T.: Symplectic integration of space debris motion considering several Earth's shadowing models. Adv. Space Res. **49**, 1472–1486 (2012)
22. Hubaux, Ch., Lemaître, A.: The impact of Earth's shadow on the long-term evolution of space debris. Celest. Mech. Dyn. Astron. **116**, 79–95 (2013)
23. Jacchia, L.G.: Revised static models of the thermosphere and exosphere with empirical temperature profiles. SAO Special Report **332** (1971)
24. Kaula, W.M.: Theory of Satellite Geodesy. Blaisdell Publishing Company, Waltam, Toronto (1966)
25. Laskar, J.: Frequency analysis of a dynamical system. Celest. Mech. Dyn. Astron. **56**, 191–196 (1993)
26. Laskar, J., Robutel, P.: High order symplectic integrators for perturbed Hamiltonian systems. Celest. Mech. Dyn. Astron. **80**, 39–62 (2001)
27. Lemaître, A., Delsate, N., Valk, S.: A web of secondary resonances for large A/m geostationary debris. Celest. Mech. Dyn. Astron. **104**, 383–402 (2009)
28. McMahon, J., Scheeres, D.: Secular orbit variation due to solar radiation effects: a detailed model for BYORP. Celest. Mech. Dyn. Astron. **106**, 261–300 (2010)
29. Metris, G., Exertier, P.: Semi- analytical theory of the mean orbital motion. Astron. Astrophys. **294**, 278–286 (1995)
30. Milani, A., Gronchi, G.F.: Theory of Orbital Determination. Cambridge University Press, Cambridge (2010)
31. Moons, M.: Averaging approaches. In: Proceedings of the Artificial Satellite Theory Workshop, U.S.N.O. Washington D.C. (1993)

32. Murawiecka, M., Lemaître, A.: Yarkovsky-Schach effect on space debris motion. Adv. Space Res. **61**(3), 935–940 (2017)
33. Petit, A., Lemaitre, A.: The impact of the atmospheric model and of the space weather data on the dynamics of clouds of space debris. Adv. Space Res. **57**, 2245–2258 (2016)
34. Petit A., Casanova, D., Dumont M., Lemaitre A.: Design of a synthetic population of geostationary space debris by statistical means. In: AAS-AAS 17-363 (2017)
35. Rosengren, A.J., Scheeres, D.J.: Long-term dynamics of high area-to-mass ratio objects in high-Earth orbit. Adv. Space Res. **52**, 1545–1560 (2013)
36. Rosengren, A.J., Daquin, J., Tsiganis, K., Alessi, E.M., Deleflie, F., Rossi, A., Valsecchi, G.B.: Galileo disposal strategy: stability, chaos and predictability. Mon. Not. R. Astron. Soc. **464**, 4063–4076 (2017)
37. Rossi, A.: Resonant dynamics of Medium Earth Orbits: space debris issues. Celest. Mech. Dyn. Astr., **100**, 267–286 (2008)
38. Schildknecht, T., Musci, R., Ploner, M., Beutler, G., Flury, W., Kuusela J., Leon Cruz, J., de Fatima Dominguez Palmero, L.: Optical observations of space debris in GEO and in highly-eccentric orbits. Adv. Space Res. **34**, 901–911 (2004)
39. Schildknecht, T., Musci, R., Flohrer, T.: Properties of the high area-to-mass ratio space debris population at high altitudes. Adv. Space Res. **41**, 1039–1045 (2007)
40. Schildknecht, T., Früh, C., Herzog, A., Hinze, J., Vananti, A.: AIUB efforts to survey, track, and characterize small-size objects at high altitudes. In: Proceedings of 2010 AMOS Technical Conference, 14–17 September, Maui, HI (2010)
41. Sehnal, L.: Thermospheric total density model TD. Bull. Astron. Inst. Czechoslovakia **39**, 120–127 (1988)
42. Valk, S., Lemaître, A.: Semi-analytical investigations of high area-to-mass ratio geosynchronous space debris including Earth's shadowing effects. Adv. Space Res. **42**, 1429–1443 (2008)
43. Valk, S., Lemaître, A., Anselmo, L.: Analytical and semi-analytical investigations of geosynchronous space debris with high area-to-mass ratios. Adv. Space Res. **41**, 1077–1090 (2008)
44. Valk, S., Lemaître, A., Deleflie, F.: Semi-analytical theory of mean orbital motion for geosynchronous space debris, under gravitational influence. Adv. Space Res. **43**, 1070–1082 (2009)
45. Valk, S., Delsate, N., Lemaître, A., Carletti, T.: Global dynamics of high area-to-mass ratios GEO space debris by means of the MEGNO indicator. Adv. Space Res. **43**, 1509–1526 (2009)

Computing Invariant Manifolds for Libration Point Missions

Josep-Maria Mondelo

Abstract The goal of this lecture is to review several methodologies for the computation of invariant manifolds, having in mind the needs of preliminary mission design of libration point missions. Because of this, the methods reviewed are developed for and applied to the circular, spatial restricted three-body problem (RTBP), although most of them can be applied with few changes, or almost none, to general dynamical systems. The methodology reviewed covers the computation of (families of) fixed points, periodic orbits, and invariant tori, together with the stable and unstable manifolds of all these kinds of invariant objects, and also homoclinic and heteroclinic connections between them. The methods reviewed include purely numerical and semi-analytical ones. No background is assumed except for a graduate level knowledge of calculus, differential equations and basic numerical methods. In particular, the notions from the theory of dynamical systems required for the development of the methods are introduced as needed.

Keywords Libration points · Restricted Three-Body Problem · Lissajous orbits · Halo orbits · Periodic orbits · Invariant tori · Invariant manifolds · Homoclinic and heteroclinic connections · Center manifold · Parameterization method · Automatic differentiation

1 Introduction

In libration point missions, spacecraft are sent to orbits that stay close to the fixed points of the circular, spatial, restricted three-body problem (RTBP) with primaries the Sun and a planet, or a planet and a moon. The RTBP model describes the motion of an infinitesimal particle under the attraction of two massive bodies known as primaries, that are assumed to revolve uniformly in circles around their center of

J.-M. Mondelo (✉)
IEEC-CERES & Departament de matemàtiques, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
e-mail: jmm@mat.uab.cat

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_4

mass. In rotating coordinates, this model has five equilibrium points: three of them, L_1, L_2, L_3, also known as *collinear*, were discovered by Euler, and two more, L_4, L_5, also known as *triangular*, were discovered by Lagrange. Compared to orbits around the Earth or other planets, orbits around the collinear libration points provide ideal locations for space observation. Among their advantages are the absence of shadow of a celestial body, thus providing a more stable thermal environment, and continuous access to the whole celestial sphere, except for a direction, that is not fixed but rotates with the primaries. Also, the instability of the collinear libration points gives rise to a very rich dynamical structure, that can be exploited not only to search for operational nominal orbits but also to find low-energy passageways between them. These operational orbits could be either of the libration point type or around celestial bodies.

Four examples of libration point missions of different kinds are:

- SOHO, launched in Dec. 1995, to an Halo orbit around the collinear point L_1 of the Earth-Sun system. Its goal is to provide continuous observations of the Sun, and is still operational.
- WMAP, launched in June 2001, to a Lissajous orbit around the collinear point L_2 of the Earth-Sun system. Its goal was to map the temperature fluctuations of the cosmic microwave radiation.
- Genesis, launched in Aug. 2001, to an Halo orbit around the collinear point L_1 of the Earth-Sun system. Its goal was to collect solar wind samples and deliver them to Earth in daylight. For this, an additional excursion close to the collinear point L_2 of the Earth-Sun system was necessary.
- Artemis, started in Jan. 2009 as an extension of the mission of two of the spacecrafts of the Themis mission, that, using the remaining fuel, were sent from high, eccentric Earth orbits to lunar orbits using L_1 and L_2 Earth-Moon dynamics.

Illustrations of the trajectories of these four missions are shown in Fig. 1.

The nominal trajectories of these four missions can be identified among the families of periodic orbits and invariant tori related to the collinear libration points of the RTBP. In the case of SOHO and Genesis, the nominal trajectory is part of the Halo family of periodic orbits. In the case of WMAP, it is part of the Lissajous family of invariant tori. In the case of Artemis, the nominal trajectories would be the final lunar ones, but invariant tori of the L_1 and L_2 Lissajous family play a fundamental role in the transfer from Earth to lunar orbits. The invariant stable (resp. unstable) manifolds of all these periodic orbits and tori can be used to arrive to (resp. depart from) them. In the case of the P1 spacecraft of Artemis, an heteroclinic connection is closely followed in order to go from the Lissajous torus around L_2 to the Lissajous torus around L_1. Such connections are obtained as intersections of the stable manifold of the arriving object and the unstable manifold of the departing object. An heteroclinic connection is also outlined by the Genesis mission.

The preliminary mission design of these kind of missions is based in being able to compute families of trajectories, in order to be able to select the one that best satisfies the requirements of the mission. The goal of this lecture is to review some

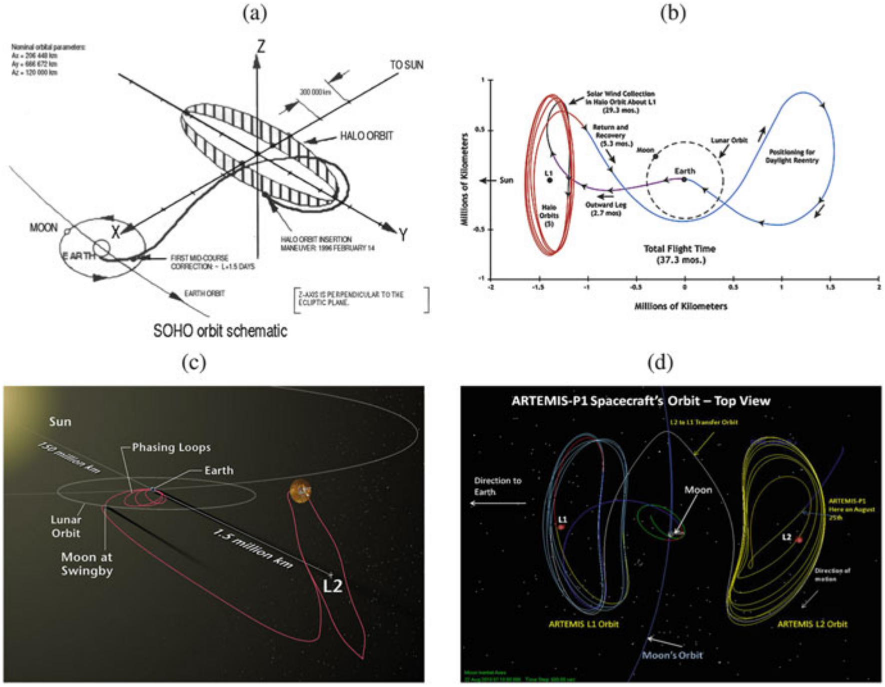


Fig. 1 Trajectories of the missions: (**a**) SOHO, (**b**) Genesis, (**c**) WMAP, (**d**) Artemis

of the numerical and semi-analytical techniques available in the literature that can be used in order to compute families of periodic orbits and invariant 2D tori of the RTBP as a dynamical system, as well as their invariant stable and unstable manifolds. Some discussion will be done also on the computation of homoclinic and heteroclinic connections. Although preliminary mission design is the main motivation for this lecture, all the techniques that will be described can be used for the numerical computation of invariant objects in other conservative dynamical systems. Many of them can be directly used in or easily adapted to the dissipative case as well.

The methods that will be described can be divided in two classes: numerical and semi-analytical. Semi-analytical methods provide expansions around a base object that must be previously known. They are more convenient than numerical methods for parametric studies of trajectory features, since a single expansion covers a family or many families of trajectories. They have as a drawback that the expansions provide a good approximation of the dynamics in a neighborhood of the base object, but not outside this neighborhood. Numerical methods are able to compute individual objects anywhere in phase space, but parametric studies with them are more tedious, since they require the previous generation of a large database of trajectories obtained by numerical continuation of one or many families of trajectories. This does not mean that parametric studies are not feasible: an example of systematic continuation families of periodic orbits and invariant tori in order to cover large regions of phase space is given in Sect. 3.9.

The numerical method chosen for the computation of the 2D tori of the RTBP is based on looking for the Fourier series of a curve in the torus invariant by the time-T flow, being T one of the periods of the quasi-periodic trajectories inside the torus [6, 14]. It is a well-established method that has proven to be among the most adequate in this context (see [3] for a review of several methods). Since its use requires starting from the normal part of periodic orbits that need to be previously obtained by continuation, this lecture also includes a discussion on the numerical computation of fixed points and periodic orbits, and develops explicit formulation of the linear approximation of their normal dynamics. On the semi-analytical side, this lecture will cover a technique based on the parameterization method [20, 21], that produces Taylor expansions of the center manifold of a collinear point and the corresponding reduction of the vector field. In the reduced field, the collinear libration point is no longer unstable, so the trajectories in a neighborhood of it can be obtained by direct numerical integration. An earlier technique known as reduction to the center manifold [15, 25] produces the same results. There is another technique (not covered here), known as the Lindstedt-Poincaré method [25, 29], which is still more convenient for parametric studies because it produces expansions of the trajectories instead of the center manifold, at the expense of a slightly smaller neighborhood of validity of the expansions.

The lecture is structured as follows. Section 2 reviews some of the common nomenclature in dynamical systems and, in doing so, introduces the relevant features of the RTBP. After that, Sect. 3 describes numerical techniques for the computation of periodic orbits and invariant tori, whereas Sect. 4 explains how to compute the same objects semi-analytically using the parameterization method. Attention is then focused on the computation of the stable and unstable manifolds of periodic orbits and tori. Section 5 reviews numerical techniques to compute their linear approximation, whereas Sect. 6 explains how to obtain trajectories in these manifolds semi-analytically via the parameterization method. Finally, Sect. 7 addresses the computation and continuation of homoclinic and heteroclinic connections.

2 Dynamical Systems and the RTBP

This section recalls some notions from the theory of dynamical systems and also introduces the circular, spatial Restricted Three-Body Problem (RTBP). Although most readers will probably be familiar with these notions, recalling them will allow us to introduce notations that will be used in the rest of the lecture.

2.1 Continuous Dynamical Systems

The theory of dynamical systems provides an abstract framework for the mathematical study of systems that evolve with time in a deterministic manner. *Continuous dynamical systems* are those in which time is considered a continuous variable, this is, $t \in \mathbb{R}$. They are usually defined in terms of a system of autonomous (time-independent) Ordinary Differential Equations (ODE)

$$\begin{cases} \dot{x}_1 = X_1(x_1, x_2, \dots, x_n), \\ \dot{x}_2 = X_2(x_1, x_2, \dots, x_n), \\ \quad \vdots \\ \dot{x}_n = X_n(x_1, x_2, \dots, x_n), \end{cases}$$

or, in short,

$$\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x}), \quad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{X} : \mathbb{R}^n \to \mathbb{R}^n.$$

Assuming that this system of ODE can be integrated for all time, for $t \in \mathbb{R}$ the *time-t flow*, $\boldsymbol{\phi}_t : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, is defined by the initial value problem

$$\left.\begin{array}{r} \frac{d}{dt}\boldsymbol{\phi}_t(\boldsymbol{x}) = \boldsymbol{X}\big(\boldsymbol{\phi}_t(\boldsymbol{x})\big) \\ \boldsymbol{\phi}_0(\boldsymbol{x}) = \boldsymbol{x} \quad\quad\quad \end{array}\right\}.$$

It can be thought as a map that “flows” initial conditions along the corresponding trajectories for t time units. The subscript notation for t is in order to stress this fact. It is also common to refer to a continuous dynamical system as “a flow”.

Given an initial condition $\boldsymbol{x}_0$, the corresponding *orbit* is $\{\boldsymbol{\phi}_t(\boldsymbol{x}_0)\}_{t\in\mathbb{R}}$. A *fixed point* of a continuous dynamical system is a point whose orbit is itself, that is, $\boldsymbol{\phi}_t(\boldsymbol{x}) = \boldsymbol{x}$, $\forall t \in \mathbb{R}$. This can only happen if $\boldsymbol{f}(\boldsymbol{x}) = 0$. An orbit $\{\boldsymbol{\phi}_t(\boldsymbol{x})\}_{t\in\mathbb{R}}$ is said to be *periodic* if there is $T > 0$ such that

$$\begin{aligned} \boldsymbol{\phi}_T(\boldsymbol{x}) &= \boldsymbol{x}, \\ \boldsymbol{\phi}_t(\boldsymbol{x}) &\neq \boldsymbol{x}, \quad \text{for} \quad 0 < t < T. \end{aligned}$$

Then T is said to be its *period.* A set of initial conditions $A \subset \mathbb{R}^n$ is said to be an *invariant set* if

$$\boldsymbol{\phi}_t(\boldsymbol{x}) \in A \quad \forall t \in \mathbb{R}, \quad \forall \boldsymbol{x} \in A.$$

A straightforward example is an orbit (in particular, a fixed point or a periodic orbit).

A *manifold* is a set of points defined (maybe piece-wise) by equations, either implicit or parametric. An *invariant manifold* is an invariant set that is a manifold (e.g. a torus). We will usually speak of general invariant manifolds as *invariant objects* and reserve “invariant manifold” to denote stable, unstable or center manifolds associated to an invariant object. Given an invariant set A, its *stable set* (resp. *unstable set*) is the set $W^s(A)$ (resp. $W^u(A)$) of initial conditions that tend to the object asymptotically through the flow forward (resp. backward) in time. In other words, the set of initial conditions that approach (resp. depart from) A. This is,

$$W^s(A) = \{\boldsymbol{x} : \operatorname{dist}\big(\boldsymbol{\phi}_t(\boldsymbol{x}), A\big) \overset{t\to+\infty}{\longrightarrow} 0\},$$
$$W^u(A) = \{\boldsymbol{x} : \operatorname{dist}\big(\boldsymbol{\phi}_t(\boldsymbol{x}), A\big) \overset{t\to-\infty}{\longrightarrow} 0\}.$$

For several cases in which A is a manifold (e.g. a fixed point, a periodic orbit, an invariant torus), $W^s(A)$ (resp. $W^u(A)$) is also a manifold, and is called the *stable manifold* (resp. *unstable manifold*) of A.

2.2 The Circular, Spatial Restricted Three-Body Problem

The *circular, spatial restricted three-body problem* (RTBP) is an example of continuous dynamical system. It can be written as a Hamiltonian system with three degrees of freedom (details on the theory of Hamiltonian systems can be found in e.g. [31]) with Hamiltonian

$$H(x, y, z, p_x, p_y, p_z) = \frac{1}{2}(p_x^2 + p_y^2 + p_z^2) - xp_y + yp_x - \frac{1-\mu}{r_1} - \frac{\mu}{r_2},$$

with $r_1^2 = (x-\mu)^2 + y^2 + z^2$, $r_2^2 = (x-\mu+1)^2 + y^2 + z^2$. The system of ODE that defines it is, therefore,

$$\dot{x} = \partial H/\partial p_x = p_x + y, \quad \dot{p}_x = -\partial H/\partial x = p_y - \frac{1-\mu}{r_1^3}(x-\mu) - \frac{\mu}{r_2^3}(x-\mu+1),$$
$$\dot{y} = \partial H/\partial p_y = p_y - x, \quad \dot{p}_y = -\partial H/\partial y = -p_x - \Big(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\Big)y,$$
$$\dot{z} = \partial H/\partial p_z = p_z, \qquad \dot{p}_z = -\partial H/\partial z = -\Big(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\Big)z.$$

The RTBP describes the motion of a massless particle (“massless” in the sense that it is considered not to influence gravitationally any other body) under the gravitational attraction of two bodies, called primaries, with masses $m_1 > m_2$. The primaries are assumed to revolve uniformly in circles around their common center of mass. The coordinate system used is a *synodic* one, that rotates with the primaries so that the primary of mass m_1 is fixed at $(\mu, 0, 0, 0, \mu, 0)$, and the primary of mass m_2 is fixed at $(\mu - 1, 0, 0, 0, \mu - 1, 0)$. The RTBP depends on the *mass parameter* $\mu = m_2/(m_1+m_2)$. As it is common with Hamiltonian systems, the (x, y, z) coordinates are called *positions*, and the (p_x, p_y, p_z) coordinates are called *momenta*. The space of positions (this is, 3D physical space) is called *configuration space*. See Fig. 2.

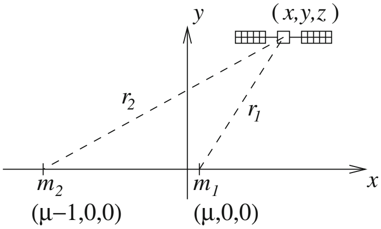


Fig. 2 Schematic description of the RTBP in configuration space

In short, the RTBP can be denoted as $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, with

$$\boldsymbol{x} = (x, y, z, p_x, p_y, p_z), \quad \boldsymbol{X}(\boldsymbol{x}) = (X_1(\boldsymbol{x}), X_2(\boldsymbol{x}), \dots, X_6(\boldsymbol{x})), \tag{1}$$

being

$$\begin{aligned} X_1(\boldsymbol{x}) &= p_x + y, & X_4(\boldsymbol{x}) &= p_y - \frac{1-\mu}{r_1^3}(x-\mu) - \frac{\mu}{r_2^3}(x-\mu+1), \\ X_2(\boldsymbol{x}) &= p_y - x, & X_5(\boldsymbol{x}) &= -p_x - \Big(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\Big)y, \\ X_3(\boldsymbol{x}) &= p_z, & X_6(\boldsymbol{x}) &= -\Big(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\Big)z. \end{aligned} \tag{2}$$

2.3 Discrete Dynamical Systems

Discrete dynamical systems are those in which time is considered as a discrete variable, this is, $t \in \mathbb{Z}$. They are defined by diffeomorphisms (smooth 1-1 maps)

$$\begin{aligned} \boldsymbol{F} : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\ \boldsymbol{x} &\longmapsto \boldsymbol{F}(\boldsymbol{x}). \end{aligned}$$

We denote by $\boldsymbol{F}^{-1}$ the inverse map of $\boldsymbol{F}$, and use superscript notation for the composition of maps:

$$\begin{aligned} &\boldsymbol{F}^0(\boldsymbol{x}) = \boldsymbol{x}, \\ &\boldsymbol{F}^1(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}), \\ &\boldsymbol{F}^2(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{F}(\boldsymbol{x})), && \boldsymbol{F}^{-2}(\boldsymbol{x}) = \boldsymbol{F}^{-1}(\boldsymbol{F}^{-1}(\boldsymbol{x})), \\ &\boldsymbol{F}^3(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{F}(\boldsymbol{F}(x))), && \boldsymbol{F}^{-3}(\boldsymbol{x}) = \boldsymbol{F}^{-1}(\boldsymbol{F}^{-1}(\boldsymbol{F}^{-1}(\boldsymbol{x}))). \\ &\quad\vdots && \quad\vdots \end{aligned}$$

In this way, $\boldsymbol{F}^n$ is “the discrete time-n flow”. Via this notion, all the previous notions from continuous dynamical systems translate to the discrete case. Given an initial condition, its related *orbit* is the set $\{\ \boldsymbol{F}^i(\boldsymbol{x})\ \}_{i\in\mathbb{Z}}$, that is,

$$\{\dots, \boldsymbol{F}^{-3}(\boldsymbol{x}), \boldsymbol{F}^{-2}(\boldsymbol{x}), \boldsymbol{F}^{-1}(\boldsymbol{x}), \boldsymbol{F}^{0}(\boldsymbol{x}), \boldsymbol{F}^{1}(\boldsymbol{x}), \boldsymbol{F}^{2}(\boldsymbol{x}), \boldsymbol{F}^{3}(\boldsymbol{x}),\ \dots\}.$$

A *fixed point* is an initial condition such that its orbit is itself, $\boldsymbol{F}(\boldsymbol{x}_0) = \boldsymbol{x}_0$. An *n-periodic point* is an initial condition $\boldsymbol{x}_0$ such that $\boldsymbol{F}^n(\boldsymbol{x}_0) = \boldsymbol{x}_0$, $\boldsymbol{F}^i(\boldsymbol{x}_0) \neq \boldsymbol{x}_0$, $\forall i = 1, \dots, n-1$. A set of initial conditions $A \subset \mathbb{R}^n$ is said to be an *invariant set* if $\boldsymbol{F}^n(\boldsymbol{x}) \in A\ \forall n \in \mathbb{Z}\ \forall \boldsymbol{x} \in A$. If A is a manifold, the stable and unstable sets of A, defined by

$$W^s(A) = \{\boldsymbol{x} : \mathrm{dist}(\boldsymbol{F}^n(\boldsymbol{x}), A) \overset{n\to+\infty}{\longrightarrow} 0\},$$

$$W^u(A) = \{\boldsymbol{x} : \mathrm{dist}(\boldsymbol{F}^n(\boldsymbol{x}), A) \overset{n\to-\infty}{\longrightarrow} 0\},$$

are usually manifolds.

An paradigmatic example of a discrete dynamical system is Chirikov's *standard map*, that, in one of its formulations is written as

$$\boldsymbol{F} : \begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \begin{pmatrix} x + a\sin(x+y) \\ x + y \end{pmatrix}. \tag{3}$$

Here a is a parameter and $x, y \in \mathbb{T} = \mathbb{R}/[0, 2\pi]$, that is, we assume that (x, y) and $(\bar{x}, \bar{y})$ are the same point if $x - \bar{x} = j2\pi$, $y - \bar{y} = k2\pi$, for $j, k \in \mathbb{Z}$. The standard map is *area-preserving*. In two dimensions, being area preserving is equivalent to being *symplectic*, which is the discrete analog of the Hamiltonian formalism (for more details see any textbook in the subject, e.g. [31]). The global dynamics (a *phase portrait*) of two-dimensional area-preserving maps on compact regions can be swiftly obtained by iteration of the map $\boldsymbol{F}$. Figure 3 is obtained by considering the initial conditions $\{\boldsymbol{p}_j := (-\pi + j2\pi/100, 0)\}_{j=0}^{100}$ and plotting the points $\{\boldsymbol{F}^k(\boldsymbol{p}_j))\}_{k=0}^{1000}$, for $j = 0, \dots, 100$. Several kinds of invariant sets (that are

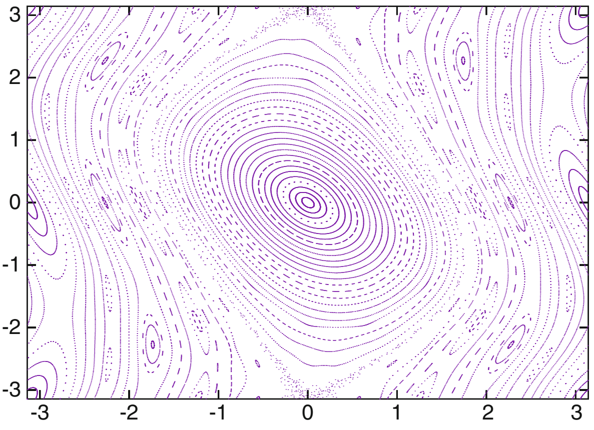


Fig. 3 Phase portrait of the standard map (3) for $a = -0.7$

manifolds) can be found, like fixed points, periodic points of different periods and invariant curves. Invariant sets with chaotic dynamics can also be observed.

2.4 Orbit Generation in a Dynamical System

Orbits in discrete dynamical systems can be generated just by iteration of the map, as it has been done in Fig. 3. In continuous dynamical systems, numerical methods for integration of ODE have to be used. In order to have error control, variable-step methods are preferred over constant-step ones. A popular family of variable-step methods are the Runge-Kutta-Fehlberg (RKF) ones, of which there are some high-order versions like RKF78 [10]. There are many alternatives (see e.g. [19, 41]). In the case of a system of ODE given by closed formulae, like the RTBP, a particularly good choice is Taylor's method, of which there are freely available implementations [1, 27]. Here we will discuss briefly the black-box usage of a one-step method with step size control.

For a system of n possibly non-autonomous ODE,

$$\dot{\boldsymbol{x}} = \boldsymbol{X}(t, \boldsymbol{x}),$$

with $\boldsymbol{x}, \boldsymbol{X}(t, \boldsymbol{x}) \in \mathbb{R}^n$, denote by $\boldsymbol{\phi}(t, t_0, \boldsymbol{x})$ its flow from time t_0 to time t, defined by the conditions

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\phi}(t, t_0, \boldsymbol{x}) &= \boldsymbol{X}\big(t, \boldsymbol{\phi}(t, t_0, \boldsymbol{x})\big), \\ \boldsymbol{\phi}(t_0, t_0, \boldsymbol{x}) &= \boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \end{aligned} \tag{4}$$

Given $t_0 \in \mathbb{R}$, $\boldsymbol{x}_0 \in \mathbb{R}^n$, $h_0 \in \mathbb{R}$ (small), and a tolerance δ, a routine implementing a one-step method with step size control will return $t_1, \boldsymbol{x}_1, h_1$ verifying

(a) $|\boldsymbol{x}_1 - \boldsymbol{\phi}(t_1, t_0, \boldsymbol{x}_0)| < \delta$,
(b) t_1 is as close to $t_0 + h_0$ as possible,
(c) h_1 is a recommended step length for the next call.

In the algorithmic descriptions that will follow, we will denote a call to such a routine as

$$(t_1, \boldsymbol{x}_1, h_1) = \text{ODEstep}(t_0, \boldsymbol{x}_0, h_0, \boldsymbol{X}, \delta).$$

In order to implement $\boldsymbol{\phi}(t_1, t_0, \boldsymbol{x}_0)$ for arbitrary $t_1, t_0, \boldsymbol{x}_0$, it is necessary to write a routine that calls ODEstep many times using as input step h_0 the recommended step h_1 of the previous call, plus a final call with h_0 the step needed to in order to reach the final time t_1 (or more than one such calls, if the step given is reduced by the step size control). In some of the algorithmic descriptions that will follow, a call to such a routine will be denoted as

$$(t_1, \boldsymbol{x}_1, h_1) = \text{ODEflow}(t_0, t_1, \boldsymbol{x}_0, h_0, \boldsymbol{X}, \delta).$$

In the systems of non-linear equations that we will solve in order to compute invariant objects, we will need to be able to numerically evaluate both the flow and its differential with respect to initial conditions, that we will denote as $D_{\boldsymbol{x}}\boldsymbol{\phi}(t, t_0, \boldsymbol{x})$, or simply $D\boldsymbol{\phi}_t(\boldsymbol{x})$ in the autonomous case. It can be found by numerically integrating the system of ODE together with its *first variational equations*:

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{X}(t, \boldsymbol{x}), \\ \dot{A} &= \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{x}}(t, \boldsymbol{x})A, \end{aligned} \tag{5}$$

where $\boldsymbol{x}$ is an n-dimensional vector and A is a $n \times n$ matrix. If $\boldsymbol{x}(t)$ and $A(t)$ are solutions of (5) with $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and $A(t_0) = I_n$ the $n \times n$ identity matrix, then $D_{\boldsymbol{x}}\boldsymbol{\phi}(t, t_0, \boldsymbol{x}_0) = A(t)$. System (5) can be written as a system of $n + n^2$ ODE as

$$\begin{aligned} \dot{x}_i &= X_i(t, x_1, \dots, x_n), & i &= 1, \dots, n, \\ \dot{a}_{i,j} &= \sum_{k=1}^{n} \Big(\frac{\partial X_i}{\partial x_k}(t, x_1, \dots, x_n)\Big) a_{k,j}, & i, j &= 1, \dots, n. \end{aligned}$$

2.5 Poincaré Maps

A way to simplify the study of a continuous dynamical systems is to consider a discrete dynamical system that has essentially the same dynamics. One way to do it is, for a fixed $T > 0$, to consider the time-T flow (or *stroboscopic map*), $\boldsymbol{\phi}_T$, which is a discrete dynamical system. In this way, for instance, T-periodic orbits are turned into fixed points. Another way to turn a continuous dynamical system into a discrete one is through a Poincaré map.

For a continuous dynamical system given by $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, let Σ be a hypersurface of $\mathbb{R}^n$, and assume it is transversal to the vector field, that is, $\boldsymbol{X}(\boldsymbol{x})$ is not tangent to Σ for all $\boldsymbol{x} \in \Sigma$. Let $\boldsymbol{x}_0$ be such that $\boldsymbol{\phi}_{T_0}(\boldsymbol{x}_0) \in \Sigma$ for some $T_0 > 0$. The implicit function theorem ensures the existence of a neighborhood $U \ni \boldsymbol{x}_0$ and a map $\tau : U \to \mathbb{R}$, known as *time-return map*, such that $\tau(\boldsymbol{x}_0) = T_0$ and

$$\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x}) \in \Sigma \qquad \forall \boldsymbol{x} \in U.$$

The map $\boldsymbol{P}(\boldsymbol{x}) := \boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x})$ is called *Poincaré map* corresponding to Σ. If $\boldsymbol{x}_0 \in \Sigma$ and $\boldsymbol{P}(\Sigma \cap U) = \Sigma \cap U$, the restriction of $\boldsymbol{P}$ to $\Sigma \cap U$ defines a discrete dynamical system. In going to the starting continuous dynamical system to the discrete one defined by $\boldsymbol{P}$, periodic orbits are turned into fixed points, and invariant tori are turned into invariant curves. In general, invariant objects lose one dimension, which is an advantage both from the theoretical and the computational point of view.

Orbit generation in this discrete dynamical system requires the numerical evaluation of a Poincaré map, which has as a difficulty that the time-return map $\tau(\boldsymbol{x})$ is unknown. It can be adjusted by Newton iterations once we get close to the section Σ. This is done in Algorithm 1.

Algorithm 1 *Evaluation of the Poincaré map* $\boldsymbol{P}$ *associated to a section* $\Sigma = \{g(\boldsymbol{x}) = 0\}$ *for* $g : \mathbb{R}^n \to \mathbb{R}$*, which is assumed to be traversed from* $g(\boldsymbol{x}) < 0$ *to* $g(\boldsymbol{x}) > 0$*. At the end of the algorithm,* $\boldsymbol{y} = \boldsymbol{P}(\boldsymbol{x})$ *and* $t = \tau(\boldsymbol{x})$*.*

```
input:    $\boldsymbol{x}$, $g$, $\boldsymbol{X}$, tol, tol$_2$, $h_0$
 do:      $t := 0$,   $\boldsymbol{y} := \boldsymbol{x}$,   $h := h_0$
          while $(g(\boldsymbol{y}) \geq -\text{tol})$
              $(t, \boldsymbol{y}, h) := \text{ODEstep}(t, \boldsymbol{y}, h, \boldsymbol{X}, \text{tol}_2)$
          while $(g(\boldsymbol{y}) < 0)$
              $(t, \boldsymbol{y}, h) := \text{ODEstep}(t, \boldsymbol{y}, h, \boldsymbol{X}, \text{tol}_2)$
          while $(|g(\boldsymbol{y})| > \text{tol})$
              $\delta := -\dfrac{g(\boldsymbol{y})}{Dg(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{y})}$
              $(t, \boldsymbol{y}, h) := \text{ODEflow}(t, t + \delta, \boldsymbol{y}, h, \boldsymbol{X}, \text{tol}_2)$
output:   $t$, $\boldsymbol{y}$.
```

If the differential of the Poincaré map, $D\boldsymbol{P}(\boldsymbol{x})$, is also needed, it can be computed as

$$D\boldsymbol{P}(x) = \boldsymbol{X}\big(\boldsymbol{P}(\boldsymbol{x})\big)D\tau(\boldsymbol{x}) + D\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x}),$$

where $D\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x})$ is to be understood as $D_{\boldsymbol{y}}\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{y})|_{\boldsymbol{y}=\boldsymbol{x}}$. An expression for $D\tau(\boldsymbol{x})$ can be obtained by implicit differentiation on $g(\boldsymbol{P}(\boldsymbol{x})) \equiv 0$. After substitution in the previous equation,

$$D\boldsymbol{P}(x) = -\boldsymbol{X}\big(\boldsymbol{P}(x)\big)\frac{Dg\big(\boldsymbol{P}(x)\big)D\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x})}{Dg\big(\boldsymbol{P}(x)\big)\boldsymbol{X}\big(\boldsymbol{P}(x)\big)} + D\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x}). \tag{6}$$

In a routine implementing Algorithm 1 for the evaluation of a Poincaré map, it is useful to keep as an option the integration of the system of ODE defining our continuous dynamical system together with its first variational equations (5), in order to have available $D\boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x})$ to be used in (6).

3 Numerical Computation of Periodic Orbits and 2D Tori

The goal of this section is to review some numerical methods for the numerical computation of the periodic orbits and invariant 2D tori related to collinear libration points. Since the RTBP is Hamiltonian, both periodic orbits and tori are not isolated but embedded in families. Once an invariant object (periodic orbit or torus) has been computed, the remaining objects of its family can be obtained by continuation. The first object of a family is usually computed from the linear approximation of the dynamics around a simpler object (e.g. a torus from the linear dynamics around a periodic orbit, or a periodic orbit from the linear dynamics around a fixed point). This approach can be followed hierarchically in order to do a systematic study of the dynamics around a collinear point.

This section starts recalling the predictor-corrector or pseudo-arclength continuation method as described by standard references (e.g. [2]). After that, Sect. 3.2 provides an strategy for the numerical solution of not necessarily square non-linear systems of equations, that simplifies considerably the practical implementation of the methods described later. The subsections that follow (from Sects. 3.3 to 3.8) provide methods for the computation of invariant objects and formulation for the linear dynamics around them, necessary to implement a systematic numerical exploration of the dynamics around a collinear libration point. This is actually done in Sect. 3.9 for the L_1 collinear point of the Earth-Moon RTBP.

3.1 Numerical Continuation

A classical way to introduce numerical continuation is as a technique to find a (unknown) solution of a system of non-linear equations $\boldsymbol{G}(\boldsymbol{x}) = 0$ from a known solution of another system $\boldsymbol{F}(\boldsymbol{x}) = 0$, that is close to $\boldsymbol{G}(\boldsymbol{x}) = 0$ in some sense. In order to look for a zero of $\boldsymbol{G}$, a one-parametric family of intermediate systems $\boldsymbol{H}(\lambda, \boldsymbol{x})$ is considered with $\boldsymbol{H}(0, \boldsymbol{x}) = \boldsymbol{F}(x)$ and $\boldsymbol{H}(1, \boldsymbol{x}) = \boldsymbol{G}(x)$. For instance, the convex homotopy between $\boldsymbol{F}$ and $\boldsymbol{G}$,

$$\boldsymbol{H}(\lambda, \boldsymbol{x}) = (1 - \lambda)\boldsymbol{F}(\boldsymbol{x}) + \lambda\boldsymbol{G}(\boldsymbol{x}),$$

Then we can try to continue the known solution $\boldsymbol{x}_0$ of $\boldsymbol{H}(0, \boldsymbol{x}) = 0$ up to a solution of $\boldsymbol{H}(1, \boldsymbol{x}) = 0$ with respect to the parameter λ. The algorithm below provides a straightforward approach.

Algorithm 2 *Continuation of* $\boldsymbol{H}(\lambda, \boldsymbol{x}) = 0$ *with respect to the parameter* λ.

```
input:   $\boldsymbol{x}_0 \in \mathbb{R}^n$ such that $\boldsymbol{H}(0, \boldsymbol{x}_0) = 0$, $m \in \mathbb{N}$
do:      $\boldsymbol{x} := \boldsymbol{x}_0$
         $\Delta\lambda := 1/m$
         $\forall i = 1, \ldots, m$
             $\lambda := i\Delta\lambda$
             solve $\boldsymbol{H}(\lambda, \boldsymbol{y}) = 0$ iteratively for $\boldsymbol{y}$ taking $\boldsymbol{x}$ as
                       starting value
             $\boldsymbol{x} := \boldsymbol{y}$
output:  $\boldsymbol{x}$
```

Algorithm 2 breaks down if there is a turning point with respect to λ along the continuation curve. An alternative that can cope with this case is the *predictor-corrector* or *pseudo-arclength* continuation method (see e.g. [2]). Its basic idea is to consider arclength instead of λ as the continuation parameter. "Pseudo" stands for the fact that the actual parameter is not truly arclength but distance along a line tangent to the continued curve. Define $\boldsymbol{y} = (\lambda, \boldsymbol{x}) \in \mathbb{R}^{n+1}$. Then $\boldsymbol{H}(\boldsymbol{y}) := \boldsymbol{H}(\lambda, \boldsymbol{x}) = 0$ defines implicitly a curve in $\mathbb{R}^{n+1}$ as long as $\operatorname{rank} D\boldsymbol{H}(\boldsymbol{y}) = n$, which is a condition we will assume. The continuation can be done according to the algorithm stated next.

Algorithm 3 *Predictor-corrector or pseudo-arclength continuation of* $\boldsymbol{H}(\boldsymbol{y}) = 0$, *for* $\boldsymbol{y} = (\lambda, \boldsymbol{x})$, *from* $\lambda = 0$ *to* $\lambda = 1$.

```
input:    $\boldsymbol{y} = (\lambda, \boldsymbol{x}) \in \mathbb{R}^{n+1}$ such that $\Pi_1 \boldsymbol{y} := \lambda = 0$, $\boldsymbol{H}(\boldsymbol{y}) = 0$.
do:       while $(\Pi_1 \boldsymbol{y} < 1)$
              let $\boldsymbol{v} \in \ker D\boldsymbol{H}(\boldsymbol{y})$, $\|\boldsymbol{v}\|_2 = 1$, pointing ahead
              take $\boldsymbol{z} := \boldsymbol{y} + \gamma \boldsymbol{v}$, for suitable $\gamma$ (see the comments below)
              if $(\Pi_1 \boldsymbol{z} < 1)$
                  solve $\boldsymbol{H}(\boldsymbol{z}) = 0$ iteratively for $\boldsymbol{z}$ by a modified Newton's
                          method taking minimum-norm corrections
              else
                  $\gamma := (1 - \Pi_1 \boldsymbol{y}) / \Pi_1 \boldsymbol{v}$
                  $\boldsymbol{z} := \boldsymbol{y} + \gamma \boldsymbol{v}$
                  solve $\boldsymbol{H}(\boldsymbol{z}) = 0$ by Newton iterations keeping $\Pi_1 \boldsymbol{z}$ constant
              $\boldsymbol{y} := \boldsymbol{z}$
output:   $\boldsymbol{y}$
```

A convenient way to control the step length γ is in order to keep constant the number of Newton iterations when solving $\boldsymbol{H}(\boldsymbol{z}) = 0$ at each continuation step. A simple rule is to assume that this number of iterations is a linear function of the step length chosen: if n_{old} is the number of iterations performed in the last continuation step, γ_{old} is the last step length used and n_{des} is the desired number of Newton iterations, we can take

$$\gamma = \frac{n_{des}}{n_{old}} \gamma_{old}. \tag{7}$$

Note that, except for the start and stop criteria, in the pseudo-arclength method there is no distinguished coordinate to be considered a parameter. It can therefore be applied to any system of non-linear equations $\boldsymbol{H}(\boldsymbol{y}) = 0$, as long as its solution is a curve.

3.2 Numerical Solution of Non-square, Non-linear Systems of Equations

In Sects. 3.6 and 3.8, the computation of periodic orbits and invariant 2D tori will be done in terms of solving non-linear systems of equations. In the case of the computation of a single object, the system to be solved will have (locally) unique solution. It is standard practice in this case to require such a system to be square, this is, of the form $\boldsymbol{G}(\boldsymbol{y}) = 0$ with $\boldsymbol{G} : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ for some N, and to use Newton's method (see any textbook on numerical analysis, e.g. [41]). In the case of the continuation of a family, the system to be solved will not have unique solution but a curve of solutions. It is standard practice in this case to require such a system to have one more unknown than equations, this is, to be of the form $\boldsymbol{G}(\boldsymbol{y}) = 0$ with $\boldsymbol{G} : \mathbb{R}^{N+1} \to \mathbb{R}^N$ for some N, and to use Newton's method with some modification to account for non-uniqueness (see e.g. [2, 39]).

In order to keep the systems of equations of Sects. 3.6 and 3.8 simple, it will be convenient not to require them to be either $N \times N$ or $(N+1) \times N$. A way to be able to solve these systems is to consider a modified Newton method $\boldsymbol{y}_{n+1} = \boldsymbol{y}_n - (\Delta \boldsymbol{y})_n$ in which the linear system to be solved for the correction, $D\boldsymbol{G}(\boldsymbol{y}_n)(\Delta \boldsymbol{y})_n = \boldsymbol{G}(\boldsymbol{y}_n)$, is solved for its minimum-norm, least-squares solution. The minimum-norm, least-squares solution always exists and is unique for any linear system of equations, square or not. Assuming that the starting non-linear system $\boldsymbol{G}(\boldsymbol{y}) = 0$ has a solution (perhaps non-unique) and that the initial guess $\boldsymbol{y}_0$ is close to a solution, this strategy will converge to a nearby solution using minimum-norm corrections.

We discuss briefly how to compute the minimum-norm, least-squares solution of a linear system $A\boldsymbol{y} = \boldsymbol{b}$ using QR decomposition with column pivoting.[1] Assume that A is an arbitrary $m \times n$ matrix with $r := \operatorname{rank} A \le \min(m, n)$. A least squares solution of $A\boldsymbol{y} = \boldsymbol{b}$,

$$\boldsymbol{y}^* \in \mathbb{R}^n : \quad \|\boldsymbol{b} - A\boldsymbol{y}^*\|_2 = \min_{\boldsymbol{y} \in \mathbb{R}^n} \|\boldsymbol{b} - A\boldsymbol{y}\|_2,$$

always exists. If $r = n$, there is an unique least-squares solution. If $r < n$, there is a linear subspace of least-squares solutions of dimension $d := n - r$. Nevertheless, as mentioned previously,

$$\boldsymbol{y}_{LS} : \quad \|\boldsymbol{y}_{LS}\|_2 = \min\Big\{ \|\boldsymbol{y}^*\|_2 : \ \|\boldsymbol{b} - A\boldsymbol{y}^*\|_2 = \min_{\boldsymbol{y} \in \mathbb{R}^n} \|\boldsymbol{b} - A\boldsymbol{y}\|_2 \Big\},$$

[1] Singular value decomposition (see e.g. [12]) is an alternative that provides more information but is also computationally more costly.

is always unique. By applying to A Householder transformations with column pivoting [12], we obtain a decomposition

$$Q^\top AP = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

with Q an $m \times m$ orthogonal matrix, R_{11} an $r \times r$ upper-triangular matrix with non-zero diagonal elements, and P a $n \times n$ permutation matrix. In order to perform this decomposition, r (or, equivalently, $d = n - r$) must be known. If we denote

$$P^\top \boldsymbol{y} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{s} \end{pmatrix}, \qquad Q^\top \boldsymbol{b} = \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix},$$

with $\boldsymbol{z}, \boldsymbol{c} \in \mathbb{R}^r$, $\boldsymbol{s} \in \mathbb{R}^d$, $\boldsymbol{d} \in \boldsymbol{R}^{m-r}$, then the least-squares solutions of $A\boldsymbol{y} = \boldsymbol{b}$ are

$$P^\top \boldsymbol{y} = \left\{ \begin{pmatrix} R_{11}^{-1}\boldsymbol{c} \\ 0 \end{pmatrix} + \begin{pmatrix} -R_{11}^{-1}R_{12} \\ I_d \end{pmatrix} \boldsymbol{s} \right\}_{\boldsymbol{s}\in\mathbb{R}^d},$$

where I_d is the $d \times d$ identity matrix. Finding the minimum-norm element of the previous set is an standard full-rank least-squares problem, that can be solved via a standard (without column pivoting) QR decomposition.

In order to solve the systems of equations of Sects. 3.6 and 3.8, it is convenient to write a routine that, for a general $m \times n$ linear system of equations, finds the minimum-norm least-squares solution and, optionally, a basis of the kernel of A, which can be obtained from

$$\ker A = \left\{ P \begin{pmatrix} -R_{11}^{-1}R_{12} \\ I_d \end{pmatrix} \boldsymbol{s} \right\}_{\boldsymbol{s}\in\mathbb{R}^d}.$$

3.3 Computation of Fixed Points of Flows and Maps

For the computation of a fixed point of a flow $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, we look for $\boldsymbol{p}$ such that $\boldsymbol{G}(\boldsymbol{p}) := \boldsymbol{X}(\boldsymbol{p}) = 0$. For the computation of a fixed point of a map $\boldsymbol{x} \mapsto \boldsymbol{F}(\boldsymbol{x})$, we look for $\boldsymbol{p}$ such that $\boldsymbol{G}(\boldsymbol{p}) := \boldsymbol{F}(\boldsymbol{p}) - \boldsymbol{p} = 0$. In any case, we can use Newton's method in several variables in order to look for a zero of $\boldsymbol{G}$.

Algorithm 4 *Newton's method in order to find a zero of a function* $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^n$ *with tolerance* tol*, starting from a first guess* $\boldsymbol{p}_0$*, allowing for a maximum of* maxit *iterations.*

```
input:   $\boldsymbol{p}_0$, $\boldsymbol{G}$, tol, maxit
do:      $\boldsymbol{p} := \boldsymbol{p}_0$
         for it from 1 to maxit do
```

```
            if ($|\boldsymbol{G}(\boldsymbol{p})| <$ tol) return $\boldsymbol{p}$
            solve $D\boldsymbol{G}(\boldsymbol{p})\Delta\boldsymbol{p} = \boldsymbol{G}(\boldsymbol{p})$ for $\Delta\boldsymbol{p}$
            $\boldsymbol{p} := \boldsymbol{p} - \Delta\boldsymbol{p}$
        error (maxit exceeded)
output: $\boldsymbol{p}$ (if OK)
```

In the RTBP, it can be analytically seen (see e.g. [42]) that the distance from L_j, $j = 1, 2, 3$ to the closest primary, that will be denoted γ_j, is given by the only positive root of the corresponding Euler's quintic equation:

$$\begin{aligned} \gamma_j^5 \mp (3-\mu)\gamma_j^4 + (3-2\mu)\gamma_j^3 - \mu\gamma_j^2 \pm 2\mu\gamma_j - \mu = 0, &\quad j = 1, 2, \\ \gamma_j^5 + (2+\mu)\gamma_j^4 + (1+2\mu)\gamma_j^3 - (1-\mu)\gamma_j^2 - 2(1-\mu)\gamma_j - (1-\mu) = 0, &\quad j = 3. \end{aligned}$$

Therefore, in this case it is enough to use Newton's method in one dimension. Good guesses are $(\mu/3)^{1/3}$ for $L_{1,2}$ and $1 - (7/12)\mu$ for L_3.

3.4 Linear Behavior Around Fixed Points of Flows

For a flow $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$ with a fixed point $\boldsymbol{p}$, since $\boldsymbol{X}(\boldsymbol{x}) = \boldsymbol{X}(\boldsymbol{p}) + D\boldsymbol{X}(\boldsymbol{p})(\boldsymbol{x}-\boldsymbol{p}) + O(\|\boldsymbol{x}-\boldsymbol{p}\|^2)$ and $\boldsymbol{X}(\boldsymbol{p}) = 0$, its linearization around $\boldsymbol{p}$ is

$$\dot{\boldsymbol{x}} = A(\boldsymbol{x}-\boldsymbol{p}), \tag{8}$$

with $A := D\boldsymbol{X}(\boldsymbol{p})$. The eigenvalues of A are known as the *exponents* of the fixed point $\boldsymbol{p}$.

Assume that $\lambda \in \mathbb{R}$, $\lambda \neq 0$ is an eigenvalue of A, and $\boldsymbol{v}$ is a corresponding eigenvector. Then

$$\boldsymbol{\varphi}(t) := \boldsymbol{p} + e^{\lambda t}\boldsymbol{v}$$

is a solution of the linearized flow (8). If $\lambda < 0$, $\boldsymbol{\varphi}(t) \to \boldsymbol{p}$ as $t \to +\infty$, so $\{\boldsymbol{\varphi}(t)\}_{t\in\mathbb{R}}$ is a trajectory in the stable manifold of $\boldsymbol{p}$ in the linearized flow. If $\lambda > 0$, $\boldsymbol{\varphi}(t) \to \boldsymbol{p}$ as $t \to -\infty$, so $\{\boldsymbol{\varphi}(t)\}_{t\in\mathbb{R}}$ is a trajectory in the unstable manifold of $\boldsymbol{p}$ in the linearized flow. The *stable manifold theorem* for flows (see e.g. [18, 35]) ensures the existence of a stable (resp. unstable) manifold of the full non-linear flow $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$ that contains $\boldsymbol{p}$ and is tangent to the linear subspace spanned by the eigenvectors of A corresponding to eigenvalues with strictly negative (resp. positive) real part.

Assume now $\lambda = \boldsymbol{i}\omega$ for $\omega \in \mathbb{R}$, $\omega \neq 0$, where $\boldsymbol{i}$ denotes the imaginary unit. In this case, $-\boldsymbol{i}\omega$ is also an eigenvalue, so we can assume that $\omega > 0$. Let $\boldsymbol{v}_1 + \boldsymbol{i}\boldsymbol{v}_2$ be a corresponding eigenvector, with $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$. Define

$$\boldsymbol{\varphi}_\gamma(t) := \boldsymbol{p} + \gamma\big((\cos\omega t)\boldsymbol{v}_1 - (\sin\omega t)\boldsymbol{v}_2\big). \tag{9}$$

By using $A\boldsymbol{v}_1 = -\omega\boldsymbol{v}_2$ and $A\boldsymbol{v}_2 = \omega\boldsymbol{v}_1$ (that follows from $A(\boldsymbol{v}_1 + \boldsymbol{i}\boldsymbol{v}_2) = \boldsymbol{i}\omega(\boldsymbol{v}_1 + \boldsymbol{i}\boldsymbol{v}_2)$), it is seen that $\boldsymbol{\varphi}_\gamma(t)$ is a solution of the linearized flow. Therefore, by varying γ, $\boldsymbol{\varphi}_\gamma(t)$ provides a one-parametric family of periodic orbits of period $2\pi/\omega$ of the linearized flow. If the remaining eigenvalues λ_j of A satisfy that $\lambda_j/(\boldsymbol{i}\omega)$ is not an integer, then *Lyapunov's center theorem* (see e.g. [31, 38]) ensures the existence of a one-parametric family of periodic orbits of the non-linear flow with periods that tend to $2\pi/\omega$ as the periodic orbits collapse to $\boldsymbol{p}$. These periodic orbits are part of the *center manifold* of $\boldsymbol{p}$, which is an invariant manifold tangent to the linear subspace spanned by eigenvectors corresponding to eigenvalues of A with zero real part. The existence of the this manifold is ensured by the *center manifold theorem for flows* (see e.g. [18, 35]).

Expressions for trajectories of the linear flow in the case $\lambda = a + \boldsymbol{i}\omega$ $a, \omega \in \mathbb{R}$, $a, \omega \neq 0$, can be obtained similarly. They will not be necessary in what follows. These trajectories would be close to trajectories of the non-linear flow in the stable or unstable manifold, according to whether $a < 0$ or $a > 0$, respectively.

3.5 Linear Behaviour Around Fixed Points of Maps

For a discrete dynamical system given by $\boldsymbol{x} \mapsto \boldsymbol{F}(\boldsymbol{x})$ with a fixed point $\boldsymbol{p}$, since $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{p} + \boldsymbol{DF}(\boldsymbol{p})(\boldsymbol{x} - \boldsymbol{p}) + O(\|\boldsymbol{x} - \boldsymbol{p}\|^2)$ and $\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{p}$, its linearization around $\boldsymbol{p}$ is

$$\boldsymbol{x} \mapsto L_{\boldsymbol{F}}(\boldsymbol{x}) := \boldsymbol{p} + A(\boldsymbol{x} - \boldsymbol{p}), \tag{10}$$

with $A = \boldsymbol{DF}(\boldsymbol{p})$. The eigenvalues of A are also called the *multipliers* of $\boldsymbol{p}$.

Assume that $\lambda \in \mathbb{R}$, $\lambda \neq 0$ is an eigenvalue of A, and $\boldsymbol{v}$ is a corresponding eigenvector. Define

$$\boldsymbol{\varphi}(\xi) := \boldsymbol{p} + \xi\boldsymbol{v}.$$

Since $L_{\boldsymbol{F}}\big(\boldsymbol{\varphi}(\xi)\big) = \boldsymbol{\varphi}(\lambda\xi)$, $\{\boldsymbol{\varphi}(\xi)\}_{\xi\in\mathbb{R}}$ is an invariant set of the linearized map (10). If $|\lambda| < 1$, $L^n_{\boldsymbol{F}}\big(\boldsymbol{\varphi}(\xi)\big) \to \boldsymbol{\varphi}(0) = \boldsymbol{p}$ as $n \to +\infty$, so $\{\boldsymbol{\varphi}(\xi)\}_{\xi\in\mathbb{R}}$ is a trajectory in the stable manifold of $\boldsymbol{p}$ in the linearized map. If $|\lambda| > 1$, $L^n_{\boldsymbol{F}}\big(\boldsymbol{\varphi}(\xi)\big) \to \boldsymbol{\varphi}(0) = \boldsymbol{p}$ as $n \to -\infty$, so $\{\boldsymbol{\varphi}(\xi)\}_{\xi\in\mathbb{R}}$ is a trajectory in the unstable manifold of $\boldsymbol{p}$ in the linearized map. The *stable manifold theorem for maps* (see e.g. [18, 35]) ensures the existence of a stable (resp. unstable) manifold of the full non-linear map $\boldsymbol{F}$ that contains $\boldsymbol{p}$ and is tangent to the linear subspace spanned by the eigenvectors of A corresponding to eigenvalues with modulus strictly smaller (resp. larger) than one.

Assume now $\lambda \in \mathbb{C}$, $|\lambda| = 1$, $\lambda = \cos\rho + \boldsymbol{i}\sin\rho$ and let $\boldsymbol{v}_1 + \boldsymbol{i}\boldsymbol{v}_2$ be an associated eigenvector, with $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$. Then

$$\boldsymbol{\varphi}_\gamma(\xi) := \boldsymbol{p} + \gamma\big((\cos\xi)\boldsymbol{v}_1 - (\sin\xi)\boldsymbol{v}_2\big) \tag{11}$$

satisfies $L_{\boldsymbol{F}}\big(\boldsymbol{\varphi}(\xi)\big) = \boldsymbol{\varphi}(\xi + \rho)$, so $\{\boldsymbol{\varphi}_\gamma(\theta)\}_{\theta\in[0,2\pi]}$ is an invariant closed curve of the linearized map. Therefore, by varying γ, $\boldsymbol{\varphi}_\gamma(\theta)$ provides a one-parametric family of invariant curves of the linearized map with *rotation number* ρ. Under number-theoretical hypotheses of ρ and non-degeneracy ones of $\boldsymbol{F}$, KAM theory (see e.g. [26]) ensures the existence of a Cantorian[2] one-parametric family of invariant curves of the full non-linear map $\boldsymbol{F}$, with rotation numbers that tend to ρ as the invariant curves collapse to $\boldsymbol{p}$.

3.6 Computation of Periodic Orbits of Flows

The computation of periodic orbits is a classical and well-known subject. There are publicly available software packages, like AUTO-07p [9], that are capable of both computing individual periodic orbits and performing continuation of families. Nevertheless, the simplicity of the methodology that will follow makes its implementation worthwhile, both for computational efficiency and for easier interaction with the methods of computation of invariant tori of Sect. 3.8. We discuss in this section how to compute initial conditions for periodic orbits by solving non-linear systems of equations stated in terms of the flow. The discussion will partially follow [39].

An initial condition for a periodic orbit of a flow can be thought as a fixed point of a discrete dynamical system. In order to turn this idea into a numerical method, consider first a non-autonomous T-periodic system of n ODE,

$$\dot{\boldsymbol{x}} = \boldsymbol{X}(\omega t, \boldsymbol{x}), \tag{12}$$

with $\omega = 2\pi/T$ and $\boldsymbol{X}(\theta, \boldsymbol{x})$ 2π-periodic in θ.[3] Denote its flow by $\boldsymbol{\phi}(t, t_0, \boldsymbol{x}_0)$, defined as in (4). Consider the map $\boldsymbol{F}(\boldsymbol{x}) := \boldsymbol{\phi}(t_0 + T, t_0, \boldsymbol{x})$, with t_0 fixed. An initial condition for a T-periodic orbit of (12) is a fixed point of the discrete dynamical system defined by $\boldsymbol{F}$, which is found as a zero of $\boldsymbol{G}(\boldsymbol{x}) := \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{x}$, as discussed in Sect. 3.3. The differential of $\boldsymbol{\phi}(t_0 + T, t_0, \boldsymbol{x})$ with respect to $\boldsymbol{x}$ is computed by integrating the first variational equations, as discussed in Sect. 2.4.

[2] This means that the parameter does not move on a real interval but in a Cantor set. KAM theory also ensures that the parameter spans a sufficiently small interval up to nearly full measure.

[3] Such a system of ODE can be considered a continuous dynamical system given by the autonomous system of ODE $\dot{\boldsymbol{x}} = \boldsymbol{X}(\theta, \boldsymbol{x})$, $\dot{\theta} = \omega$, where θ is an additional dependent variable defined modulo 2π.

Assume now that we have an autonomous system of ODE

$$\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x}), \tag{13}$$

with flow $\boldsymbol{\phi}_t$, defined as in Sect. 2.1. If we wanted to apply the previous approach, we would look for a fixed point of the discrete dynamical system defined by $\boldsymbol{F}(\boldsymbol{x}) := \boldsymbol{\phi}_T(\boldsymbol{x})$. A direct application of Newton's method to look for a zero of $\boldsymbol{G}(\boldsymbol{x}) := \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{x}$ would fail: since $\{\boldsymbol{G}(\boldsymbol{x}) = 0\}$ defines the whole periodic orbit as a manifold, $D\boldsymbol{G}(\boldsymbol{x})$ is singular at every point of the periodic orbit. We could still use the modified Newton strategy of Sect. 3.2, but that would introduce difficulties for continuation.[4] A better strategy is to get rid of the singularity by considering a different discrete dynamical system: a Poincaré map. If Σ is a surface of section transversal to the flow and intersected by the periodic orbit we are looking for, denote as $\boldsymbol{P}(\boldsymbol{x}) = \boldsymbol{\phi}_{\tau(\boldsymbol{x})}(\boldsymbol{x})$ the corresponding Poincaré map, where $\tau(\boldsymbol{x})$ is the time-return map (see Sect. 2.5). Then, by looking for a fixed point of $\boldsymbol{P}$ as a zero of $\boldsymbol{G}(\boldsymbol{x}) := \boldsymbol{P}(\boldsymbol{x}) - \boldsymbol{x}$, we would be looking for an initial condition of the periodic orbit in the Poincaré section Σ, which is locally unique.

The previous approach works as long as the periodic orbit we are looking for is isolated, which is usual in generic dynamical systems. But in Hamiltonian systems like the RTBP, periodic orbits are embedded in families. Assume that we are given a Hamiltonian continuous dynamical system with Hamiltonian $H(\boldsymbol{x})$. The intersections of the periodic orbits of a family with a Poincaré section Σ define locally a curve. On all the points of this curve, $D\boldsymbol{G}(\boldsymbol{x})$ is singular. A way to get rid of this singularity would be to first reduce our starting dynamical system (13) to an energy manifold $\{H(\boldsymbol{x}) = h\}$. Then, an initial condition of a periodic orbit (of energy h) as a fixed point of $\boldsymbol{P}$ would be locally unique. Nevertheless, instead of modifying (13), it is simpler to add an energy equation to the fixed point condition on the Poincaré map. In this way, we would solve for $\boldsymbol{x}$ the non-linear system

$$\left.\begin{aligned} H(\boldsymbol{x}) &= h \\ \boldsymbol{P}(\boldsymbol{x}) &= \boldsymbol{x} \end{aligned}\right\}.$$

This system is not square, so a standard approach using Newton's method would not work, but it can be solved by the modified Newton approach of Sect. 3.2 with $d = 0$. In doing this, $\boldsymbol{P}$ and $D\boldsymbol{P}$ can be evaluated as discussed in Sect. 2.5.

3.6.1 Practical Implementation

An strategy for the computation of periodic orbits still simpler to implement than the one just discussed is to add the Poincaré section as an additional equation. In

[4] We would need to choose a direction tangent to the family of periodic orbits within the two-dimensional kernel of $D\boldsymbol{G}(\boldsymbol{x})$.

this way, at the cost of one extra equation, the evaluation of $\boldsymbol{P}$ and $D\boldsymbol{P}$ is avoided. Assuming that the Poincaré section is $\Sigma = \{g(\boldsymbol{x}) = 0\}$, we would solve the $(n + 2) \times (n + 1)$ system

$$\left.\begin{aligned} H(\boldsymbol{x}) &= h \\ g(\boldsymbol{x}) &= 0 \\ \boldsymbol{\phi}_T(\boldsymbol{x}) &= \boldsymbol{x} \end{aligned}\right\} \tag{14}$$

for $(T, \boldsymbol{x})$. This can be done using the modified Newton strategy of Sect. 3.2 with $d = 0$, $\boldsymbol{y} = (T, \boldsymbol{x})$ and

$$\boldsymbol{G}(\boldsymbol{y}) = \begin{pmatrix} H(\boldsymbol{x}) - h \\ g(\boldsymbol{x}) \\ \boldsymbol{\phi}_T(\boldsymbol{x}) - \boldsymbol{x} \end{pmatrix}. \tag{15}$$

An additional advantage of this approach is that the period of the periodic orbit appears explicitly.

The system of equations (14) can also be used for the continuation of a family of periodic orbits. This would be done using Algorithm 3 with $\boldsymbol{H} := \boldsymbol{G}$ defined as in (15) but for $\boldsymbol{y} = (h, T, \boldsymbol{x})$. In an implementation of Algorithm 3, the routine proposed at the end of Sect. 3.2 would be able to compute $\ker D\boldsymbol{H}(\boldsymbol{y})$ and to solve $\boldsymbol{H}(\boldsymbol{y}) = 0$ with minimum-norm corrections.

From the system of equations (14), other systems of interest for the computation and continuation of periodic orbits can be obtained by eliminating equations and unknowns. For instance, if we eliminate the unknown h, keep T constant and eliminate the energy equation $H(\boldsymbol{x}) = 0$, the resulting system of equations,

$$\left.\begin{aligned} g(\boldsymbol{x}) &= 0 \\ \boldsymbol{\phi}_T(\boldsymbol{x}) &= \boldsymbol{x} \end{aligned}\right\}, \tag{16}$$

that is to be solved only for $\boldsymbol{x}$, would allow us to compute a periodic orbit of a given period. A routine implementing the evaluation of $\boldsymbol{G}(\boldsymbol{y})$ (as defined in (15)) and $D\boldsymbol{G}(\boldsymbol{y})$ can also be used in order to solve a system like (16) by giving it the option of eliminating components of $\boldsymbol{G}(\boldsymbol{y})$ and files and/or columns of $D\boldsymbol{G}(\boldsymbol{y})$.

3.6.2 Multiple Shooting

As we will see in Sect. 3.9.1 (e.g. in Fig. 4), for many periodic orbits of the neighborhood of the collinear points of the Earth-Moon RTBP, the maximum absolute value of the eigenvalues of $D\boldsymbol{\phi}_T(\boldsymbol{x})$ can be larger than 2000. This means that, after numerical integration for T time units, any error in the initial condition can be amplified by this factor. Even with exact data, the local truncation error of the first step of numerical integration could be amplified by this factor.[5] Then, for example, if the tolerance of numerical integration is set to 10^{-14}, we cannot expect an error smaller than 10^{-11}. Because of this, initial conditions for Newton's method need to be very accurate in order to obtain convergence, and continuation steps become very small.

We can reduce these amplification factors by making use of *multiple shooting*. Multiple shooting is classically introduced as a way to overcome dynamical instability in the solution of boundary value problems (see e.g. [41]). As a general idea, the multiple shooting strategy can be thought as introducing intermediate objects as unknowns in order to reduce integration time. In our case, we would need to consider points $\boldsymbol{x}_0 := \boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m-1}$ along the periodic orbit and add the corresponding matching equations to the system to be solved. In this way, system (14) would become

$$\left.\begin{aligned} H(\boldsymbol{x}_0) &= h \\ g(\boldsymbol{x}_0) &= 0 \\ \boldsymbol{\phi}_{T/m}(\boldsymbol{x}_j) &= x_{j+1}, \quad j = 0, \ldots, m-2 \\ \boldsymbol{\phi}_{T/m}(\boldsymbol{x}_{m-1}) &= x_0 \end{aligned}\right\} . \tag{17}$$

In order to compute a single periodic orbit, the unknowns to consider would be $(T, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{m-1})$. In order to continue a family of periodic orbits, the unknowns would be $(h, T, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{m-1})$. As commented before, other systems of interest can be obtained from this one by eliminating equations and unknowns.

By using multiple shooting with m points, the amplification factors are typically reduced to the m-th root of the starting ones, at the cost of multiplying by m the dimension of the system of non-linear equations to be solved.

3.7 Linear Behaviour Around a Periodic Orbit of a Hamiltonian Autonomous System

An initial condition $\boldsymbol{x}_0$ of a T-periodic orbit is also fixed point of $\boldsymbol{\phi}_T$. In the case of a Hamiltonian autonomous system $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, this fact by itself was not enough in order to find $\boldsymbol{x}_0$ numerically, but it is useful to study the linear behavior of the flow around $\boldsymbol{x}_0$. Let $M := D\boldsymbol{\phi}_T(\boldsymbol{x}_0)$ be the monodromy matrix of our periodic orbit. Because of the autonomous character of our system and the fact that it has a first integral (the Hamiltonian), M has 1 as a double eigenvalue (for a proof see, e.g., [31]). Moreover, M is a *symplectic matrix* (see e.g. also [31]), which implies that, if

[5] Actually, even the error of the first floating point operation, which can be as large as the machine epsilon, can be amplified by this factor.

λ is an eigenvalue of M, then $1/\lambda$ is also eigenvalue. Now assume that our system is, as the RTBP, of three degrees of freedom, this is, $\boldsymbol{x} \in \mathbb{R}^6$. Then the eigenvalues of M are

$$\{1, 1, \lambda_1, \lambda_1^{-1}, \lambda_2, \lambda_2^{-1}\}.$$

In the remaining discussion, we will assume that $|\lambda_i| \leq |\lambda_i^{-1}|$.

The linear behaviour around a periodic orbit in our 3-degrees-of-freedom Hamiltonian system is better studied in terms of Hénon's *stability parameters* [22], that are defined as

$$s_1 = \lambda_1 + 1/\lambda_1, \qquad s_2 = \lambda_2 + 1/\lambda_2. \tag{18}$$

A calculation shows that

$$\begin{aligned} s_i \in \mathbb{R}, \quad |s_i| > 2 &\Longleftrightarrow \lambda_i \in \mathbb{R}\backslash\{-1, 1\}, \\ s_i \in \mathbb{R}, \quad |s_i| \leq 2 &\Longleftrightarrow \lambda_i \in \mathbb{C}, \quad |\lambda_i| = 1, \\ s_i \in \mathbb{C}\backslash\mathbb{R} &\Longleftrightarrow \lambda_i \in \mathbb{C}\backslash\mathbb{R}, \quad |\lambda_i| \neq 1. \end{aligned}$$

From the discussion of Sect. 3.5, if $s_i \in \mathbb{R}$, $|s_i| > 2$ (*hyperbolic case*), the stable (resp. unstable) manifold of $\boldsymbol{x}_0$ as fixed point of $\boldsymbol{\phi}_T$ is tangent to the λ_i (resp. λ_i^{-1}) eigendirection. This means that the periodic orbit has a stable (resp. unstable) manifold, and its section through the $\lambda_i, \lambda_i^{-1}$ eigenplane is tangent to the λ_i (resp. λ_i^{-1}) eigendirection. If $s_i \in \mathbb{R}$, $|s_i| \leq 2$ (*elliptic case*), assume $\lambda_i = \cos\rho + \boldsymbol{i}\sin\rho$ and that $\boldsymbol{v}$ is an eigenvector of eigenvalue λ_i. As we have seen, there is a continuous, one-parametric family of closed curves invariant by the linearization of $\boldsymbol{\phi}_T$ around $\boldsymbol{x}_0$ in the $\{\boldsymbol{x}_0 + \alpha_1 \operatorname{Re}\boldsymbol{v} + \alpha_2 \operatorname{Im}\boldsymbol{v}\}_{\alpha_1,\alpha_2\in\mathbb{R}}$ plane, with rotation number ρ. According to the discussion in Sect. 3.5, the full non-linear flow $\boldsymbol{\phi}_T$ possesses a Cantorian family of invariant curves around $\boldsymbol{x}_0$, with limiting rotation number ρ. When transported by the flow, these invariant curves generate two-dimensional invariant tori. Rotation numbers of the form $\rho = 2\pi n/m$ give rise to bifurcated families by multiplication of the period by m (further details on this kind of bifurcations can be found in [37]). The particular values $\rho = 0$ and $\rho = \pi$, which correspond to $s_i = 2$ and $s_i = -2$, respectively, are known as the *parabolic case*.

Note that, if a stability parameter s_i satisfies $|s_i| < 2$ on a range of energies, since for each energy in this range a one-parametric family of tori is born, across energies this family of tori becomes two-parametric.

3.8 Computation of 2D Invariant Tori

This subsection is devoted to the computation of 2D invariant tori. The method discussed, first introduced in [6], consists in looking for a curve inside the torus invariant by the time-T flow, where T is one of the periods of the torus. The formulation will be made explicit for an autonomous Hamiltonian system with three degrees of freedom (as the RTBP), but it can be modified to account for systems with a different number of degrees of freedom, non-autonomous[6] or not Hamiltonian ones.

3.8.1 Looking for a Parameterization of an Invariant Curve

According to KAM theory (see, e.g., [26]), a 2D torus born around the collinear points of the RTBP can be parameterized by a function $\boldsymbol{\psi}(\theta_1, \theta_2)$, 2π-periodic in θ_1, θ_2, satisfying an invariance equation of the form

$$\boldsymbol{\psi}(\theta_1 + t\omega_1, \theta_2 + t\omega_2) = \boldsymbol{\phi}_t\big(\boldsymbol{\psi}(\theta_1, \theta_2)\big), \quad \forall t \in \mathbb{R}, \quad \forall \theta_1, \theta_2 \in [0, 2\pi], \tag{19}$$

where (ω_1, ω_2) is the vector of frequencies of the torus. Looking for a torus can be reduced to looking for an invariant curve inside it by observing that $\varphi(\xi) := \psi(\xi, 0)$ parameterizes a curve invariant by $\boldsymbol{\phi}_{2\pi/\omega_2}$, and satisfies

$$\boldsymbol{\varphi}(\xi + \rho) = \boldsymbol{\phi}_{\Delta}\big(\boldsymbol{\varphi}(\xi)\big), \tag{20}$$

for $\rho = 2\pi\omega_1/\omega_2$ and $\Delta = 2\pi/\omega_2$. Once we have $\boldsymbol{\varphi}$, we can recover $\boldsymbol{\psi}$ by

$$\boldsymbol{\psi}(\theta_1, \theta_2) = \boldsymbol{\phi}_{\frac{\theta_2}{2\pi}\Delta}\Big(\boldsymbol{\varphi}(\theta_1 - \frac{\theta_2}{2\pi}\rho)\Big). \tag{21}$$

A calculation shows that, if $\boldsymbol{\varphi}$ is a 2π-periodic function satisfying (20), then$\boldsymbol{\psi}$ as defined by (21) is 2π-periodic in each variable and satisfies the invariance equation (19) for $\omega_1 := \rho/\Delta$, $\omega_2 := 2\pi/\Delta$. In order to turn (20) into a finite system of non-linear equations, we can take $\boldsymbol{\varphi}$ as a truncated Fourier series,

$$\boldsymbol{\varphi}(\xi) = \boldsymbol{A}_0 + \sum_{k=1}^{N_f}\Big(\boldsymbol{A}_k \cos(k\xi) + \boldsymbol{B}_k \sin(k\xi)\Big), \tag{22}$$

with $\{\boldsymbol{A}_k\}_{k=0}^{N_f}, \{\boldsymbol{B}_k\}_{k=1}^{N_f} \subset \mathbb{R}^6$, and impose (20) at as many values of ξ as the number of Fourier coefficients needed. This is, we will look for $\boldsymbol{\varphi}$ defined as in

[6] The non-autonomous case is actually simpler, because the indeterminacies discussed in Sect. 3.8.1 are not present.

(22) satisfying

$$\boldsymbol{\varphi}(\xi_j + \rho) = \boldsymbol{\phi}_\Delta\big(\boldsymbol{\varphi}(\xi_j)\big), \qquad j = 0, \ldots, 2N_f, \tag{23}$$

with $\xi_j = j2\pi/(1 + 2N_f)$.

The fact that our dynamical system is autonomous gives rise to two indeterminacies:

- An invariant curve inside a torus is not unique: if $\boldsymbol{\varphi}(\xi)$ satisfies (20) or (23), then $\boldsymbol{\phi}_t\big(\boldsymbol{\varphi}(\xi)\big)$ also does, for any $t \in \mathbb{R}$.
- The origin of ξ is free: if $\boldsymbol{\varphi}(\xi)$ satisfies (20) or (23), then $\boldsymbol{\varphi}(\xi - \xi_0)$ also does, for any $\xi_0 \in \mathbb{R}$.

The first indeterminacy can be eliminated by prescribing the value of a coordinate of $\boldsymbol{A}_0$ (the value chosen must be valid for the torus we are looking for). The second indeterminacy can be eliminated by prescribing a coordinate of $\boldsymbol{A}_1$ to be zero: if we denote $\boldsymbol{A}_1 = (A_{1,1}, \ldots, A_{1,6})$, $\boldsymbol{B}_1 = (B_{1,1}, \ldots, B_{1,6})$, and assume that $j \in \{1, \ldots, 6\}$ is such that $(A_{1,j}, B_{1,j}) \neq (0, 0)$, since

$$\begin{aligned} &A_{1,j} \cos(\xi - \xi_0) + B_{1,j} \sin(\xi - \xi_0) \\ &\quad = (A_{1,j} \cos\xi_0 - B_{1,j} \sin\xi_0) \cos\xi + (A_{1,j} \sin\xi_0 + B_{1,j} \cos\xi_0) \sin\xi, \end{aligned}$$

we can always choose ξ_0 such that $A_{1,j} \cos\xi_0 - B_{1,j} \sin\xi_0 = 0$. With the two indeterminacies removed in this way, there is a one-to-one correspondence between (approximate) Fourier coefficients of parameterizations of invariant curves $\boldsymbol{\varphi}$ solution of (23) and invariant 2D tori of our dynamical system.

3.8.2 The System of Equations

By solving system (23) with its two indeterminacies removed, we could compute an invariant curve $\boldsymbol{\varphi}$ of a torus with "longitudinal period" Δ and rotation number ρ, that via (21) would correspond to a torus with frequencies $\omega_1 = \rho/\Delta$, $\omega_2 = 2\pi/\Delta$. We could also use this system in order to do continuation with respect to Δ and/or ρ and, in this way, obtain the corresponding 2-parametric family. Nevertheless, we make two more considerations before stating the final system of equations that we will solve:

- We want to be able to prescribe values for the energy, so we will add an extra equation for this.
- We want to overcome the effects of instability, so we will use multiple shooting, as we did in Sect. 3.6.2.

We will, therefore, look for $\boldsymbol{\varphi}_0, \ldots, \boldsymbol{\varphi}_{m-1}$ satisfying

$$\left\{\begin{array}{rll} H\big(\boldsymbol{\varphi}_0(0)\big) - h = 0 & & \\ \boldsymbol{\varphi}_{j+1}(\xi_i) - \boldsymbol{\phi}_{\Delta/m}\big(\boldsymbol{\varphi}_j(\xi_i)\big) = 0, & j = 0, \ldots, m-2, & i = 0, \ldots, 2N_f, \\ \boldsymbol{\varphi}_0(\xi_i + \rho) - \boldsymbol{\phi}_{\Delta/m}\big(\boldsymbol{\varphi}_{m-1}(\xi_i)\big) = 0, & i = 0, \ldots, 2N_f, & \end{array}\right. \tag{24}$$

where

$$\xi_i = i\frac{2\pi}{1+2N_f}, \qquad i = 0, \ldots, N_f,$$

with unknowns

$$h, \Delta, \rho, \boldsymbol{A}_0^0, \boldsymbol{A}_1^0, \boldsymbol{B}_1^0, \ldots, \boldsymbol{A}_{N_f}^0, \boldsymbol{B}_{N_f}^0, \ \ldots\ , \boldsymbol{A}_0^{m-1}, \boldsymbol{A}_1^{m-1}, \boldsymbol{B}_1^{m-1}, \ \ldots\ , \boldsymbol{A}_{N_f}^{m-1}, \boldsymbol{B}_{N_f}^{m-1},$$

(except for a coordinate of $\boldsymbol{A}_0^0$ and another one of $\boldsymbol{A}_1^0$, according to the previous subsection) with $h, \Delta, \rho \in \mathbb{R}$, $\boldsymbol{A}_j^l, \boldsymbol{B}_j^l \in \mathbb{R}^6$ and

$$\boldsymbol{\varphi}_l(\xi) = \boldsymbol{A}_0^l + \sum_{j=0}^{N_f}\Big(\boldsymbol{A}_j^l\cos(j\xi) + \boldsymbol{B}_j^l\sin(j\xi)\Big). \tag{25}$$

In order to compute a single torus, we can solve system (24) keeping constant, in addition to the coordinates given by the considerations of the previous subsection, two parameters among h, ρ, T. This will fix a torus within its two-parametric family. In order to continue this torus via the pseudo-arclength method, only one of the parameters h, ρ, T must be keep fixed. Two interesting cases are:

- To fix ρ to a number with good Diophantine properties. For instance, a noble number (a number with continued fraction expansion equal to one from a point on).
- To fix h, in order to follow an iso-energetic family of tori. In this case, care must be taken because the family is not continuous but Cantorian: the pseudo-arclength method will work as long as the gaps due to resonances are small.

Note that, both in the computation of a single torus and in the continuation of a one-parametric subfamily, we end up with a system of non-linear equations with more equations than unknowns. Namely, in the first case the system is $\big(1+6m(1+2N_f)\big)\times\big(-1+6m(1+2N_f)\big)$, whereas in the second case is $\big(1+6m(1+2N_f)\big)\times 6m(1+2N_f)$. This is not a problem as long as we use the modified Newton method of Sect. 3.2. Note that, when solving the linear system for the Newton correction, d must be set to zero in the case of the computation of a single torus, whereas it must be set to one in the case of continuation of a one-parametric subfamily.

Once a torus has been computed (either individually of by continuation), an estimate of its error can be obtained by evaluating the invariance equation in a refinement of the discretization in ξ used for its computation. In this way, we can use the estimate

$$\max_{j=0,\dots,M}\left\|\begin{pmatrix}\Big(\boldsymbol{\varphi}_{l+1}(\tilde{\xi}_j)-\boldsymbol{\phi}_{\Delta/m}\big(\boldsymbol{\varphi}_l(\tilde{\xi}_j)\big)\Big)_{l=0}^{m-2}\\ \boldsymbol{\varphi}_0(\tilde{\xi}_j+\rho)-\boldsymbol{\phi}_{\Delta/m}\big(\boldsymbol{\varphi}_{m-1}(\tilde{\xi}_j)\big)\end{pmatrix}\right\| \tag{26}$$

for $\tilde{\xi}_j = j2\pi/M$ and $M \gg 1+2N_f$. The value of this estimate can be used to choose the number of Fourier coefficients N_f. When doing continuation, N_f can be increased or decreased in order to keep this error estimate within a prescribed interval. Observe that large values of N_f will give rise to large systems of equations. The time needed for their solution, which requires $O((6m(1+2N_f))^3)$ operations, will overcome the time needed for numerical integration and become the computational bottleneck of the procedure.

3.8.3 Starting from the Central Part of a Periodic Orbit

According to the discussion of Sect. 3.7, a family of periodic orbits with an elliptic stability parameter in a range of energies gives rise to a 2-parametric family of invariant tori. Here we will develop formulae from the linear flow around the backbone periodic orbit in order to obtain initial conditions to start the continuation of such a family of tori using system (24).

For an arbitrary function $\boldsymbol{G}$, let us denote the linearization of $\boldsymbol{G}$ around $\boldsymbol{y}_0$ as

$$L_{\boldsymbol{G}}^{\boldsymbol{y}_0}(\boldsymbol{y})=\boldsymbol{G}(\boldsymbol{y}_0)+D\boldsymbol{G}(\boldsymbol{y}_0)(\boldsymbol{y}-\boldsymbol{y}_0).$$

Let $\boldsymbol{x}_0$ be an initial condition of a T-periodic orbit, with a stability parameter $s_i=\lambda_i+\lambda_i^{-1}$ satisfying $|s_i|\le 2$ and $\lambda_i=\cos\nu+\boldsymbol{i}\sin\nu$. If we define $\boldsymbol{F}:=\boldsymbol{\phi}_T$, then $\boldsymbol{x}_0$ is a fixed point of $\boldsymbol{F}$, and Eq. (11) provides an expression for an invariant curve of the linearized flow. In this expression, ξ can be substituted by $\xi-\xi_0$, and then we have that

$$\bar{\boldsymbol{\varphi}}(\xi):=\boldsymbol{x}_0+\gamma\Big((\boldsymbol{v}_1\cos\xi_0+\boldsymbol{v}_2\sin\xi_0)\cos\xi+(\boldsymbol{v}_1\sin\xi_0-\boldsymbol{v}_2\cos\xi_0)\sin\xi\Big)$$

also satisfies

$$L_{\boldsymbol{\phi}_T}^{\boldsymbol{x}_0}\big(\bar{\boldsymbol{\varphi}}(\xi)\big)=\bar{\boldsymbol{\varphi}}(\xi+\nu), \tag{27}$$

which is the linearized-flow version of Eq. (20). Therefore, as initial seed to get a torus around the o.p., we can take

$$\begin{aligned}
h &= H(\boldsymbol{x}_0), & \boldsymbol{A}_0^l &= \boldsymbol{\phi}_{lT/m}(\boldsymbol{x}_0),\\
\Delta &= T, & \boldsymbol{A}_1^l &= D\boldsymbol{\phi}_{lT/m}(\boldsymbol{x}_0)\big(\boldsymbol{v}_1\cos\xi_0 + \boldsymbol{v}_2\sin\xi_0\big),\\
\rho &= \nu, & \boldsymbol{B}_1^l &= D\boldsymbol{\phi}_{lT/m}(\boldsymbol{x}_0)\big(\boldsymbol{v}_1\sin\xi_0 - \boldsymbol{v}_2\cos\xi_0\big),\\
& & \boldsymbol{A}_j^l &= \boldsymbol{B}_j^l = 0, \quad j \ge 2,\ l = 0,\dots,m-1.
\end{aligned}$$

The parameter γ should be chosen small enough for Eq. (27) to be a good approximation of Eq. (20). All the computations of Sect. 3.9 have been done with either $\gamma = 10^{-3}$ or $\gamma = 10^{-4}$. The free parameter ξ_0 can be chosen in order to make zero a coordinate of $\boldsymbol{A}_1^0$, and in this way avoid the second indeterminacy discussed in Sect. 3.8.1. An additional problem when computing a first torus around a periodic orbit is that the periodic orbit has a large basin of attraction and is also a (singular) solution of system (24). A way to prevent falling back to it during the Newton iterations is to keep constant a coordinate of $\boldsymbol{A}_1^0$ or $\boldsymbol{B}_1^0$. A good choice is $B_{1,j}^0$, for j such that $A_{1,j}^0$ is being kept equal to zero in order to prevent the second indeterminacy of Sect. 3.8.1.

When we obtain a first invariant curve around a periodic orbit in this way we will say that we are "starting longitudinally to the periodic orbit", because we obtain a tiny invariant curve around $\boldsymbol{x}_0$ for which, in the evaluation of the flow in (24), numerical integration in order to come back to it is "along the periodic orbit". It will be convenient later to be able to obtain a first invariant curve not tiny but approximately of the same size of the periodic orbit and close to it. We will call this second strategy "starting transversally to the periodic orbit".

In order to develop formulae for this second case, we first globalize the invariant curve $\bar{\boldsymbol{\varphi}}$ of the linearized flow to a whole 2D torus by

$$\bar{\boldsymbol{\psi}}(\theta_1,\theta_2) := L^{\boldsymbol{x}_0}_{\boldsymbol{\phi}_{T\theta_2/(2\pi)}}\Big(\bar{\boldsymbol{\varphi}}(\theta_1 - \frac{\theta_2}{2\pi}\rho)\Big).$$

A calculation shows that $\bar{\boldsymbol{\psi}}$ satisfies the linearized-flow equivalent of the invariant equation (19), namely

$$L^{\boldsymbol{\phi}_{T\theta_2/(2\pi)}(\boldsymbol{x}_0)}_{\boldsymbol{\phi}_t}\big(\bar{\boldsymbol{\psi}}(\theta_1,\theta_2)\big) = \bar{\boldsymbol{\psi}}\Big(\theta_1 + t\frac{\nu}{T}, \theta_2 + t\frac{2\pi}{T}\Big).$$

The invariant curve we are looking for will be close to $\bar{\boldsymbol{\psi}}(0,\theta_2)$. In order to find it, and since ν can be substituted by $\pm\nu + j2\pi$ in all the previous expressions, we can take as initial seed

$$h = H(\boldsymbol{x}_0), \quad \Delta = \frac{2\pi}{\pm\nu + j2\pi}T, \quad \rho = \frac{(2\pi)^2}{\pm\nu + j2\pi} + k2\pi, \tag{28}$$

and $\boldsymbol{A}^l_k$, $\boldsymbol{B}^l_k$ coming from a Discrete Fourier transform (DFT) of $\{\bar{\boldsymbol{\psi}}(0, j\frac{2\pi}{N})\}_{j=0}^{N-1}$. Some notation for the DFT and its relation with Fourier coefficients is developed in Sect. 5.2.

3.9 Numerical Exploration of the Dynamics Around the L_1 Point of the Earth-Moon RTBP

The goal of this subsection is to implement the hierarchical approach mentioned at the beginning of this section in order to systematically compute families of periodic orbits and tori around a collinear libration point of the RTBP. This can be also seen as numerically growing the center manifold of the collinear libration point. The numerical results shown, which are a subset of the ones in [14], will be for the L_1 point and the Earth-Moon mass ratio. In all the computations of this subsection, the flow of the RTBP and its differential with respect to initial conditions have been evaluated according to the discussion of Sect. 2.4, using as one-step method with step size control for numerical integration a Runge-Kutta-Fehlberg one of orders 7 and 8 [10] with tolerance 10^{-14}. The value used for the Earth-Moon mass ratio is

$$\mu = 1.215\,0585\,6096\,2404 \cdot 10^{-2}, \tag{29}$$

as obtained from the DE406 JPL ephemeris file [40].

3.9.1 Periodic Orbits

The linear behavior around the L_1 point for the Earth-Moon mass ratio is of the type center×center×saddle [42]. This is, if we denote by $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$ the vector field of the RTBP, as in Eqs. (1), (2), we have

$$\operatorname{Spec} D\boldsymbol{X}(L_1) = \{\boldsymbol{i}\omega_p, -\boldsymbol{i}\omega_p, \boldsymbol{i}\omega_v, -\boldsymbol{i}\omega_v, \lambda, -\lambda\}, \tag{30}$$

for $\omega_p, \omega_v, \lambda > 0$. As discussed in Sect. 3.4, Lyapunov's center theorem ensures that each center gives rise to a family of periodic orbits. In the expression for $\operatorname{Spec} D\boldsymbol{X}(L_1)$ above, ω_p (resp. ω_v) can be chosen in such a way that the eigenplane corresponding to the eigenvalues $\pm\boldsymbol{i}\omega_p$ (resp. $\pm\boldsymbol{i}\omega_v$) is contained in $\{z = p_z = 0\}$ (resp. $\{x = p_x = y = p_y = 0\}$). Because of this, the family of periodic orbits related to the $\pm\boldsymbol{i}\omega_p$ (resp. $\pm\boldsymbol{i}\omega_v$) eigenvalues is known as the *planar* (resp. *vertical*) *Lyapunov family*. Initial guesses to start the numerical continuation (see Sect. 3.1) of these families can be obtained from (9). When doing Newton iterations on system (17) to find the first periodic orbit, a convenient way to avoid falling back to the L_1 point (which is a singular solution of system (17) with a large basin of attraction) is to keep constant a coordinate of $\boldsymbol{x}_0$.

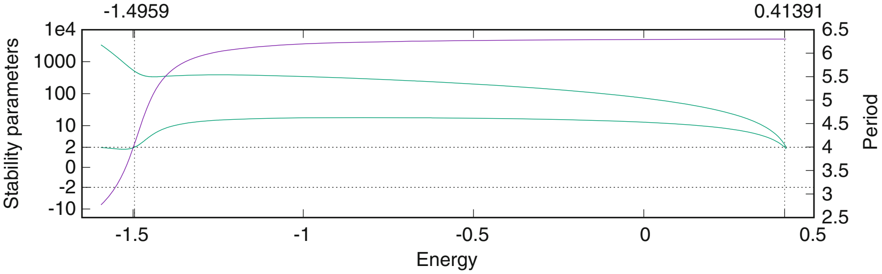


Fig. 4 Characteristic curve (in violet) and stability parameters (in green) of the vertical Lyapunov family around L_1 of the Earth-Moon RTBP

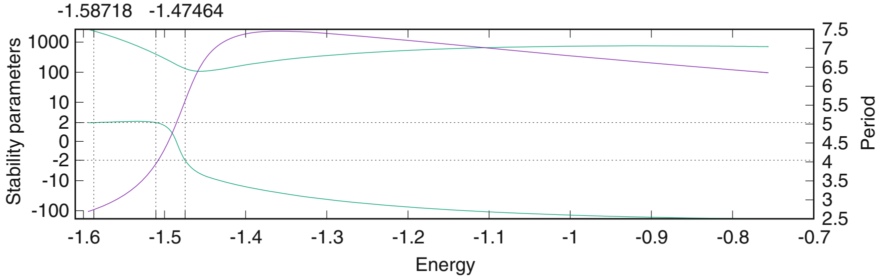
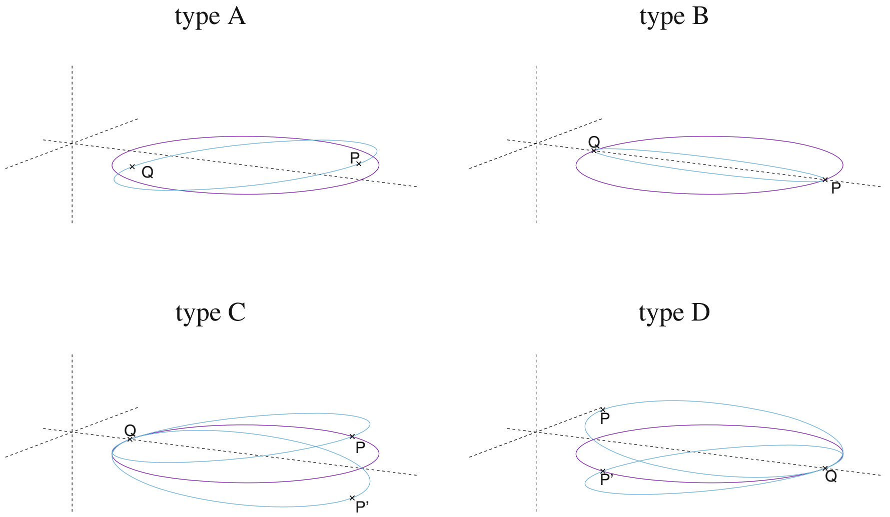


Fig. 5 Characteristic curve (in violet) and stability parameters (in green) of the planar Lyapunov family around L_1 of the Earth-Moon RTBP

A convenient way to represent a family of periodic orbits that has been obtained by numerical continuation is by plotting the period and the stability parameters (18) of its orbits with respect to energy. The period vs. energy curve is known as *characteristic curve*. Figure 4 represents the characteristic curve and stability parameters of the vertical Lyapunov family. This family starts at energy -1.59417 (the one of L_1), has a bifurcation at energy -1.49590, that will be commented later, and ends at a large planar orbit with energy 0.41391, that surrounds the Earth and the collinear points L_1, L_3. Plots of sample orbits of this family and all the other families of periodic orbits that we will consider can be found in [32].

In Fig. 5 we have represented the characteristic curve and stability parameters of the planar Lyapunov family. This family starts at energy -1.59417 (the one of L_1), has several bifurcations and ends at a collision with the Earth.[7] According to [23], the only possible kinds of bifurcation from the planar Lyapunov family to a family of three-dimensional orbits are the ones sketched in Fig. 6. Types A and B correspond to a stability parameter crossing 2, whereas types C and D correspond to a stability parameter crossing -2 (and thus are period-doubling bifurcations). In cases A, B

[7] By using regularization (see e.g. [42]), the planar Lyapunov family could be continued for energies past this collision. We do not continue the family further because this collision is already outside of the range of energies of the invariant tori that we will compute.


type A

type B

type C

type D

Fig. 6 Possible bifurcation types of a bifurcating planar Lyapunov orbit to a non-planar family

Table 1 Bifurcations of the planar Lyapunov family around L_1 of the Earth-Moon RTBP

#Bif.	Energy	Type
1	-1.58718	A
2	-1.51070	B
3	-1.47464	C

not one but two families of periodic orbits bifurcate from the planar family. The two bifurcated families are symmetric with respect to $\{z = 0\}$. Assuming that the Poincaré section used in the continuation of the planar Lyapunov family is $\{y = 0\}$,[8] an initial condition for one of such bifurcated orbits can be obtained by doing a small displacement in the z coordinate for types A, C, D, and in the p_z coordinate for type B. The displaced coordinate can be kept constant during Newton iterations on system (17) in order to avoid falling back to an orbit in the planar Lyapunov family. The bifurcations found for the planar Lyapunov family, together with its classification according to [23], are given in Table 1.

The first bifurcation of the planar Lyapunov family gives rise to the two symmetric families of periodic orbits known as halo orbits. The corresponding characteristic curve and stability parameters[9] are shown in Fig. 7. Both families end at a large planar orbit that surrounds the Earth, the Moon and the collinear points L_1, L_2. For a large range of energies halo orbits have complex (non-real) stability parameters; Fig. 8 zooms Fig. 7 in order to show the transition from real to complex stability parameters and vice-versa. In Fig. 8 left, it is also shown how the

[8] This is, $g(x, y, z, p_x, p_y, p_z) = y$ in systems (14) or (17).

[9] Given one periodic orbit of a halo family, the symmetric periodic orbit of the symmetric family has the same period and stability parameters.

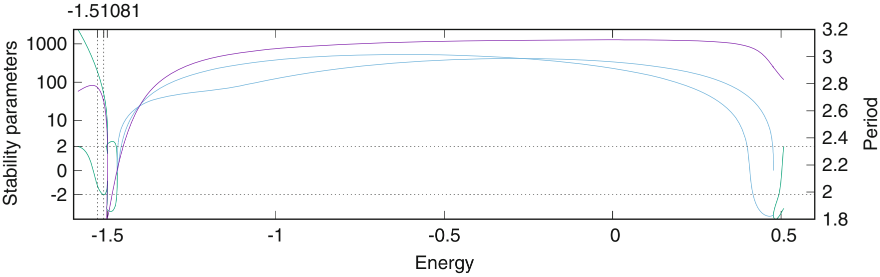


Fig. 7 Characteristic curve (in violet) and stability parameters (in green and blue) of the halo family around L_1 of the Earth-Moon RTBP. In the case $s_1 \in \mathbb{C} \backslash \mathbb{R}$, $s_2 = \bar{s}_1$ (complex saddle), $\mathrm{Re}\, s_1$ and $\mathrm{Im}\, s_1$ are represented in blue

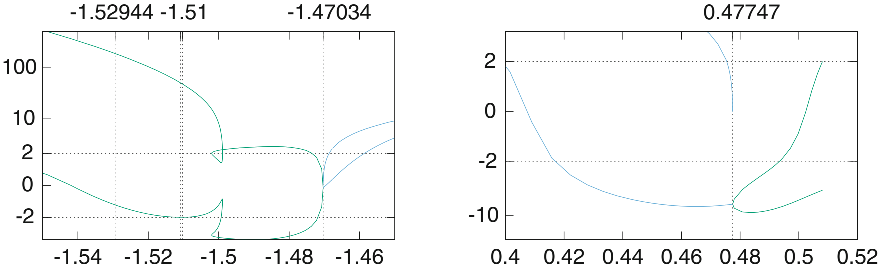


Fig. 8 Magnifications of Fig. 7 showing the transitions to and from complex saddle

small stability parameter goes across $2\cos(2\pi/3)$ once, at energy -1.52944, and across $-2 = 2\cos(2\pi/2)$ twice, at energies -1.51081, -1.51033. The first case gives rise to two period-triplicated bifurcated families of periodic orbits, one with elliptic-hyperbolic normal behaviour and the other one with elliptic-elliptic normal behaviour. The second case gives rise to a period-duplicated family of periodic orbits with elliptic-elliptic normal behaviour. The third case gives rise to another period-duplicated family of periodic orbits but with elliptic-hyperbolic normal behaviour. These three bifurcations take place for each of two symmetric halo families. As discussed in Sect. 3.7, there are many more bifurcations, but these three will play a role in the computations of invariant tori of the next subsection. The actual initial conditions used to find orbits of these families have been found by shooting from invariant tori nearby.

The second bifurcation of the planar Lyapunov family gives rise to two families, symmetric with respect to $z = 0$, that can be thought as a two-lane bridge that connects the planar Lyapunov family with the vertical one at its bifurcation at energy -1.49590. Some orbits of this family are shown in Fig. 9. Table 1 still reflects a third bifurcation of the planar family that we do not follow, because it takes place at an energy outside the range of energies that will be reached by the continuation of invariant tori of the next subsection.

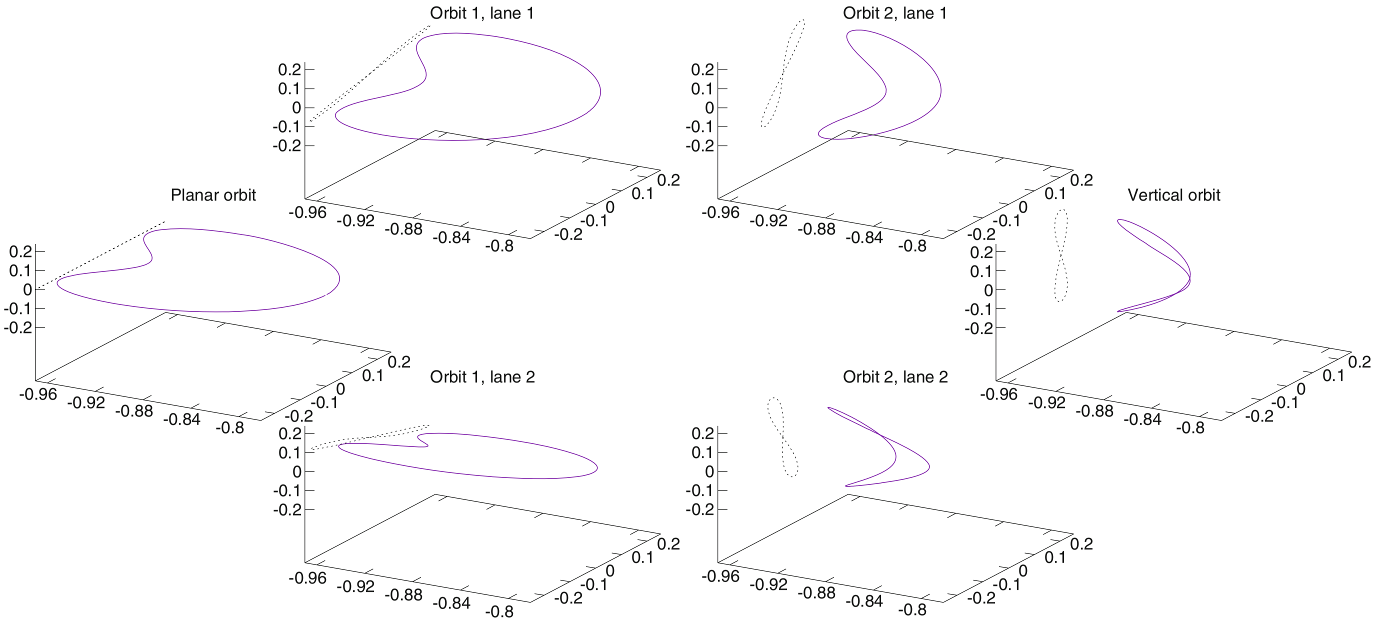


Fig. 9 Some orbits in the two-lane bridge joining the planar Lyapunov family around L_1 of the Earth-Moon RTBP and the vertical one

3.9.2 Invariant Tori

The first families of tori around the libration point L_1 that we will compute will be the ones of constant rotation number ρ starting longitudinally from the vertical Lyapunov family of periodic orbits. The range of values of ρ to be considered is thus provided by the values of $\nu > 0$ such that, according to Sect. 3.8.3, $2\cos\nu$ is one of the stability parameters of the base vertical Lyapunov orbit. Therefore, initiating the continuation of each constant ρ family requires to find a initial condition of a vertical periodic orbit corresponding to a specific value of ν. This initial condition is obtained by continuation of system (17). Since ν is not a continuation variable, it must be considered a function of a continuation variable, for instance $\nu = \nu(h)$. The value of h providing a prescribed value of $\nu(h)$ can be found by a numerical one-dimensional zero-finding method. A good choice is Brent's (see e.g. [36]), since it has fast, global convergence and does not require computing derivatives.

If we represent the value of ν with respect to energy along the vertical Lyapunov family of periodic orbits for the range of energies in which they have central part (see Fig. 4), we obtain the curve labeled β in Fig. 10. This curve goes from the point P_2, that corresponds to the collinear point L_1, to the point P_3, that corresponds the bifurcation of the vertical Lyapunov family at energy -1.4959 (see Fig. 4). Our first continuation of families of tori, with constant rotation number, has been done by choosing an approximately equally spaced grid of noble values of ρ (in order to stay away from resonances, as discussed in Sect. 3.8.2), ranging from the ordinate value of the point P_2 of Fig. 10 to the maximum value of ν along the β curve, and

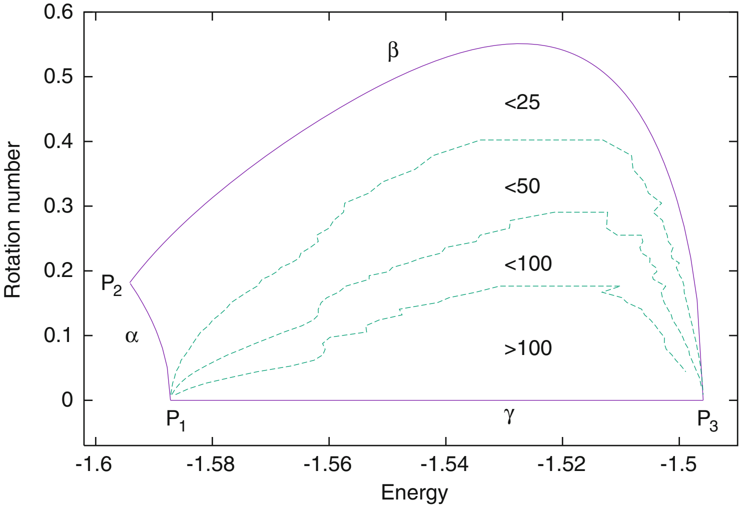


Fig. 10 Energy-rotation number representation of the tori computed around the Lyapunov families of periodic orbits around L_1 of the Earth-Moon RTBP. The region delimited by the curves α, β, γ, which contains the tori, is divided in subregions according to the values of N_f used in the computation of each torus

starting longitudinally from the leftmost planar Lyapunov periodic orbit of the β curve with stability parameter $2\cos\rho$. Each of the obtained families of tori, with constant ρ, that would be seen in Fig. 10 as a horizontal line, collapses at a vertical Lyapunov orbit of higher energy, as the shape of the β curve suggests. With this first continuation we cover the region in Fig. 10 delimited by the curves α, β, γ with $\rho \geq \rho(P_2)$. This continuation of families of invariant tori, and all the remaining continuations that we will describe, have been done by solving system (24) with $m = 2$, a tolerance of 10^{-11} for Newton iterations, and with continuation step size control with $n_{des} = 4$ in (7). In the continuation of each family, the number of harmonics N_f of the Fourier expansions (25) has been determined in order to keep the estimate (26) under 10^{-10}. In addition to this, an upper limit of $N_f = 100$ has been set. When this limit is reached, the error estimate (26) is allowed to grow up to 10^{-8} and, when this happens, the continuation is stopped. This has never happened in this first exploration.

In order to cover the region within the curves α, β, γ with $\rho < \rho(P_2)$, a possibility would be to start from the β curve and go downwards. This means to perform continuation of families of tori with h constant. If h is close to the energy of L_1, the iso-energetic family of tori obtained should end by collapsing to a planar Lyapunov orbit, because this is what happens linearly. The actual tori of such a continuation are shown in Fig. 11, for $h = -1.59$. Although the tori do collapse to a planar orbit, the corresponding invariant curves $\boldsymbol{\varphi}_0$ obtained by solving system (24) do not collapse to a point but tend to the whole ending planar Lyapunov orbit. The limiting value of ρ is numerically checked to be

$$\frac{(2\pi)^2}{2\pi - \nu} - 2\pi, \tag{31}$$

where ν is such that $2\cos\nu$ is a stability parameter of the ending planar Lyapunov periodic orbit. Therefore, according to (28), the same invariant curves within the tori of Fig. 11 could be obtained by starting transversally from this ending planar Lyapunov orbit. The α curve of Fig. 10 is obtained by plotting expression (31) as a function of h, with ν such that $2\cos\nu$ is a stability parameter of the planar Lyapunov orbit of energy h. The point with label P_1 in this curve corresponds to the bifurcation of the planar Lyapunov family of periodic orbits to the halo families (see Fig. 5 and Table 1). The family of tori of Fig. 11 would be seen in Fig. 10 as a vertical line with $h = -1.59$ that goes from the curve β to the curve α.

In order to complete the computation of invariant tori within the curves α, β, γ, and in order to avoid “jumping over resonances”, we go back to the constant ρ continuation strategy. From the discussion in the last paragraph, the remaining tori within the α, β, γ curves can be computed by starting transversally from the family of planar Lyapunov periodic orbits, for an approximately equally spaced grid of noble values of ρ of the form (31), for the range of values of ν that produced the α curve. When doing so, some of the corresponding constant-ρ families of invariant tori with largest ρ value have reached a vertical Lyapunov periodic orbit of higher energy. The remaining ones have stopped due to the $N_f = 100$ computational limit. For each value of ρ in which this has happened, we have also continued

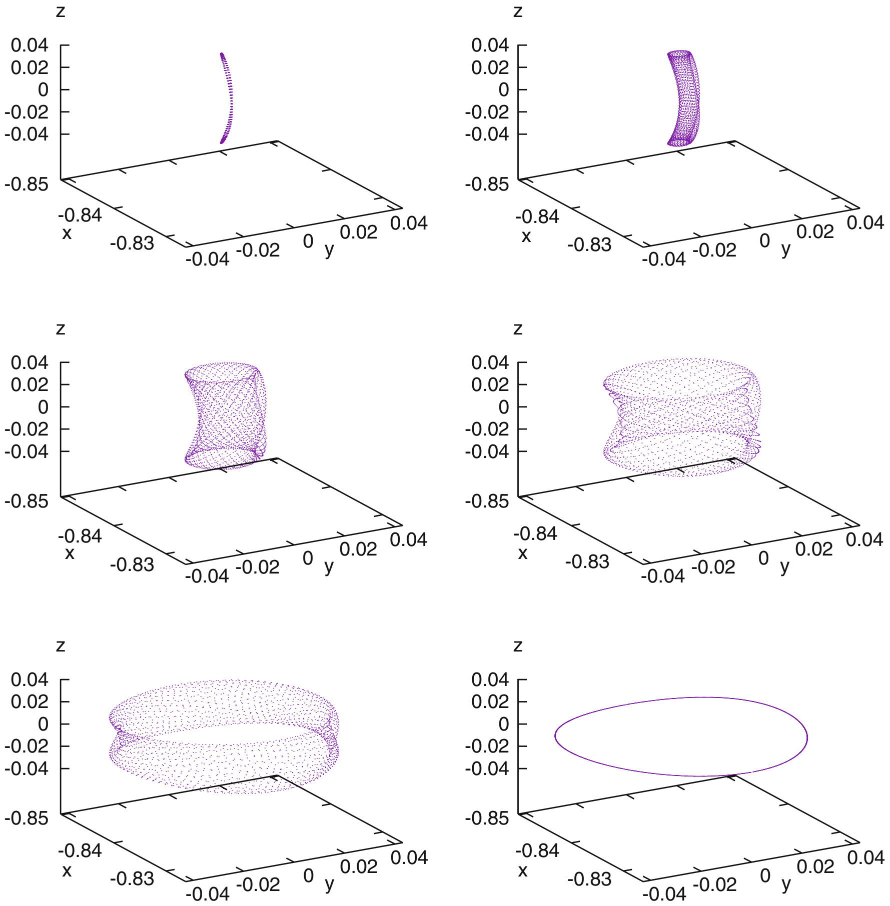


Fig. 11 Sample tori of the iso-energetic family starting from the vertical Lyapunov periodic orbit around L_1 of the Earth-Moon RTBP of energy -1.59

for decreasing energies the family with constant ρ starting longitudinally from the rightmost vertical Lyapunov periodic orbit of the β curve with this ρ value. In this way, we have covered with invariant tori all the region within the α, β, γ curves of Fig. 10 except for the one labeled as > 100. By allowing for $N_f > 100$, some of the tori of this last region could be computed. Many of them, however, simply do not exist, because, as we will see later, as ρ goes to zero for fixed energies larger than the one of the point P_1 in Fig. 10, we approach homoclinic connections of periodic orbits.

The α, β, γ curves of Fig. 10 delimit a set of tori that can be considered a single family, since all of these tori can be reached by numerical continuation starting from L_1. Close to L_1, the tori of this family are the ones given by KAM theory. Trajectories in them are known as Lissajous trajectories by the astrodynamics community. We will thus denote this family as the Lissajous family of invariant tori. Tori in this family can be considered to have "natural" frequencies $\omega_v(T, \rho)$, $\omega_p(T, \rho)$, obtained by continuation from the frequencies ω_v, ω_p of the collinear point L_1 in Eq. (30). An application of Lyapunov's center theorem shows that

$$T = \frac{2\pi}{\omega_v(T, \rho)}, \quad \rho = 2\pi\left(\frac{\omega_p(T, \rho)}{\omega_v(T, \rho)} - 1\right).$$

Following the strategy of choosing an approximately equally spaced grid of noble values of ν along a family of periodic orbits with $2\cos\nu$ a central stability parameter and starting longitudinally the family of invariant tori with constant rotation number $\rho = \nu$, we have also performed numerical continuation of several additional families of invariant tori. These additional families are:

- Invariant tori around halo orbits, from the beginning of the family up to its first turning point in the energy (see Figs. 7 and 8 left).
- Invariant tori around the elliptic-hyperbolic period-triplicated halo-type family of periodic orbits, from the beginning of the family up to the energy in which the central stability parameter crosses -2.
- Invariant tori around the elliptic-hyperbolic period-duplicated halo-type family of periodic orbits, in an energy range analogous to the previous one.
- Invariant tori around planar Lyapunov orbits, in a short energy range starting at the bifurcation of the two-lane bridge joining it with the vertical one, in order to complete the Poincaré sections of Fig. 14.

Except for the last one, these families are represented in Fig. 12 in h-ρ plots analogous to Fig. 10. Contrary to the Lissajous family of invariant tori, none of these new families has been described completely. The numerical continuations have been stopped when the $N_f = 100$ computational limit has been reached. How these families further evolve is an open question.

3.9.3 Iso-Energetic Poincaré Sections

Since the center manifold of L_1, $W^c(L_1)$, is four-dimensional, its restriction to an energy value, $W^c(L_1) \cap \{H = h\}$, would be three-dimensional, and a Poincaré section in this restriction, $W^c(L_1) \cap \{H = h\} \cap \Sigma$, would be two-dimensional. Following [15, 25], it is convenient to visualize $W^c(L_1)$ by a sequence of iso-energetic Poincaré sections. This is done in Figs. 13 and 14, using the Poincaré section $\Sigma := \{z = 0, p_z > 0\}$. In order to be able to produce these figures, in the continuation of each constant ρ family of tori of the previous subsection, the tori of the energies of the plots of Fig. 13 have been obtained by doing Newton iterations keeping h constant, starting from pseudo-arclength predictions from nearby tori (see Algorithm 3).

All the plots of Fig. 13 have a similar structure. The exterior curve in each plot is a Lyapunov planar orbit of the energy level corresponding to the plot. Strictly speaking, the Poincaré section is not valid for this orbit, so it should not have been plotted. Nevertheless, it is useful to use it as boundary of $W^c(L_1) \cap \Sigma$ at the energy

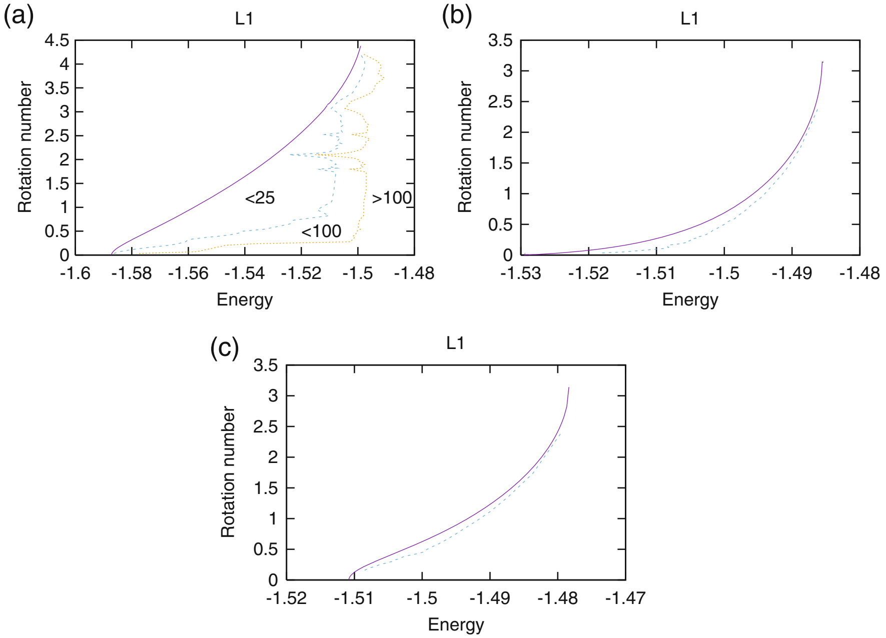


Fig. 12 Representation of the invariant tori computed around (**a**) halo orbits, (**b**) period-triplicated halo-type orbits and (**c**) period-duplicated ones. The outermost dotted curve represents reaching the computational limit $N_f = 100$

of the plot. The closed curves inside the region bounded by the Lyapunov planar orbit are the intersections with Σ of the invariant tori computed in the previous subsection. These intersections are computed through Algorithm 1, starting from the invariant curve $\boldsymbol{\varphi}_0$ (see system (24)) of each torus.

In all the plots there is a fixed point on the x axis associated to the vertical Lyapunov orbit. This point is not represented, but outlined by the smallest blue curves. For small energy values, the whole picture is formed by invariant curves surrounding this fixed point. They are associated to the intersections with Σ of Lissajous-type trajectories around the vertical periodic orbit, whose evolution from the vertical Lyapunov periodic orbits to the planar one is similar to the one displayed in Fig. 11. At the energy levels in which halo orbits have bifurcated from the planar Lyapunov family, there appear two additional fixed points, again not represented but outlined by the smallest violet invariant curves. Increasing the values of the energy, the halo family undergoes the two bifurcations mentioned in Sect. 3.9.1, by period triplication and duplication. Within the bifurcated families there are some with central part, which are surrounded by invariant tori, also computed in the previous subsection, whose Poincaré sections provide here the red invariant curves. These invariant curves give rise to the "island chain" structure typical of two-dimensional area-preserving maps (compare with Fig. 3). To display more clearly this behaviour,

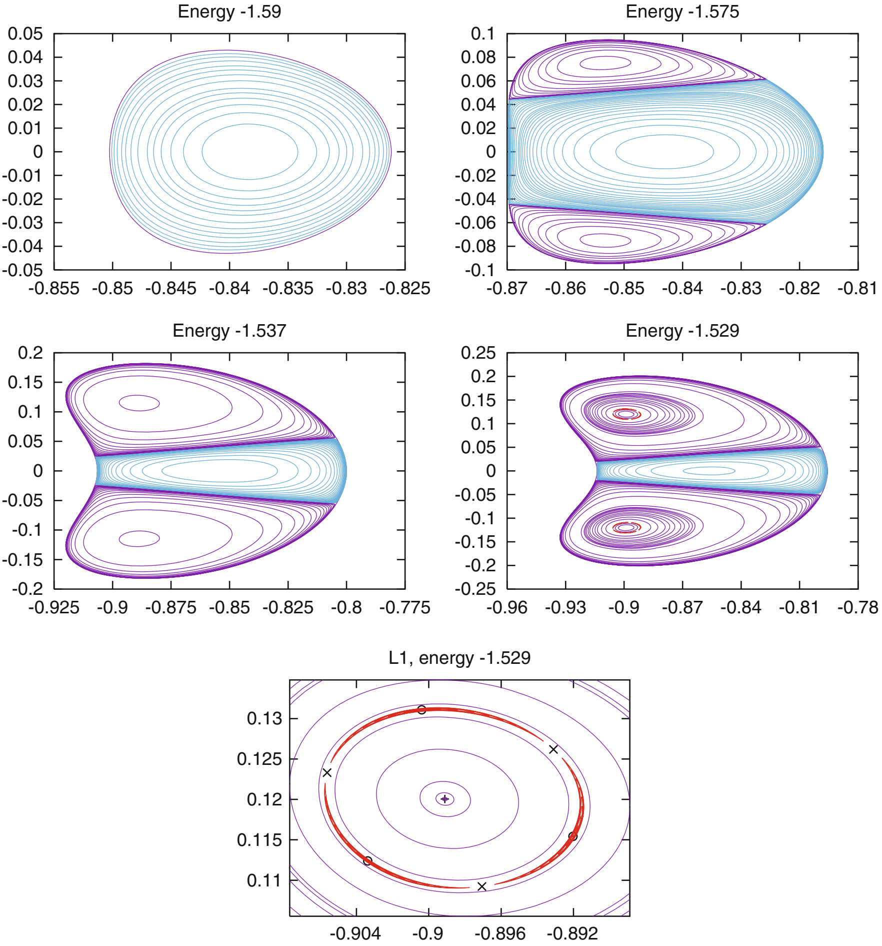


Fig. 13 Iso-energetic Poincaré sections with $\Sigma = \{z = 0, p_z > 0\}$ of the families of periodic orbits and invariant tori computed. The last plot is a magnification around the period-triplicated halo-type family of periodic orbits

the last plot of Fig. 13 displays a magnification of the bifurcated periodic orbits and its surrounding invariant tori.

The region between the tori around the vertical Lyapunov orbit and the tori around the halo orbits is not empty, as it appears in the above figures. It should contain, at least, the traces on the surface of section of the invariant manifolds of the Lyapunov planar orbit. These manifolds act as separatrices between both kinds of tori. The same thing happens between the islands of the bifurcated halo-type orbits and the tori around halo orbits. In this case, the region between both kinds of tori is filled with the traces of the invariant manifolds of the bifurcated hyperbolic halo-type orbits. In all these boundary regions, the motion should have chaotic behaviour.

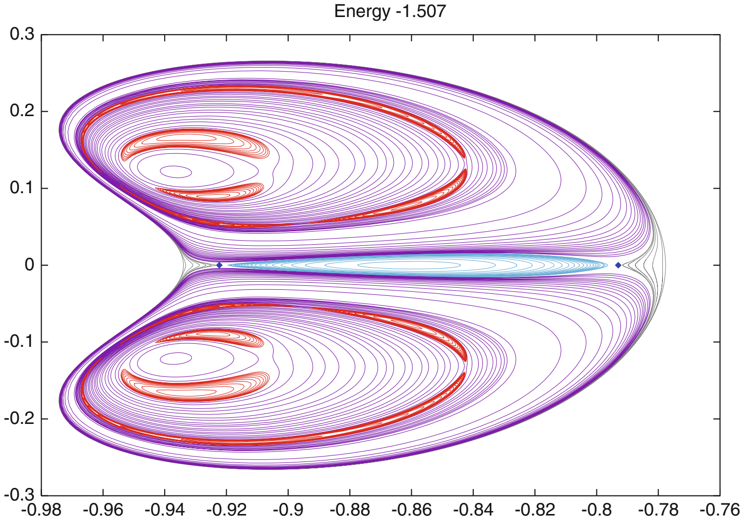


Fig. 14 Poincaré section corresponding to energy -1.507

The numerical methods of this section are not able to capture this chaotic motion, but the semi-analytical methods of the next section can capture it.

The plot corresponding to energy -1.507, shown in Fig. 14, has more structure. For this energy level, the two-lane bridge between the planar and vertical Lyapunov families of periodic orbits has already bifurcated, so the planar family has gained central part, and its periodic orbits are again surrounded by invariant tori. The $\{z = 0\}$ sections of these tori are the outermost curves that appear in Fig. 14 (in this case, the planar Lyapunov periodic orbit, that surrounds all these curves, is not represented). In the figure, the two diamond points are the fixed points corresponding to the intersections of the two orbits of the bridge with the surface of section. The invariant manifolds of these bifurcated periodic orbits are the ones that must act as separatrices between the tori around the halo orbits and the tori around the vertical Lyapunov orbit of this energy.

4 Semi-Analytical Computation of Invariant Objects Using the Parameterization Method

The *parameterization method* is an approach to the study of invariant manifolds, whose general idea is to seek for parameterizations of invariant manifolds as solution of invariance equations, that are simplified through changes of variables that exploit geometrical properties. It is a strong point of this approach that "theoretical" and "numerical" are two aspects of the same philosophy. On the one hand, the proofs are constructive and can be turned into algorithms. On the other hand, these algorithms, when implemented with rigorous numerics based on interval arithmetic, can be turned into computer assisted proofs. Since its introduction in [5], it has been used by many authors. A recent review, that also has some original developments, can be found in [21].

Here we will be concerned with the use of the parameterization method for the (non-rigorous) computation of Taylor expansions of invariant manifolds around fixed points of flows. It will be applied to the computation of the center manifold of the collinear points L_1, L_2, of the Earth-Moon RTBP. In this way, this variant of the parameterization method can be seen as a semi-analytical technique for the computation of the invariant objects inside the center manifold of the collinear points of the RTBP. An earlier technique, known as reduction to the center manifold [15, 25], produces essentially the same results. The parameterization method has some advantages in computational speed, generality (the implementation is independent of the dynamical system under study, the RTBP in our case) and flexibility, since the coordinates of the manifold can be adapted to the dynamics, as we will see in Sect. 4.4.2.

The discussion will follow chapter 2 of [21] except for some notational changes, additional computations and plots. The software package in http://www.maia.ub.edu/dsg/param/ includes a C routine that computes expansions of invariant manifolds of fixed points of flows as described below.

4.1 The Method

Assume we are given a continuous, n-dimensional dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$ with a fixed point $\boldsymbol{p}$ at which the differential of the vector field is diagonalizable. We would like to compute a d-dimensional manifold that contains the fixed point and is tangent to a d-dimensional eigenspace of the differential of the vector field. By a change of variables of the form

$$\boldsymbol{x} = \boldsymbol{p} + P\boldsymbol{y},$$

our original system can be turned into $\dot{\boldsymbol{y}} = \boldsymbol{Y}(\boldsymbol{y})$, $\boldsymbol{y} = (y_1, \dots, y_n)$, with $D\boldsymbol{Y}(0) = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_i \in \mathbb{C}$, in such a way that the eigenspace of interest is $\{\boldsymbol{y} \in \mathbb{R}^n : y_{d+1} = \cdots = y_n = 0\}$. Then our goal is to compute an expansion of a d-dimensional manifold that contains the origin, is invariant by the flow, and is tangent to the $y_1, \dots, y_d$ coordinates.

To do this, we look for $\boldsymbol{W} : \mathbb{C}^d \longrightarrow \mathbb{C}^n$, parameterization of the manifold, and for $\boldsymbol{f} : \mathbb{C}^d \longrightarrow \mathbb{C}^d$, the vector field reduced to the manifold. In this way, if we denote by $\boldsymbol{s} \in \mathbb{C}^d$ the parameters describing the manifold, then the differential equations in parameter space are $\dot{\boldsymbol{s}} = \boldsymbol{f}(\boldsymbol{s})$. From the parameterization of the manifold $\boldsymbol{W}(\boldsymbol{s})$ in the $\boldsymbol{y}$ variables, a parameterization of the manifold in the original

$\boldsymbol{x}$ variables can be recovered as

$$\bar{\boldsymbol{W}}(\boldsymbol{s}) = \boldsymbol{p} + P\boldsymbol{W}(\boldsymbol{s}). \tag{32}$$

In order to find $\boldsymbol{W}$, $\boldsymbol{f}$ we need to solve the invariance equation:

$$\boldsymbol{Y}\big(\boldsymbol{W}(\boldsymbol{s})\big) = D\boldsymbol{W}(\boldsymbol{s})\boldsymbol{f}(\boldsymbol{s}). \tag{33}$$

Assume that $\boldsymbol{W}$, $\boldsymbol{f}$ are expanded as power series in $\boldsymbol{s}$,

$$\boldsymbol{W} = \sum_{k\geq 1} \boldsymbol{W}_k, \quad \boldsymbol{f} = \sum_{k\geq 1} \boldsymbol{f}_k,$$

with $\boldsymbol{W}_k$ n-vector and $\boldsymbol{f}_k$ d-vector of homogeneous polynomials of degree k in $\boldsymbol{s} = (s_1, \dots, s_d)$,

$$\boldsymbol{W}_k = (W_k^1, \dots, W_k^n), \quad W_k^i = \sum_{m_1+\cdots+m_d=k} W_{k,m}^i s_1^{m_1} \dots s_d^{m_d},$$

for $m = (m_1, \dots, m_d) \in \mathbb{N}^d$. With these notations, we can solve the invariance equation order by order. Orders 0, 1 are satisfied by taking:

$$\begin{aligned} \boldsymbol{W}_0 &= 0, & \boldsymbol{W}_1 &= (s_1, \dots, s_d, 0, \dots, 0), \\ \boldsymbol{f}_0 &= 0, & \boldsymbol{f}_1 &= (\lambda_1 s_1, \dots, \lambda_d s_d). \end{aligned}$$

Now assume that

$$\begin{aligned} \boldsymbol{W}_{<k} &:= \boldsymbol{W}_1 + \cdots + \boldsymbol{W}_{k-1}, \\ \boldsymbol{f}_{<k} &:= \boldsymbol{f}_1 + \cdots + \boldsymbol{f}_{k-1} \end{aligned}$$

are known. If we restrict Eq. (33) to its terms of order k, we obtain the order-k cohomological equation. By putting all the unknown terms in the left-hand side and all the known terms in the right-hand one, we obtain as right-hand side

$$\boldsymbol{R}_k := [\boldsymbol{Y}(\boldsymbol{W}_{<k}(\boldsymbol{s}))]_k - \sum_{l=2}^{k-1} D\boldsymbol{W}_{k-l+1}(\boldsymbol{s})\boldsymbol{f}_l(\boldsymbol{s}), \tag{34}$$

where $[\,]_k$ stands for "terms of order k". The evaluation of the second term in the previous expression involves products of homogeneous polynomials. The first term, which consists in plugging the known part of $\boldsymbol{W}$ into the vector field and obtaining the terms of degree k, is computationally more costly. High efficiency is achieved through the use of automatic differentiation, as will be discussed below.

The expression for the left-hand side of the order-k cohomological equation depends on the component. The whole order-k cohomological equation reads

$$(\langle\bar{\lambda}, m\rangle - \lambda_i)W^i_{k,m} + f^i_{k,m} = R^i_{k,m}, \qquad \text{for } i \in \{1, \dots, d\}, \tag{35}$$

$$(\langle\bar{\lambda}, m\rangle - \lambda_i)W^i_{k,m} = R^i_{k,m}, \qquad \text{for } i \in \{d+1, \dots, n\}, \tag{36}$$

where $\bar{\lambda} := (\lambda_1, \dots, \lambda_d)$, $\langle\bar{\lambda}, m\rangle := \lambda_1 m_1 + \dots \lambda_d m_d$. The manifold can be computed as long as (36) can be solved, this is, there are no $m \in \mathbb{N}^d, i \in \{d+1, \dots, n\}$ such that $\lambda_i = \langle\bar{\lambda}, m\rangle$, which would be a *cross-resonance*. The solution of (35) can be done in several ways, that give rise to different *styles* of parameterization:

- The *graph style*, that consists in taking $W^i_{k,m} = 0$, $f^i_{k,m} = R^i_{k,m}$, as to obtain $W^1(s) = s_1, \dots, W^d(s) = s_d$, so that, in $\boldsymbol{y}$ coordinates, the manifold is the graph of the function $(W^{d+1}, \dots, W^d)$.
- The *normal form style*, in which the expansion of $\boldsymbol{f}$ is taken as simple as possible:

$$\begin{aligned} &W^i_{k,m} = 0, && f^i_{k,m} = R^i_{k,m}, && \text{if } \langle\bar{\lambda}, m\rangle - \lambda_i = 0, \\ &W^i_{k,m} = R^i_{k,m}/(\langle\bar{\lambda}, m\rangle - \lambda_i), && f^i_{k,m} = 0, && \text{otherwise.} \end{aligned}$$

  When $\lambda_i = \langle\bar{\lambda}, m\rangle$ for $i \in \{1, \dots, d\}$, one speaks of an *internal resonance*.
- The following *mixed style*, that, given sets of indexes $I_1, \dots, I_N \subset \{1, \dots, n\}$, turns the sets $\{s_i = 0, i \in I_l\}$, $l = 1, \dots, N$, into invariant submanifolds:

$$\begin{aligned} &W^i_{k,m} = R^i_{k,m}, && f^i_{k,m} = 0, && \text{if } \exists l : \ i \in I_l \text{ and } m_j = 0 \ \forall j \in I_l, \\ &W^i_{k,m} = 0, && f^i_{k,m} = R^i_{k,m}, && \text{otherwise.} \end{aligned}$$

  This mixed style allows adapting the parameterization to the dynamics, as will be shown in the examples that follow.

Note that, as a whole, the order-k cohomological equation is linear and diagonal: each unknown monomial of the left-hand side is computed as a constant times the corresponding monomial of the right-hand side. All the computational effort goes in the evaluation of $\boldsymbol{R}_k$.

4.2 Efficiency Considerations

Once the $\boldsymbol{R}_k$ term is computed, the solution of the order-k cohomological equation with any of the styles previously mentioned is very fast. Assuming that we have explicit formulae for the vector field, as is the case in the RTBP, the evaluation of $\boldsymbol{R}_k$

as given in (34) depends on both being able to perform sums and products of dense[10] multivariate polynomials and being able to compose truncated (multivariate) power series into elementary functions such as sine or square root.

An strategy for an efficient implementation of the product of homogeneous polynomials is to represent them recursively with respect to the number of variables. A d-variate homogeneous polynomial of degree k can be represented as a linear combination of $(d-1)$-variate polynomials of degrees $k, k-1, \dots, 0$: for $s = (s_1, \dots, s_d)$, $\hat{s} = (s_1, \dots, s_{d-1})$,

$$f_k(s) = f_k^d(\hat{s}) + f_{k-1}^d(\hat{s})s_d + \cdots + f_0^d(\hat{s})s_d^k.$$

The memory representation can be made to mimic this recursive definition. The use of this strategy avoids the need for hash tables and reduces the product of homogeneous polynomials to dot products of vectors of coefficients.

With respect to the composition of truncated Taylor expansions into elementary functions, an efficient strategy is the use of a form of *automatic differentiation*[11] based on the notion of *radial derivative*. The radial derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$Rf(x) := \nabla f(x) \cdot x = \sum_{i=1}^{d} \frac{\partial f(x)}{\partial x_i} x_i$$

On an homogeneous polynomial of degree k, it satisfies

$$Rf_k(x) = kf_k(x).$$

It also satisfies a form of chain rule: for a function $\varphi : \mathbb{R} \to \mathbb{R}$

$$R(\varphi \circ f)(x) = \varphi'(f(x))\, Rf(x).$$

Now, if φ satisfies a differential equation, the previous two properties can be used to deduce a recurrence that relates the series expansions of f and $\varphi \circ f$. For instance, for

$$\varphi(x) = x^\alpha, \quad f = \sum_{k=0}^{k_{\max}} f_k, \quad f_0 \neq 0, \quad [\varphi \circ f]_{\le k_{\max}} =: p = \sum_{k=0}^{k_{\max}} p_k,$$

[10] As opposed to sparse.

[11] Here “automatic” is used in the sense of computing Taylor expansions in which the different terms are obtained through recurrences, instead of doing symbolic differentiation.

from $R(\varphi \circ f)(x) = \varphi'(f(x))\, Rf(x)$ and $x\varphi'(x) = \alpha\varphi(x)$, it follows that $p_0 = f_0^\alpha$ and

$$p_k(x) = \frac{1}{kf_0}\sum_{j=0}^{k-1}(\alpha(k-j)-j)\, f_{k-j}(x)p_j(x).$$

Using this recurrence, p_k can be computed from $f_1, \dots, f_{k-1}$ *and* $p_0, \dots, p_{k-1}$. This is, the terms of order $< k$ of $\varphi \circ f$ are also needed. Because of this, in order to proceed order by order in the computation $\boldsymbol{W}$, $\boldsymbol{f}$, we need to store the power series expansions of all the intermediate operations in the evaluation of $[\boldsymbol{F}(\boldsymbol{W}_{<k}(s))]_k$ that involve the composition of a power series with an elementary function. The software package in http://www.maia.ub.edu/dsg/param/ includes a C library for the manipulation of multivariate, truncated power series that implements all these ideas.

4.3 Error Estimation

Once we have computed

$$\boldsymbol{W}_{k\le k_{\max}} := \boldsymbol{W}_1 + \boldsymbol{W}_2 + \cdots + \boldsymbol{W}_{k_{\max}}, \quad \boldsymbol{f}_{k\le k_{\max}} := \boldsymbol{f}_1 + \boldsymbol{f}_2 + \cdots + \boldsymbol{f}_{k_{\max}}$$

up to a maximum order $k_{\max}$, we need to check the quality of these truncated expansions. For notational simplicity, we denote $\boldsymbol{W}_{k\le k_{\max}}$, $\boldsymbol{f}_{k\le k_{\max}}$ as $\boldsymbol{W}$, $\boldsymbol{f}$. For a specific initial condition $\boldsymbol{s}_0$ in parameter space, the following three error estimates are straightforward to check. Denote as $\boldsymbol{s}(t)$ the solution of $\dot{\boldsymbol{s}} = \boldsymbol{f}(\boldsymbol{x})$, $\boldsymbol{s}(0) = \boldsymbol{s}_0$, denote as $\boldsymbol{x}(t)$ the solution of $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, $\boldsymbol{x}(0) = \bar{\boldsymbol{W}}(\boldsymbol{s}_0)$, where $\bar{\boldsymbol{W}}$ is the parameterization of the manifold in original coordinates, as in (32), and choose an integration time T adequate for the problem under study. We can consider:

- The *error in the invariance equation*,

$$e_I(T, \boldsymbol{s}_0) = \sup_{t\in[0,T]} \|\boldsymbol{X}\big(\bar{\boldsymbol{W}}(\boldsymbol{s}(t))\big) - D\bar{\boldsymbol{W}}\big(\boldsymbol{s}(t)\big)\boldsymbol{f}\big(\boldsymbol{s}(t)\big)\|.$$

- The *error in the orbit*,

$$e_O(T, \boldsymbol{s}_0) = \sup_{t\in[0,T]} \|\bar{\boldsymbol{W}}\big(\boldsymbol{s}(t)\big) - \boldsymbol{x}(t)\|.$$

- If $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$ has a first integral H, the *error in the reduced first integral* $H \circ \bar{\boldsymbol{W}}$,

$$e_H(T, \boldsymbol{s}_0) = \sup_{t\in[0,T]} \|H\big(\bar{\boldsymbol{W}}(\boldsymbol{s}(t))\big) - H\big(\bar{\boldsymbol{W}}(\boldsymbol{s}_0)\big)\|$$

In the following we will use $e_O(T, s_0)$ for varying s_0 in order to determine *neighborhoods of validity* of the expansions obtained.

4.4 Expansions of the Center Manifold of the $L_{1,2}$ Collinear Points of the Earth-Moon RTBP

This subsection shows sample results on the RTBP for the Earth-Moon mass ratio given Eq. (29).

4.4.1 Using the Graph Style

The first example will be the computation of $W^c(L_1)$ using the graph style. Denote the vector field of the RTBP in Hamiltonian form as $\dot{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{x})$, and denote the eigenvalues of $D\boldsymbol{X}(L_1)$ as in Eq. (30). Denote as P a matrix having as columns eigenvectors of eigenvalues $\boldsymbol{i}\omega_p$, $-\boldsymbol{i}\omega_p$, $\boldsymbol{i}\omega_v$, $-\boldsymbol{i}\omega_v$, λ, $-\lambda$, in this order. For this example, we apply the procedure of Sect. 4.1 with $n = 6$, $d = 4$ to

$$\boldsymbol{Y}(\boldsymbol{y}) := P^{-1}\big(\boldsymbol{X}(L_1 + P\boldsymbol{y})\big),$$

using the graph style. In this way, we obtain a parameterization of the 4D center manifold of L_1 as

$$\boldsymbol{s} \longmapsto \bar{\boldsymbol{W}}(\boldsymbol{s}) := L_1 + P\boldsymbol{W}(\boldsymbol{s}), \tag{37}$$

with $W^i(\boldsymbol{s}) = s_i$, $i = 1, 2, 3, 4$. Expansions of $\boldsymbol{W}$ have been computed for several orders. Table 2 shows some computing times. Note that a 4-variate series truncated to order 70 has $\binom{4+70}{4} = 1, 150, 626$ coefficients.

Figure 15 shows the $\{s_4 = 0\}$ Poincaré sections of several trajectories at fixed energies. Note that each point in these plots uniquely determines a trajectory: s_3 is computed from s_1, s_2 and the (fixed) value of the energy. The Poincaré sections in Fig. 15 are analogous to the ones computed in [15, 25]. Since through the

Table 2 For several orders, computing times (in seconds) of the expansions of $W^c(L_1)$ for the Earth-Moon RTBP using the graph style, on a Mac with Intel Core Duo @ 2.16 GHz

10	20	30	40	50	60	70
7.790e-03	4.048e-01	5.497e+00	3.921e+01	1.900e+02	7.104e+02	2.207e+03

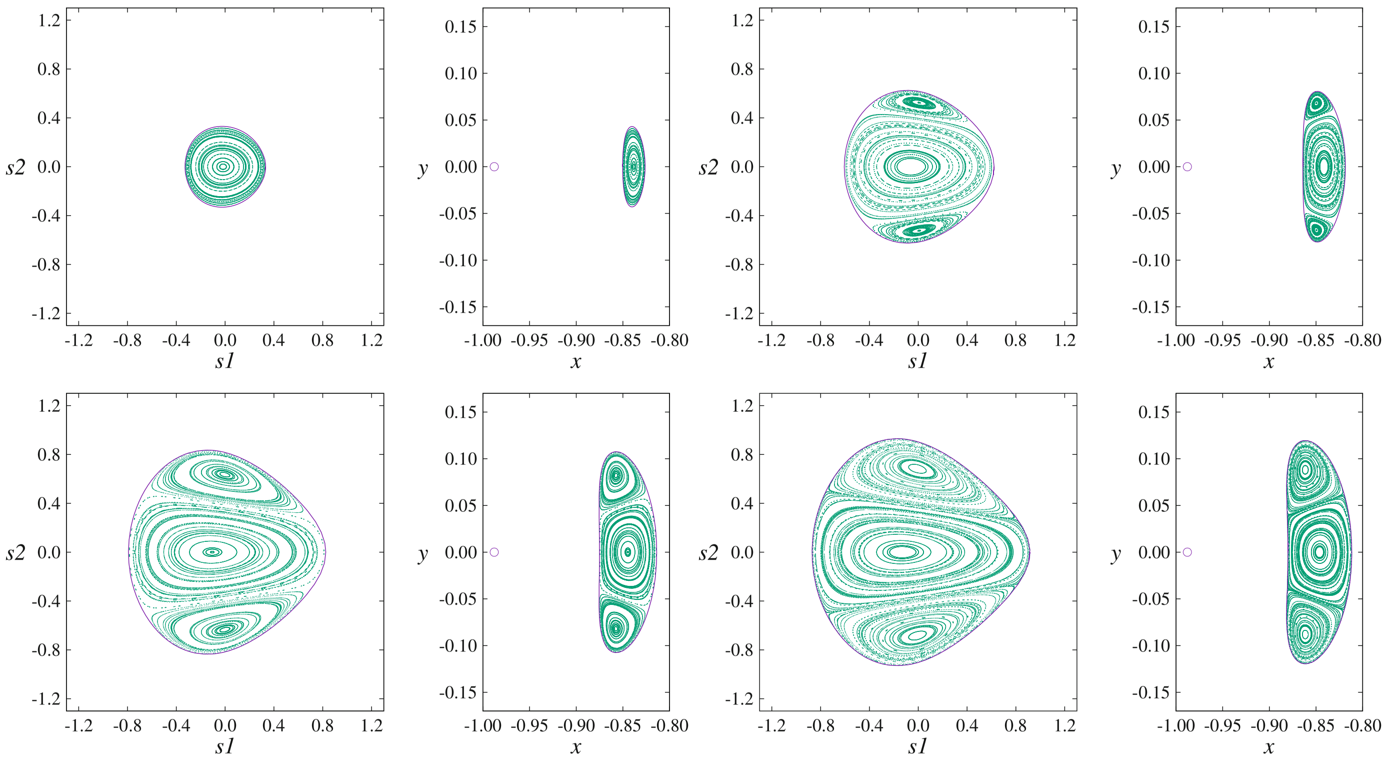


Fig. 15 Poincaré $\{s_4 = 0\}$ sections of $W^c(L_1)$ at several energies of the Earth-Moon RTBP, for the energies -1.59, -1.58, -1.570, -1.565. For each energy, we show both a plot in parameter space of the corresponding Poincaré section and a plot of the points of the previous plot converted to the original (synodic) coordinates through (37). Figure courtesy of A. Haro

parameterization (37) points with $s_4 = 0$ go to points with $z = 0$, the Poincaré sections in Fig. 15 are also analogous to the ones in Fig. 13. Note that they are obtained in completely different ways: here by direct numerical integration of $\dot{\boldsymbol{s}} = \boldsymbol{f}(\boldsymbol{s})$; there by computing individually every torus represented. Figure 13 can reach higher energies because of the numerical approach. Here, the use of the expansions is limited to their domain of validity. An estimation of this domain is shown next. Here, on the other hand, the numerical integration of $\dot{\boldsymbol{s}} = \boldsymbol{f}(\boldsymbol{s})$ allows us to capture all the dynamics in the center manifold at each energy level. In the numerical approach of Fig. 13, we can only display the objects that we individually compute.

4.4.2 Using Mixed Styles

In the next example, we have recomputed $W^c(L_1)$ with a mixed style parameterization with $N = 2$, $I_1 = \{1, 2\}$, $I_2 = \{3, 4\}$ The choice of P, n, d is the same as in the previous example. With this mixed style, due to the ordering of eigenvalues in (30), $\{\bar{\boldsymbol{W}}(s_1, s_2, 0, 0)\}_{s_1,s_2}$ describes the 2D manifold spanned by the family of planar Lyapunov orbits, whereas $\{\bar{\boldsymbol{W}}(0, 0, s_3, s_4)\}_{s_3,s_4}$ describes 2D manifold spanned by the family of vertical Lyapunov orbits. In particular, at each $\{s_4 = 0\}$ Poincaré section at a fixed energy level, the vertical Lyapunov orbit corresponds to the point with $s_1 = s_2 = 0$. One of such Poincaré sections is shown in Fig. 16.

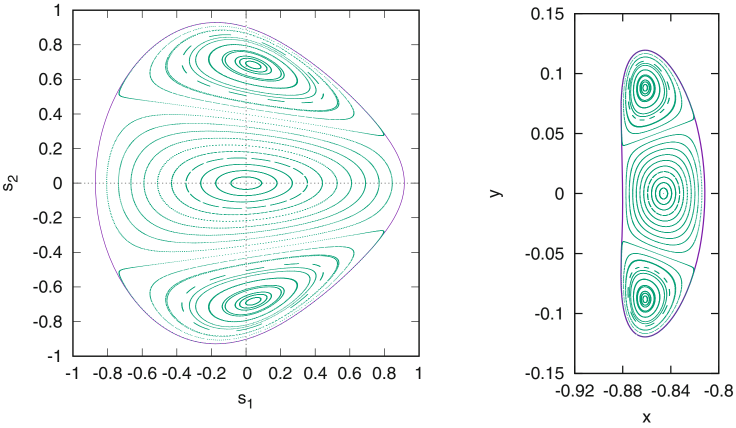


Fig. 16 Left: Poincaré $\{s_4 = 0\}$ section (in parameter space) of $W^c(L_1)$ for the fixed energy -1.565 of the Earth-Moon RTBP, computed with a mixed style parameterization with $I_1 = \{1, 2\}$, $I_2 = \{3, 4\}$. Observe that the vertical Lyapunov periodic orbit of this energy corresponds to $s_1 = s_2 = 0$. Right: conversion to the points in the left plot to the original (synodic) coordinates through (37)

This adaptation of the s_1, s_2, s_3, s_4 parameters to the dynamics allows us to choose easily initial conditions for a numerical exploration in order to determine the domain of validity of the expansions using the e_O estimate. For $s_2 > 0$, denote as $h(s_2)$ the energy of the planar Lyapunov orbit with this "s_2 amplitude", this is, $h(s_2) := H\big(\bar{\boldsymbol{W}}(0, s_2, 0, 0)\big)$. Then, for $s_2 > 0$ and $\alpha \in [-1, 1]$, define $\boldsymbol{s}(s_2, \alpha) := (0, \alpha s_2, s_3, 0)$, with s_3 chosen as to have $H\big(\boldsymbol{s}(s_2, \alpha)\big) = h(s_2)$. Denote also as T_{s_2} the maximum of the periods of the planar and vertical Lyapunov periodic orbits of energy $h(s_2)$. Then, for a trajectory with initial condition $\boldsymbol{s}(s_2, \alpha)$, we consider the error estimate

$$\varepsilon(s_2, \alpha) := e_O\big(T_{s_2}, \boldsymbol{s}(s_2, \alpha)\big). \tag{38}$$

Figure 17 shows the results on the evaluation of $\varepsilon(s_2, \alpha)$ at 100 values of s_2 and 100 values of α, for different orders of the expansions. In this figure it can be seen that there is not much improvement from order 30 on. Order 20 provides a precision of about 10^{-6} up to energy -1.57, whereas order 30 provides a precision of about 10^{-10} up to the same energy, and of about 10^{-6} up to energy -1.565.

As a final example, we have also computed the expansions of $W^c(L_2)$ with the same mixed style strategy. Figure 18 displays the Poincaré $\{s_4 = 0\}$ section of $W^c(L_2)$ at the fixed energy -1.570. Figure 19 displays the $\varepsilon(s_2, \alpha)$ error estimate for the expansions of orders 10, 20, 30. Compared to the expansions around L_1, the domain of validity is smaller, but the precision is about the same for the same energies. This is coherent with the fact that the energy of L_2 is larger than the one of L_1.

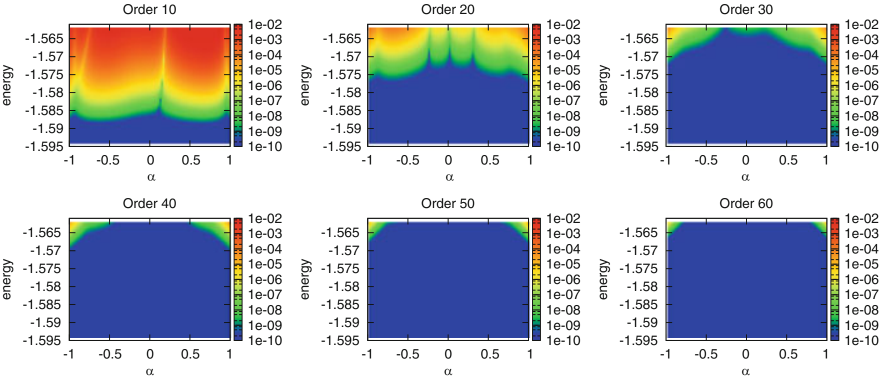


Fig. 17 For the expansions of $W^c(L_1)$ for the Earth-Moon RTBP, computed up to orders indicated, evaluation of the error estimate $\varepsilon(s_2, \alpha)$ of (38). Each plot has been generated for 100 values of s_2 (represented in the vertical axis as $h(s_2)$) and 100 values of α

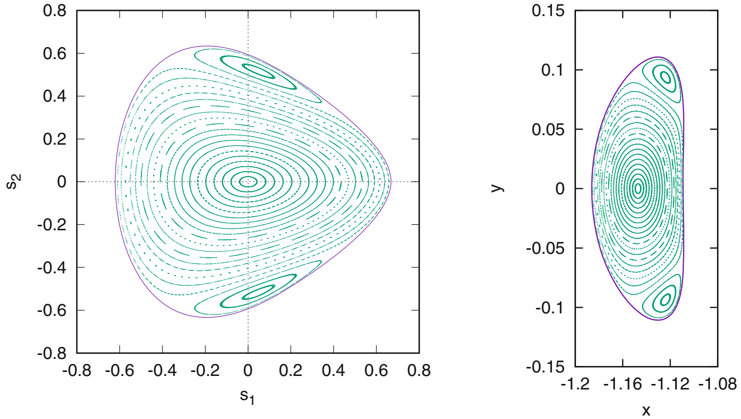


Fig. 18 Left: Poincaré section $\{s_4 = 0\}$ of $W^c(L_2)$ for the fixed energy -1.570 of the Earth-Moon RTBP, computed with a mixed style parameterization with $I_1 = \{1, 2\}$, $I_2 = \{3, 4\}$. Right: conversion to the points in the left plot to the original (synodic) coordinates through (37)

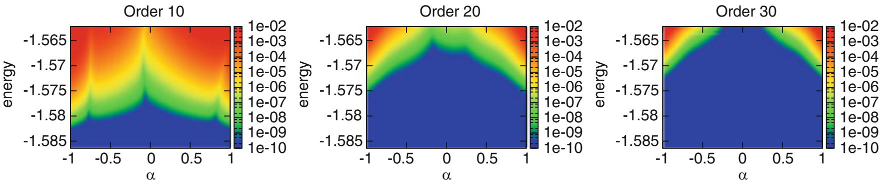


Fig. 19 For the expansions of $W^c(L_2)$ for the Earth-Moon RTBP, computed up to orders indicated, evaluation of the error estimate $\varepsilon(s_2, \alpha)$ of (38). The number of points in each plot and the interpretation of the axes is the same as in Fig. 17

5 Numerical Computation of Stable and Unstable Manifolds of Periodic Orbits and 2D Tori

In this section we will see how to compute numerically the linear approximation of stable and unstable manifolds of periodic orbits and tori. The linear approximation provides a local approximation with an error that is quadratic in the distance to the base object, which is adequate for many applications, including preliminary mission design. Approximations of higher order can be obtained through semi-analytical techniques, including the parameterization method, as will be discussed in Sect. 6. The Lindstedt-Poincaré method [29] is another semi-analytical alternative.

5.1 Invariant Manifolds of Periodic Orbits

Let $\boldsymbol{x}_0$ be an initial condition of a periodic orbit of period T, this is, $\boldsymbol{\phi}_T(\boldsymbol{x}_0) = \boldsymbol{x}_0$. A parameterization of the periodic orbit as an invariant manifold is given by the

2π-periodic function $\boldsymbol{\varphi} : [0, 2\pi] \longrightarrow \mathbb{R}^6$ defined as

$$\boldsymbol{\varphi}(\theta) := \boldsymbol{\phi}_{\frac{\theta}{2\pi}T}(\boldsymbol{x}_0).$$

Let $\Lambda \in \mathbb{R}$, $|\Lambda| \neq 1$ be an eigenvalue of the monodromy matrix of the periodic orbit with eigenvector $\boldsymbol{v}$, this is

$$D\boldsymbol{\phi}_T(\boldsymbol{x}_0)\boldsymbol{v} = \Lambda\boldsymbol{v}.$$

An eigenvalue Λ with $|\Lambda| > 1$ (resp. $|\Lambda| < 1$) would correspond to an unstable (resp. stable) manifold. For brevity, let us assume for the rest of the discussion that $\Lambda > 0$; a comment will be made on the case $\Lambda < 0$ at the end. Therefore, $\Lambda > 1$ (resp. $\Lambda < 1$) would correspond to an unstable (resp. stable) manifold. A parameterization of a set of vectors tangent to the unstable (resp. stable) manifold, also know as *unstable bundle* (resp. stable bundle), is given by the 2π-periodic function

$$\boldsymbol{v}(\theta) := \Lambda^{-\frac{\theta}{2\pi}} D\boldsymbol{\phi}_{\frac{\theta}{2\pi}T}(\boldsymbol{x}_0)\boldsymbol{v}.$$

By combining the two previous expressions, we can obtain a parameterization of the linear approximation of the unstable (resp. stable) manifold:

$$\bar{\boldsymbol{\psi}}(\theta, \xi) := \boldsymbol{\varphi}(\theta) + \xi\boldsymbol{v}(\theta). \tag{39}$$

It satisfies the approximate invariance equation

$$\boldsymbol{\phi}_t\big(\bar{\boldsymbol{\psi}}(\theta, \xi)\big) = \bar{\boldsymbol{\psi}}(\theta + t\omega, e^{t\lambda}\xi) + O(\xi^2),$$

for $\omega = \frac{2\pi}{T}$, $\lambda = \frac{\omega \ln \Lambda}{2\pi}$. It can thus be evaluated for small ξ only. Nevertheless, $\bar{\boldsymbol{\psi}}$ can be used to globalize the manifold by numerical integration while still providing a cylinder-like parameterization: for ξ not necessarily small, we can take an integer $m > 0$ (resp. $m < 0$) such that $\Lambda^{-m}\xi$ is small and compute

$$\bar{\boldsymbol{\Psi}}(\theta, \xi) = \boldsymbol{\phi}_{mT}\big(\bar{\boldsymbol{\psi}}(\theta, \Lambda^{-m}\xi)\big).$$

Figure 20 displays the Moon branch of the 2D unstable manifold of a Halo orbit globalized in this way until past its first periselene. Note that it is not represented as a set of trajectories but as a surface parameterized by the (θ, ξ) variables.

In the case $\Lambda < 0$, all the previous discussion is valid if we substitute T by $2T$ and Λ by Λ^2. In this way, $\boldsymbol{v}$ is 2π-periodic and the expressions for $\bar{\boldsymbol{\psi}}$, $\bar{\boldsymbol{\Psi}}$ still provide cylinder-like parameterizations.

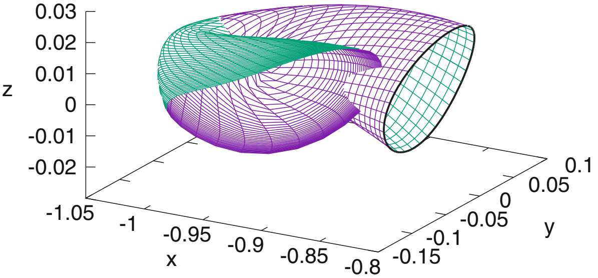


Fig. 20 Moon branch of the 2D unstable manifold of a halo orbit around L_1 of the Earth-Moon RTBP, represented as a surface. The halo orbit is shown in black

5.2 *Invariant Manifolds of 2D Tori*

We follow the discussion of [24] with a slightly modified computational strategy. Assume that $\boldsymbol{\varphi}$ parameterizes an invariant curve inside a 2D torus, as in Sect. 3.8.1,

$$\boldsymbol{\phi}_\Delta\big(\boldsymbol{\varphi}(\theta)\big) = \boldsymbol{\varphi}(\theta + \rho). \tag{40}$$

We want to find $\Lambda \in \mathbb{R}$, $|\Lambda| \neq 1$ and $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^6$, 2π-periodic, s.t.

$$D\boldsymbol{\phi}_\Delta\Big(\boldsymbol{\varphi}(\theta - \rho)\Big)\boldsymbol{u}(\theta - \rho) = \Lambda \boldsymbol{u}(\theta), \tag{41}$$

this is, an invariant bundle associated to the eigenvalue Λ. It will be an *unstable* (resp. stable) invariant bundle if $|\Lambda| > 1$ (resp. $|\Lambda| < 1$), that will be tangent to the unstable (resp. stable) manifold of the torus on the invariant curve parameterized by $\boldsymbol{\varphi}$.

Equation (41) can be compactly written as

$$\mathscr{C}\boldsymbol{u} = \Lambda \boldsymbol{u}, \tag{42}$$

with

$$(\mathscr{C}\boldsymbol{u})(\theta) = D\boldsymbol{\phi}_\Delta\big(\boldsymbol{\varphi}(\theta - \rho)\big)\boldsymbol{u}(\theta - \rho).$$

Assuming that $\boldsymbol{u}$ is expanded as a truncated Fourier series, the eigenvalue problem (42) can be discretized and thus converted in a finite-dimensional matrix-vector eigenvalue problem by approximating the Fourier coefficients of $\mathscr{C}\boldsymbol{u}$ by their Discrete Fourier Transform (DFT).

We use the following notation for the DFT: for N even, given real data $\{f_j\}_{j=0}^{N-1}$, we denote

$$F_{\{f_j\}_{j=0}^{N-1}}(k) := \frac{1}{N}\sum_{j=0}^{N-1} f_j e^{-\boldsymbol{i}2\pi \frac{k}{N} j}, \qquad k = 0, \dots, N-1,$$

$$A_{\{f_j\}_{j=0}^{N-1}}(k) := \frac{\delta_k}{N}\sum_{j=0}^{N-1} f_j \cos(2\pi \frac{k}{N} j), \qquad k = 0, \dots, N/2,$$

$$B_{\{f_j\}_{j=0}^{N-1}}(k) := \frac{2}{N}\sum_{j=0}^{N-1} f_j \sin(2\pi \frac{k}{N} j), \qquad k = 1, \dots, N/2-1,$$

with $\delta_0 = \delta_{\frac{N}{2}} = 1$, $\delta_k = 2$ for $k = 1, \dots, \frac{N}{2} - 1$. If the data comes from the regular sampling of a 2π-periodic function, this is, $f_j = f(\theta_j)$ for $\theta_j = j2\pi/N$ and f is 2π-periodic,

$$f(\theta) \approx A_{\{f_j\}_{j=0}^{N-1}}(0) + \sum_{k=0}^{N/2}\Big(A_{\{f_j\}_{j=0}^{N-1}}(k)\cos(k\theta) + B_{\{f_j\}_{j=0}^{N-1}}(k)\sin(k\theta)\Big) + A_{\{f_j\}_{j=0}^{N-1}}(N/2)\cos((N/2)\theta),$$

where the approximation is an equality if $\theta = \theta_j$, $0 \le j \le N-1$. In this way, the DFT coefficients provide an approximation of the Fourier coefficients (for a bound on the difference, see e.g. [11, 17]).

Now, for

$$\boldsymbol{u}(\theta) = \boldsymbol{A}_0 + \sum_{k=1}^{N/2-1}\Big(\boldsymbol{A}_k\cos(k\theta) + \boldsymbol{B}_k\sin(k\theta)\Big) + \boldsymbol{A}_{N/2}\cos((N/2)\theta),$$

let us denote the DFT coefficients of $(\mathscr{C}\boldsymbol{u})(\theta)$ by $\{\bar{\boldsymbol{A}}_k\}_{k=0}^{N/2}$, $\{\bar{\boldsymbol{B}}_k\}_{k=1}^{N/2-1}$, this is,

$$(\mathscr{C}\boldsymbol{u})(\theta) \approx \bar{\boldsymbol{A}}_0 + \sum_{k=1}^{N/2-1}\Big(\bar{\boldsymbol{A}}_k\cos(k\theta) + \bar{\boldsymbol{B}}_k\sin(k\theta)\Big) + \bar{\boldsymbol{A}}_{N/2}\cos((N/2)\theta).$$

If we denote

$$\boldsymbol{x} = \big(\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{B}_1, \dots, \boldsymbol{A}_{N/2-1}, \boldsymbol{B}_{N/2-1}, \boldsymbol{A}_{N/2}\big),$$
$$\bar{\boldsymbol{x}} = \big(\bar{\boldsymbol{A}}_0, \bar{\boldsymbol{A}}_1, \bar{\boldsymbol{B}}_1, \dots, \bar{\boldsymbol{A}}_{N/2-1}, \bar{\boldsymbol{B}}_{N/2-1}, \bar{\boldsymbol{A}}_{N/2}\big),$$

then, for a suitable (finite-dimensional) matrix C,

$$\bar{\boldsymbol{x}} = C\boldsymbol{x}. \tag{43}$$

The columns of C can be found as the DFT coefficients of the operator $\mathscr{C}$ applied to the canonical basis elements in $\boldsymbol{x}$ space, this is, to the functions $\boldsymbol{w}_k$, $\boldsymbol{w}_k \cos(\theta)$, $\boldsymbol{w}_k \sin(\theta)$, $\boldsymbol{w}_k \cos(2\theta)$, $\boldsymbol{w}_k \sin(2\theta)$, etc., being $\boldsymbol{w}_k \in \mathbb{R}^6$ the k-th element of the canonical basis, $k = 1, \dots, 6$. Since all these functions can be written in terms of complex exponentials of the form $e^{ik\theta}$, the coefficients of the C matrix can be computed from

$$F_{\{D\boldsymbol{\phi}_{\Delta}(\varphi(\theta_l-\rho))\boldsymbol{w}_j e^{ik(\theta_l-\rho)}\}_{l=0}^{N-1}}(m),$$

which, after a few calculations, is found to be

$$e^{-ik\rho} F_{\{D\boldsymbol{\phi}_{\Delta}(\varphi(\theta_l-\rho))\boldsymbol{w}_j\}_{l=0}^{N-1}}(m-k),$$

for $j = 1, \dots, 6$ and $k, m = 0, \dots, N/2$. Since $D\boldsymbol{\phi}_{\Delta}(\varphi(\theta_l - \rho))\boldsymbol{w}_j$ is a 6-vector for each j, the computation of all the needed values of the previous expression is reduced to 36 DFT, which, by using FFT, are computed in $O(N \log N)$ operations each.

Some knowledge on the structure of the spectrum of the invariant bundle we are looking for is necessary in order to choose the right eigenvalues of the C matrix of (43). The eigenvalues of C appear grouped in circles. Since the tori we are looking for are reducible, there are as many circles as eigenvalues of the reduced matrix (which can be considered analogous to the monodromy matrix of a periodic orbit). Assuming that (41) has a solution, from the fact that the RTBP is a Hamiltonian system, apart from unit circles there will be a circle containing Λ and another circle containing Λ^{-1}. These are the ones we are interested in. The corresponding eigenvectors provide the Fourier coefficients of the invariant bundles we are looking for. More details on this discussion and some additional considerations on the accuracy of the computed eigenvalues can be found in [24].

Now, from an invariant stable or unstable bundle $\boldsymbol{u}(\theta)$, tangent to the stable or unstable manifold of the torus on the invariant curve $\boldsymbol{\varphi}(\theta)$, we can obtain the invariant bundle tangent to the stable or unstable manifold of the torus on the whole torus through

$$\boldsymbol{v}(\theta_1, \theta_2) = \Lambda^{-\frac{\theta_2}{2\pi}} D\boldsymbol{\phi}_{\frac{\theta_2}{2\pi}\Delta}\Big(\boldsymbol{\varphi}\big(\theta_1 - \frac{\theta_2}{2\pi}\rho\big)\Big)\boldsymbol{u}\big(\theta_1 - \frac{\theta_2}{2\pi}\rho\big).$$

This expression assumes $\Lambda > 0$. If this is not the case, Δ needs to be changed to 2Δ, so Eqs. (40) and (41) are satisfied with ρ substituted by 2ρ and Λ by Λ^2. Defined as above, the $\boldsymbol{v}$ function is 2π-periodic in θ_1, θ_2 and satisfies

$$D\boldsymbol{\phi}_t\big(\boldsymbol{\psi}(\theta_1, \theta_2)\big)\boldsymbol{v}(\theta_1, \theta_2) = \Lambda^{\frac{t\omega_2}{2\pi}} \boldsymbol{v}(\theta_1 + t\omega_1, \theta_2 + t\omega_2),$$

where $\boldsymbol{\psi}$ is the parameterization of the 2D torus defined in Eq. (21), $\omega_1 = \rho/\Delta$ and $\omega_2 = 2\pi/\Delta$. From the parameterization of the stable or unstable bundle defined on the whole torus, we can write a parameterization of the linear approximation of the stable or unstable manifold of the torus as

$$\bar{\boldsymbol{\psi}}(\theta_1, \theta_2, \xi) = \boldsymbol{\psi}(\theta_1, \theta_2) + \xi \boldsymbol{v}(\theta_1, \theta_2), \tag{44}$$

which is 2π-periodic in θ_1, θ_2 and satisfies the approximate invariance equation

$$\boldsymbol{\phi}_t\big(\bar{\boldsymbol{\psi}}(\theta_1, \theta_2, \xi)\big) = \bar{\boldsymbol{\psi}}(\theta_1 + t\omega_1, \theta_2 + t\omega_2, e^{t\lambda}\xi) + O(\xi^2),$$

for $\omega_1 = \rho/\Delta$, $\omega_2 = 2\pi/\Delta$, $\lambda = \omega_2 \ln \Lambda/(2\pi)$, and thus Eq. (44) can be evaluated for small ξ only. For ξ not necessarily small, we can consider an integer m such that $\Lambda^{-m}\xi$ is small ($m > 0$ for the unstable manifold, $m < 0$ for the stable manifold) and compute

$$\bar{\boldsymbol{\Psi}}(\theta_1, \theta_2, \xi) = \boldsymbol{\phi}_{m\Delta}\Big(\bar{\boldsymbol{\psi}}(\theta_1 - m\rho, \theta_2, \Lambda^{-m}\xi)\Big).$$

6 Semi-Analytical Computation of Stable and Unstable Manifolds Using the Parameterization Method

We have seen how the parameterization method can be used as a semi-analytical technique in order to find the periodic orbits and tori in the center manifold of a collinear libration point. Without any modification, it can also be used to find the invariant stable and unstable manifolds of these trajectories. All the unstable manifolds of the invariant objects of $W^c(L_1)$ are contained in the center-unstable manifold of L_1, $W^{cu}(L_1)$, which is an invariant manifold tangent to the directions given by the eigenvectors with eigenvalues

$$\boldsymbol{i}\omega_p, -\boldsymbol{i}\omega_p, \boldsymbol{i}\omega_v, -\boldsymbol{i}\omega_v, \lambda,$$

where we have recovered the notation of Eq. (30). All the stable manifolds of the invariant objects of $W^c(L_1)$ are contained in the center-stable manifold of L_1, $W^{cs}(L_1)$, which is the invariant manifold tangent to the directions given by the eigenvectors with eigenvalues

$$\boldsymbol{i}\omega_p, -\boldsymbol{i}\omega_p, \boldsymbol{i}\omega_v, -\boldsymbol{i}\omega_v, -\lambda.$$

The parameterization method does not need any modification to compute $W^{cu}(L_1)$ or $W^{cs}(L_1)$ instead of $W^c(L_1)$.

As an example, we can apply the procedure described in Sect. 4.1 with the same choice of P as in Sect. 4.4, $n = d = 6$ and choosing a mixed style parameterization

Table 3 Sets of indexes used for the mixed style reparameterization of the neighborhood of L_1 of the Earth-Moon RTBP

l	I_l	Submanifold described by $s_i = 0, i \in I_l$
1	$\{1, 2, 3, 4, 6\}$	The unstable manifold of L_1
2	$\{1, 2, 3, 4, 5\}$	The stable manifold of L_1
3	$\{1, 2, 3, 4\}$	The hyperbolic normal part of L_1
4	$\{3, 4, 5, 6\}$	The planar Lyapunov family of periodic orbits
5	$\{3, 4, 6\}$	The unstable manifold of the planar Lyapunov family
6	$\{3, 4, 5\}$	The stable manifold of the planar Lyapunov family
7	$\{3, 4\}$	The normal hyperbolic part of the planar Lyapunov family
8	$\{1, 2, 5, 6\}$	The vertical Lyapunov family of periodic orbits
9	$\{1, 2, 6\}$	The unstable manifold of the vertical Lyapunov family
10	$\{1, 2, 5\}$	The stable manifold of the vertical Lyapunov family
11	$\{1, 2\}$	The normal hyperbolic part of the vertical Lyapunov family
12	$\{5, 6\}$	The center manifold of L_1
13	$\{6\}$	The center-unstable manifold of L_1
14	$\{5\}$	The center-stable manifold of L_1

Table 4 For several orders, computing times of the expansions of the mixed style reparameterization of the neighborhood of L_1 of the Earth-Moon RTBP

$k_{\max}$	10	15	20	25	30
Time (s)	0.48	6.66	64.83	470.37	1311.42

with the sets of indexes $I_1, \ldots, I_{14}$ defined by Table 3. In this way, we obtain a reparameterization of a whole neighborhood of L_1 that is completely adapted to the dynamics. Table 3 is the recipe to choose initial conditions on the different kind of objects. For example, points of the form $\bar{\boldsymbol{W}}(0, 0, 0, 0, s_5, 0)$ are in the unstable manifold of L_1 because of the use of I_1, whereas points of the form $\bar{\boldsymbol{W}}(0, 0, s_3, s_4, 0, s_6)$ are in the stable manifold of the vertical Lyapunov family of periodic orbits because of the use of I_{10}. Table 4 shows the computing times of the expansions for several orders. These times are now larger than the ones of Sect. 4.4 because the truncated power series have 6 variables instead of 4.

As before, it is necessary to determine a neighborhood of validity of the expansions. This has been done in Fig. 21, by an exploration similar to the one done in Sect. 4.4.2, but now taking initial conditions with $s_5, s_6 \neq 0$ in evaluation of the e_O estimate, and also integrating both forward and backward in time, in order to test both the stable and the unstable manifold. The maximum e_O of the trajectories of each energy tested are represented by a point in Fig. 21. The pairs of green-violet curves correspond, from left to right, to orders 10, 15, 20, 25, 30. The full details of the exploration can be found in [21].

A sample application of the use of these expansions is the generation of what are known as transit and non-transit trajectories [7, 8]. With the choice of the eigenvectors corresponding to $\pm\lambda$ shown schematically in Fig. 22, orbits with $s_5 s_6 > 0$ are *transit* in the sense that go from the Earth to the Moon or vice-versa.

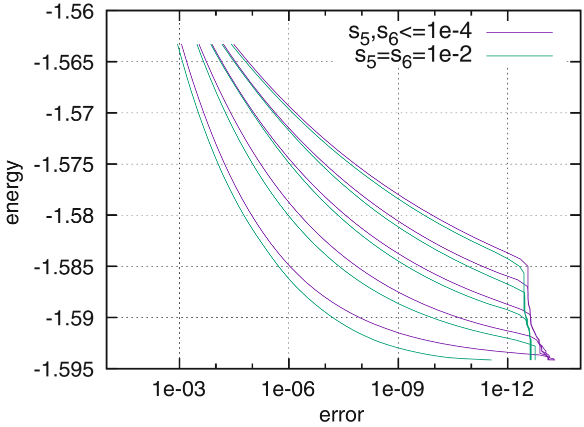


Fig. 21 Error estimates for the mixed-style reparameterization of the neighborhood of L_1 of the Earth-Moon RTBP

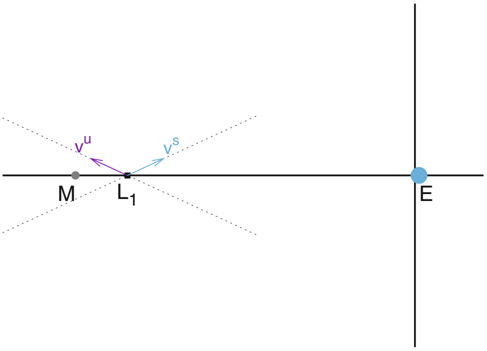


Fig. 22 Schematic representation of the choice of the hyperbolic eigenvectors of L_1, in order to produce transit and non-transit trajectories

Orbits with $s_5 s_6 < 0$, however, are *non-transit* in the sense that after departing from a primary they "bounce back" to it. Figure 23 shows some trajectories used in the evaluation of the error estimate of Fig. 21, which are all transit because they were chosen with $s_5 = s_6 > 0$. For clarity, the trajectories are not integrated as in the evaluation of the error estimate, but forward in time up to the first cut with $x = \mu - 1 + R_M$, where R_M is radius of the Moon in dimensionless units (red trajectories), and backward in time up to the second cut with $y = 0$ after the first passage behind the Earth (blue trajectories). Looking at each blue curve followed by the red one as a single trajectory, the plots show that all of them are Earth-Moon transit.

7 Computation of Homoclinic and Heteroclinic Connections

An homoclinic connection of a object (with itself) is a trajectory that tends to the object both forward and backward in time. An heteroclinic connection of a departing object and an arrival object is a trajectory that tends to the departing object backward

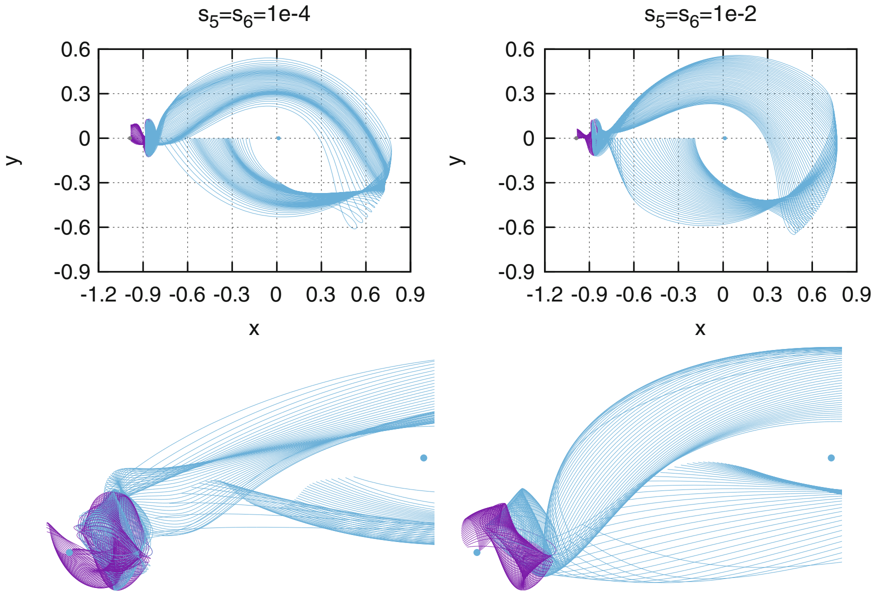


Fig. 23 Some transit trajectories associated to the L_1 collinear point of the Earth-Moon RTBP. The plot of the second line are the $3D$ views of the ones of the first line

in time and to the arrival object forward in time. From the dynamical systems point of view, these connections play a fundamental role in the global organization of the dynamics. In the RTBP, they also provide low-energy transfers between objects [13] and resonance transitions [16, 28]. Using Conley-McGehee tubes [7, 30] inside Hill's regions, they allow to prescribe itineraries between the interior and exterior regions of a moon-planet system, as in [16, 28, 34].

7.1 Computing Individual Connections

Consider $\boldsymbol{\psi}^u(\theta, \xi)$ a parameterization of an approximation of the unstable manifold of a departure object, and $\boldsymbol{\psi}^s(\theta, \xi)$ a parameterization of an approximation the stable manifold of an arrival object. These approximations can be the linear ones, or of higher order. Let $\Sigma = \{g(\boldsymbol{x}) = 0\}$ be a Poincaré section intersected by the manifolds, and consider two associated Poincaré maps: P^+_Σ, computed integrating forward in time, and P^-_Σ, integrating backward in time. This is,

$$\boldsymbol{P}^+_\Sigma(\boldsymbol{x}) = \boldsymbol{\phi}_{\tau^+(\boldsymbol{x})}(\boldsymbol{x}), \quad \boldsymbol{P}^-_\Sigma(\boldsymbol{x}) = \boldsymbol{\phi}_{\tau^-(\boldsymbol{x})}(\boldsymbol{x}), \tag{45}$$

where the functions $\tau^+(\boldsymbol{x})$, $\tau^-(\boldsymbol{x})$ are time-return maps with $\tau^+(\boldsymbol{x}) > 0$, $\tau^-(\boldsymbol{x}) < 0$ defined implicitly by the conditions

$$g\big(\boldsymbol{\phi}_{\tau^+(\boldsymbol{x})}(\boldsymbol{x})\big) = g\big(\boldsymbol{\phi}_{\tau^-(\boldsymbol{x})}(\boldsymbol{x})\big) = 0.$$

The intersections of homoclinic (if the departure and arrival objects are the same) or heteroclinic (in the case of different departure and arrival objects) connections with the section Σ are given by the zeros of the function

$$\boldsymbol{F}(\boldsymbol{\theta}^u, \boldsymbol{\theta}^s) = \boldsymbol{P}^+_\Sigma(\boldsymbol{\psi}^u(\boldsymbol{\theta}^u, \xi)) - \boldsymbol{P}^-_\Sigma(\boldsymbol{\psi}^s(\boldsymbol{\theta}^s, \xi)). \tag{46}$$

In this function, ξ is a fixed parameter, that needs to be taken small if $\boldsymbol{\psi}^u$, $\boldsymbol{\psi}^s$ are linear approximations, or not necessarily, if $\boldsymbol{\psi}^u$, $\boldsymbol{\psi}^s$ are approximations of higher order. The $\boldsymbol{\theta}^u, \boldsymbol{\theta}^s$ parameters are vectors of phases of the same dimension of the connecting objects (scalars for periodic orbits, 2-vectors for 2D tori).

In the case of periodic orbits, their stable and unstable manifolds are locally diffeomorphic to 2D cylinders. As long as this remains true when globalizing their manifolds, the computation of connections is reduced to intersect 2D tubes, which can be visualized without much difficulty. Their visualization is particularly simple if the orbit is planar: the planar RTBP has 2 degrees of freedom, and therefore a Poincaré section of fixed energy is 2D. Figure 24 shows the manifold tubes of a planar Lyapunov orbit around L_1 of the Earth-Moon RTBP, and also their intersection with $\Sigma := \{x = \mu - 1\}$. The two points of intersection of the two curves coming from the sections of the manifold tubes with Σ (Fig. 24 right) give rise to two homoclinic connections. Initial conditions in order to find zeros of the function $\boldsymbol{F}$ of Eq. (46) via Newton iterations can be obtained from this plot. Care must be taken with the number of cuts of the manifold that define the Poincaré maps: according to Fig. 24 left, $\boldsymbol{P}^+_\Sigma$ is defined as the second cut with Σ, whereas $\boldsymbol{P}^-_\Sigma$ is defined as the first cut.

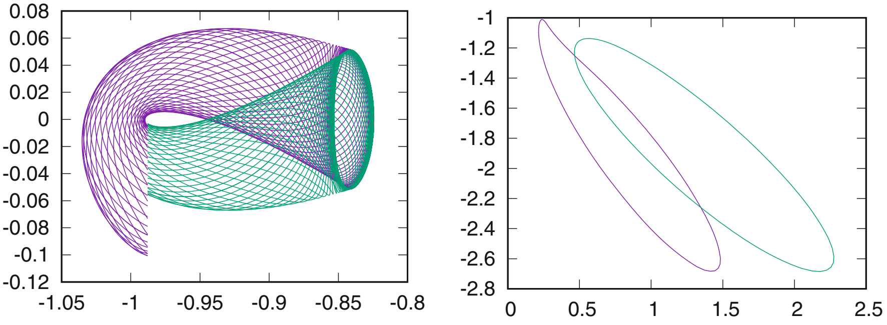


Fig. 24 Left: manifold tubes (green: stable, violet: unstable) of a planar Lyapunov orbit around L_1 of the Earth-Moon RTBP. Right: intersection of the manifold tubes with the section $\Sigma = \{x = \mu - 1\}$. The coordinates are: x, y in the left plot, p_x, p_y in the right one

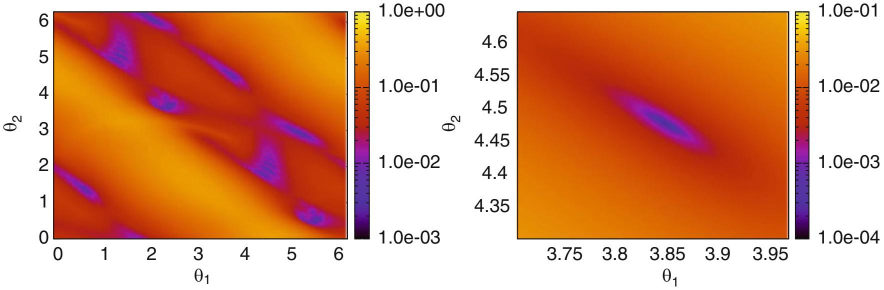


Fig. 25 Plot in order to locate heteroclinic connections between the Lissajous torus of the Earth-Moon RTBP around L_1 with $h = -1.58$, $\bar{\rho} := 0.2800082$ and the one around L_2 with the same energy and $\tilde{\rho} := 0.1700025$. The right plot is a zoom of the left one

In the cases in which the sections of the manifold tubes with Σ are not easy to visualize, other approaches need to be followed. As an example, consider searching for heteroclinic connections between:

- a Lissajous torus around L_1 of the Earth-Moon RTBP with energy $\bar{h} := -1.58$ and rotation number $\bar{\rho} := 0.2800082$, and
- a Lissajous torus around L_2 with the same energy and rotation number $\tilde{\rho} := 0.1700025$.

Denote as $\boldsymbol{\Psi}^u(\theta_1, \theta_2, \xi)$ (resp. $\boldsymbol{\Psi}^s(\theta_1, \theta_2, \xi)$) a parameterization of the linear approximation of the unstable (resp. stable) manifold of the departing (resp. arrival) torus. Denote also as P^+_Σ, P^-_Σ the Poincaré sections defined in Eq. (45) after the needed number of cuts with the section. Then, in order to look for connections, we can plot in terms of θ_1, θ_2 the function

$$\min_{\bar{\theta}_1, \bar{\theta}_2 \in [0, 2\pi]} \operatorname{dist}\big(P^+_\Sigma(\boldsymbol{\Psi}^u(\theta_1, \theta_2)),\ P^-_\Sigma(\boldsymbol{\Psi}^s(\bar{\theta}_1, \bar{\theta}_2))\big).$$

This is done in Fig. 25. The heteroclinic connection corresponding to the zoom in the right plot of this figure is shown in Fig. 26.

7.2 *Continuation of Connections*

Since the RTBP is a Hamiltonian system, periodic orbits and tori are not isolated but part of families. As a consequence, the connections between them are part of families too. If we want to compute several connections along a family, it is a tedious procedure to compute them individually as described before.

The process of computing homoclinic or heteroclinic connections along families can be automated by the use of continuation on Eq. (46), by letting $\boldsymbol{\psi}^u(\theta^u, \xi)$ and $\boldsymbol{\psi}^s(\theta^s, \xi)$ evolve freely along the families of departing and arrival objects. The

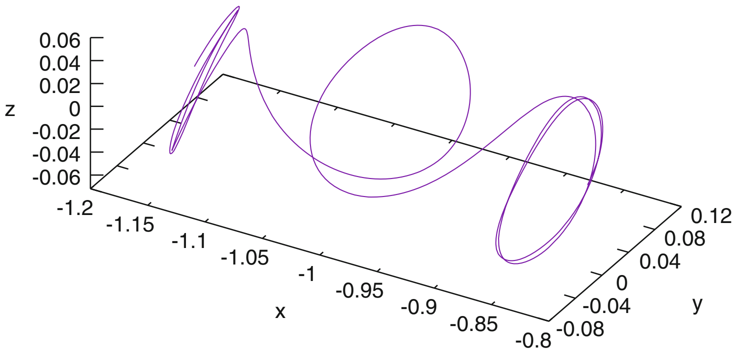


Fig. 26 Heteroclinic connection corresponding to the zoom in the right plot of Fig. 25

actual way to do it depends on the way that $\boldsymbol{\psi}^u$, $\boldsymbol{\psi}^s$ have been obtained, that can be semi-analytical or numerical. In the following we will focus in the numerical approach.

Assume we wanted to numerically compute a family of homoclinic connections of periodic orbits of the RTBP by continuation. Let $\Sigma_1 = \{g_1(\boldsymbol{x}) = 0\}$ be a Poincaré section for the initial conditions of the periodic orbit, and $\Sigma_2 = \{g_2(\boldsymbol{x}) = 0\}$ a Poincaré section used to match the invariant manifolds of the periodic orbit. Assume these Poincaré sections are valid along the portion of the family we want to compute. We need to consider as unknown everything necessary to determine a periodic orbit of the family and its homoclinic connection: the value of the energy, h, the initial condition of the periodic orbit, $\boldsymbol{x}_0$, the eigenvalue of the monodromy matrix related to the unstable (resp. stable) manifold, $\boldsymbol{v}^u$ (resp. $\boldsymbol{v}^s$), the departing (resp. arriving) phase on the linear approximation of the unstable (resp. stable) manifold, θ^u (resp. θ^s), and, finally, the time of flight from the linear approximation of the unstable (resp. stable) manifold to the surface of section in which the manifolds are intersected, T^u (resp. T^s). The system of equations needs to impose all the conditions for h, $\boldsymbol{x}$, T, Λ^u, $\boldsymbol{v}^u$, Λ^s, $\boldsymbol{v}^s$, θ^u, T^u, θ^s, T^s to determine a periodic orbit and an homoclinic connection of it. It would thus be

$$
\begin{gathered}
H(\boldsymbol{x}) - h = 0,\\
g_1(\boldsymbol{x}) = 0,\\
\boldsymbol{\phi}_T(\boldsymbol{x}) - \boldsymbol{x} = 0,\\
\|\boldsymbol{v}^u\|^2 - 1 = 0, \qquad\qquad \|\boldsymbol{v}^s\|^2 - 1 = 0,\\
D\boldsymbol{\phi}_T(\boldsymbol{x})\boldsymbol{v}^u - \Lambda^u\boldsymbol{v}^u = 0, \qquad D\boldsymbol{\phi}_T(\boldsymbol{x})\boldsymbol{v}^s - \Lambda^s\boldsymbol{v}^s = 0,\\
g_2\Big(\boldsymbol{\phi}_{T^u}\big(\boldsymbol{\psi}^u(\theta^u,\xi)\big)\Big) = 0,\\
g_2\Big(\boldsymbol{\phi}_{T^s}\big(\boldsymbol{\psi}^s(\theta^s,\xi)\big)\Big) = 0,\\
\boldsymbol{\phi}_{T^u}\big(\boldsymbol{\psi}^u(\theta^u,\xi)\big) - \boldsymbol{\phi}_{T^s}\big(\boldsymbol{\psi}^s(\theta^s,\xi)\big) = 0,
\end{gathered}
\tag{47}
$$

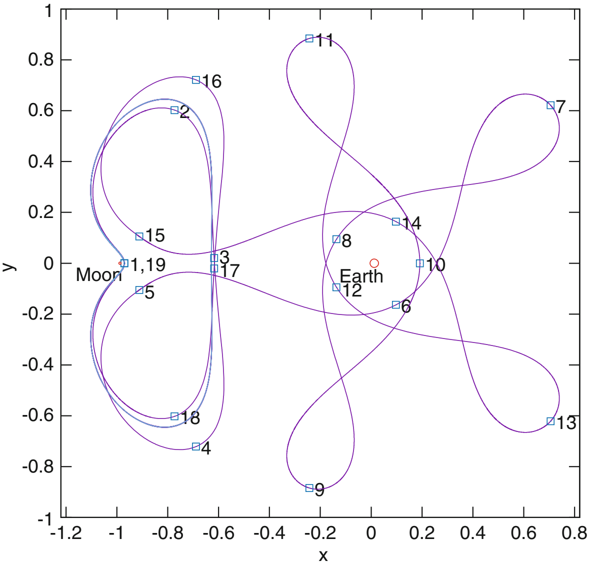


Fig. 27 An homoclinic connection (in violet) of a large planar Lyapunov orbit (in blue) of the Earth-Moon RTBP, obtained by numerical continuation

with, according to (39),

$$\boldsymbol{\psi}^j(\theta, \xi) = \boldsymbol{\phi}_{\frac{\theta}{2\pi}T}(\boldsymbol{x}) + \xi(\Lambda^j)^{\frac{\theta}{2\pi}} D\boldsymbol{\phi}_{\frac{\theta}{2\pi}T}(\boldsymbol{x})\boldsymbol{v}^j,$$

for $j = u, s$. Note that the system (47) includes a normalization condition on $\boldsymbol{v}^u$, $\boldsymbol{v}^s$, in order to make them to be locally unique. Also observe that, since we use the linear approximation of the manifolds, ξ is a parameter that must be kept fixed at a small value, e.g. 10^{-6}. An actual implementation requires multiple shooting, both in the periodic orbit and in the connection. Additional details can be found in [4]. Figure 27 displays a homoclinic connection (in violet) of a large planar Lyapunov orbit (in blue) around L_1 of the Earth-Moon RTBP that has been reached by such a continuation procedure. In order to aid visualization, all the perigees, apogees, periselenes, and aposelenes have been numbered as their appear along the connection.

The same ideas can be used in order to perform continuation of connections of tori. Assume we wanted to perform continuation of heteroclinic connections of tori of the RTBP. Consider a Poincaré section Σ in order to match the stable and unstable manifolds, and assume that it is valid along all the portion of the family of connections we want to continue. As unknowns, we would need to consider all the data determining the departing and arrival tori and the connection. This would be:

- The energy, h.
- The data of the departing torus: its "longitudinal period", Δ^u; its rotation number, ρ^u; the Fourier coefficients of the parameterization of its invariant curve, $\boldsymbol{\varphi}^u$; the eigenvalue of its unstable bundle, Λ^u; its unstable bundle, $\boldsymbol{u}^u$; the departing phases of the connection, θ_1^u, θ_2^u; and the time of flight from the manifold to the Poincaré section Δ_*^u.
- The analogous data for the arrival torus: Δ^s, ρ^s, $\boldsymbol{\varphi}^s$, Λ^s, $\boldsymbol{u}^s$, θ_1^s, θ_2^s, Δ_*^s.

As before, the system of equations needs to impose all the conditions for h, Δ^u, ρ^u, $\boldsymbol{\varphi}^u$, Λ^u, $\boldsymbol{u}^u$, θ_1^u, θ_2^u, Δ_*^u, Δ^s, ρ^s, $\boldsymbol{\varphi}^s$, Λ^s, $\boldsymbol{u}^s$, θ_1^s, θ_2^s, Δ_*^s to determine two invariant tori and an heteroclinic connection between them. It would thus be

$$
\begin{aligned}
H(\boldsymbol{\varphi}^u(0)) - h &= 0, & H(\boldsymbol{\varphi}^s(0)) - h &= 0,\\
\boldsymbol{\phi}_{\Delta^u}(\boldsymbol{\varphi}^u(\theta)) - \boldsymbol{\varphi}^u(\theta + \rho^u) &= 0, & \boldsymbol{\phi}_{\Delta^s}(\boldsymbol{\varphi}^s(\theta)) - \boldsymbol{\varphi}^s(\theta + \rho^s) &= 0,\\
\boldsymbol{v}^u(0) \cdot \boldsymbol{v}^u(0) - 1 &= 0, & \boldsymbol{v}^s(0) \cdot \boldsymbol{v}^s(0) - 1 &= 0,\\
D\boldsymbol{\phi}_{\Delta^u}(\boldsymbol{\varphi}^u(\theta))\boldsymbol{v}^u(\theta) - \Lambda^u \boldsymbol{v}^u(\theta + \rho^u) &= 0, & D\boldsymbol{\phi}_{\Delta^s}(\boldsymbol{\varphi}^s(\theta))\boldsymbol{v}^s(\theta) - \Lambda^s \boldsymbol{v}^s(\theta + \rho^s) &= 0,\\
g\big(\boldsymbol{\phi}_{\Delta_*^u}(\boldsymbol{\Psi}^u(\theta_1^u, \theta_2^u))\big) &= 0, & g\big(\boldsymbol{\phi}_{\Delta_*^s}(\boldsymbol{\Psi}^s(\theta_1^s, \theta_2^s))\big) &= 0,\\
\boldsymbol{\phi}_{\Delta_*^u}(\boldsymbol{\Psi}^u(\theta_1^u, \theta_2^u)) - \boldsymbol{\phi}_{\Delta_*^s}(\boldsymbol{\Psi}^s(\theta_1^s, \theta_2^s)) &= 0
\end{aligned}
\tag{48}
$$

for as many discrete values of θ as Fourier coefficients needed to be determined in the corresponding equation, and with

$$
\boldsymbol{\Psi}^i(\theta_1^i, \theta_2^i) = \boldsymbol{\phi}_{\frac{\theta_2^i}{2\pi}\Delta^i}\Big(\boldsymbol{\varphi}^i(\theta_1^i - \frac{\theta_2^i}{2\pi}\rho^i) + (\Lambda^i)^{-\frac{\theta_2^i}{2\pi}}\xi^i \boldsymbol{v}^i(\theta_1^i - \frac{\theta_2^i}{2\pi}\rho^i)\Big), \tag{49}
$$

for $i = u, s$. Note that a Taylor expansion of the previous expression around $\boldsymbol{\varphi}^i\big(\theta_1^i - (\theta_2^i/(2\pi))\rho^i\big)$ up to first order in ξ^i turns it into an expression analogous to (44) except for an error $O((\xi^i)^2)$, which is already the error of the linear approximation of the manifold. Compared to (44), expression (49) has as an advantage the fact that it does not contain the differential of the flow. The comments made for system (47) also apply here: system (48) also includes a normalization condition for the invariant bundles $\boldsymbol{v}^u$, $\boldsymbol{v}^s$ to be locally unique, ξ^i is a parameter that must be kept fixed at a small value (e.g. 10^{-6}), and an actual implementation requires multiple shooting, both in the tori and the connection. Additional details can be found in [33]. Figure 28 shows some connections obtained by continuation forward and backward in energy with fixed rotation numbers ρ^u, ρ^s.

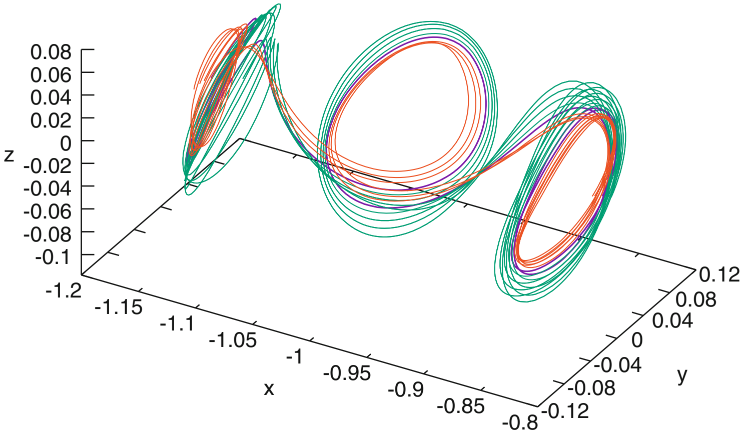


Fig. 28 Continuation with fixed $\rho^u := \bar{\rho}$, $\rho^s := \tilde{\rho}$ of the connection of Fig. 26, both forward (green) and backward (orange) in energy

References

1. Abad, A., Barrio, R., Blesa, F., Rodriguez, M.: Algorithm 924: TIDES, a Taylor series integrator for differential equations. ACM Trans. Math. Softw. **39**(1), Article No. 5 (2012)
2. Allgower, E.L., Georg, K.: Numerical Continuation Methods. Springer Series in Computational Mathematics, vol. 13. Springer, Berlin (1990). An introduction
3. Baresi, N., Olikara, Z.P., Scheeres, D.J.: Fully numerical methods for continuing families of quasi-periodic invariant tori in astrodynamics. J. Astronaut. Sci. **65**(2), 157–182 (2018)
4. Barrabés, E., Mondelo, J.-M., Ollé, M.: Numerical continuation of families of homoclinic connections of periodic orbits in the RTBP. Nonlinearity **22**(12), 2901–2918 (2009)
5. Cabré, X., Fontich, E., de la Llave, R.: The parameterization method for invariant manifolds. I. Manifolds associated to non-resonant subspaces. Indiana Univ. Math. J. **52**(2), 283–328 (2003)
6. Castellà, E., Jorba, À.: On the vertical families of two-dimensional tori near the triangular points of the bicircular problem. Celest. Mech. Dyn. Astron. **76**(1), 35–54 (2000)
7. Conley, C.C.: Low energy transit orbits in the restricted three-body problem. SIAM J. Appl. Math. **16**, 732–746 (1968)
8. Conley, C.C.: On the ultimate behavior of orbits with respect to an unstable critical point. I. Oscillating, asymptotic, and capture orbits. J. Differ. Equ. **5**, 136–158 (1969)
9. Doedel, E.J.: Continuation and bifurcation software for ordinary differential equations. Technical report, Concordia University (2007)
10. Fehlberg, E.: Runge-kutta formulas of high order with stepsize control through leading truncation error term. Technical report, NASA, (1968). NASA-TR-R-287
11. Gasquet, C., Witomski, P.: Fourier Analysis and Applications. Texts in Applied Mathematics, vol. 30. Springer, New York, (1999). Filtering, numerical computation, wavelets, Translated from the French and with a preface by R. Ryan
12. Golub, G.H., van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore and London (1996)
13. Gómez, G., Masdemont, J.J.: Some zero cost transfers between Libration Point orbits. Adv. Astronaut. Sci. **105**, 1199–1216 (2000)
14. Gómez, G., Mondelo, J.-M.: The dynamics around the collinear equilibrium points of the RTBP. Phys. D **157**(4), 283–321 (2001)

15. Gómez, G., Jorba, À., Masdemont, J., Simó, C.: Dynamics and Mission Design Near Libration Point Orbits – Volume 3: Advanced Methods for Collinear Points. World Scientific, Singapore (2001). Reprint of ESA Report *Study Refinement of Semi–Analytical Halo Orbit Theory*, 1991
16. Gómez, G., Koon, W.S., Lo, M.W., Marsden, J.E., Masdemont, J., Ross, S.D.: Connecting orbits and invariant manifolds in the spatial restricted three-body problem. Nonlinearity **17**(5), 1571–1606 (2004)
17. Gómez, G., Mondelo, J.-M., Simó, C.: A collocation method for the numerical Fourier analysis of quasi-periodic functions. II. Analytical error estimates. Discrete Contin. Dyn. Syst. Ser. B **14**(1), 75–109 (2010)
18. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, New York (1983)
19. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd edn. Springer, Berlin (1993)
20. Haro, À.: Automatic differentiation tools in computational dynamical systems (2008, preprint). Available at www.maia.ub.edu/dsg/
21. Haro, A., Canadell, M., Figueras, J.-L., Luque, A., Mondelo, J.M.: The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations. Applied Mathematical Sciences, vol. 195. Springer, Cham (2016)
22. Hénon, M.: Exploration numérique du problème restreint. II.- masses égales, stabilité des orbites periódiques. Ann. Astrophys. **28**, 992–1007 (1965)
23. Hénon, M.: Vertical stability of periodic orbits in the restricted problem. I. Equal masses. Astron. Astrophys. **28**, 415–426 (1973)
24. Jorba, À.: Numerical computation of the normal behaviour of invariant curves of n-dimensional maps. Nonlinearity **14**(5), 943–976 (2001)
25. Jorba, À., Masdemont, J.J.: Dynamics in the center manifold of the Restricted Three-Body Problem. Phys. D **132**, 189–213 (1999)
26. Jorba, À., Villanueva, J.: On the normal behaviour of partially elliptic lower-dimensional tori of Hamiltonian systems. Nonlinearity **10**(4), 783–822 (1997)
27. Jorba, A., Zou, M.: A software package for the numerical integration of odes by means of high-order Taylor methods. Exp. Math. **14**(1), 99–117 (2005)
28. Koon, W.S., Lo, M.W., Marsden, J.E., Ross, S.D.: Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics. Chaos **10**(2), 427–469 (2000)
29. Masdemont, J.: High-order expansions of invariant manifolds of libration point orbits with applications to mission design. Dyn. Syst. Int. J. **20**(1), 59–113 (2005)
30. McGehee, R.P.: Some homoclinic orbits for the restricted three-body problem. PhD thesis, University of Wisconsin (1969)
31. Meyer, K.R., Hall, G.R., Offin, D.: Introduction to Hamiltonian Dynamical Systems and the N-Body Problem, 2nd edn. Springer, New York (2009)
32. Mondelo, J.-M.: Contribution to the Study of Fourier Methods for Quasi-Periodic Functions and the Vicinity of the Collinear Libration Points. PhD thesis, Universitat de Barcelona (2001). Available at www.tdx.cat/TDX-0117102-130456
33. Mondelo, J.-M., Ollé, M., de Sousa-Silva, P., Terra, M.: Families of heteroclinic connections between quasi-periodic libration point trajectories. Paper IAC-14, C1, 1, 9, x25029, 65th International Astronautical Congress 2014, September 29–October 3, Toronto
34. Parker, J.S., Lo, M.W.: Shoot the moon 3D. Adv. Astronaut. Sci. **123**, 2067–2086 (2006)
35. Perko, L.: Differential Equations and Dynamical Systems, 3rd edn. Springer, Berlin (2001)
36. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C++: The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge (2002)
37. Rimmer, R.J.: Generic bifurcations for involutory area preserving maps. Mem. Am. Math. Soc. **41**(272), v+165 (1983)
38. Siegel, C.L., Moser, J.K.: Lectures on Celestial Mechanics. Classics in Mathematics. Springer, Berlin (1995). Translated from the German by C. I. Kalme. Reprint of the 1971 translation.
39. Simó, C.: On the analytical and numerical approximation of invariant manifolds. In: Benest, D., Froeshlé, C. (eds.) Modern Methods in Celestial Mechanics, pp. 285–330. Editions Frontières, Gif-sur-Yvette (1990)

40. Standish, E.M.: JPL planetary and lunar ephemerides, DE405/LE405. Technical Report IOM 312.F.98-048, Jet Propultion Laboratory (1998)
41. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis, 3rd edn. Springer, Berlin (2002)
42. Szebehely, V.: Theory of orbits. The Restricted Problem of Three Bodies. Academic, New York (1967)

Celestial Mechanics of Rubble Pile Bodies

Daniel J. Scheeres

Abstract This chapter derives the general equations for interacting rigid bodies accounting for their rolling motion on each other. The derivation of the non-rolling motions has been given previously, but to accommodating rolling and slipping motion it is necessary to develop a non-holonomic form of the equations of motion. The resulting derivation shows that the key analysis parameters for a collection of grains that gravitationally attract and rest on each other are preserved in this more advanced formulation. The chapter ends with a simple application of these results to a series of bodies that can roll on each other, satisfying a non-holonomic constraint.

Keywords Rubble pile asteroid · N rigid body problem · Granular celestial mechanics

1 Introduction

The motion of rigid bodies that are resting on and orbiting about each other, attracted by mutual gravity, is a topic of study that has become more relevant in recent decades with the discovery of "rubble pile" asteroids [2, 7]. These are small bodies that are comprised of a size distribution of hard grains, mutually resting on each other. Their mechanics can be studied using principles of continuum mechanics, as has been done by Holsapple [4], or through discrete element particle mechanics simulations [8, 16]. They can also be analyzed from a discrete body Celestial Mechanics perspective which is more focused on finding analytical constraints on the orbital and rotational dynamics of these bodies [11]. This chapter presents an

D. J. Scheeres (✉)
Smead Department of Aerospace Engineering Sciences, The University of Colorado at Boulder, Boulder, CO, USA
e-mail: scheeres@colorado.edu

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_5

introduction to this problem from a Celestial Mechanics point of view with an emphasis on how contact constraints are dealt with dynamically, and provides a detailed example calculation of the stability of a certain resting configuration of bodies.

The outline of the paper is as follows. First, the basic mass distribution properties of rigid bodies and their mutual attraction are outlined. Next, the states of the system are defined, along with their constraints, with a special focus on the dynamical constraints when the bodies are in contact. Following this the kinematical equations and related quantities are given, leading up to the definition of the full set of Lagrange's equations for the system. A special version of the Lagrange equations are derived using Routh reduction, culminating with the definition of the amended potential and the full set of equilibrium conditions. Following this a special case of stability for a rubble pile body is studied, focusing on the stability conditions for an arbitrary number of bodies resting on each other placed in a straight line.

2 Problem Specification

Consider the mass distributions of a set of N rigid bodies that interact with each other through gravitation and surface contact forces. We specify these bodies in general as mass distributions, denoted as $\mathscr{B}_i, i = 1, 2, \ldots, N$. Each body has its own center of mass location, velocity and body orientation and rotation. Thus in total there are $6N$ degrees of freedom for the system. We note that the rigid body assumption places specific constraints on these degrees of freedom which have been noted in previous discussions of the problem, however these have not been fully detailed as of yet in terms of the appropriate equations of motion [12]. Thus this contribution will provide a more rigorous definition of these constraints and note some specific results once the equations of motion are properly formulated.

2.1 Density Distributions and Body States

Consider an arbitrary collection of N mass distributions, denoted as $\mathscr{B}_i, i = 1, 2, \ldots, N$, following the derivation in [9]. Each body $\mathscr{B}_i$ is defined by a differential mass distribution dm_i that is assumed to be a finite density distribution, denoted as

$$dm_i = \rho_i(\mathbf{r})dV \tag{1}$$

$$m_i = \int_{\mathscr{B}_i} dm_i \tag{2}$$

where m_i is the total mass of body $\mathscr{B}_i$, ρ_i is the density of body $\mathscr{B}_i$ (possibly constant), $\mathbf{r}$ is a spatial position vector variable relative to a given frame and dV is the differential volume element. If $\mathscr{B}_i$ is described by a point mass density distribution, the body itself is just defined as a single point $\mathbf{r}_i$. Instead, if the body is defined as a finite density distribution, $\mathscr{B}_i$ is defined as a compact set in $\mathbb{R}^3$ over which $\rho_i(\mathbf{r}) > 0$. In either case the $\mathscr{B}_i$ are defined as compact sets. This notation can be further generalized by defining the general mass differential

$$dm(\mathbf{r}) = \sum_{i=1}^{N} dm_i(\mathbf{r}) \tag{3}$$

and the total mass distribution $\mathscr{B} = \{\mathscr{B}_i, i = 1, 2, \dots, N\}$. Then the above definitions can be reduced to integrals over $\mathscr{B}$:

$$M = \int_{\mathscr{B}} dm \tag{4}$$

where $M = \sum_{i=1}^{N} m_i$ is the total mass of the system.

Each differential mass element $dm_i(\mathbf{r})$ has a specified position and an associated velocity. For components within a given body $\mathscr{B}_i$ a rigid body assumption is made so that the state of the entire body can be defined by the position and velocity of its center of mass,

$$\mathbf{r}_i = \frac{1}{m_i} \int_{\mathscr{B}_i} \mathbf{r} dm \tag{5}$$

$$\mathbf{v}_i = \frac{1}{m_i} \int_{\mathscr{B}_i} \mathbf{v} dm , \tag{6}$$

where velocities are assumed measured relative to an inertial frame, and its attitude and angular velocity (see below). Finally, we assume that these positions and velocities are defined relative to the system barycenter, which is chosen as the origin, so

$$\int_{\mathscr{B}} \mathbf{r} dm(\mathbf{r}) = 0 \tag{7}$$

$$\int_{\mathscr{B}} \mathbf{v} dm(\mathbf{r}) = 0 . \tag{8}$$

Thus, the individual bodies are located by their position vector $\mathbf{r}_i$ relative to the origin, with the additional constraint that $\sum_{i=1}^{N} m_i \mathbf{r}_i = 0$. Their relative position to each other is also defined as $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$, and their relative velocity as $\dot{\mathbf{r}}_{ij} = \dot{\mathbf{r}}_j - \dot{\mathbf{r}}_i$.

2.2 Body Orientations and Inertias

Each rigid body has a unique orientation relative to an inertial frame. The relevant mass distribution parameter of a rigid body then expands to also include its moments of inertia (or inertia tensor/dyad), and the relevant orientation degrees of freedom must also be defined.

Within each body a unique set of orthogonal axes can be defined which enable the orientation of the rigid body. Then we can use a transformation matrix, or direction cosine matrix, to define the orientation of these axes relative to an inertial frame and thus define the orientation of the rigid body. Denote the dyad $\overline{\mathbf{A}}_i$ as mapping a vector in an inertial frame into the body frame of body i. We note that such a dyad must be orthonormal, meaning that $\det(\overline{\mathbf{A}}_i) = 1$ and that $\overline{\mathbf{A}}_i^T \cdot \overline{\mathbf{A}}_i = \overline{\mathbf{A}}_i \cdot \overline{\mathbf{A}}_i^T = \overline{\mathbf{U}}$, where $\overline{\mathbf{U}}$ is the identity dyad and the $(\cdot)$ operator stands for the dot product between dyads.

We note that the dyad $\overline{\mathbf{A}}_i$ suffices to define the attitude of the body, however it is over-constrained due to its above properties, with 9 numbers needed to specify but 6 constraints that must be satisfied. Thus one can always choose a unique set of three Euler angles to represent the dyad and orientation of the body, although any such representation will always have singularities associated with it. Alternatively, one can also define the rotation axis and rotation angle of the dyad $\overline{\mathbf{A}}_i$, which are the eigenvector of the unity eigenvalue and the angle associated with its complex conjugate eigenvalues. This representation does not have any singularities, although it still has at least one constraint equation. Closely related to the axis-angle variables are the quaternion representation, which still has 4 numbers with 1 constraint.

Due to the simplicity of notation, we will rely on the dyad $\overline{\mathbf{A}}_i$ to define the inertial orientation of our bodies. Similarly, it is important to define the orientation of two bodies relative to each other. To do this we define $\overline{\mathbf{A}}_{ij} = \overline{\mathbf{A}}_j \cdot \overline{\mathbf{A}}_i^T$ as the dyad that maps from the body i frame to the body j frame.

To specify the kinematics of a rotating body define the angular velocity $\boldsymbol{\omega}_i$ as the angular velocity of body i relative to an inertial frame. Then the orientation dyad $\overline{\mathbf{A}}_i$ has the following kinematics

$$\dot{\overline{\mathbf{A}}}_i = -\tilde{\boldsymbol{\omega}}_i \cdot \overline{\mathbf{A}}_i \tag{9}$$

where $\tilde{\mathbf{a}}$ is the cross-product dyad associated with a vector $\mathbf{a}$, such that $\tilde{\mathbf{a}} \cdot \mathbf{b} = \mathbf{a} \cdot \tilde{\mathbf{b}} = -\mathbf{b} \cdot \tilde{\mathbf{a}} = \mathbf{a} \times \mathbf{b}$. Similarly, if the angular velocity of body j relative to body i is $\boldsymbol{\omega}_{ij} = \boldsymbol{\omega}_j - \boldsymbol{\omega}_i$, then the time rate of change of $\overline{\mathbf{A}}_{ij}$ is

$$\dot{\overline{\mathbf{A}}}_{ij} = -\tilde{\boldsymbol{\omega}}_{ij} \cdot \overline{\mathbf{A}}_{ij} \ . \tag{10}$$

The counterpart to a body's mass for its translational motion is a body's inertia tensor/dyad for its rotational motion. The inertia dyad for a body is defined as the mass integral

$$\overline{\mathbf{I}}_i = -\int_{\mathscr{B}_i} (\widetilde{\boldsymbol{\rho}} \cdot \widetilde{\boldsymbol{\rho}})\, dm_i(\boldsymbol{\rho}) \;, \tag{11}$$

where we assume that the position of the mass element $\boldsymbol{\rho}$ is defined with respect to the body's center of mass, and that the dyad is nominally defined in the body-fixed frame. Thus, to transform the inertia dyad to an inertial frame requires the use of the orientation dyad as $\overline{\mathbf{A}}_i^T \cdot \overline{\mathbf{I}}_i \cdot \overline{\mathbf{A}}_i$. A separate distinction is the body moment of inertia relative to the origin of the system, which we define as

$$\overline{\mathbf{J}}_i(\mathbf{r}_i) = \overline{\mathbf{I}}_i - m_i \widetilde{\mathbf{r}}_i \cdot \widetilde{\mathbf{r}}_i \;, \tag{12}$$

explicitly calling out that this is a function of the position of body i in the coordinate frame. The use of $\overline{\mathbf{J}}$ denotes that the inertia matrix is not specified relative to a center of mass and $\mathbf{r}_i$ is the position vector of body i relative to the system coordinate origin, meaning that $\overline{\mathbf{J}}_i$ is the total moment of inertia of body i relative to the coordinate origin.

In addition to the inertia dyad of a single body, we also need to express the system moment of inertia. Fundamentally, this equals

$$\overline{\mathbf{J}} = \sum_{i=1}^{N} \overline{\mathbf{J}}_i \;. \tag{13}$$

Using the general mass differential we can express the entire system inertia dyad relative to the system coordinate origin as a single integral

$$\overline{\mathbf{J}} = -\int_{\mathscr{B}} (\widetilde{\mathbf{r}} \cdot \widetilde{\mathbf{r}})\, dm(\mathbf{r}) \;. \tag{14}$$

Assuming that the coordinate origin is the system center of mass, we can rewrite the inertia dyad to be in terms of the relative orientation of mass elements through an application of Lagrange's Identity to find

$$\overline{\mathbf{I}} = -\frac{1}{2M} \int_{\mathscr{B}} \int_{\mathscr{B}} \left(\widetilde{\mathbf{r}} - \widetilde{\mathbf{r}}'\right) \cdot \left(\widetilde{\mathbf{r}} - \widetilde{\mathbf{r}}'\right) dm(\mathbf{r}) dm(\mathbf{r}') \;, \tag{15}$$

where we note then that $\overline{\mathbf{J}} = \overline{\mathbf{I}}$. From this version it is easy to note that the system inertia is then defined only in terms of the relative positions and orientations of the system's rigid bodies, with these vectors being specified in the given inertial frame.

2.3 Degrees of Freedom and Constraints

We recall that for N bodies there are $3N$ translational degrees of freedom and $3N$ rotational degrees of freedom for a total of $6N$ DOF. In our formulation we have already removed 3 DOF by setting the center of the system at the barycenter, reducing the total to $3(2N - 1)$. The degrees of freedom are split between three general classes, the relative positions of the bodies, the relative orientations of the bodies to each other, and the overall orientation of the system with respect to the inertial frame. We note that the relative positions and orientations will be independent of the overall orientation of the system relative to inertial space. This is practically realized by choosing the reference frame for relative orientation and position to be fixed in one of the bodies.

It is instructive to review these degrees of freedom. For $N = 1$ there are no relative position or attitude degrees of freedom, and thus there is only the inertial orientation degrees of freedom for the system, yielding a total of 3 DOF in agreement with the general rule. For $N = 2$ we start with one central body with no degrees of freedom. The position of a second body relative to this has 3 DOF and its relative attitude has 3 DOF. Finally, we add the inertial orientation to get a total of 9 DOF. Each additional body then adds 6 DOF again, reproducing our general rule.

We distinguish between the internal, relative degrees of freedom and the inertial orientation degrees of freedom. For the current system we represent the internal degrees of freedom as $q_i : i = 1, 2, \ldots, 6(N-1)$. These are specifically the relative positions of the centers of mass and the orientations of the rigid bodies relative to each other. For convenience we can imagine these to be Cartesian position vectors and Euler angles in a common frame fixed in one of the bodies. We note that their time derivatives are expressed with respect to an inertial frame. The additional 3 DOF that orient the system relative to inertial space is represented as the rotation dyad $\overline{\mathbf{A}}$ which takes the relative frame into inertial space.

Note again that in our general statements, the final 3 DOF that orient the system relative to inertial space do not change any of our fundamental integral quantities except that of the total angular momentum $\mathbf{H}$ and total system moment of inertia $\overline{\mathbf{I}}$. This is because each of these orientations acts on the entire system but do not change the relative orientations or speeds. This invariance is tied to the existence of the angular momentum integral. Despite this, since the kinetic energy is defined relative to an inertial frame there remains a fundamental connection between the inertial and relative frames.

2.3.1 Unilateral Constraints

Unilateral constraints exist between any two bodies, when their surfaces touch each other. This is driven by their relative position and attitude and arise from the rigid body and finite density constraints. Ultimately this means that the body centers of mass cannot get arbitrarily close to each other due to their shapes. The constraint exists between every pair of bodies, i and j, and has the general form

$$r_{ij} \geq d_{ij}(\hat{\mathbf{r}}_{ij}, \overline{\mathbf{A}}_{ij}) , \tag{16}$$

where $\mathbf{r}_{ij}$ is the relative position of the two bodies, r_{ij} is its magnitude, $\hat{\mathbf{r}}_{ij}$ is the unit vector and $\overline{\mathbf{A}}_{ij}$ is their relative attitude. The function d_{ij} is defined for a particular pair of bodies. If we assume both bodies are convex, then only the relative distance is affected. If either of the bodies are not convex, once they are in contact there can be additional unilateral attitude constraints that can arise.

For a 2-body system which has 6 relative degrees of freedom (3 position and 3 angular), once the bodies are in contact this is reduced to 2 position degrees of freedom, but still 3 angular degrees of freedom. When the bodies are in contact, this constraint is reduced to a holonomic constraint, as it only applies to the geometry between the bodies as a function of their relative attitude and position, specifically $r_{ij} = d_{ij}(\hat{\mathbf{r}}_{ij}, \overline{\mathbf{A}}_{ij})$. There are additional physics that arises when in contact, depending on the relative friction between the two bodies. The easiest to model is to assume that the two bodies roll without slipping on each other. This creates additional constraints, however, reviewed below.

2.3.2 Non-holonomic Constraints

When the unilateral constraints are active, if we assume that the bodies roll without slipping on each other, then non-holonomic constraints arise. In general a non-holonomic constraint is one that exists on the relative velocities of a system, and restricts the direction of motion as a function of the other states of the system. The distinguishing feature of non-holonomic constraints is that they are not integrable, meaning that they cannot be reduced to a purely geometric constraint. Another way to view this is that there are no inherent restrictions of the relative geometry of two bodies subject to non-holonomic constraints, although at every relative geometry there are constraints on how the two bodies can move relative to each other.

A general statement of a set of m non-holonomic constraints on a system with n degrees of freedom q_i can be reduced to [3]

$$\sum_{i=1}^{n} a_{ji}(\mathbf{q}, t)\dot{q}_i + a_{jt}(\mathbf{q}, t) = 0 , \tag{17}$$

where $j = 1, 2, \ldots, m$. We note that the constraints depend on the degrees of freedom of the system and potentially on time (although for our systems time will not be present). The non-integrability of the constraints can be checked explicitly by showing that $\partial a_{ji}/\partial q_k \neq \partial a_{jk}/\partial q_i$ for some indices k and i. If these cross-partials are equal for all degrees of freedom, then the constraint is integrable and can in principle be reduced to a holonomic constraint.

A simple example is when two spheres rest on each other. Then, assuming that they only roll without slipping, the travel of the spheres relative to each other is constrained by the direction in which they rotate relative to each other. A simple rotation of the sphere about a given axis then results in a specific path of their contact point. Given that the sphere can rotate in three directions, one perpendicular to the tangent plane through the point of contact (which does not result in motion), and two about orthogonal axes whose combination can cause the sphere to roll in a specified direction on the surface once the rotation angles are chosen. Due to this, the system has two additional constraints, meaning that only the relative attitude between the spheres is free to change—the relative position on the sphere being constrained by the non-holonomic constraint. However, despite these constraints the relative location and orientation of the spheres can take on all possible values of the original 5-DOF system under the active unilateral constraint.

Specifically, when the spheres are in contact, we can state the motion constraints as

$$\dot{u} - \cos(\theta_3) R\dot{\theta}_1 - \sin(\theta_3) R\dot{\theta}_2 = 0\,, \tag{18}$$

$$\dot{v} + \sin(\theta_3) R\dot{\theta}_1 - \cos(\theta_3) R\dot{\theta}_2 = 0\,. \tag{19}$$

Here $\dot{u}$ and $\dot{v}$ are the velocity of the center of the rolling sphere relative to the stationary sphere, defined in orthogonal directions, θ_3 is the rotation of the spheres relative to each other about their point of contact, and θ_1 and θ_2 are rotations of the sphere in two mutually perpendicular directions which are in turn perpendicular to the axis associated with θ_3. Non-integrability can be shown by noting that the coefficient of $\dot{\theta}_3$ is zero for both constraints, meaning that its partial with respect to any of the other angles is zero. However, the coefficients of $\dot{\theta}_1$ and $\dot{\theta}_2$ are functions of θ_3 and thus their partials are non-zero in general, showing that the integrability conditions are not satisfied.

2.3.3 General Constraints

When the unilateral constraints are active the non-holonomic constraints are also active and lead to restrictions on the relative motion of the two bodies in contact. The non-holonomic constraints will also affect the statement of the equations of motion, to be reviewed later. See Fig. 1 for an illustration of the general case.

If the two bodies are both locally smooth and convex about their point of contact, the above "spherical" model is qualitatively similar to the actual situation, with the generalization being that the surface can have different local curvatures in different directions. If the surfaces are not convex, or not smooth, then it is possible to for additional unilateral constraints to become active in additional angular directions.

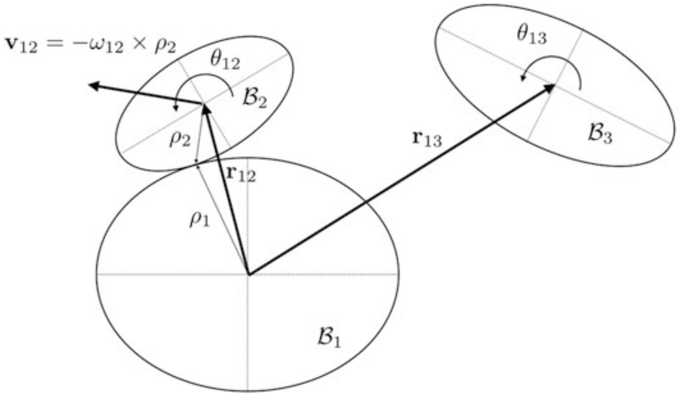


Fig. 1 Geometry of bodies resting and orbiting each other

When two bodies i and j are in contact we will generally represent the constraints as follows, using the indices $i = 1$ and $j = 2$. When in contact the relative position vector is $\mathbf{r}_{12} = d_{12}(\hat{\mathbf{r}}_{12}, \overline{\mathbf{A}}_{12})\hat{\mathbf{r}}_{12}$, with the magnitude of the relative position being a holonomic constraint. For our convex body assumption, there will be a unique point of contact between the bodies as a function of their relative state, which we denote as $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$, and which are shown in Fig. 1. We note that the surfaces will have their normals anti-parallel at this contact point. These contact position vectors are defined in their relative body-fixed frames, and are smooth functions of the relative state allowing us to write them functionally as $\boldsymbol{\rho}_i(\hat{\mathbf{r}}_{12}, \overline{\mathbf{A}}_{12})$. We then have the identity $\mathbf{r}_{12} = \boldsymbol{\rho}_1 - \boldsymbol{\rho}_2$.

If we take body 1 as the reference frame for the relative orientation of the system and assume that body 2 rolls without slipping on its surface, then the velocity of body 2 relative to body 1 can be found as a non-holonomic constraint

$$\dot{\mathbf{r}}_{12} + \boldsymbol{\omega}_{12} \times \boldsymbol{\rho}_2 = \mathbf{0}\,, \tag{20}$$

where the velocity and angular velocity are relative to body 1, and do not directly consider the spin or motion of body 1. It can be noted that the projection of the velocity along the direction $\hat{\mathbf{r}}_{12}$ corresponds to the integrable direction, meaning that the distance along this direction is a constraint defined by Eq. (16). The non-holonomic constraints are then the directions perpendicular to this line, and correspond to the two possible directions of rolling motion that body 2 can take relative to body 1. It is also possible to formulate the non-holonomic constraints locally about the contact point between the two bodies, however we do not pursue that approach here.

It is instructive to carry out the computation for two spheres of radius R_1 and R_2 rolling on each other (see Fig. 2). Let us take sphere 1 as fixed and sphere 2 as rolling on it, constraining motion to lie in a plane. From our computation the velocity of

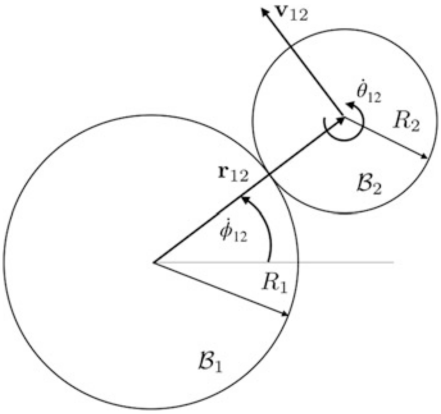


Fig. 2 Geometry of two spheres rolling on each other

sphere 2 relative to sphere 1 will be

$$\dot{\mathbf{r}}_{12} = R_2\dot{\theta}_{12}\hat{\mathbf{r}}_\perp \ , \tag{21}$$

where $\dot{\theta}_{12}$ is the relative angular velocity of the spheres (measured in a plane) and $\hat{\mathbf{r}}_\perp$ is the unit vector in the direction of motion and orthogonal to both the relative position of the sphere centers and the angular velocity vector. This provides us with an expression for the velocity. However, we can equivalently reduce this to the angular rate at which the sphere 2 center moves relative to the sphere 1 center, denoted in the figure as $\dot{\phi}_{12}$. Note that the distance from sphere 1 to sphere 2 equals $R_1 + R_2$, and the angular rate of sphere 2 relative to sphere 1, $\dot{\phi}_{12}$, must yield the same velocity, thus

$$\dot{\mathbf{r}}_{12} = (R_1 + R_2)\dot{\phi}_{12}\hat{\mathbf{r}}_\perp \ . \tag{22}$$

Equating these two we find the constraint

$$\dot{\phi}_{12} = \frac{R_2}{R_1 + R_2}\dot{\theta}_{12} \ . \tag{23}$$

In this way the relative translational motion of the spheres are coupled with their relative rotational motion.

If we wish, a further step can be taken, allowing for unilateral constraints on the relative attitude. For the above angle θ_{12} this would express itself in general as

$$\theta_{12} - d_\theta(\mathbf{r}, \overline{\mathbf{A}}) \geq 0 \tag{24}$$

and indicates that the angle would be stopped at a specific relative orientation. These would arise if the surfaces of either body have concavities or discontinuous slopes. One such unilateral constraint being active should reduce the dimensionality of the non-holonomic constraint by one, while two such active constraints would remove the non-holonomic constraints altogether, and just provide a geometrical constraint between the bodies.

2.4 Mutual Gravitational Potential

The mutual gravitational potential of the system is comprised of the pairwise mutual potentials between the different rigid bodies.

$$\mathscr{U} = -\mathscr{G} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{\mathscr{B}_i} \int_{\mathscr{B}_j} \frac{dm_i \, dm_j}{|\mathbf{r}_{ij} + \boldsymbol{\rho}_j - \boldsymbol{\rho}_i|} , \tag{25}$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and the $\boldsymbol{\rho}_i$ denote the integration variable over the mass distribution. Note that the definition of $\mathscr{U}$ in Eq. (25) eliminates the self-potentials of these bodies from consideration. As the finite density mass distributions are assumed to be rigid bodies this elimination is reasonable. However, the more general expression of the gravitational potential

$$\mathscr{U}' = -\frac{\mathscr{G}}{2} \int_{\mathscr{B}} \int_{\mathscr{B}} \frac{dm(\mathbf{r}) \, dm(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} , \tag{26}$$

naturally includes the self-potentials. For the rigid body assumption, the gradient of either of these forms with respect to the degrees of freedom in the system will be equal, however.

The mutual potential of two bodies can be reduced to a convenient expression that only involves integration over the different surface areas. Following the summary in [17] of the classical results by Liebenthal [5] for constant density bodies (which we can nominally assume at this point) we find

$$\mathscr{U}_{ij} = -\mathscr{G} \int_{\mathscr{B}_i} \int_{\mathscr{B}_j} \frac{dm_i \, dm_j}{|\mathbf{r}_{ij} + \boldsymbol{\rho}_j - \boldsymbol{\rho}_i|} \tag{27}$$

$$= \frac{\sigma_i \sigma_j}{2} \mathscr{G} \int_{\partial\mathscr{B}_i} \int_{\partial\mathscr{B}_j} |\mathbf{r}_{ij} + \overline{\mathbf{A}}_{ij}^T \cdot \boldsymbol{\rho}_j - \boldsymbol{\rho}_i| \, d\mathbf{S}_i \cdot d\mathbf{S}_j , \tag{28}$$

where the σ_i are the densities of the respective bodies and we explicitly call out the necessary orientation of the body j frame to the body i frame, tacitly assuming that all of the other vectors in the magnitude operator are in the body i frame. Note that the mutual potential has been reduced to the mutual integration of the differential, oriented surface areas $d\mathbf{S}_i$ dotted with each other over the surface of the distinct mutual bodies, denoted by $\partial\mathscr{B}_i$.

When in this form, the general expressions for the force and moment acting between two bodies is especially simple to describe, even though these integrals cannot be evaluated in closed form except for the simplest systems. Then the force between these two bodies is evaluated as the negative gradient of this expression with respect to the mutual distance between their centers of mass

$$\mathscr{F}_{ij} = -\frac{\partial \mathscr{U}_{ij}}{\partial \mathbf{r}_{ij}} \tag{29}$$

$$= -\frac{\sigma_i \sigma_j}{2}\mathscr{G} \int_{\partial \mathscr{B}_i} \int_{\partial \mathscr{B}_j} \hat{\mathbf{r}}\, d\mathbf{S}_i \cdot d\mathbf{S}_j \ , \tag{30}$$

where $\mathbf{r} = \mathbf{r}_{ij} + \overline{\mathbf{A}}_{ij}^T \cdot \boldsymbol{\rho}_j - \boldsymbol{\rho}_i$ and $\hat{\mathbf{r}} = \mathbf{r}/r$. We note that $\mathscr{F}_{ij}$ is implicitly written with respect to body frame 1.

For the torque between the bodies, we must take the gradient of the potential with respect to the orientation of body j relative to the body i frame. An easy way to specify this is to start with the current orientation between these bodies, denoted as $\overline{\mathbf{A}}_{ij}$ and consider a small additional variation consisting of Euler angle rotations about each axis of body j, denoted as $\delta\boldsymbol{\theta}_{ij}$. When the variations are small we can assume that $\delta\boldsymbol{\theta}_{ij} = \boldsymbol{\omega}_{ij}\delta t$ and thus the orientation of body j relative to body i with a small additional variation can be expressed as $\delta\overline{\mathbf{A}}_{ij} = \widetilde{\delta\boldsymbol{\theta}}_{ij} \cdot \overline{\mathbf{A}}_{ij}$. Given this, the general variation of the relative vector $\mathbf{r} = \mathbf{r}_{ij} + \overline{\mathbf{A}}_{ij}^T \cdot \boldsymbol{\rho}_j - \boldsymbol{\rho}_i$ due to a variation in the relative attitude can be expressed as $\delta_{\boldsymbol{\theta}}\mathbf{r} = \left(\delta\overline{\mathbf{A}}_{ij}\right)^T \cdot \boldsymbol{\rho}_j = -\overline{\mathbf{A}}_{ij}^T \cdot \widetilde{\delta\boldsymbol{\theta}}_{ij} \cdot \boldsymbol{\rho}_j = \overline{\mathbf{A}}_{ij}^T \cdot \tilde{\boldsymbol{\rho}}_j \cdot \delta\boldsymbol{\theta}_{ij}$. Then, as $\delta_{\boldsymbol{\theta}}\mathscr{U}_{ij} = \frac{\partial \mathscr{U}_{ij}}{\partial \boldsymbol{\theta}} \cdot \delta\boldsymbol{\theta}$ we then find the general equation for the torque of body i on body j as

$$\mathscr{M}_{ij} = -\frac{\partial \mathscr{U}_{ij}}{\partial \boldsymbol{\theta}_{ij}} \tag{31}$$

$$= -\frac{\sigma_i \sigma_j}{2}\mathscr{G} \int_{\partial \mathscr{B}_i} \int_{\partial \mathscr{B}_j} \hat{\mathbf{r}} \cdot \overline{\mathbf{A}}_{ij}^T \cdot \tilde{\boldsymbol{\rho}}_j \, d\mathbf{S}_i \cdot d\mathbf{S}_j \ . \tag{32}$$

We note that $\mathscr{M}_{ij}$ is implicitly written with respect to body frame 1.

From these definitions of the force and moment between two bodies it is relatively easy to prove the identities

$$\mathscr{F}_{ij} + \overline{\mathbf{A}}_{ij}^T \cdot \mathscr{F}_{ji} = 0 \ , \tag{33}$$

$$\mathscr{M}_{ij} + \overline{\mathbf{A}}_{ij}^T \cdot \mathscr{M}_{ji} = \mathbf{r}_{ij} \times \mathscr{F}_{ij} \ . \tag{34}$$

2.5 Kinetic Energy and Angular Momentum

Finally, the integral form of the kinetic energy and angular momentum vector can be stated as [9]

$$T = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathscr{B}_i} (\mathbf{v}\cdot\mathbf{v})\, dm_i(\mathbf{r}) \;, \tag{35}$$

$$\mathbf{H} = \sum_{i=1}^{N}\int_{\mathscr{B}_i} (\mathbf{r}\times\mathbf{v})\, dm_i(\mathbf{r}) \;, \tag{36}$$

where we assume that $\mathbf{v}$ is the velocity measured relative to an inertial frame. This notation can be further generalized by again defining the single and joint general mass differentials

$$T = \frac{1}{2}\int_{\mathscr{B}} (\mathbf{v}\cdot\mathbf{v})\, dm(\mathbf{r}) \;, \tag{37}$$

$$\mathbf{H} = \int_{\mathscr{B}} (\mathbf{r}\times\mathbf{v})\, dm(\mathbf{r}) \;, \tag{38}$$

where we assume that these vectors are specified relative to the inertial frame. These quantities can also be expressed in terms of relative coordinates only—similar to the gravitational potential

$$T = \frac{1}{4M}\int_{\mathscr{B}}\int_{\mathscr{B}} \left(\mathbf{v}-\mathbf{v}'\right)\cdot\left(\mathbf{v}-\mathbf{v}'\right) dm(\mathbf{r})dm(\mathbf{r}') \;, \tag{39}$$

$$\mathbf{H} = \overline{\mathbf{A}}\cdot\frac{1}{2M}\int_{\mathscr{B}}\int_{\mathscr{B}} \left(\mathbf{r}-\mathbf{r}'\right)\times\left(\mathbf{v}-\mathbf{v}'\right) dm(\mathbf{r})dm(\mathbf{r}') \;. \tag{40}$$

For the angular momentum, we note that it must be mapped from the relative frame into the inertial frame where it is naturally conserved for a closed system.

3 Equations of Motion

3.1 The Lagrange Equations for a Set of Rigid Bodies

We first consider the Lagrange equations for the system in their simplest form. Define the coordinates of the system to be the N position vectors $\mathbf{r}_i$ of the body centers of mass relative to the barycenter of the system, and the corresponding attitudes of these bodies, nominally signified by Euler angles $\boldsymbol{\theta}_i$, where we note that the specific angles used can be changed at will. The specific form of the equations will vary depending on whether any of the unilateral constraints are active.

For definiteness we will take the independent $6(N-1)+3$ degrees of freedom as follows. Assume that we take body 1 as the reference body and its orientation with respect to inertial space is denoted as $\boldsymbol{\theta}_1$. Then we take the other bodies, $j = 2, 3, \ldots, N$, with their position $\mathbf{r}_{1j}$, time rate of position with respect to an inertial frame $\dot{\mathbf{r}}_{1j}$ and their Euler angles $\boldsymbol{\theta}_{1j}$ relative to body 1.

The Lagrangian is $L = T - \mathscr{U}$ and when there are no active constraints the equations take the general form

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\mathbf{r}}_{1i}} = \frac{\partial L}{\partial \mathbf{r}_{1i}} , \tag{41}$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}_{1i}} = \frac{\partial L}{\partial \boldsymbol{\theta}_{1i}} , \tag{42}$$

plus the equation for the attitude of body 1

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}_{1}} = \frac{\partial L}{\partial \boldsymbol{\theta}_{1}} . \tag{43}$$

We could also choose different independent degrees of freedom of the system, so long that they consist of a minimal set of body positions and orientations.

Let us consider the situation when there is an active constraint between two of the bodies. We can renumber the system and take one of the bodies as being "central" with index $i = 1$ and the other as resting on the central body with index $i = 2$. Then their relative position vector $\mathbf{r}_{12}$ is constrained in magnitude as $r_{12} = d_{12}(\hat{\mathbf{r}}_{12}, \overline{\mathbf{A}}_{12})$. Associated with this holonomic constraint will be two additional non-holonomic constraints of the form $\mathbf{a}_{1r} \cdot \dot{\mathbf{r}}_{12} + \mathbf{a}_{1\theta} \cdot \dot{\boldsymbol{\theta}}_{12} = 0$ where $\mathbf{a}_{1x} \in \mathbb{R}^{2\times 3}$. We note that we can also write the holonomic constraint in differential form, resulting in a form $\mathbf{a}_{2r} \cdot \dot{\mathbf{r}}_{12} + \mathbf{a}_{2\theta} \cdot \dot{\boldsymbol{\theta}}_{12} = 0$ where $\mathbf{a}_{2x} \in \mathbb{R}^{1\times 3}$. The driving reason to do this is that it allows the constraint force to be solved for explicitly, as described below. We can then combine these three constraints together into one general form

$$[\mathbf{a}_r] \cdot \dot{\mathbf{r}}_{12} + [\mathbf{a}_\theta] \cdot \dot{\boldsymbol{\theta}}_{12} = 0 , \tag{44}$$

where $[\mathbf{a}_x] \in \mathbb{R}^{3\times 3}$, which represents three constraint equations.

Given that this constraint is active, the equations of relative motion between bodies 1 and 2 becomes

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\mathbf{r}}_{12}} = \frac{\partial L}{\partial \mathbf{r}_{12}} + \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_r] , \tag{45}$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}_{12}} = \frac{\partial L}{\partial \boldsymbol{\theta}_{12}} + \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_\theta] , \tag{46}$$

where the Lagrange multipliers $\boldsymbol{\lambda}_{12} \in \mathbb{R}^3$ are the same between the two equations, and they are solved for in concert with Eq. (44). Once the Lagrange multipliers are found, the constraint force and moment are found as

$$\mathbf{F}_{12} = \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_r] \ , \tag{47}$$

$$\mathbf{M}_{12} = \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_\theta] \ . \tag{48}$$

See [3] for a more general introduction to non-holonomic dynamics.

3.2 Transformation into a Rotating System

Of specific interest to us is the overall rotation of the system due to a non-zero but constant angular momentum. A specific goal is to remove this integral of motion, sometimes termed the elimination of the nodes. In our analysis we can remove one degree of freedom quite simply, and by doing so define the amended potential that we use to discuss relative equilibrium and their stability.

We define a very specific rotating frame from which we will measure motion. This is done by defining a system angular velocity which is a function of the angular momentum integral. Before we do this, however we initially define an angular velocity vector that will be shown to be derived from the angular momentum later,

$$\boldsymbol{\omega} = \frac{\mathbf{H}}{I_H} \tag{49}$$

$$= \dot{\theta}\hat{\mathbf{H}} \ , \tag{50}$$

where $I_H = \hat{\mathbf{H}} \cdot \bar{\mathbf{I}} \cdot \hat{\mathbf{H}}$ is the moment of inertia of the system about a fixed direction in inertial space and $\mathbf{H}$ is an arbitrarily chosen constant vector in inertial space. We note that I_H is a function of both the internal system described in terms of the relative positions and orientations of the bodies and its orientation relative to $\hat{\mathbf{H}}$, but not to rotations around this unit vector which we denote by the angle θ. Since the axis of rotation is fixed in space the angular velocity $\boldsymbol{\omega} = \dot{\theta}\hat{\mathbf{H}}$ is a true velocity and can be expressed as the time derivative of an angle.

This defines an overall rotation rate for the system that is directly tied to its total angular momentum and the distribution of its mass. The system can then be rewritten relative to this rotating frame, noting that the rotation rate $\dot{\theta}$ is not necessarily constant as the moment I_H is not a constant in general. We note that our initial definition of this rotation is independent of angular momentum, however we can show that the proper choice of θ and its spin direction will relate back to this conserved quantity.

First the system kinetic energy is expressed in a rotating frame, with rotation vector defined by $\boldsymbol{\omega}$, meaning that the time derivatives will be expressed relative to a rotating frame. In the following we use the shorthand notation $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r}'$, and similar for other quantities, where both of these vectors will be integrated over the total mass distribution. Then given an inertial velocity $\Delta\mathbf{v}$, it can be expressed relative to the rotating frame as

$$\Delta\mathbf{v} = \Delta\dot{\mathbf{r}} + \boldsymbol{\omega} \times \Delta\mathbf{r} \; , \tag{51}$$

where $\Delta\mathbf{v}$ represents the velocity relative to the inertial frame, $\Delta\dot{\mathbf{r}}$ the velocity relative to the rotating frame and $\Delta\mathbf{r}$ is the location of the mass elements in question. The dot product of this with itself, which is the kinetic energy integrand, then becomes

$$\Delta\mathbf{v} \cdot \Delta\mathbf{v} = (\Delta\dot{\mathbf{r}} + \widetilde{\boldsymbol{\omega}} \cdot \Delta\mathbf{r}) \cdot (\Delta\dot{\mathbf{r}} + \widetilde{\boldsymbol{\omega}} \cdot \Delta\mathbf{r}) \tag{52}$$

$$= \Delta\dot{\mathbf{r}} \cdot \Delta\dot{\mathbf{r}} + 2\boldsymbol{\omega} \cdot \widetilde{\Delta\mathbf{r}} \cdot \Delta\dot{\mathbf{r}} - \boldsymbol{\omega} \cdot \widetilde{\Delta\mathbf{r}} \cdot \widetilde{\Delta\mathbf{r}} \cdot \boldsymbol{\omega} \; , \tag{53}$$

where we have used the properties of the cross product dyad and rearranged the terms.

Now consider the double integration over each of these terms. The first term is the kinetic energy relative to the rotating frame

$$T_r = \frac{1}{4M} \int_{\mathscr{B}} \int_{\mathscr{B}} \Delta\dot{\mathbf{r}} \cdot \Delta\dot{\mathbf{r}} dm(\mathbf{r}) dm(\mathbf{r}') \; . \tag{54}$$

The final term takes on a simple form as well, once one recalls the definition of the inertia dyad $\bar{\mathbf{I}}$ (see Eq. (15))

$$\frac{1}{2}\boldsymbol{\omega} \cdot \bar{\mathbf{I}} \cdot \boldsymbol{\omega} = -\frac{1}{4M}\boldsymbol{\omega} \cdot \int_{\mathscr{B}} \int_{\mathscr{B}} \widetilde{\Delta\mathbf{r}} \cdot \widetilde{\Delta\mathbf{r}} dm(\mathbf{r}) dm(\mathbf{r}') \cdot \boldsymbol{\omega} \; . \tag{55}$$

From the definition of $\boldsymbol{\omega} = \mathbf{H}/I_H = \dot{\theta}\hat{\mathbf{H}}$, we find that

$$\frac{1}{2}\boldsymbol{\omega} \cdot \bar{\mathbf{I}} \cdot \boldsymbol{\omega} = \frac{1}{2} I_H \dot{\theta}^2 \; . \tag{56}$$

Finally consider the middle term, which we represent as:

$$\boldsymbol{\omega} \cdot \frac{1}{2M} \int_{\mathscr{B}} \int_{\mathscr{B}} \widetilde{\Delta\mathbf{r}} \cdot \Delta\dot{\mathbf{r}} dm(\mathbf{r}) dm(\mathbf{r}') = \boldsymbol{\omega} \cdot \mathbf{H}_r \; , \tag{57}$$

where $\mathbf{H}_r$ is the angular momentum relative to the rotating frame. We will eventually show that this is zero, however we technically cannot make this substitution until the equations of motion are fully defined, meaning that this term participates in the equations of motion, as will be seen.

The result is that the kinetic energy becomes

$$T = T_r + \frac{1}{2} I_H \dot{\theta}^2 + \dot{\theta}\hat{\mathbf{H}} \cdot \mathbf{H}_r \; . \tag{58}$$

3.2.1 Reduced Lagrangian Function

The Lagrangian of the original system is just $L = T - \mathscr{U}$. In this rotating coordinate system it is

$$L = T_r + \dot{\theta}\hat{\mathbf{H}} \cdot \mathbf{H}_r + \frac{1}{2} I_H \dot{\theta}^2 - \mathscr{U} \ . \tag{59}$$

We note that all of the terms are independent of the angle θ, and thus $\partial L / \partial \theta = 0$ leading to the momentum integral

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}} = 0 \ , \tag{60}$$

$$I_H \dot{\theta} + \hat{\mathbf{H}} \cdot \mathbf{H}_r = H \ . \tag{61}$$

If $\hat{\mathbf{H}}$ is chosen along the total angular momentum vector of the system this quantity equals the total angular momentum of the system.

We can apply Routhian reduction to this system (see [1, 3, 15] for a rigorous application of this approach). The Routhian function is then defined as

$$L_R = L - \dot{\theta}\frac{\partial L}{\partial \dot{\theta}} \tag{62}$$

and we can solve for the angular rate $\dot{\theta}$ as

$$\dot{\theta} = \frac{1}{I_H}\left[H - \hat{\mathbf{H}} \cdot \mathbf{H}_r\right] \ . \tag{63}$$

Substituting this back into the newly defined Routhian function and simplifying yields

$$L_R = T_r - \left(\frac{H^2}{2I_H} + \mathscr{U}\right) + \frac{1}{I_H}\mathbf{H} \cdot \mathbf{H}_r - \frac{\left(\hat{\mathbf{H}} \cdot \mathbf{H}_r\right)^2}{2I_H} \ . \tag{64}$$

We define the amended potential $\mathscr{E}$ as

$$\mathscr{E} = \frac{H^2}{2I_H} + \mathscr{U} \ . \tag{65}$$

3.2.2 Equations of Motion

First consider the general equations of motion when no constraints are active. The following results are noted, where we only focus on the translational motion for detailed description

$$
\begin{aligned}
\frac{\partial T_r}{\partial \dot{\mathbf{r}}_{1i}} &= \frac{m_1 m_i}{m_1 + m_i} \dot{\mathbf{r}}_{1i} \ , && (66) \\
\frac{\partial \mathbf{H}_r}{\partial \dot{\mathbf{r}}_{1i}} &= \frac{m_1 m_i}{m_1 + m_i} \tilde{\mathbf{r}}_{1i} \ , && (67) \\
\frac{\partial \mathbf{H}_r}{\partial \mathbf{r}_{1i}} &= -\frac{m_1 m_i}{m_1 + m_i} \tilde{\dot{\mathbf{r}}}_{1i} \ , && (68) \\
i &= 2, 3, \ldots, N \ .
\end{aligned}
$$

In addition, for the current equations of motion we note that H is constant and that $\mathbf{H}_r = \mathbf{0}$, although the partials of this quantity are not necessarily equal to zero and must be incorporated into the equations of motion. This results in

$$
\begin{aligned}
\frac{m_1 m_i}{m_1 + m_i} \left[\ddot{\mathbf{r}}_{1i} + 2\tilde{\boldsymbol{\omega}} \cdot \dot{\mathbf{r}}_{1i} + \dot{\boldsymbol{\omega}} \cdot \tilde{\mathbf{r}}_{1i} \right] &= -\frac{\partial \mathscr{E}}{\partial \mathbf{r}_{1i}} \ , && (69) \\
\frac{d}{dt} \frac{\partial \left(T_r + \boldsymbol{\omega} \cdot \mathbf{H}_r \right)}{\partial \dot{\boldsymbol{\theta}}_{1i}} - \frac{\partial \left(T_r + \boldsymbol{\omega} \cdot \mathbf{H}_r \right)}{\partial \boldsymbol{\theta}_{1i}} &= -\frac{\partial \mathscr{E}}{\partial \boldsymbol{\theta}_{1i}} \ , && (70)
\end{aligned}
$$

where we note that $\dot{\boldsymbol{\omega}} = -\left(H / I_H^2 \right) \frac{dI_H}{dt}$.

Now consider the case when the unilateral constraints, and hence the non-holonomic constraints, between bodies 1 and 2 are active. The equations will take the same form, except for the additional constraint force terms

$$
\begin{aligned}
\frac{m_1 m_2}{m_1 + m_2} \left[\ddot{\mathbf{r}}_{12} + 2\tilde{\boldsymbol{\omega}} \cdot \dot{\mathbf{r}}_{12} + \dot{\boldsymbol{\omega}} \cdot \tilde{\mathbf{r}}_{12} \right] &= -\frac{\partial \mathscr{E}}{\partial \mathbf{r}_{12}} + \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_r] \ , && (71) \\
\frac{d}{dt} \frac{\partial \left(T_r + \boldsymbol{\omega} \cdot \mathbf{H}_r \right)}{\partial \dot{\boldsymbol{\theta}}_{12}} - \frac{\partial \left(T_r + \boldsymbol{\omega} \cdot \mathbf{H}_r \right)}{\partial \boldsymbol{\theta}_{12}} &= -\frac{\partial \mathscr{E}}{\partial \boldsymbol{\theta}_{12}} + \boldsymbol{\lambda}_{12} \cdot [\mathbf{a}_\theta] \ , && (72)
\end{aligned}
$$

along with the constraints

$$
[\mathbf{a}_r] \cdot \dot{\mathbf{r}}_{12} + [\mathbf{a}_\theta] \cdot \dot{\boldsymbol{\theta}}_{12} = 0 \ . \qquad (73)
$$

3.2.3 Jacobi Integral of Motion

As the Lagrangian as defined is time invariant, there should exist a Jacobi integral of motion (i.e., conservation of Energy) in the case where none of the unilateral constraints are active. Under our rolling without slipping assumption for the contact non-holonomic constraints when active, we can show that this integral will also exist. It will not exist should we include slipping motion, which would lead to non-conservative work occurring.

To derive explicitly, take the dot product of Eqs. (69)–(72) with $\dot{\mathbf{r}}_{1i}$ and $\dot{\boldsymbol{\theta}}_{1i}$, respectively, and sum to find

$$\begin{aligned}
&\sum_{i=2}^{N}\left\{\frac{m_1 m_i}{m_1+m_i}\left[\ddot{\mathbf{r}}_{1i}+2\widetilde{\boldsymbol{\omega}}\cdot\dot{\mathbf{r}}_{1i}+\dot{\boldsymbol{\omega}}\cdot\widetilde{\mathbf{r}}_{1i}\right]\right\}\cdot\dot{\mathbf{r}}_{1i}+\\
&\sum_{i=2}^{N}\left\{\frac{d}{dt}\frac{\partial\left(T_r+\boldsymbol{\omega}\cdot\mathbf{H}_r\right)}{\partial\dot{\boldsymbol{\theta}}_{1i}}-\frac{\partial\left(T_r+\boldsymbol{\omega}\cdot\mathbf{H}_r\right)}{\partial\boldsymbol{\theta}_{1i}}\right\}\cdot\dot{\boldsymbol{\theta}}_{1i}\\
=-&\sum_{i=2}^{N}\left\{\frac{\partial\mathscr{E}}{\partial\mathbf{r}_{1i}}\cdot\dot{\mathbf{r}}_{1i}+\frac{\partial\mathscr{E}}{\partial\boldsymbol{\theta}_{1i}}\cdot\dot{\boldsymbol{\theta}}_{1i}\right\}+\boldsymbol{\lambda}_{1i}\cdot\left\{[\mathbf{a}_r]\cdot\dot{\mathbf{r}}_{1i}+[\mathbf{a}_\theta]\cdot\dot{\boldsymbol{\theta}}_{1i}\right\}\ . \qquad (74)
\end{aligned}$$

Going in reverse order, we note that under the no-slip condition the constraint terms should be identically zero. Next, the term involving the amended potential is easily recognized as the total time derivative of $-\mathscr{E}$. Finally, the leading terms can be shown to equal $d/dt\,(T_r+\omega\cdot\mathbf{H}_r)$, however under the equations of motion we note the identity $\mathbf{H}_r=\mathbf{0}$, meaning that its time derivative is also zero.

Thus we explicitly find the energy integral of this system from

$$\frac{d}{dt}\left[T_r+\mathscr{E}\right]=0\,, \qquad (75)$$

$$E=T_r+\mathscr{E}\ . \qquad (76)$$

There are several conclusions we can draw from this analysis. First, from the energy equation we see that

$$E-\mathscr{E}=T_r \qquad (77)$$

$$\geq 0 \qquad (78)$$

and thus we have

$$E\geq\mathscr{E}\,, \qquad (79)$$

with equality occurring when the relative kinetic energy is $T_r=0$. It can be shown that this minimum kinetic energy can be achieved with a system with a given total angular momentum H [12].

We note that the condition $E=\mathscr{E}$ at some instant in time is not sufficient for the system to be in a relative equilibrium (defined below), as the forces acting within the system may not be balanced and thus may cause the system to evolve in time.

If the rolling without slipping assumption is violated, for example if the lateral force is greater than the friction limit, the energy is no longer conserved and is reduced by

$$\frac{d}{dt}\left[T_r + \mathscr{E}\right] = \mathbf{F}_{12} \cdot \dot{\mathbf{r}}_{12} + \mathbf{M}_{12} \cdot \dot{\boldsymbol{\theta}}_{12} \, . \tag{80}$$

For a rigid body system in contact this is likely the main way in which it can reduce its overall energy while maintaining its total angular momentum.

3.3 Equilibrium and Stability Conditions

With the equations of motion specified we can determine conditions for relative equilibrium and conditions for stability. In fact, given the classical form of the energy, split into a quadratic and a potential part, the derivation of stability conditions is simple. The only catch involves the presence of the uni-lateral constraints which exist when the rigid bodies are in contact. We consider the cases separately. First we present some definitions.

Definition 1 (Relative Equilibrium) A given configuration is said to be in "Relative Equilibrium" if its internal kinetic energy is null ($T_r = 0$), meaning that $\mathscr{E} = E$ at an instant, and that it remains in this state over at least a finite interval of time.

Definition 2 (Energetic Stability) A given relative equilibrium is said to be "Energetically Stable" if any equi-energy deviation from that relative equilibrium requires a negative internal kinetic energy, $T_r < 0$, meaning that this motion is not allowed.

Note that energetic stability is different than Lyapunov or spectral stability, which are the usual notions of stability in astrodynamics (these distinctions are discussed in detail for the Full Body Problem in [10]). Energetic stability is stronger in general, as it is robust to any energy dissipation and in fact—if it applies—means that a given relative equilibrium configuration cannot shed any additional energy and thus is static without the injection of exogenous energy, a condition we refer to as being in a (local) minimum energy state.

3.3.1 No Contact Case

When there are no contacts between the bodies, there are necessarily no active unilateral constraints and all of the degrees of freedom are unconstrained. We also note that the kinetic energy is quadratic in the generalized coordinate rates and has the form of a natural system ([3], pg 72). Then the condition for a relative equilibrium is that all of the $\dot{\mathbf{q}} = 0$ (yielding $T_r = 0$) and $\partial\mathscr{E}/\partial\mathbf{q} = \mathbf{0}$.

Energetic stability of the configuration occurs when the Hessian of the amended potential is positive definite, or $\partial^2\mathscr{E}/\partial\mathbf{q}^2 > 0$, meaning that it has only positive eigenvalues. Neutral stability can occur when $\partial^2\mathscr{E}/\partial\mathbf{q}^2 \geq 0$, meaning that at least one eigenvalue is equal to zero. In this case it is possible for the system to drift at a constant rate relative to the equilibrium, ultimately destroying the configuration. If the configuration is not positive definite or semi-definite, then it is unstable and there exists at least one negative eigenvalue and the system can escape from the equilibrium configuration while conserving energy. Another way to consider the unstable case is that the system can still dissipate energy, and thus can evolve to a lower energy state. We note that this is a stronger form of stability than is sometimes used in celestial mechanics and astrodynamics, where spectral stability of linearized motion can sometimes be stable (as in the Lagrange configurations of the 3-body problem that satisfy the Routh criterion).

It is a remarkable fact of celestial mechanics that in the point mass n-body case for $n \geq 3$, the Hessian of any relative equilibrium configuration has at least one negative eigenvalue and is unstable [6]. Thus for the point mass n-body problem all central configurations are always energetically unstable except for the 2-body problem. For the $n = 2$ body problem there is only a single relative equilibrium and it is positive definite and thus stable. If we consider the 3-body problem, we note that while the Lagrange configurations may be spectrally stable when they satisfy the Routh criterion, they are not at a minimum of the amended potential and thus if energy dissipation occurs they can progressively escape from these configurations. We note that for the finite density cases there are always stable configurations at any angular momentum [11].

In keeping with a variational notation, in the no-contact case (i.e., when there are no constraints on the coordinates), the equilibrium condition is

$$\delta\mathscr{E} = 0\,, \tag{81}$$

where $\delta\mathscr{E} = \sum_{i=1}^{n} (\partial\mathscr{E}/\partial q_i)\,\delta q_i$ which corresponds to $\partial\mathscr{E}/\partial q_i = 0$ for all i. The stability condition is

$$\delta^2\mathscr{E} > 0\,, \tag{82}$$

which corresponds to the Hessian $\left[\partial^2\mathscr{E}/\partial q_i \partial q_j\right]$ being positive definite.

3.3.2 Contact Case

The equilibrium and stability conditions must be modified if there are constraints which are activated. We assume, without loss of generality, that generalized coordinates are chosen to correspond to each contact constraint, such that in the vicinity of their being active the unilateral constraint can be restated as

$$\delta q_j \geq 0\,, \tag{83}$$

for $j = 1, 2, \ldots, m$ constraints. We note that these constrained generalized coordinates may either be relative positions or Euler angles between bodies. Assume we have the system at a configuration $\mathbf{q}$ with m active constraints as just enumerated and $T_r = 0$. Further, assume that the $n - m$ unconstrained states satisfy $\mathscr{E}_{q_i} = 0$ for $i = m + 1, \ldots, n$. For this system to be at rest the principle of virtual work and energy states that the variation of the m constrained states are such that the amended potential only increases, or

$$\delta\mathscr{E} \geq 0\,, \tag{84}$$

$$\delta\mathscr{E} = \sum_{j=1}^{m} \mathscr{E}_{q_j}\delta q_j\;, \tag{85}$$

which, for our assumed constraints on the states, is the same as $\mathscr{E}_{q_j} \geq 0$ for $j = 1, 2, \ldots, m$. The derivation of this just notes that, as defined, if the amended potential can only increase in value then motion is not allowed, as this corresponds to a decrease in kinetic energy from its zero value, which of course is non-physical.

For stability, we require the $n - m$ unconstrained variables to satisfy the same positive definite condition as derived earlier. For the constrained states we only need to tighten the condition to $\delta\mathscr{E} > 0$ or $\mathscr{E}_{q_j} > 0$ for $j = 1, 2, \ldots, m$. This last assertion demands some more specific proof and motivates the following general theorem on necessary and sufficient conditions for a relative equilibrium.

4 Application to Euler Resting Configurations

Now let us apply these results to better understand the stability of the N-body Euler Resting configuration, recently studied in [13, 14]. We take a more general version of that problem, allowing each of the bodies to have a different diameter. We do study that same general configuration, however, which has the grains resting on each other in a straight line (see Fig. 3). We note that there are then $N - 1$ active unilateral constraints, and hence twice as many non-holonomic constraints existing between the bodies in contact. The important point to make, however, is that the equilibrium and stability of this configuration can be studied by a pure focus on the amended potential.

In [14] a detailed discussion is given of this system and its constraints, which we summarize here. Mainly, we can describe the system in terms of the relative distance between any two grains, and define an independent set of variations in these distances by considering the displacement between any two grains in

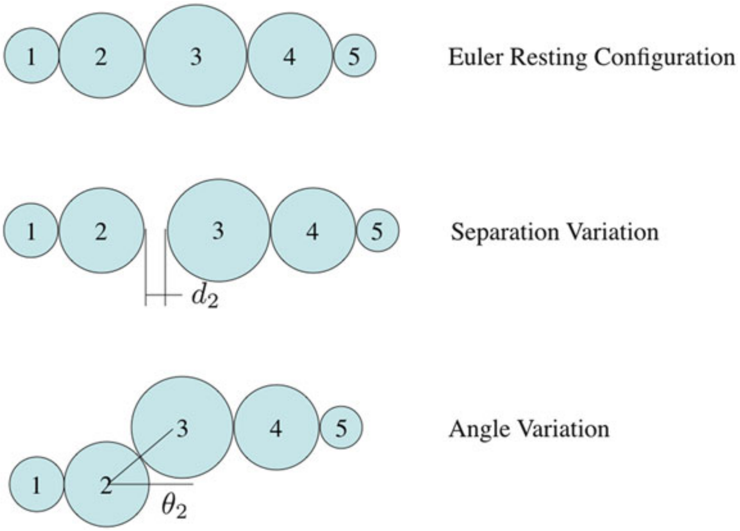


Fig. 3 Definition of the Euler Resting configuration and the distance and angular variations

contact. Similarly, we can restrict ourselves to planar motion and define the relative orientation of the grains by specifying the angle between neighboring grain centers and the initial straight line configuration.

4.1 Specifying the Euler Resting Configuration

We assume that we have N spherical bodies in mutual contact with a neighbor on each side, except at the ends. Each body has mass m_i and diameter D_i, in general we assume they share an equal density (although this can be trivially generalized too). Then the moment of inertia of each body is $I_i = \frac{1}{10} m_i D_i^2$. The relative position vector between any two bodies i and j is then

$$\mathbf{d}_{ij} = \left[\sum_{k=i}^{j-1} \frac{1}{2} (D_k + D_{k+1}) \cos \theta_k \right] \hat{\mathbf{x}} + \left[\sum_{k=i}^{j-1} \frac{1}{2} (D_k + D_{k+1}) \sin \theta_k \right] \hat{\mathbf{y}} \, , \quad (86)$$

where $\hat{\mathbf{x}}$ is taken along the initial aligned configuration and $\hat{\mathbf{y}}$ is in the plane of rotation and perpendicular to it.

$$d_{ij}^2 = \frac{1}{4} \left[\left(\sum_{k=i}^{j-1} (D_k + D_{k+1}) \cos \theta_k \right)^2 + \left(\sum_{k=i}^{j-1} (D_k + D_{k+1}) \sin \theta_k \right)^2 \right] . \quad (87)$$

The system moment of inertia about the axis perpendicular to the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ plane and about the system center of mass, and the gravitational potential are then

$$I_H = \sum_i^N \frac{1}{10} m_i D_i^2 + \frac{1}{\sum_k^N m_k} \sum_{i<j} m_i m_j d_{ij}^2 \ , \tag{88}$$

$$\mathscr{U} = -\mathscr{G} \sum_{i<j} \frac{m_i m_j}{d_{ij}} \ , \tag{89}$$

which combine together in the amended potential

$$\mathscr{E} = \frac{H^2}{2 I_H} + \mathscr{U} \ , \tag{90}$$

where the angular momentum H is a free parameter of the system.

4.2 Equilibrium Conditions

The equilibrium conditions are found by checking that the first variation of the amended potential is zero for the unconstrained variables (which are the angles), and that it is positive for the constrained variables (the distances between any two grains). See Fig. 3 for the geometrical definition of these variations. We consider the angular variables first, meaning that we need to solve for $\partial \mathscr{E} / \partial \theta_m = 0$, yielding

$$-\frac{H^2}{2 I_H^2} \frac{\partial I_H}{\partial \theta_m} + \frac{\partial \mathscr{U}}{\partial \theta_m} = 0 \ , \tag{91}$$

where we note that

$$\frac{\partial I_H}{\partial \theta_m} = \frac{1}{\sum_k^N m_k} \sum_{i<j} m_i m_j \frac{\partial d_{ij}^2}{\partial \theta_m} \ , \tag{92}$$

$$\frac{\partial \mathscr{U}}{\partial \theta_m} = \frac{\mathscr{G}}{2} \sum_{i<j} \frac{m_i m_j}{d_{ij}^3} \frac{\partial d_{ij}^2}{\partial \theta_m} \ , \tag{93}$$

meaning that we just need to solve for $\frac{\partial d_{ij}^2}{\partial \theta_m}$. Carrying out this partial we find

$$\frac{\partial d_{ij}^2}{\partial \theta_m} = \frac{D_m + D_{m+1}}{2} \begin{cases} 0 & m \leq i-1 \\ \Big[-\sin\theta_m \sum_{k=i}^{j-1} (D_k + D_{k+1}) \cos\theta_k + \\ \cos\theta_m \sum_{k=i}^{j-1} (D_k + D_{k+1}) \sin\theta_k \Big] & i \leq m \leq j-1 \\ 0 & m \geq j \end{cases} . \quad (94)$$

It is clear that this equals 0 when all of the angles $\theta_k = 0$, or if they are all equal to each other. This establishes the Euler resting configuration as a possible equilibrium, so long as the unilateral constraints are all active.

To test whether all of the unilateral constraints are active, or more precisely that the constraint forces are all opposing the compression, we must determine whether $\partial \mathscr{E}/\partial d_m \geq 0$, where d_m is defined as the distance between particles m and $m+1$. This is evaluated at the equilibrium condition, meaning that we can set $\theta_k = 0$ and then evaluate the partial.

$$-\frac{H^2}{2I_H^2}\frac{\partial I_H}{\partial d_m} + \frac{\partial \mathscr{U}}{\partial d_m} \geq 0 , \quad (95)$$

where

$$\frac{\partial I_H}{\partial d_m} = \frac{2}{\sum_k^N m_k} \sum_{i<j} m_i m_j d_{ij} \frac{\partial d_{ij}}{\partial d_m} , \quad (96)$$

$$\frac{\partial \mathscr{U}}{\partial d_m} = \mathscr{G} \sum_{i<j} \frac{m_i m_j}{d_{ij}^2} \frac{\partial d_{ij}}{\partial d_m} . \quad (97)$$

Now we just need to solve for $\frac{\partial d_{ij}}{\partial d_m}$. Carrying out this partial we find

$$\frac{\partial d_{ij}}{\partial d_m} = \begin{cases} 0 & m \leq i-1 \\ 1 \; i \leq m \leq j-1 \\ 0 & m \geq j \end{cases} . \quad (98)$$

Substituting this into the partials for I_H and $\mathscr{U}$ gives a simplified set of expressions

$$\frac{\partial I_H}{\partial d_m} = \frac{2}{\sum_k^N m_k} \sum_{i=1}^{m} \sum_{j=m+1}^{N} m_i m_j d_{ij} \ , \tag{99}$$

$$\frac{\partial \mathscr{U}}{\partial d_m} = \mathscr{G} \sum_{i=1}^{m} \sum_{j=m+1}^{N} \frac{m_i m_j}{d_{ij}^2} \ , \tag{100}$$

$$d_{ij} = \sum_{k=i}^{j} D_k - \frac{1}{2}\left(D_i + D_j\right) \ . \tag{101}$$

The existence of the equilibrium then depends on the level of angular momentum of the system. Specifically, we find a limit on H for each variation in d_m

$$\left(\frac{H}{I_H}\right)^2 \leq \frac{2\partial \mathscr{U}/\partial d_m}{\partial I_H/\partial d_m} \tag{102}$$

$$= \mathscr{G} M \frac{\sum_{i=1}^m \sum_{j=m+1}^N \frac{m_i m_j}{d_{ij}^2}}{\sum_{i=1}^m \sum_{j=m+1}^N m_i m_j d_{ij}} \tag{103}$$

$$= \Omega_{F,m}^2 \ . \tag{104}$$

where we note that the ratio H/I_H is just the total spin rate of the system, and the spin rate $\Omega_{F,m}$ is the spin rate at which the Euler resting configuration will separate between bodies m and $m + 1$. Thus the limiting spin rate for the system to exist as a relative equilibrium is then

$$\Omega_F = \min_m \Omega_{F,m} \ , \tag{105}$$

where we call this limiting spin rate the "fission" spin rate.

4.3 Stability Conditions

If $H^2 \leq I_H^2 \Omega_F^2$ then the Euler resting configuration is a relative equilibria. To test stability then requires us to compute the Hessian $\mathscr{E}_{mn} = \left[\partial^2 \mathscr{E}/\partial \theta_m \partial \theta_n\right]$ evaluated at the equilibrium condition and test whether it is positive definite.

The second partial of the amended potential with respect to an angle θ_n is

$$\frac{\partial^2 \mathscr{E}}{\partial\theta_n \partial\theta_m} = \frac{H^2}{I_H^3}\frac{\partial I_H}{\partial\theta_m}\frac{\partial I_H}{\partial\theta_n} - \frac{H^2}{2I_H^2}\frac{\partial^2 I_H}{\partial\theta_n\partial\theta_m} + \frac{\partial^2 \mathscr{U}}{\partial\theta_n\partial\theta_m} . \tag{106}$$

However when evaluated at the equilibrium the partial $\frac{\partial I_H}{\partial\theta_m} = 0$. In addition, the second partials of I_H and $\mathscr{U}$ are

$$\frac{\partial^2 I_H}{\partial\theta_n\partial\theta_m} = \frac{1}{M}\sum_{i<j} m_i m_j \frac{\partial^2 d_{ij}^2}{\partial\theta_n\partial\theta_m} , \tag{107}$$

$$\frac{\partial^2 \mathscr{U}}{\partial\theta_n\partial\theta_m} = \frac{\mathscr{G}}{2}\sum_{i<j} \frac{m_i m_j}{d_{ij}^3} \frac{\partial^2 d_{ij}^2}{\partial\theta_n\partial\theta_m} , \tag{108}$$

with an additional term in the partial of the gravitational potential being equal to zero at the equilibrium condition. Finally, we note that the second partial of the term d_{ij}^2 evaluated at the equilibrium is

$$\frac{\partial^2 d_{ij}^2}{\partial\theta_n\partial\theta_m} = \frac{1}{2}\begin{cases} 0 & m \le i-1 \\ \left(D_m + D_{m+1}\right)\left(D_n + D_{n+1}\right) & i \le m < n \le j-1 \\ \left(D_m + D_{m+1}\right)\left[D_m + D_{m+1} - \sum_{k=i}^{j-1}\left(D_k + D_{k+1}\right)\right] & i \le m = n \le j-1 \\ 0 & n \ge j \end{cases} ,\tag{109}$$

where we assume $m \le n$. These can all be combined into a general form of the second partial of the amended potential

$$\frac{\partial^2 \mathscr{E}}{\partial\theta_n\partial\theta_m} = \frac{\mathscr{G}}{2}\sum_{i=1}^{m}\sum_{j=n+1}^{N} \frac{m_i m_j}{d_{ij}^3}\left[1 - \left(\frac{H}{I_H}\right)^2 \frac{d_{ij}^3}{\mathscr{G}M}\right]\frac{\partial^2 d_{ij}^2}{\partial\theta_n\partial\theta_m} . \tag{110}$$

This provides a detailed algorithm for the computation of the matrix entries. It is also useful to develop a matrix-level notation for this system, where we can note that when evaluated at the relative equilibrium the spin rate is equal to the ratio $H/I_H = \Omega$ again. Denote $\mathscr{E}_{\theta\theta} = \left[\frac{\partial^2 \mathscr{E}}{\partial\theta_n\partial\theta_m}\right]$, and similarly denote $\mathscr{U}_{\theta\theta} = \left[\frac{\partial^2 \mathscr{U}}{\partial\theta_n\partial\theta_m}\right]$ and $I_{\theta\theta} = \left[\frac{\partial^2 I_H}{\partial\theta_n\partial\theta_m}\right]$. Then the Hessian of the amended potential has the general form

$$\mathscr{E}_{\theta\theta} = \mathscr{U}_{\theta\theta} - \frac{1}{2}\Omega^2 I_{\theta\theta} \tag{111}$$

and stability occurs when $\mathscr{E}_{\theta\theta}$ is positive definite, meaning that all of its eigenvalues are positive. We note that in general the matrix $I_{\theta\theta}$ is non-singular, thus allowing us to rewrite the equation as

$$\mathscr{E}_{\theta\theta} = \frac{1}{2} I_{\theta\theta} \left[2 I_{\theta\theta}^{-1} \mathscr{U}_{\theta\theta} - \Omega^2 I \right] , \tag{112}$$

where I is the identity matrix. Thus we can reduce this to a classic eigenvalue problem by finding the eigenvalues of $2 I_{\theta\theta}^{-1} \mathscr{U}_{\theta\theta}$, denoted as λ_k^2. Then the condition for stability is that $\Omega^2 > \max_k \lambda_k^2 = \Omega_S^2$, defining the minimum spin rate for complete stabilization of the system.

We do note one detail from the discussion in [14] about the size of this matrix. Although there are $N - 1$ degrees of freedom in the angular variables, we only need to test the system for $N - 2$ variations, as there is always one null eigenvalue of the system corresponding to the uniform rotation of the system. In other words, for a 2-body system there is no need to test for stability, while for a 3 body system there is only one unique angle of variation between the bodies, etc.

4.4 Results

Given these detailed results we can explicitly test the stability and fission spin rates for a range of N and relative sizes and masses of the system. The key parameter to test for a given configuration is whether the interval $\Delta\Omega = \Omega_F - \Omega_S$ is positive, meaning that there will be a range of spin rates for which the system is in a relative equilibrium and is stable.

Stability of Equal Size Configurations as a Function of N We first recreate the topic studied in [14] and corrected in [13]. Here we assume that all the bodies are of equal size and density, and determine whether $\Delta\Omega > 0$ as a function of N. The difference is plotted in Fig. 4 as a function of N, also shown are the stabilization and fission spin rates. The spin rates are normalized by the critical spin rate of a single sphere, defined as the mean motion at the surface. Here we can clearly see that the transition occurs at $N = 21$, meaning that for $N \geq 22$ an equal mass and size Euler resting configuration can never be stable.

Robustness to Relative Variations in Particle Size Next we consider the robustness of the configurations as the relative size of a single grain is varied. First, we just consider the effect of varying the size of the center grain for an odd number of particles, and investigate the stability gap $\Delta\Omega$ as the center grain varies from 0.1 to 10 times the other grains. Figure 5 shows the evolution of the stability gap as a function of the center grain size for a number of different odd N. We note that the

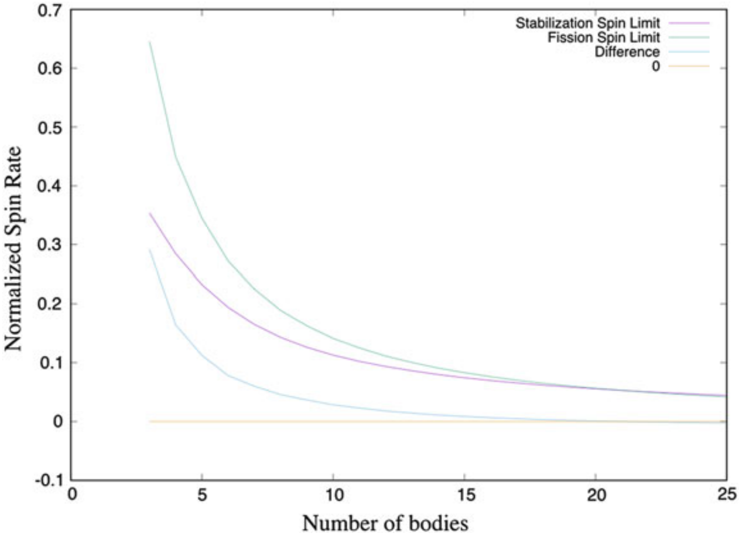


Fig. 4 Normalized spin rates as a function of N for equal-sized Euler Resting configuration. The stabilization rate exceeds the fission rate for $N \geq 22$

Euler resting configuration can be stable above $N = 22$ if the central grain is slightly larger than the other grains. However, by $N = 29$ this is no longer the case and the Euler resting configuration is not stable for any size of the central body. Most likely, growing additional central grains could stabilize this, however the number of free parameters is too large to systematically study here.

As a case study Fig. 6 shows the different stabilization and fission spin rates for different modes as a function of the central body size for the case of $N = 21$. On the bottom shows the difference between the relevant limiting spin rates, defining a region of positive stability.

We note that the case $N = 5$ has a non-zero stability gap across the entire spectrum of size variations, indicating a level of robustness that is not present at the $N = 7$ level, which cannot be stabilized once the center grain is 3.3 times larger than the others. Thus we explore the robustness of $N = 5$ to variations in its other grains as well. We find that any of the grains in the 5 body system can be shrunk to arbitrarily small size. The end grain can be increased up to 2.8 times the others before stability ceases. The next grain can be increased to 3.2 times the other grains before the stability gap closes. And as seen in Fig. 4 the central grain can be increased to an arbitrary size without the stability gap disappearing.

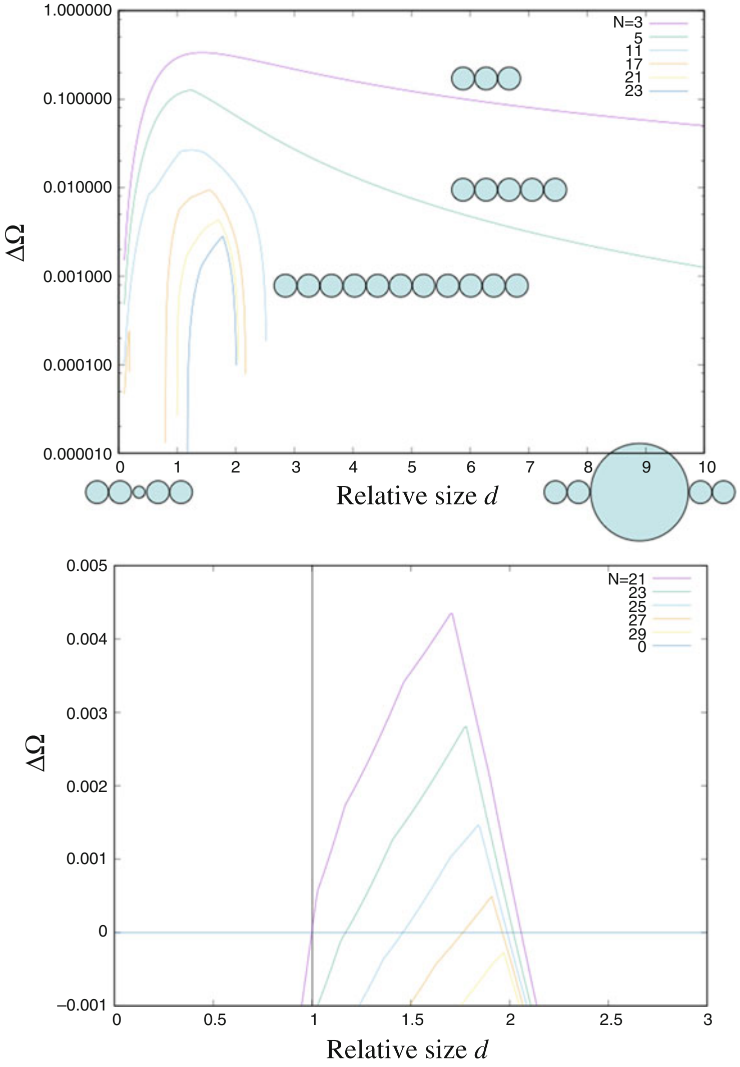


Fig. 5 Top: Spin rate stability gap as a function of the relative size of the central body for a number of odd Euler Resting configurations. Bottom: Detail showing the loss of stability as a function of central body size for higher numbers of bodies

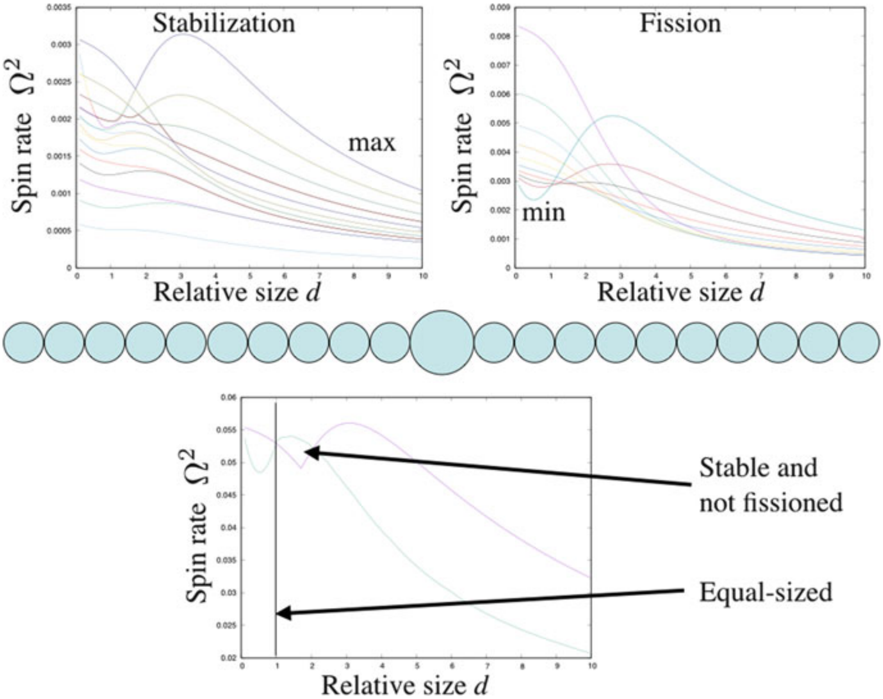


Fig. 6 Stabilization and fission spin rates as a function of central body size for $N = 21$. The stabilization rates plotted are the different eigenvalues of the matrix. The fission rates are for separation of different members of the Euler Resting configuration. The bottom plot shows the maximum stability envelope and minimum fission envelope

5 Conclusions

The Celestial Mechanics of a finite density N-body system is studied, with a focus on accounting for the contact constraints when the components are resting on each other. It is shown that, under no-slip assumptions, the equilibria and stability of a resting configuration can be studied using only the amended potential. A detailed example is given for the case of N bodies resting on each other in a line, serving as an application of the theory to a contact case.

Acknowledgements This chapter was derived from a set of lectures given at the SDSM 2017 School entitled: "Satellite Dynamics and Space Missions: Theory and Applications of Celestial Mechanics," held August 28–September 2, 2017, in San Martino al Cimino (VT), Italy. The author acknowledges support from NASA's SSERVI program (Institute for the Science of Exploration Targets) through institute grant number NNA14AB03A. The author acknowledges Dr. Giulio Baù for his insightful and comprehensive review of this contribution, which greatly helped to improve its discussion and correctness.

References

1. Cendra, H., Marsden, J.E.: Geometric mechanics and the dynamics of asteroid pairs. Dyn. Syst. Int. J. **20**, 3–21 (2005)
2. Fujiwara, A., Kawaguchi, J., Yeomans, D.K., Abe, M., Mukai, T., Okada, T., Saito, J., Yano, H., Yoshikawa, M., Scheeres, D.J., et al.: The rubble-pile asteroid Itokawa as observed by Hayabusa. Science **312**(5778), 1330–1334 (2006)
3. Greenwood, D.T.: Classical Dynamics. Dover, Mineola (1997)
4. Holsapple, K.A.: On YORP-induced spin deformations of asteroids. Icarus **205**(2), 430–442 (2010)
5. Liebenthal, E.: Untersuchungen uber die Attraction zweier homogener Korper. PhD thesis, Universitat Greifswald (1880)
6. Moeckel, R.: On central configurations. Math. Z. **205**(1), 499–517 (1990)
7. Pravec, P., Harris, A.W.: Fast and slow rotation of asteroids. Icarus **148**, 12–20 (2000)
8. Sánchez, P., Scheeres, D.J.: Simulating asteroid rubble piles with a self-gravitating soft-sphere distinct element method model. Astrophys. J. **727**, 120 (2011)
9. Scheeres, D.J.: Stability in the full two-body problem. Celest. Mech. Dyn. Astron. **83**(1), 155–169 (2002)
10. Scheeres, D.J.: Relative equilibria for general gravity fields in the sphere-restricted full 2-body problem. Celest. Mech. Dyn. Astron. **94**, 317–349 (2006)
11. Scheeres, D.J.: Minimum energy configurations in the N-body problem and the celestial mechanics of granular systems. Celest. Mech. Dyn. Astron. **113**(3), 291–320 (2012)
12. Scheeres, D.J.: Relative equilibria in the full N-body problem with applications to the equal mass problem. In: Recent Advances in Celestial and Space Mechanics, pp. 31–81. Springer, Cham (2016)
13. Scheeres, D.J.: Correction to: stability of the Euler resting N-body relative equilibria. Celest. Mech. Dyn. Astron. **130**, 55–56 (2018)
14. Scheeres, D.J.: Stability of the Euler resting N-body relative equilibria. Celest. Mech. Dyn. Astron. **130**(3), 26 (2018)
15. Simo, J.C., Lewis, D., Marsden, J.E.: Stability of relative equilibria. Part I: the reduced energy-momentum method. Arch. Ration. Mech. Anal. **115**(1), 15–59 (1991)
16. Walsh, K.J., Richardson, D.C., Michel, P.: Spin-up of rubble-pile asteroids: disruption, satellite formation, and equilibrium shapes. Icarus **220**(2), 514–529 (2012)
17. Werner, R.A., Scheeres, D.J.: Mutual potential of homogeneous polyhedra. Celest. Mech. Dyn. Astron. **91**, 337–349 (2005)

Multi-Objective Optimal Control: A Direct Approach

Massimiliano Vasile

Abstract The chapter introduces an approach to solve optimal control problems with multiple conflicting objectives. The approach proposed in this chapter generates sets of Pareto optimal control laws that satisfy a set of boundary conditions and path constraints. The chapter starts by introducing basic concepts of multi-objective optimisation and optimal control theory and then presents a general formulation of multi-objective optimal control problems in scalar form using the Pascoletti-Serafini scalarisation method. From this scalar form the chapter derives the first order necessary conditions for local optimality and develops a direct transcription method by Finite Elements in Time (DFET) that turns the infinite dimensional multi-objective optimal control problem into a finite dimensional multi-objective nonlinear programming problem (MONLP). The transcription method is proven to be locally convergent under some assumptions on the nature of the optimal control problem. A memetic agent-based optimisation approach is then proposed to solve the MONLP problem and return a partial reconstruction of the globally optimal Pareto set. An illustrative example concludes the chapter.

Keywords Multi-objective optimisation · Optimal control · Finite elements · Trajectory optimisation

1 Introduction

Optimal control theory is a branch of mathematical optimisation that searches for control laws, or policies, that optimise (minimise or maximise) a given cost function and drive a dynamical system from an initial to a final state. Methods for the solution

M. Vasile (✉)
University of Strathclyde, Glasgow, UK
e-mail: massimiliano.vasile@strath.ac.uk

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_6

of optimal control problems are generally divided in two categories: direct and indirect. Indirect methods derive and solve a set of differential-algebraic equations that satisfy Pontryagin's maximum principle, while direct methods transcribe the infinite dimensional optimal control problem into a finite dimensional nonlinear programming (NLP) problem and solve it with a numerical optimisation method. Optimal control theory and most existing numerical methods for the solution of optimal control problems generally consider a single scalar cost function. However, in many real scenarios one is interested in optimising several, often conflicting, performance indexes.

In this case, the problem is to find the set of solutions such that none of the objective functions can be improved in value without degrading some of the other objective values. These solutions are called Pareto optimal or Pareto efficient after the economist Vilfrido Pareto who first introduced the concept of Pareto efficiency or Pareto optimality.

Methods for the solution of multi-objective nonlinear programming problems exist and are well supported by theory. However, only a few methods have been proposed in the literature for the solution of multi-objective optimal control problems [1–4].

This chapter presents a methodology to numerically solve general multi-objective optimal control problems. The main difference with respect to traditional single objective optimal control problems is that the solution corresponds to a set of optimal control laws, rather than a single optimal one. The solution approach proposed in this chapter transcribes the original control problem into a Multi-Objective Non-Linear Programming (MONLP) problem using a particular technique based on Finite Elements in Time. The resulting NONLP is then solved with a memetic algorithm that combines a population-based exploration of the search space with a gradient-based strategy for local convergence.

The chapter is structured as follows: after stating the problem under investigation and introducing some basic concepts of optimal control and multi-objective optimisation, the chapter will present the transcription method based on Finite Elements in Time and the solution of the resulting NLP problem. Along with the solution methodology, the chapter proposes two theoretical developments that provide a set of necessary conditions for the local optimality of the solutions. A simple example will demonstrate the applicability of the proposed approach.

2 Definitions and Preliminary Ideas

The first section of this chapter will introduce some basic concepts and ideas and the notation that will be used in the remainder of the chapter.

2.1 Multi-Objective Optimal Control Problem

Consider the following multi-objective optimal control problem (MOCP):

$$\begin{aligned} &\min_{\mathbf{u}} \mathbf{F} \\ &s.t. \\ &\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \\ &\mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0 \\ &\boldsymbol{\psi}(\mathbf{x}_0, \mathbf{x}_f, t_0, t_f) \leq 0 \\ &t \in [t_0, t_f] \end{aligned} \tag{MOCP}$$

where $\mathbf{F} = [f_1, \ldots, f_i, \ldots, f_m]^T$ is, in general, a vector function of the state variables $\mathbf{x} : [t_0, t_f] \to \mathbb{R}^n$, control variables $\mathbf{u} \in L^\infty(U \subset \mathbb{R}^{n_u})$ and time t. Functions $\mathbf{x}$ belong to the Sobolev space $W^{1,\infty}$, objective functions are $f_i : \mathbb{R}^{2n+n} \times \mathbb{R}^{n_u} \times [t_0, t_f] \longrightarrow \mathbb{R}$, $\mathbf{h} : \mathbb{R}^n \times \mathbb{R}^{n_u} \times [t_0, t_f] \longrightarrow \mathbb{R}^n$, algebraic constraint function $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}^{n_u} \times [t_0, t_f] \longrightarrow \mathbb{R}^q$, and boundary condition functions $\boldsymbol{\psi} : \mathbb{R}^{2n+2} \longrightarrow \mathbb{R}^{n_\psi}$. Note that problem (MOCP) is generally non-smooth and can have many local minima.

2.2 Pareto Dominance and Efficiency

At the beginning of the twentieth century, Vilfredo Pareto, an Italian engineer, sociologist, economist, political scientist, and philosopher, introduced a revolutionary concept in economics: the notion of what is now known as Pareto-optimality, or the idea that maximum economic satisfaction can be achieved when no one can be made better off without making someone else worse off. In mathematical terms this can be written in the following way. Consider the vector functions $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$, with $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_i(\mathbf{x}), \ldots, f_m(\mathbf{x})]^T$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^q$, with $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_j(\mathbf{x}), \ldots, g_q(\mathbf{x})]^T$ and problem

$$\begin{aligned} &\min_{\mathbf{x}} \mathbf{F} \\ &s.t. \\ &\mathbf{g}(\mathbf{x}) \leq 0 \end{aligned} \tag{MOP}$$

Given the feasible set $X = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{g}(\mathbf{x}) \leq 0\}$ and two feasible vectors $\mathbf{x}, \hat{\mathbf{x}} \in X$, we say that $\mathbf{x}$ is dominated by $\hat{\mathbf{x}}$ if $f_i(\hat{\mathbf{x}}) \leq f_i(\mathbf{x})$ for all $i = 1, \ldots, m$ and there exists a k so that $f_k(\hat{\mathbf{x}}) \neq f_k(\mathbf{x})$. We use the relation $\hat{\mathbf{x}} \prec \mathbf{x}$ that states that $\hat{\mathbf{x}}$ dominates $\mathbf{x}$. A vector $\mathbf{x}^* \in X$ will be said to be Pareto efficient, or optimal, with respect to

Problem (MOP) if there is no other vector $\mathbf{x} \in X$ dominating $\mathbf{x}^*$ or:

$$\mathbf{x} \not\prec \mathbf{x}^*, \qquad \forall \mathbf{x} \in X - \{\mathbf{x}^*\} \tag{1}$$

All non-dominated decision vectors in X form the Pareto set X_P and the corresponding image in criteria space is the Pareto front.

2.3 Karush-Khun-Tucker Optimality Conditions for Multi-Objective Problems

The notion of Pareto-optimality and dominance do not immediately translate into a criterion for a solution to be optimal. In this section we introduce a set of necessary conditions for local optimality for generic constrained optimisation problems with multiple objectives. A set of necessary conditions for local optimality of scalar problems were first stated by William Karush in his master's thesis in 1939 and then, independently, published by Harold W. Kuhn, and Albert W. Tucker in 1951. A treatment of vector objective functions can be found in [5] and is reported here in the form of a theorem on the local optimality of a solution $\mathbf{x}^*$ of problem MOP.

Theorem 1 (KKT) *If $\mathbf{x}^* \in X$ is an efficient solution to problem MOP and a regular point of the constraints $\mathbf{g}$, then there exist vectors $\eta \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}^q$ such that:*

$$\sum_i^m \eta_i \nabla f_i(\mathbf{x}^*) + \sum_j^q \lambda_j \nabla g_j(\mathbf{x}^*) = 0 \tag{2}$$

$$\sum_j^q \lambda_j g_j(\mathbf{x}^*) = 0 \tag{3}$$

$$g_j(\mathbf{x}^*) \leq 0, \quad j = 1, \ldots, q \tag{4}$$

$$\lambda_j \geq 0, j = 1, \ldots, q \tag{5}$$

$$\eta_i \geq 0, i = 1, \ldots, m \tag{6}$$

$$\exists \eta_i > 0 \tag{7}$$

In the unconstrained case KKT optimality conditions reduce to:

$$\sum_i^m \eta_i \nabla f_i(\mathbf{x}^*) = 0 \tag{8}$$

$$\eta_i \geq 0, i = 1, \ldots, m \tag{9}$$

$$\exists \eta_i > 0 \tag{10}$$

Condition (8) leads to an interesting result (see [6]) that the Pareto set is an $m - 1$ dimensional manifold. This implies that the Pareto set has zero measure in $\mathbb{R}^n$ with $m \leq n$, which means that simply sampling the n dimensional domain of a vector function is not sufficient to reconstruct the Pareto set. The development of a solution algorithm, therefore, requires more sophisticated heuristics.

2.4 *Pascoletti-Serafini Scalarisation*

One way to address the solution of multi-objective optimisation problems is to translate the original vector objective function into a scalar and then use any method for single objective optimisation. The KKT conditions (1) already suggest that a weighted sum of the objective functions could work. However, this approach does not work in practice if the Pareto front is not convex. A better way to solve the problem is to define a descent cone, in criteria space, that eventually converges to the Pareto front.

In 1984 Adriano Pascoletti and Paolo Serafini introduced a scalarisation method [7] based on this idea of descent cones. According to the Pascoletti-Serafini method, an optimal (or K-minimal) solution to problem (MOP) is also solving the following constrained single objective optimisation problem:

$$\begin{aligned} & \min_{t \in \mathbb{R}} t \\ & s.t. \\ & \mathbf{a}t - \mathbf{F}(\mathbf{x}) + \mathbf{r} \in K \\ & \mathbf{g}(\mathbf{x}) \leq 0 \\ & \mathbf{a} \in \mathbb{R}^m \\ & \mathbf{r} \in \mathbb{R}^m \end{aligned} \tag{11}$$

where K identifies a descent cone whose vertex slides along the rectilinear line parameterised in t and defined by the vectors $\mathbf{a}$ and $\mathbf{r}$. In a more computationally friendly form, problem (11) can be written as:

$$\begin{aligned} & \min_{s \geq 0} s \\ & s.t. \\ & \boldsymbol{\omega}_i (f_i(\mathbf{x}) - z_i) \leq \alpha \quad \forall i = 1, \ldots, m \\ & \mathbf{g}(\mathbf{x}) \leq 0 \\ & \mathbf{z} \in \mathbb{R}^m \\ & \boldsymbol{\omega} \in \mathbb{R}^m_+ \end{aligned} \tag{PS}$$

where the points $\mathbf{z} = [z_1, \ldots, z_i, \ldots, z_m]^T$ and the vectors of positive weights $\boldsymbol{\omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_i, \ldots, \boldsymbol{\omega}_m]^T$ define rectilinear lines (or descent directions) in the m dimensional criteria space. The K-cone in problem (PS) and the associated descent directions can be represented as in Fig. 1a. When the cone reaches the Pareto front

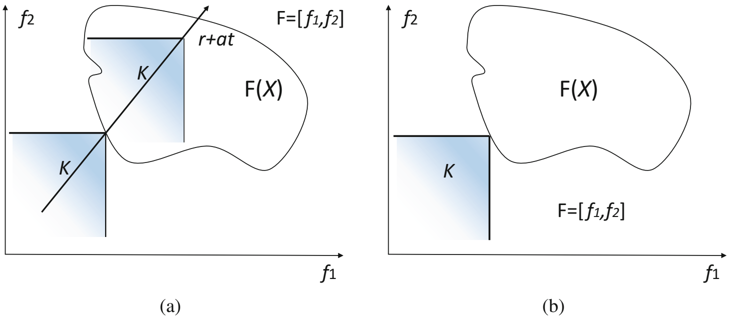


Fig. 1 Convergence of the Pascoletti-Serafini scalarisation: (**a**) Descent K-cone, (**b**) K-efficient point

the vertex of the cone identifies a Pareto efficient point (see Fig. 1b). More formally a point is K-minimal when:

$$(\bar{\mathbf{F}} - K) \cap \mathbf{F}(X) = \{\bar{\mathbf{F}}\}$$

From this definition and from Fig. 1, one can understand that a K-minimal point is Pareto efficient.

2.5 Chebyshev Scalarisation

A different approach to treat vector objective functions is known as Chebyshev scalarisation. This method is here presented as in [8] and similar to Pascoletti-Serafini sclarisation it is based on the idea of descent directions ζ identified by the vector of weights $\boldsymbol{\omega}$:

$$\begin{aligned} &\min_{\mathbf{x}\in X} \max_{i\in\{1,\dots,m\}} \omega_i (f_i(\mathbf{x}) - z_i) \\ &s.t. \\ &\mathbf{g}(\mathbf{x}) \leq 0 \end{aligned} \tag{CS}$$

This form of scalarisation is not introducing any constraints on the objective functions and can be directly used with a sampling or population-based method (like any Evolutionary Computation technique [9, 10]) as it represents a simple method to accept or reject a sample. The major difficulty, as for (PS), is to properly define the descent directions. Figure 2 illustrates the logic of Chebyshev scalarisation and the region that satisfies condition (CS) where $\chi = \max_{i\in\{1,\dots,m\}} \omega_i (f_i(\mathbf{x}) - z_i)$. In [8], the author analyses the relationship among different scalarisation methods and presents the following important theoretical results on the equivalence between the solution of problem (CS) and (PS).

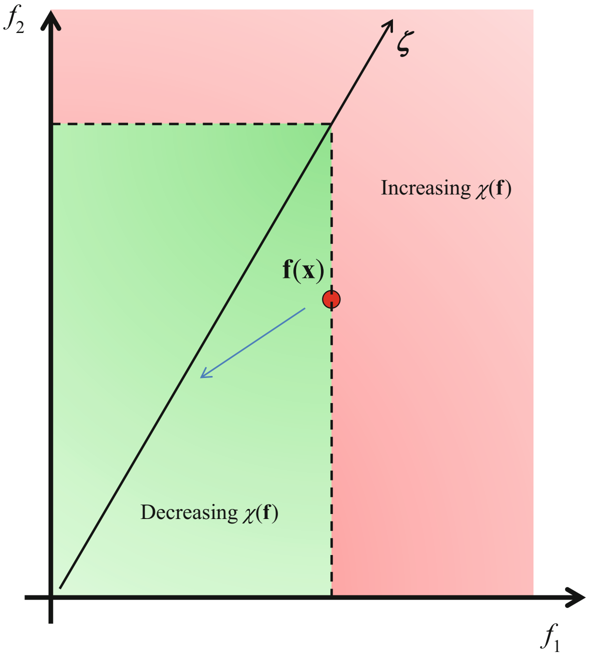


Fig. 2 Logic of Chebyshev's scalarisation: at each step, the green region is where a new solution is accepted. The direction defined by ζ is the descent direction defined by the weights ω_j

Theorem 2 (CS) *A point* $(s, \mathbf{x}) \in \mathbb{R} \times X$ *is a minimal solution of problem (PS) with* $\mathbf{z} \in \mathbb{R}^m$, $z_i < \min_{\mathbf{x} \in X} f_i(\mathbf{x})$, $i = 1, \ldots, m$, *and* $\boldsymbol{\omega} \in int(\mathbb{R}^m_+)$, *if and only if* $\mathbf{x}$ *is a solution of problem (CS).*

From theorem (CS) one can expect that the solution of problem (PS) and (CS) are equivalent. This is an important property when designing algorithms because, in some cases (as it will be shown later in this chapter), the solution of (PS) practically translates into the solution of (CS) or a partial solution to (CS) might need to be improved by solving (PS).

2.6 (Scalar) Pontryagin Maximum Principle

In this chapter we focus on optimal control problems that are formulated in the so called Mayer's form. Other forms, Lagrange and Bolza, express the cost function respectively either as the time integral of a function that depends on states, controls and time or a mix of time integral and scalar function of the terminal conditions [11]. It is easy to verify that the three forms are equivalent and lead to the same solution. However, for the discussion that follows, Mayer's form is more easily applicable to a scalarised MOCP. In Mayer's form the cost function depends on the terminal conditions on states and time. The resulting optimal control problem reads:

$$\begin{array}{l} \min_{\mathbf{u}} f(\mathbf{x}_f, t_f) \\ s.t \\ \dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \\ \mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0 \\ \boldsymbol{\psi}(\mathbf{x}_0, \mathbf{x}_f, t_0, t_f) \leq 0 \\ t \in [t_0, t_f] \end{array} \tag{12}$$

where $\mathbf{x}$ is the state vector, $\mathbf{u}$ the control vector, $\mathbf{h}$ is the dynamic function, $\mathbf{g}$ a set of path constraints, $\boldsymbol{\psi}$ the boundary constraints and t is the time. If $\mathbf{u}^*$ is a locally optimal solution for problem (12), then Potryagin's minimum (maximum) principle says that there exist a vector $\lambda \in \mathbb{R}^n$, $\nu \in \mathbb{R}^{n_\psi}$ and a vector $\mu \in \mathbb{R}^q$ such that:

$$\begin{array}{l} \mathbf{u}^* = \text{argmin}_{\mathbf{u} \in U} (\lambda^T \mathbf{h}(\mathbf{x}^*, \mathbf{u}, t) + \mu^T \mathbf{g}(\mathbf{x}^*, \mathbf{u}, t)) \\ U = \{\mathbf{u} | \mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0\} \\ \lambda^T \nabla_x \mathbf{h}(\mathbf{x}^*, \mathbf{u}^*, t) + \mu^T \nabla_x \mathbf{g}(\mathbf{x}^*, \mathbf{u}^*, t) + \dot{\lambda} = 0 \\ \mu \geq 0 \end{array} \tag{13}$$

with transversality conditions:

$$\begin{array}{l} \nabla_x f + \nu^T \nabla_x \boldsymbol{\psi} = \lambda(t_f) \\ \lambda^T \mathbf{h}(\mathbf{x}^*, \mathbf{u}, t_f) + \mu^T \mathbf{g}(\mathbf{x}^*, \mathbf{u}, t_f) + (\nabla_t f + \nu^T \nabla_t \boldsymbol{\psi})_{t_f} = 0 \quad [\text{if } t_f \text{ is free}] \\ \nu \geq 0 \end{array} \tag{14}$$

Equations (13) represent a system of Differential Algebraic Equations (DAE) that needs to be solved with boundary conditions (14). Equations (13) and (14) are necessary condition for optimality and in case a locally optimal control is sought (and both $\mathbf{h}$ and $\mathbf{g}$ are locally differentiable with respect to the controls $\mathbf{u} \in U$) one can express the first of (13) as:

$$\nabla_{\mathbf{u}} (\lambda^T \mathbf{h}(\mathbf{x}^*, \mathbf{u}, t) + \mu^T \mathbf{g}(\mathbf{x}^*, \mathbf{u}, t)) = 0 \tag{15}$$

2.7 Pascoletti-Serafini Scalarised MOCP

Problem (12) has a single scalar objective function. If function f in (12) was replaced by the vector function $\mathbf{F} = [f_1, \ldots, f_i, \ldots f_m]^T$ one could use scalarisation approach (PS) to obtain:

$$
\begin{aligned}
&\min_{\alpha_f \geq 0} \alpha_f \\
&s.t. \\
&\omega_i(f_i(\mathbf{x}_f, t_f) - z_i) - \alpha_f \leq 0 \qquad \forall i = 1, \ldots, m \\
&\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \\
&\mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0 \\
&\boldsymbol{\psi}(\mathbf{x}_0, \mathbf{x}_f, t_0, t_f) \leq 0 \\
&t \in [t_0, t_f]
\end{aligned}
\tag{PSOCP}
$$

If s is a slack variable with final condition α_f and zero time variation $\dot{\alpha} = 0$, then problem (PSOCP) presents itself in a form similar to Mayer's problem. The major difference is the mixed boundary constraint on x_f, t_f and α_f for every $i = 1, \ldots, m$.

3 Solution Approach

We are now ready to approach problem (MOCP). We are interested in finding the set of control laws that are globally Pareto optimal. Recalling the definition in Sect. 2.2, this means that we want the family of control laws such that the vectors $\mathbf{F}$ corresponding to each control law belong to the global Pareto front of problem (MOCP). In the case of a multi-modal function it is possible that there exist multiple sets that satisfy the Pareto efficiency criterion only locally, i.e. in a subset of U. Here the interest is to devise a method that allows convergence to solutions that satisfy the Pareto efficiency criterion globally in U.

The solution approach proposed in this chapter, first translates the optimal control problem into a non-linear programming problem. The NLP is then scalarised and a so called memetic approach, that combines a population-based search and a gradient method, is used to find an approximation to the Pareto set. The following sections first briefly describe the transcription method from optimal control problem to NLP problem, followed by the solution approach of the multi-objective NLP problem.

3.1 Direct Transcription of Multi-Objective Optimal Control Problems

The approach proposed in this chapter, for the transcription of problem (MOCP), falls under the class of direct approaches and is based on a numerical method called Finite Elements in Time (FET) [12]. Direct FET transcription for scalar optimal control problems was first introduced by Vasile [13] in 2000 and uses finite elements in time on spectral bases to transcribe the differential equations into a set of algebraic equations. Finite Elements in Time (FET) for the indirect solution of optimal control problems were initially proposed in 1991 by Hodges and collaborators [14], and during the late 1990s evolved to the bi-discontinuous version that will be presented in this section. As a numerical integration scheme for ordinary differential equations, FET are equivalent to some classes of implicit Runge-Kutta integration schemes [15], can be extended to arbitrary high-orders, are very robust and allow full h-p adaptivity, where h-adaptivity means adapting the size of each element and p-adaptivity means adapting the order of the polynomials for each element. In the past decade, direct transcription with FET on spectral bases has been successfully used to solve a range of difficult problems: from the design of low-thrust multi-gravity assist trajectories to Mercury [16] and to the Sun [17], to the design of weak stability boundary transfers to the Moon, low-thrust transfers in the restricted three body problem and optimal landing trajectories to the Moon [13].

The first transcription step is to recast the differential constraints in weak form as follows:

$$\int_{t_0}^{t_f} \dot{\mathbf{w}}^T \mathbf{x} + \mathbf{w}^T \mathbf{h}(\mathbf{x}, \mathbf{u}, t) dt - \mathbf{w}_f^T \mathbf{x}_f^b + \mathbf{w}_0^T \mathbf{x}_0^b = 0 \tag{16}$$

where $\mathbf{w}$ are generalised weight functions and $\mathbf{x}^b$ are the boundary values of the states, that may be either imposed or free. Now one can decompose the time domain D into N finite elements such that

$$D = \bigcup_{j=1}^{N} D_j(t_{j-1}, t_j) \tag{17}$$

and parametrise, over each D_j, the states, controls and weight functions as

$$\mathbf{x}_j(t) = \sum_{s=0}^{l} \varphi_{s,j}(t)\, \mathbf{x}_{s,j} \tag{18a}$$

$$\mathbf{u}_j(t) = \sum_{s=0}^{l} \gamma_{s,j}(t)\, \mathbf{u}_{s,j} \tag{18b}$$

$$\mathbf{w}_j(t) = \sum_{s=0}^{l+1} \theta_{s,j}(t)\, \mathbf{w}_{s,j} \tag{18c}$$

where the functions $\varphi_{s,j}$, $\gamma_{s,j}$ and $\theta_{s,j}$ are chosen among the space of polynomials of degree l, and $(l + 1)$ respectively. Note that, in general, the controls $\mathbf{u}$ can be collocated on a number of nodes different from the states $\mathbf{x}$. It is practical to define each D_j over the normalised interval $[-1,\ 1]$ through the transformation,

$$\tau = 2\frac{t - \frac{t_j - t_{j-1}}{2}}{t_j - t_{j-1}} \tag{19}$$

This way the domain of the basis function is constant and irrespective of the size of the element and also overlaps with the interval of the Gauss nodes that will be employed for the integration of the dynamics. The objective function is simply:

$$J_i = \phi_i(\mathbf{x}_0^b, \mathbf{x}_f^b, t_0, t_f) \tag{20}$$

while, after substituting the definitions of the polynomials into the variational constraints and integrating with Gauss quadrature formulas, one gets the following system for each finite element:

$$\sum_{k=1}^{l+1} \sigma_k \left[\dot{\mathbf{w}}_j(\tau_k)^T \mathbf{x}_j(\tau_k) + \mathbf{w}_j(\tau_k)^T \mathbf{h}_j(\tau_k) \frac{\Delta t_j}{2} \right] - \mathbf{w}_j(1)^T \mathbf{x}_j^b + \mathbf{w}_j(-1)^T \mathbf{x}_{j-1}^b = 0 \tag{21}$$

where τ_k and σ_k are the Gauss nodes and weights, and $\mathbf{h}_j(\tau_k)$ is the shorthand notation for $\mathbf{h}\left(\mathbf{x}_j(\tau_k), \mathbf{u}_j(\tau_k), t(\tau_k)\right)$, τ_k and σ_k are Gauss nodes and weights, $\Delta t_j = (t_j - t_{j-1})$. Since Eq. (21) must be valid for every arbitrary $\mathbf{w}_{s,j}$, Eq. (21) gives rise to a system of $(l_x + 1)$ vector equations for each element:

$$\begin{aligned}
&\textstyle\sum_{k=0}^{l_u} \sigma_k \left[\dot{\theta}_{1,j}(\tau_k)\, \mathbf{x}_j(\tau_k) + \theta_{1,j}(\tau_k) \mathbf{h}_j(\tau_k) \frac{\Delta t_j}{2} \right] + \mathbf{x}_{j-1}^b = 0 \\
&\vdots \\
&\textstyle\sum_{k=0}^{l_u} \sigma_k \left[\dot{\theta}_{s,j}(\tau_k) \mathbf{x}_j(\tau_k) + \theta_{s,j}(\tau_k) \mathbf{h}_j(\tau_k) \frac{\Delta t_j}{2} \right] = 0 \\
&\vdots \\
&\textstyle\sum_{k=0}^{l_u} \sigma_k \left[\dot{\theta}_{l_x+1,j}(\tau_k) \mathbf{x}_j(\tau_k) + \theta_{l_x+1,j}(\tau_k) \mathbf{h}_j(\tau_k) \frac{\Delta t_j}{2} \right] - \mathbf{x}_j^b = 0
\end{aligned} \tag{22}$$

Path constraints are evaluated at Gauss nodes for each element:

$$\mathbf{g}\left(\mathbf{x}_j(\tau_k), \mathbf{u}_j(\tau_k), t(\tau_k)\right) \leq 0 \tag{23}$$

All the elements are then assembled together, by imposing the continuity relation:

$$\mathbf{x}_j^b = \mathbf{x}_{j-1}^b \tag{24}$$

The assembly process, therefore, removes all the boundary values except for the initial and final ones, at time t_0 and t_f. The result is that, optimal control problem in Eq. (MOCP) is transcribed into the following non-linear programming (NLP) problem:

$$\begin{aligned}&\min_{\mathbf{p}\in\Pi,\mathbf{y}\in Y} \mathbf{J}(\mathbf{y},\mathbf{p})\\ &s.t.\\ &\mathbf{c}(\mathbf{y},\mathbf{p}) \leq 0\end{aligned} \tag{25}$$

where **c** contains all constraints (22) and (23) and all boundary constraints and the vector **y** contains all the nodal values for the states, $\mathbf{p} = [\mathbf{u}_{1,0}, \ldots, \mathbf{u}_{s,j}, \ldots, \mathbf{u}_{l,N}, \mathbf{x}_0, \mathbf{x}_f, t_0, t_f]^T$ collects all control variables and $\Pi \subseteq \mathbb{R}^{2n+2} \times \mathbb{R}^{n_s}$, $Y \subseteq \mathbb{R}^{n_y}$, with $n_s = n_u \cdot l \cdot N$ and $n_y = n \cdot l \cdot N$. It is worth noting that the DFET transcription is very flexible and allows one to choose any basis for states, controls and test functions, and the basis could also be different for every variable. Similarly it is possible to employ several choices for the type of quadrature nodes [18].

3.2 Solution of the Transcribed MOCP

It is proposed to solve problem (25) with a memetic multi-objective optimisation algorithm that combines a stochastic agent-based search with a local (gradient in this case) refinement of the solutions [19–23]. The version of the algorithm that will be presented in this section is here called MACSoc (Multi-Agent Collaborative Search for optimal control).

Multi-Agent Collaborative Search is a meta-heuristics to combine local and global search strategies. A set of agents is endowed with a list of possible actions that can involve other agents or simply collect information on a neighbourhood of each agent. MACSoc incorporates the idea of search directions and descent cones within the decision logic of the agents: each agent can select new candidate solutions according to either dominance or Chebyshev scalarisation. Furthermore, each agent can start a local search directly solving problem (PSOCP). The ability of the agents to incorporate local gradient-based actions are here exploited to solve problem (25). The general MACSoc scheme is summarised in Algorithm 1. The individualistic and social actions are described in the following section and are related to the solution of two different problems. The population P_0 (Line 1 in Algorithm 1) is initialised randomly with Latin Hypercube sampling [24], while the weights ω (Line 2 in Algorithm 1) are generated as in Sect. 3.2.2. After performing individualistic and social actions (lines 4 and 7 in Algorithm 1) both the population and the archive are updated. The filtering process (Lines 6 and 9 in Algorithm 1) that updates the global archive A_g, where all Pareto optimal solutions are stored, is redistributing solutions so that a pseudo-electric potential function (function of the reciprocal distance of the elements in the archive) is minimised (see [20] for further details). Finally, at each iteration, the descent

Algorithm 1 MACSoc framework

```
 1: Initialise population $P_0$ and global archive $A_g$
 2: Initialise search directions $\mathbf{d}$ and weights $\boldsymbol{\omega}$
 3: while $nfeval < max\_fun\_eval$ do
 4:     Run individualistic heuristics on $P_k$
 5:     $P_k \to P_k^+$
 6:     Update archive $A_g$ with potential field filter
 7:     Run social heuristics combining $P_k^+$ and $A_g$
 8:     $P_k^+ \to P_{k+1}$
 9:     Update archive $A_g$ with potential field filter (see [20])
10:     Update subproblem allocation
11: end while
```

direction (or scalar subproblem) allocated to each agent is updated (line 10 in Algorithm 1).

3.2.1 Problem Formulation in the MACS Framework

In order to solve (25), MACSoc makes use of both Pascoletti-Serafini and Chebyshev scalarisations with either the individualistic or social actions. When agents search for a local Pareto efficient solution, each agent j uses its own Pascoletti-Serafini scalarisation of the problem in the form:

$$\begin{aligned} &\min_{\alpha_f \geq 0} \alpha_f \\ &s.t. \\ &\omega_{ij}\vartheta_{ij}(\overline{\mathbf{x}}, \overline{\mathbf{p}}) \leq \alpha_f \qquad i = 1, .., m \\ &\mathbf{c}(\overline{\mathbf{x}}, \overline{\mathbf{p}}) \leq 0 \end{aligned} \tag{26}$$

where $\boldsymbol{\omega}_j$ is the vector of weights associated to agent j, ϑ_{ij} is the i-th component of the rescaled objective vector of the j-th agent and α_f is a slack variable. This reformulation of the problem is constraining the j-th agent move, in criteria space, within the descent cone defined by the point $\alpha_f \mathbf{d}_j + \boldsymbol{\zeta}_j$ along the direction $\mathbf{d}_j = [1/\omega_{1j}, \ldots, 1/\omega_{ij}, \ldots, 1/\omega_{mj}]^T$. The rescaled objective vector is

$$\vartheta_{ij}(\overline{\mathbf{x}}, \overline{\mathbf{p}}) = \frac{J_{ij}(\overline{\mathbf{x}}, \overline{\mathbf{p}}) - \tilde{z}_i}{z_{ij}^* - \tilde{z}_i} \qquad i = 1, .., m \tag{27}$$

where $\mathbf{z}_j^*$ is equal to $\mathbf{J}_j(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ and $(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ is the initial guess for the solution of (26). This way the components of $\boldsymbol{\vartheta}_j(\overline{\mathbf{x}}, \overline{\mathbf{p}})$ have value 1 at the beginning of the local search and if the agent converges to the utopia point $\tilde{\mathbf{z}}$, the components of $\boldsymbol{\vartheta}_j(\overline{\mathbf{x}}, \overline{\mathbf{p}})$

Algorithm 2 Individualistic action

```
Set $\tilde{\mathbf{z}} = 2\mathbf{z} - \mathbf{z}_A^*$
if current agent is solving problem $i$ only then
    $\boldsymbol{\omega}_j = (0, 0, i, .., 0, 0)$
else
    $\boldsymbol{\omega}_j = \frac{[1,1,1,\cdots,1]^T}{\|[1,1,1,\cdots,1]^T\|}$
end if
Pick a point $(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ in $B_j$
Run local search from $(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ to solve Problem (26) and find solution $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j$
if $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j$ feasible then
    Return $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j, \mathbf{J}_j(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)$ and increase $\rho_j < 1$
else
    if number of times $\rho_j$ is reduced $> max\_contr\_ratio$ then
        $\rho_j = 1$
    else
        Reduce $\rho_j$ and return $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j = (\overline{\mathbf{x}}, \overline{\mathbf{p}})_j, \mathbf{J}_j = M + \|\mathbf{c}\|$
    end if
end if
```

become all equal to 0. The choice of $\boldsymbol{\omega}_j$ and $\tilde{\mathbf{z}}$ will be discussed in the following subsection. From the normalisation one can derive the components of the vector $\boldsymbol{\zeta}_j$:

$$\zeta_{ij} = \frac{z_i}{z_{ij}^* - \tilde{z}_i} \qquad i = 1, .., m \tag{28}$$

The presence of the rescaling of the objectives, together with the choice of $\boldsymbol{\omega}_j$ and $\tilde{\mathbf{z}}_j$, are the elements that distinguish the proposed approach from others in the literature [2]. Note that solving problem (26) already provides a non-dominated solution that can be potentially inserted in A_g and used to update P_k. The pseudocode can be found in Algorithm 2. The local search starts from a point $(\overline{\mathbf{x}}, \mathbf{p}^c)$ taken at random in a neighbourhood B_j of the current agent, if the location of the curretn agent $(\overline{\mathbf{x}}, \mathbf{p})_j$ did not change from the previous iteration, otherwise $(\overline{\mathbf{x}}, \mathbf{p}^c) = (\overline{\mathbf{x}}, \mathbf{p})_j$. The neighbourhood B_j is a hypercube with edge of size $2\rho_j$. If the local search returns an infeasible solution a penalty value M (which is higher than the highest objective function in the population) plus the norm of the constraint violation is assigned to all cost functions.

When agents explore the search space, either implementing individualistic actions or as a population, they use a bi-level formulation of problem (25) in which the upper level is handling only the objective functions and the lower level the constraint functions:

$$\begin{aligned} &\min_{\mathbf{p}*} \mathbf{J}_j(\overline{\mathbf{x}}^*, \mathbf{p}^*) \\ &s.t. \\ &(\overline{\mathbf{x}}^*, \mathbf{p}^*)_j = \operatorname{argmin}\{\delta(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c) | \mathbf{c}(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c) \le 0\} \end{aligned} \tag{29}$$

Algorithm 3 Social action

```
Select weight $\boldsymbol{\omega}$
Select agents associated with $\boldsymbol{\omega}$ and elements of the archive $A_g$
Apply DE operator to selected agents and elements of the archive and generate candidate solution $\mathbf{u} = (\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$
Run inner level on $\mathbf{u}$
```

Algorithm 4 Inner level

```
Run local search from $(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ to find solution $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j$
if $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j$ feasible then
    Return $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j, \mathbf{J}_j(\overline{\mathbf{p}}^*, \overline{\mathbf{x}}^*)$
else
    Return $(\overline{\mathbf{x}}^*, \overline{\mathbf{p}}^*)_j = (\overline{\mathbf{x}}, \overline{\mathbf{p}})_j, \mathbf{J}_j = M + \|\mathbf{c}\|$
end if
```

where δ is the distance between any new vector $\mathbf{p}$ generated at the lower level and the initial $\mathbf{p}$ that the upper level is passing to the lower level, $(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c)$ is a candidate solution generated either with a Differential Evolution (DE) operator or with a pattern search technique (refer to [20] for more details). The DE operator is applied to a mix of agents associated to a particular weight $\boldsymbol{\omega}$ and elements of the archive A_g. If the inner level returns a feasible solution, that solution is selected for possible inclusion in the population P_{k+1} using Chebyshev criterion (line 8 of Algorithm 1). In other words, Problem (29) is scalarised in the following form:

$$\begin{array}{l} \min_{\mathbf{p}^*} \max_{i \in \{1,\ldots,m\}} \omega_i (J_i(\overline{\mathbf{x}}^*, \mathbf{p}^*) - z_i) \\ s.t. \\ (\overline{\mathbf{x}}^*, \mathbf{p}^*)_j = \text{argmin}\{\delta(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c) | \mathbf{c}(\overline{\mathbf{x}}, \overline{\mathbf{p}}^c) \leq 0\} \end{array} \tag{30}$$

The pseudocode for both levels can be found in Algorithms 3 and 4. Note that if the inner level returns an infeasible solution, as before, the associated objective function is a penalty value M (which is higher than the highest objective function in the population) plus the norm of the violation of the constraints.

At this point it is worth explaining how the equivalence of (CS) and (PS), demonstrated in Theorem 2, is exploited in this framework. Suppose that Problem (26) had to be solved with an evolutionary approach. In that case the constraints on the objective functions would translate into:

$$\min_{\mathbf{p}^*} \max_{i \in \{1,\ldots,m\}} \omega_i (\vartheta_i(\overline{\mathbf{x}}^*, \mathbf{p}^*) - \alpha_f) \tag{31}$$

which, for $\alpha_f = 0$, is equivalent to the outer Problem in (30). Since from Theorem 2 one can say that problem (31), with $\alpha_f = 0$ and (26) are equivalent and lead to the same optimal solution, by combining (31) in the search phase with (26) in the refinement phase, the algorithm has a smooth transition from global search to local convergence.

3.2.2 Selection of ω and $\tilde{\mathbf{z}}$

In [2], the MOCP was tackled by first solving each of the two individual objectives, and then choosing a set of evenly spaced weights, obtaining a set of directions **d**. This approach has a two main limitations: first, since only a local strategy was employed, there is the possibility that the extreme values of the Pareto front generated are on a local Pareto front. Second, that approach is not easy to generalise for more than two objectives. The proposed approach instead consists in assigning vector $\boldsymbol{\omega}_j = [0, 0, i, .., 0, 0]^T$ to agents solving subproblem i and vector $\boldsymbol{\omega}_j = \frac{[1,1,1,\cdots,1]^T}{\|[1,1,1,\cdots,1]^T\|}$ to all the other agents. The modified utopia point $\tilde{\mathbf{z}}$ is given by

$$\tilde{\mathbf{z}} = 2\mathbf{z} - \mathbf{z}^*{}_A \tag{32}$$

where $\mathbf{z}$ and $\mathbf{z}^*{}_A$ are respectively the utopia and nadir points of the current approximation to the Pareto front that is contained in the archive A_g. When an agent j solving subproblem i has locally converged and is not displaced by any action, its subproblem is updated with $\boldsymbol{\omega}_j = \frac{[1,1,1,\cdots,1]^T}{\|[1,1,1,\cdots,1]^T\|}$, conversely an agent associated to $\boldsymbol{\omega}_j = \frac{[1,1,1,\cdots,1]^T}{\|[1,1,1,\cdots,1]^T\|}$ that has locally converged and is not displaced by any action will have its subproblem replaced with $\boldsymbol{\omega}_j = [0, 0, i, .., 0, 0]^T$ (line 10 in Algorithm 1).

3.3 First Order Necessary Optimality Conditions

The scalarised version of the transcribed problem allows one to recover some theoretical results developed for single objective optimal control problems. In fact if one applies the Pascoletti-Serafini scalarisation to the original optimal control problem the following system of differential algebraic equations is obtained:

$$\begin{aligned}
&\min_{\alpha_f \geq 0} \ \alpha_f \\
&s.t. \\
&\omega_i(J_i(\mathbf{x}) - z_i) \leq \alpha_f \qquad i = 1, .., m \\
&\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \\
&\dot{\alpha} = 0 \\
&\mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0 \\
&\boldsymbol{\psi}(\mathbf{x}_0, \mathbf{x}_f, t_0, t_f) \leq 0 \\
&t \in [t_0, t_f]
\end{aligned} \tag{33}$$

We can now introduce the following necessary optimality conditions for the scalarised problem (33) with constraints on terminal states and final time and given initial conditions $\mathbf{x}_0$ and time t_0.

Theorem 3 *Consider the function* $H = \lambda^T \mathbf{h}(\mathbf{x}, \mathbf{u}, t) + \mu^T \mathbf{g}(\mathbf{x}, \mathbf{u}, t)$*. If* $\mathbf{u}^*$ *is a locally optimal solution for problem (33), with associated state vector* $\mathbf{x}^*$*, and* H *is Frechet differentiable at* $\mathbf{u}^*$ *and a regular point of the algebraic constraints, then there exist vectors* $\eta \in \mathbb{R}^m$*,* $\lambda \in \mathbb{R}^n$*,* $\lambda_\alpha \in \mathbb{R}$*,* $\nu \in \mathbb{R}^{n_\psi}$ *and* $\mu \in \mathbb{R}^q$ *such that:*

$$\begin{aligned} &\mathbf{u}^* = \text{argmin}_{\mathbf{u}\in U}\, \lambda^T \mathbf{h}(\mathbf{x}^*, \mathbf{u}, t) + \mu^T \mathbf{g}(\mathbf{x}^*, \mathbf{u}, t) \\ &\lambda^T \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}^*, \mathbf{u}^*, t) + \mu^T \nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}^*, \mathbf{u}^*, t) + \dot{\lambda}^T = 0 \\ &\dot{\lambda}_\alpha = 0 \\ &\mu \geq 0 \end{aligned} \tag{34}$$

with transversality conditions:

$$\begin{aligned} &1 - \textstyle\sum_i^m \eta_i + \lambda_{\alpha_f}(t_f) = 0 \\ &\eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi} + \lambda(t_f)^T = 0 \\ &\eta > 0;\ \nu \geq 0 \end{aligned} \tag{35}$$

and

$$H_{t_f} - \eta^T \boldsymbol{\omega} \partial_{t_f} \mathbf{J} - \nu^T \partial_{t_f} \boldsymbol{\psi} = 0 \tag{36}$$

where $\boldsymbol{\omega}$ *is a diagonal matrix with the components* ω_i*,* $i = 1, \ldots, m$ *along the diagonal and* U *is the space of admissible controls that satisfy respectively the algebraic and differential constraints* $\mathbf{g}(\mathbf{x}, \mathbf{u}, t) \leq 0$ *and* $\dot{\mathbf{x}} - \mathbf{h}(\mathbf{x}, \mathbf{u}, t) = 0$.

Proof A possible proof comes from the direct application of Pontryagin's maximum principle to problem (33). Alternatively, at the solution, one can take the first variation of the functional:

$$L = \alpha_f + \eta^T (\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1}) + \nu^T \boldsymbol{\psi} + \int_{t_0}^{t_f} [\lambda^T (\mathbf{h} - \dot{\mathbf{x}}) + \lambda_\alpha \dot{\alpha} + \mu^T \mathbf{g}] dt \tag{37}$$

with $\mathbf{1}$ a vector of ones, which gives:

$$\begin{aligned} &\delta L = \delta\alpha_f + \delta\eta^T (\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1}) + \eta^T \boldsymbol{\omega} \delta\mathbf{J} + \eta^T \delta(\alpha_f \mathbf{1}) + \delta\nu^T \boldsymbol{\psi} + \nu^T \delta\boldsymbol{\psi} + \\ &\textstyle\int_{t_0}^{t_f} [\delta\lambda^T (\mathbf{h} - \dot{\mathbf{x}}) + \lambda^T (\delta\mathbf{h} - \delta\dot{\mathbf{x}}) + \delta\lambda_\alpha \dot{\alpha} + \lambda_\alpha \delta\dot{\alpha} + \delta\mu^T \mathbf{g} + \mu^T \delta\mathbf{g}] dt = 0. \end{aligned} \tag{38}$$

We can now collect terms with equal variation δ and d:

$$\begin{aligned} &\delta L = \delta\alpha_f (1 - \eta^T \mathbf{1}) + \delta\eta^T (\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1}) + \delta\nu^T \boldsymbol{\psi} + (\eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi}) d\mathbf{x}_{t_f} + \\ &\textstyle\int_{t_0}^{t_f} [\delta\lambda^T (\mathbf{h} - \dot{\mathbf{x}}) - \lambda^T \delta\dot{\mathbf{x}} + (\lambda^T \nabla_{\mathbf{x}} \mathbf{h} + \mu^T \nabla_{\mathbf{x}} \mathbf{g}) \delta\mathbf{x} + (\lambda^T \nabla_{\mathbf{u}} \mathbf{h} + \mu^T \nabla_{\mathbf{u}} \mathbf{g}) \delta\mathbf{u} + \\ &\delta\lambda_\alpha \dot{\alpha} + \lambda_\alpha \delta\dot{\alpha} + \delta\mu^T \mathbf{g}] dt + [\lambda^T \mathbf{h} + \mu^T \mathbf{g}]_{t_f} \delta t_f + \nu \partial_{t_f} \boldsymbol{\psi} \delta t_f + \eta^T \boldsymbol{\omega} \partial_{t_f} J \delta t_f = 0 \end{aligned} \tag{39}$$

and after integrating by parts the terms $\lambda^T \delta \dot{\mathbf{x}}$ and $\lambda_\alpha \delta \dot{\alpha}$ we get:

$$\begin{aligned}
&\delta L = \delta\alpha_f(1 - \eta^T \mathbf{1}) + \delta\eta^T(\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1}) + \\
&\delta\nu^T \boldsymbol{\psi} + (\eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi}) d\mathbf{x}_{t_f} - \lambda(t_f)^T \delta\mathbf{x}_{t_f} + \lambda_\alpha(t_f)\delta\alpha_f \\
&\int_{t_0}^{t_f} [\delta\lambda^T(\mathbf{h} - \dot{\mathbf{x}}) + \dot{\lambda}^T \delta\mathbf{x} + (\lambda^T \nabla_{\mathbf{x}} \mathbf{h} + \mu^T \nabla_{\mathbf{x}} \mathbf{g})\delta\mathbf{x} + (\lambda^T \nabla_{\mathbf{u}} \mathbf{h} + \mu^T \nabla_{\mathbf{u}} \mathbf{g})\delta\mathbf{u} + \\
&\delta\lambda_\alpha \dot{\alpha} - \dot{\lambda}_\alpha \delta\alpha + \delta\mu^T \mathbf{g}]dt + [\lambda^T \mathbf{h} + \mu^T \mathbf{g}]_{t_f} \delta t_f + \nu \partial_{t_f} \boldsymbol{\psi} \delta t_f + \eta^T \boldsymbol{\omega} \partial_{t_f} J \delta t_f = 0
\end{aligned} \tag{40}$$

Now in order for the variation to be zero for every value of the δ and d quantities the following equations must be satisfied:

$$\begin{aligned}
&\dot{\mathbf{x}} - \mathbf{h} = 0 \\
&\lambda^T \nabla_{\mathbf{u}} \mathbf{h}(\mathbf{x}^*, \mathbf{u}^*, t) + \mu^T \nabla_{\mathbf{u}} \mathbf{g}(\mathbf{x}^*, \mathbf{u}^*, t) = 0 \\
&\lambda^T \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}^*, \mathbf{u}^*, t) + \mu^T \nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}^*, \mathbf{u}^*, t) + \dot{\lambda}^T = 0 \\
&\mathbf{g}(\mathbf{x}^*, \mathbf{u}^*, t) \leq 0 \\
&\dot{\lambda}_\alpha = 0 \\
&\mu \geq 0 \\
&1 - \sum_i^m \eta_i + \lambda_\alpha(t_f) = 0 \\
&\eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi} - \lambda(t_f)^T = 0 \\
&\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1} \leq 0 \\
&\boldsymbol{\psi} \leq 0 \\
&\eta > 0; \nu \geq 0
\end{aligned} \tag{41}$$

and

$$[\lambda^T \mathbf{h} + \mu^T \mathbf{g}]_{t_f} + \eta^T \boldsymbol{\omega} \partial_{t_f} \mathbf{J} + \nu^T \partial_{t_f} \boldsymbol{\psi} = 0 \tag{42}$$

If now one introduces the function $H = \lambda^T \mathbf{h}(\mathbf{x}, \mathbf{u}, t) + \mu^T \mathbf{g}(\mathbf{x}, \mathbf{u}, t)$, Eq. (41) reduce to:

$$\begin{aligned}
&\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1} \leq 0 \\
&\dot{\mathbf{x}} = \frac{\partial H}{\partial \boldsymbol{\lambda}} \\
&\dot{\lambda}^T = -\frac{\partial H}{\partial \mathbf{x}} \\
&\frac{\partial H}{\partial \mathbf{u}} = 0 \\
&\frac{\partial H}{\partial \boldsymbol{\mu}} \leq 0 \\
&\dot{\lambda}_\alpha = -\frac{\partial H}{\partial \alpha} \\
&\boldsymbol{\psi} \leq 0 \\
&\mu \geq 0 \\
&\alpha_f \geq 0
\end{aligned} \tag{43}$$

with transversality conditions:

$$\begin{aligned}
&H_{t_f} + \eta^T \boldsymbol{\omega} \partial_{t_f} \mathbf{J} + \nu^T \partial_{t_f} \boldsymbol{\psi} = 0\\
&1 - \sum_i^m \eta_i + \lambda_\alpha(t_f) = 0\\
&\eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi} - \lambda(t_f)^T = 0\\
&\eta > 0; \; \nu \geq 0
\end{aligned} \tag{44}$$

Remark 1 At the solution $(\mathbf{x}^*, \mathbf{u}^*)$ all constraints are assumed to be active, which means that $\boldsymbol{\psi} = 0$ and $\mathbf{g}(\mathbf{x}^*, \mathbf{u}^*) = 0$. Furthermore, it is assumed that also the constraints on the objective functions are active, which means that at the solution $\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1} = 0$.

Remark 2 If initial conditions and time were not given but had to satisfy some constraint functions, one would need to add other three transversality conditions similar to (44) but at the initial time t_0.

Remark 3 The condition $\eta > 0$ implies that we cannot have a trivial solution with both $\eta = 0$ and $\nu = 0$. The same condition can be relaxed as in the KKT conditions so that $\eta \geq 0$ and exist at least a component $\eta_i > 0$.

3.3.1 Example

Consider the very simple one-dimensional controlled dynamical system with constant control acceleration and two objectives on the terminal states:

$$\begin{aligned}
&\min \alpha_f\\
&\omega_1(-x_f + 1) < \alpha_f\\
&\omega_2 v_f < \alpha_f\\
&\dot{x} = v; \; \dot{v} = -u; \; \dot{\alpha} = 0;\\
&x(t_0) = 0; \; v(t_0) = 1;
\end{aligned}$$

$$\begin{aligned}
&0 \leq u \leq 1\\
&\alpha_f \geq 0\\
&t_f = 1
\end{aligned} \tag{45}$$

with $x_f = x(t_f)$, $v_f = v(t_f)$, $\alpha_f = \alpha(t_f)$. The necessary conditions for optimality are:

$$\begin{aligned}
&H = \lambda_x v - \lambda_v u + \mu_1(u - 1) - \mu_2 u\\
&\frac{\partial H}{\partial u} = -\lambda_v + \mu_1 - \mu_2 = 0\\
&\dot{\lambda}_\alpha = 0;\\
&\dot{\lambda}_x = 0; \; \dot{\lambda}_v = -\lambda_x;
\end{aligned} \tag{46}$$

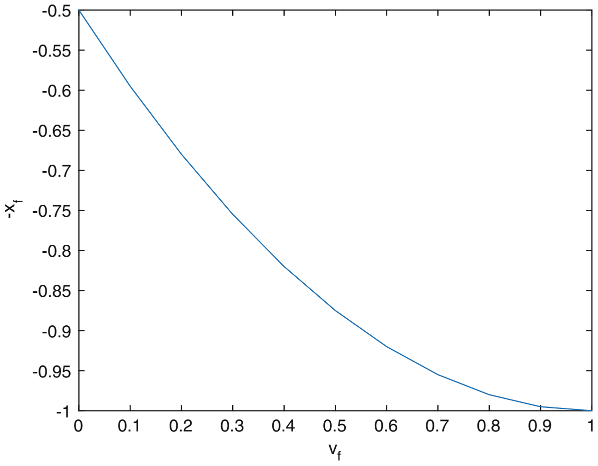


Fig. 3 Pareto front for problem (45)

with terminal conditions:

$$\begin{aligned}&\lambda_\alpha(t_f) = 1 - \eta_1 - \eta_2;\\ &\lambda_x(t_f) = -\eta_1\omega_1; \lambda_v(t_f) = \eta_2\omega_2;\end{aligned} \tag{47}$$

The weights ω_1 and ω_2 take only positive values. Note that the vector $\mathbf{z}$ is $\mathbf{z} = [-1, 0]^T$, where the two values -1 and 0 are the extreme values that one would obtain if the two objective functions were optimised individually. However, any other pair of values sufficiently low would equally work. The solution of the controlled dynamics is given by:

$$\begin{aligned}&x = -\tfrac{t^2}{2} + t \quad t \in [t_0, t_1]\\ &v = -t + 1 \quad t \in [t_0, t_1]\\ &x = v_f t + x_1 \quad t \in [t_1, t_f]\\ &v = v_1 = v_f \quad t \in [t_1, t_f]\\ &x_1 = x(t_1); v_1 = v(t_1)\end{aligned} \tag{48}$$

In this case it is easy to demonstrate that the Pareto front is given by the following second order algebraic equation (see Fig. 3):

$$x_f = -\frac{1 + 2v_f - v_f^2}{2} \tag{49}$$

We want to show that all the points along the front satisfy the optimality conditions and represent a minimum for α_f. Consider first one of the extreme values:

$$\begin{array}{l} \min_{\alpha_f \geq 0} \alpha_f \\ -x_f + 1 \leq \alpha_f \\ x = -\frac{t^2}{2} + t \quad t \in [t_0, t_1] \\ v = -t + 1 \quad t \in [t_0, t_1] \\ x = v_f t + x_1 \quad t \in [t_1, t_f] \\ v = v_1 = v_f \quad t \in [t_1, t_f] \\ x_1 = x(t_1);\, v_1 = v(t_1) \end{array} \tag{50}$$

By imposing the continuity conditions at t_1 we get a simple algebraic problem:

$$\begin{array}{l} \min_{\alpha_f \geq 0} \alpha_f \\ x_1 = -\frac{t_1^2}{2} + t_1 \\ v_f = -t_1 + 1 \\ -\alpha_f + 1 = v_f t_f + x_1 \end{array} \tag{51}$$

where we introduced the assumption that the maximum value that x_f can take is $-\alpha_f + 1$. For $t_f = 1$ the system reduces to:

$$\begin{array}{l} \min_{\alpha_f \geq 0} \alpha_f \\ -\alpha_f + 1 = 1 - \frac{t_1^2}{2} \\ 0 \leq t_1 \leq 1 \end{array} \tag{52}$$

Problem (52) has the simple solution:

$$\alpha_f = 0;\, t_1 = 0 \tag{53}$$

If one follows the same process with the other extreme solution the result is:

$$\alpha_f = 0;\, t_1 = 1 \tag{54}$$

We now need to verify that we can find a suitable set of Lagrange multipliers that satisfy the necessary conditions. The solution for λ_x is the constant $-\eta_1\omega_1$ and for λ_v is:

$$\lambda_v = -\lambda_x(t - t_f) + \lambda_v(t_f) \tag{55}$$

These equations confirm that there is a single switching point for the control u^*. For the extreme case in which $\omega_1 = 1$ and $\omega_2 = 0$ the final values are:

$$\begin{array}{l} \lambda_v(t_f) = 0 \\ \lambda_x(t_f) = -\eta_1 \end{array} \tag{56}$$

which leads to the conclusion that:

$$\lambda_v < 0 \quad \forall t \in [t_0, t_f] \tag{57}$$

that moves the switching point t_1 to t_0. The conditions on the multipliers associated to the slack variable α_f are always satisfied.

3.4 Convergence of the Transcribed Problem

We can now prove that the transcribed problem converges asymptotically to the necessary conditions for local optimality (34) and (35).

Theorem 4 *If* **c** *is Frechet differentiable at* $\mathbf{u}^*$ *and both* **x**, **h** *and are* **g** *integrable functions, then the necessary conditions for local optimality of problem (26) converge asymptotically to (34) and (35) for* $k \to \infty$ *and* $s \to \infty$.

Proof We start from the augmented Lagrangian of the related NLP problem:

$$L = \alpha_f + \eta^T(\boldsymbol{\omega}(\mathbf{J} - \mathbf{z}) - \alpha_f \mathbf{1}) + \hat{\lambda}^T \mathbf{c}_d + \hat{\mu}^T \mathbf{g} + \lambda_{\alpha_f} \alpha_f \tag{58}$$

where $\mathbf{c}_d$ is the part of the constraint vector $\mathbf{c} = [\mathbf{c}_d, \mathbf{g}]$ that does not contain path constraints. If one differentiates the Lagrangian, the result is the necessary conditions for local optimality:

$$\frac{\partial L}{\partial \mathbf{u}} = \hat{\lambda}^T \nabla_{\mathbf{u}} \mathbf{c}_d + \hat{\mu}^T \nabla_{\mathbf{u}} \mathbf{g} = 0 \tag{59}$$

$$\frac{\partial L}{\partial \mathbf{x}} = \hat{\lambda}^T \nabla_{\mathbf{x}} \mathbf{c}_d + \hat{\mu}^T \nabla_{\mathbf{x}} \mathbf{g} = 0 \tag{60}$$

$$\frac{\partial L}{\partial \mathbf{x}_f} = \eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \hat{\lambda}_f^T \nabla_{\mathbf{x}_f} \mathbf{c}_d = 0 \tag{61}$$

$$\frac{\partial L}{\partial \alpha_f} = 1 - \sum_{i=1}^{m} \eta_i + \lambda_{\alpha_f} = 0 \tag{62}$$

$$\frac{\partial L}{\partial t_f} = \eta^T \boldsymbol{\omega} \partial_{t_f} \mathbf{J} + \hat{\lambda}_f^T \partial_{t_f} \mathbf{c}_d + \hat{\mu}_f^T \partial_{t_f} \mathbf{g} = 0 \tag{63}$$

where the first equation corresponds to the transversality condition on $\lambda_\alpha(t_f)$, the fourth equation corresponds to the transversality conditions on $\lambda(t_f)$ that were derived in (35) and the fifth equation is the transversality condition (36). We used the symbol $\hat{\lambda}_f^T$ to indicate the Lagrange multipliers that correspond to the boundary constraints and to the dynamic constraints at the boundaries. In fact, one of the constraint equations $\mathbf{c}_d$ is $\boldsymbol{\psi} \leq 0$, which defines the boundary conditions on $\mathbf{x}_0$ and $\mathbf{x}_f$, and other two correspond to the first and last finite element (see Eq. (21)), which contain again the vectors $\mathbf{x}_0$ and $\mathbf{x}_f$. If we call $\boldsymbol{\nu}$ the $\hat{\boldsymbol{\lambda}}_f$ that corresponds to $\boldsymbol{\psi}$ and expand Eqs. (59), (60), and (61), we get:

$$\frac{\partial L}{\partial \mathbf{u}_{s,j}} = \sum_{k=1}^{l+1} \sigma_k \left[\hat{\lambda}^T \theta_{s,j} \nabla_{\mathbf{u}_{s,j}} \mathbf{h} + \hat{\mu}^T \nabla_{\mathbf{u}_{s,j}} \mathbf{g} \right] = 0 \tag{64}$$

$$\frac{\partial L}{\partial \mathbf{x}_{s,j}} = \sum_{k=1}^{l+1} \sigma_k \left[\hat{\lambda}^T \dot{\theta}_{s,j} \varphi_{s,j} + \hat{\theta}_{s,j} \nabla_{\mathbf{x}_{s,j}} \mathbf{h} + \hat{\mu}^T \nabla_{\mathbf{x}_{s,j}} \mathbf{g} \right] = 0 \tag{65}$$

$$\frac{\partial L}{\partial \mathbf{x}_f} = \eta^T \boldsymbol{\omega} \nabla_{\mathbf{x}_f} \mathbf{J} + \nu^T \nabla_{\mathbf{x}_f} \boldsymbol{\psi} - \hat{\lambda}_f^T = 0 \tag{66}$$

The third equation is the transversality condition on the terminal states in (35). The second equation becomes:

$$\begin{aligned} &\textstyle\sum_{k=1}^{l+1} \sigma_k \left[\hat{\lambda}^T \dot{\theta}_{s,j} \varphi_{s,j} + \theta_{s,j} \nabla_{\mathbf{x}} \mathbf{h} \frac{d\mathbf{x}}{d\mathbf{x}_{s,j}} + \hat{\mu}^T \nabla_{\mathbf{x}} \mathbf{g} \frac{d\mathbf{x}}{d\mathbf{x}_{s,j}} \right] = \\ &\textstyle\sum_{k=1}^{l+1} \sigma_k \left[\hat{\lambda}^T \dot{\theta}_{s,j} \varphi_{s,j} + \theta_{s,j} \nabla_{\mathbf{x}} \mathbf{h} \varphi_{s,j} + \hat{\mu}^T \nabla_{\mathbf{x}} \mathbf{g} \varphi_{s,j} \right] = 0 \end{aligned} \tag{67}$$

Here we made use of the fact that $\mathbf{g}(\mathbf{x}_{s,j}, \mathbf{u}_{s,j}, t_s) \leq 0 \Rightarrow \sum_{k=1}^{l+1} \sigma_k \mathbf{g}(\mathbf{x}_{s,j}, \mathbf{u}_{s,j}, t_s) \leq 0$. If one now takes the limit for an infinite number of integration points the sums become continuous integrals:

$$\int \left[\sum_s \hat{\lambda}_{s,j} \theta_{s,j} \nabla_{\mathbf{u}} \mathbf{h} + \hat{\mu}^T \nabla_{\mathbf{u}} \mathbf{g} \right] dt = 0 \tag{68}$$

$$\int \left[\sum_s \hat{\lambda}_{s,j} \dot{\theta}_{s,j} + \sum_s \hat{\lambda}_{s,j} \theta_{s,j} \nabla_{\mathbf{x}} \mathbf{h} + \hat{\mu}^T \nabla_{\mathbf{x}} \mathbf{g} \right] dt = 0 \tag{69}$$

Now if we make use of the fact that λ is approximated by the polynomial $\lambda \simeq \sum \hat{\lambda}_{s,j} \theta_{s,j}$, for an infinite number of collocation points, which would correspond to an infinite number of integration points, we have:

$$\int \left[\lambda^T \nabla_{\mathbf{u}} \mathbf{h} + \mu^T \nabla_{\mathbf{u}} \mathbf{g} \right] dt = 0 \tag{70}$$

$$\int \left[\dot{\lambda}^T + \lambda^T \nabla_{\mathbf{x}} \mathbf{h} + \mu^T \nabla_{\mathbf{x}} \mathbf{g} \right] dt = 0 \tag{71}$$

which are satisfied if the quantities in brackets are identically zero. These equations correspond to the optimality condition and to the differential equations on λ in (34).

4 Test Case

In this final section we present the application of the MOCP solution method to a simple optimal control problem. The test case, also known as the Goddard Rocket problem, has an analytical solution that is available in the literature [25]. The algorithm MACSoc was run 30 times on this problem to gather statistics on the quality of the Pareto front, given the stochastic nature of the population component of the algorithm. The local NLP solver is the Matlab function $fmincon$.

4.1 Ascent Trajectory with Constant Acceleration

The problem is to find an optimal ascent trajectory from a flat celestial body with no atmosphere to a prescribed altitude. The control variable is the thrust angle and both gravity and thrust accelerations are constant. The final altitude is assigned and the final vertical component of the velocity has to be zero. The single objective optimal control formulation of the problem and its analytical solutions for either minimum time or maximum horizontal velocity can be found in [25], while a numerical solution with DFET can be found in [12].

In this paper, the problem is reformulated as follows, to consider the two objectives simultaneously:

$$\min_{t_f,u}(J_1 = t_f, J_2 = -v_x(t_f)) \tag{72}$$

subject to the dynamic constraints:

$$\begin{cases} \dot{x} & = v_x \\ \dot{v}_x & = a \cos u \\ \dot{y} & = v_y \\ \dot{v}_y & = -g + a \sin u \end{cases} \tag{73}$$

where g is the gravity acceleration, a the thrust acceleration, x and y are the components of the position vector, v_x and v_y the components of the velocity vector and u the control. The dynamics is integrated from time $t = 0$ to time $t = t_f$. The

Table 1 MACSoc settings

max_fun_eval	10,000
pop_size	10
ρ_ini	1
F	0.9
CR	0.9
p_social	1
max_arch	10
max_contr_ratio	5

Table 2 *fmincon* settings

max_eval	100
tol_con	1e−6

boundary conditions are:

$$\begin{cases} x(0) = 0; & v_x(0) = 0 \\ y(0) = 0; & v_y(0) = 0 \\ y(t_f) = h; & v_y(t_f) = 0 \end{cases} \tag{74}$$

The parameters g, a and h were respectively set to 1.6×10^{-3}, 4×10^{-3} and 10. Following [12], the DFET method was applied splitting the time domain into 4 elements, with polynomials of order 6 for each control and state variable. The control angle was bounded between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, while total mission time was bounded between 100 and 250. This gives a total of 29 optimisation variables. Table 1 summarises the settings of the optimiser: max_fun_eval the maximum number of objective functions evaluation, pop_size the number of agents performing the search, ρ_ini the initial radius of the local neighbourhood, F and CR the standard parameters for the Differential Evolution social actions, p_social the ratio between agents performing only social actions and the total number of agents, max_arch the number of solutions to be stored in A_g, $contr\_ratio$ contraction rate of the neighbourhood radius, and max_contr_ratio the maximum number of times ρ_j can contract before it is reset (for more details on the settings of MACS, the multi-objective solver in MACSoc, please refer to [20]). Settings reported in Table 2 instead refer to the parameters of *fmincon*: max_con_eval is the maximum number of constraints evaluation (for each call to the objective functions) and tol_con is the threshold under which the solution is considered to be feasible. All other *fmincon* settings are left as default.

Algorithm 1 was run 30 times to collect some statistics on its convergence behaviour (see Table 3). The Generational Distance (GD) [26] and Inverse Generational Distance (IGD) were used as accuracy metrics and were computed on a rescaled front in the interval [0, 1]. GD and IGD were computed using the analytical solution of the minimum time problem for different maximum v_x. Figure 4 shows the cumulative front from all 30 runs, along with four representative solutions

Table 3 Convergence and spreading statistics for the two problems

Problem	Mean GD (Variance)	Mean IGD (Variance)
Goddard	2.833e-2 (1.4232e-5)	2.9449e-2 (1.5498e-5)

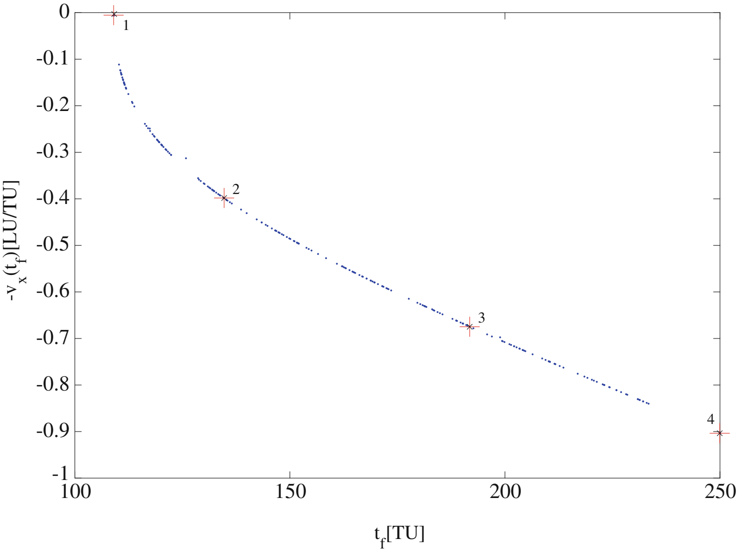


Fig. 4 Non dominated solutions of 30 different runs for the Goddard problem. Crosses indicate solutions for which trajectories, velocities and control law over time are also plotted. Circles indicate the objective values corresponding to the analytic solutions with the same time as the solutions marked with crosses

(marked with crosses) and the analytic solutions with the same ascent time of the representative solutions (marked with circles). The crosses and circles are perfectly overlapping. The trajectories and time histories of the controls and velocities for the four representative solutions are plotted in Figs. 5, 6, 7, 8 and 9 along with the single objective numerical solution and the analytic solution for the same ascent times. The solution obtained with the proposed approach is very close to both the numerical single objective and the analytic solutions (1e − 6). The discontinuities in the control laws are due to the discretisation scheme and to tolerance on the optimality of the solutions.

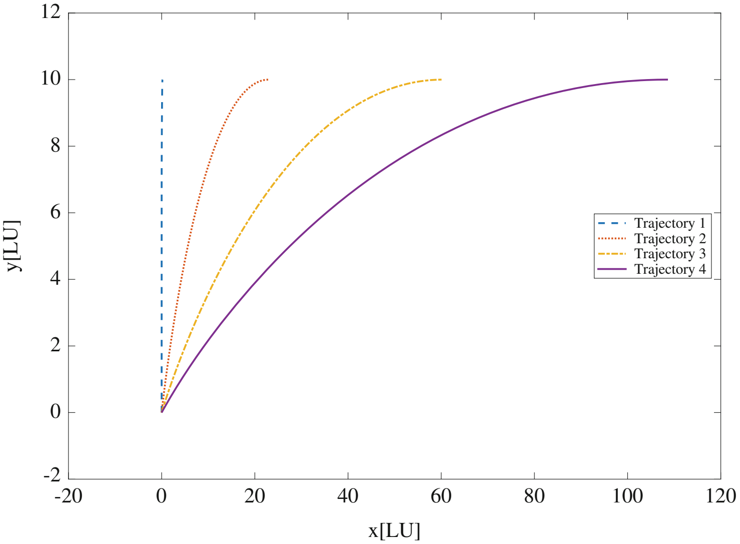
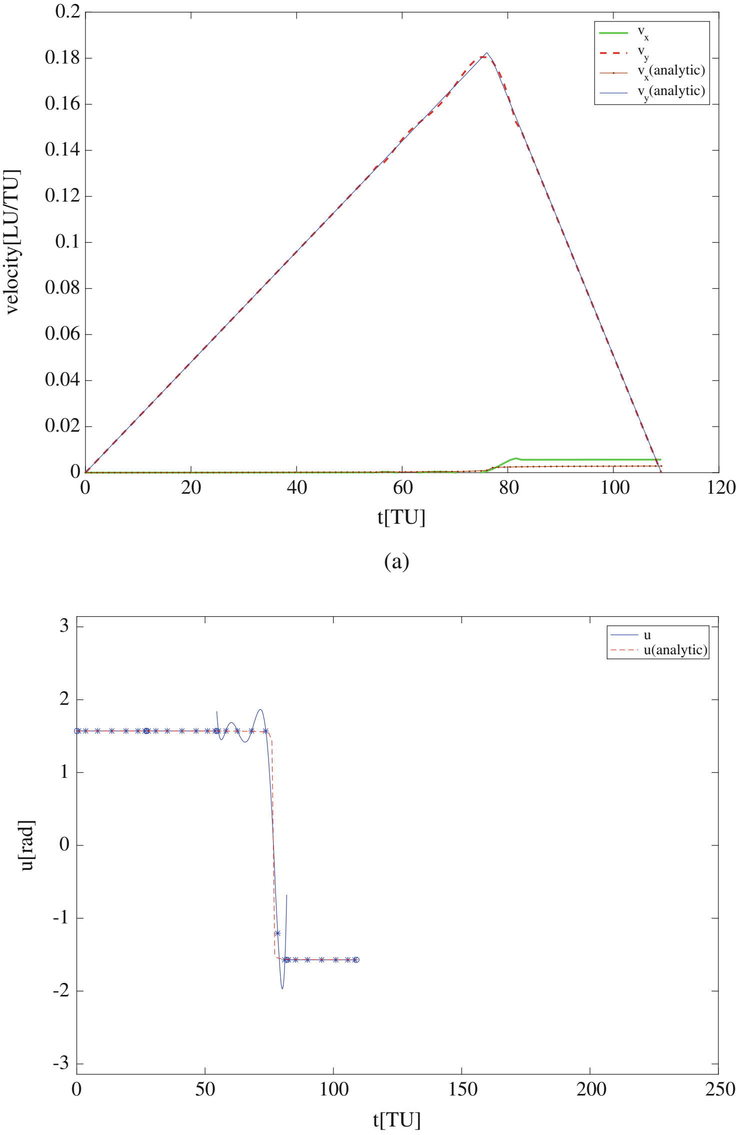


Fig. 5 Trajectories corresponding to the four selected points on the Pareto front

5 Conclusion

In this chapter we introduced some basic notions of multi-objective optimisation and optimal control and we derived an optimal control theory for multi-objective optimal control problems scalarsised with Pascoletti-Serafini scalarisation method. We then presented a possible solution approach that makes use of a direct transcription of the optimal control problem with Finite Elements in Time and solves the resulting NLP problem with a memetic algorithm.

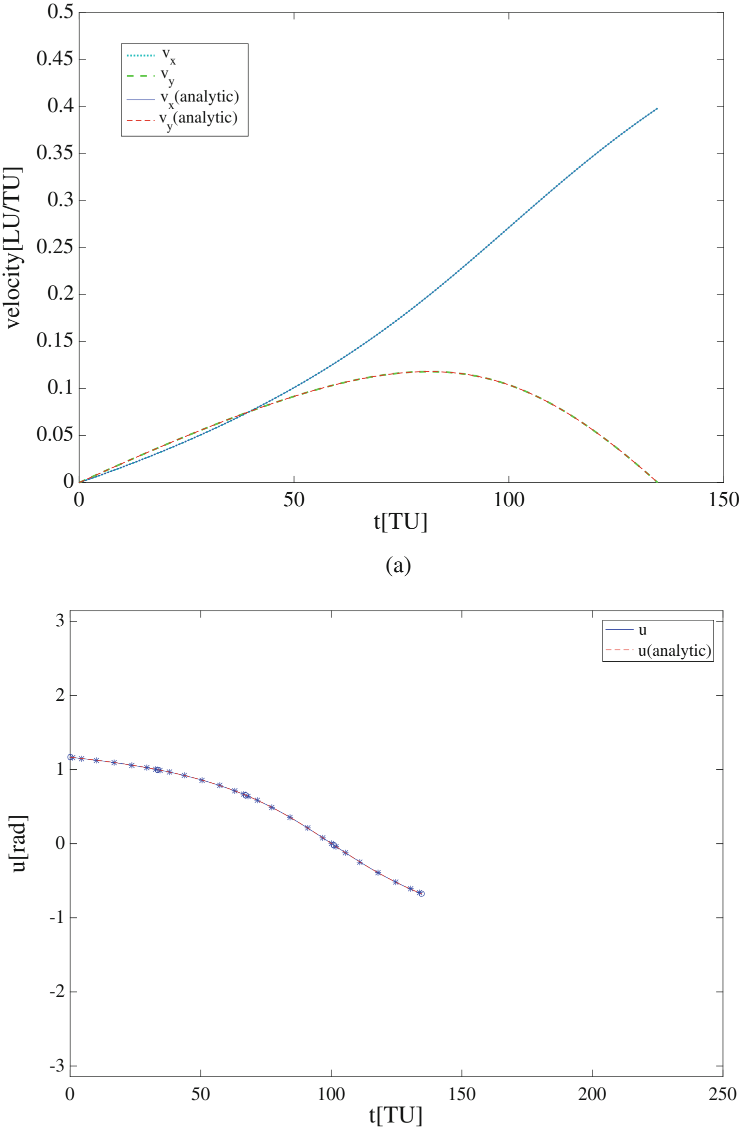
This combination provides an effective solution of multi-objective optimal control problems, as demonstrated by the simple example of the Goddard's rocket. Future direction include a more flexible treatment of the infeasible solutions in the bi-level scheme to limit the effort of the gradient-based solver and allow a faster and broader exploration of the parameter space. Also, the choice of the weights is subject to a proper normalisation of the objective functions and needs some cleaver adaptation heuristics in case of many objectives with very irregular Pareto fronts.



(a)

(b)

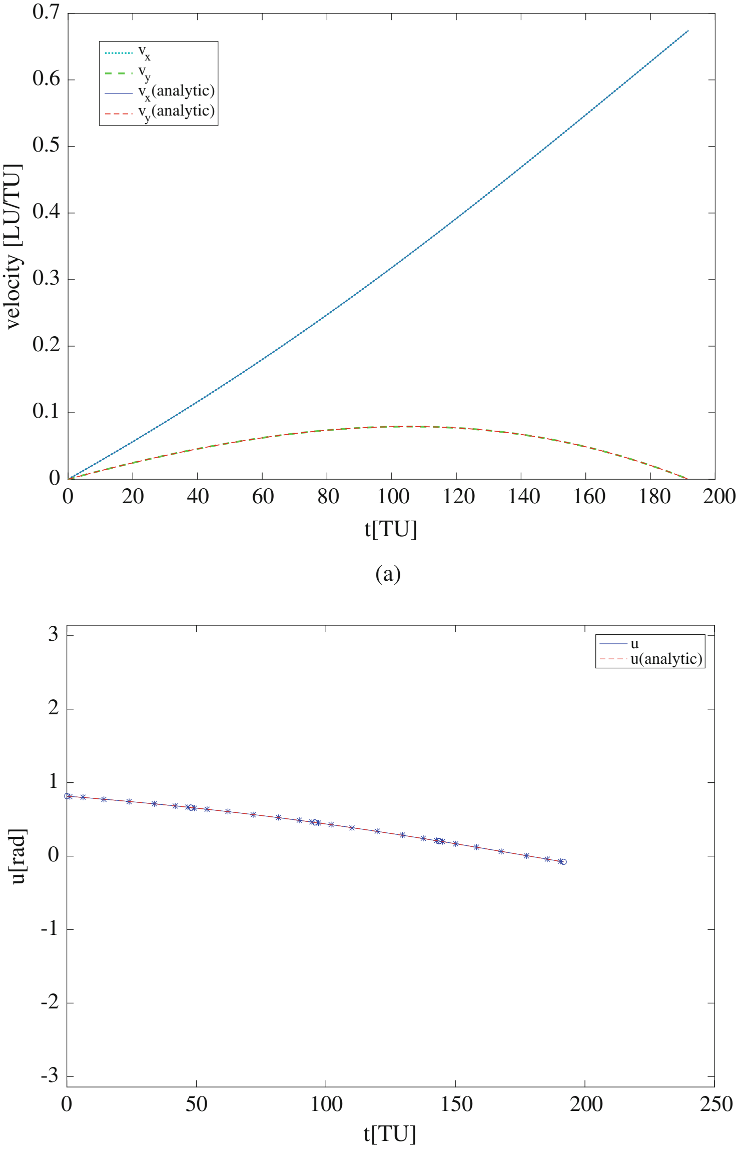
Fig. 6 Time history for velocities and controls, point 1 on the Pareto front. (**a**) Time history for the velocities. (**b**) Time history for the controls



(a)

(b)

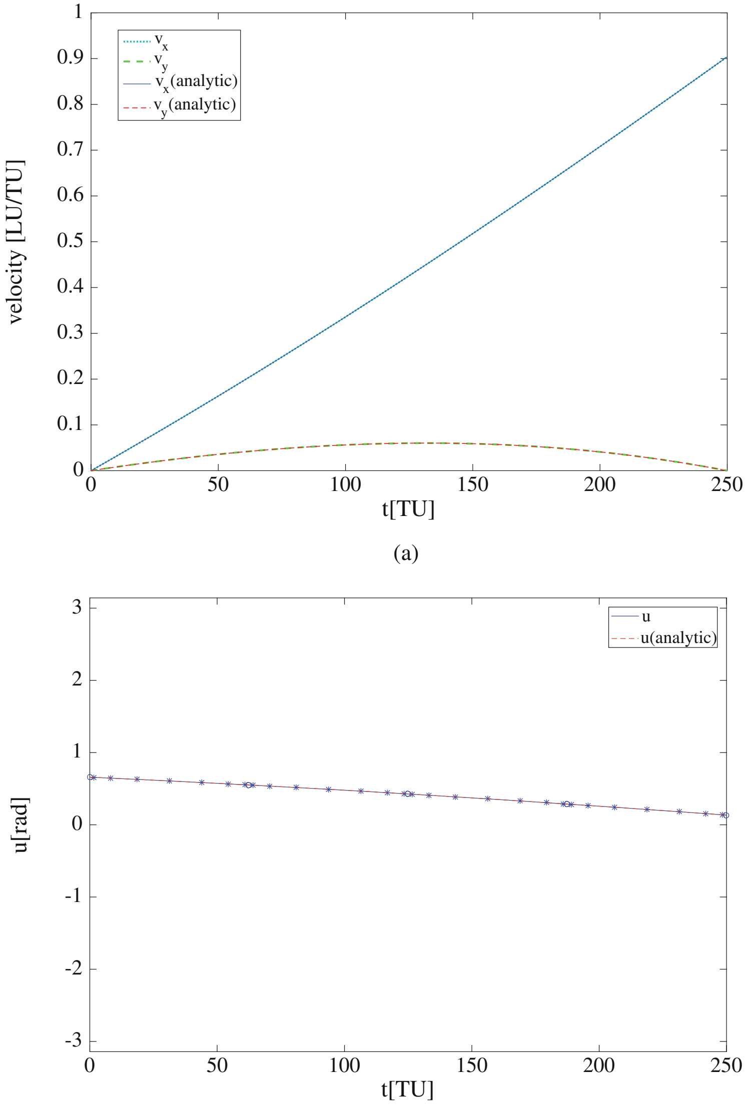
Fig. 7 Time history of velocities and controls, point 2 on the Pareto front. (**a**) Time history of the velocities. (**b**) Time history of the controls



(a)

(b)

Fig. 8 Time history of velocities and controls, point 3 on the Pareto front. (**a**) Time history of the velocities. (**b**) Time history of the controls



(a)

(b)

Fig. 9 Time history of velocities and controls, point 4 on the Pareto front. (**a**) Time history of the velocities. (**b**) Time history of the controls

Acknowledgements The research was partially funded by an ESA NPI grant (ref TEC-ECN-SoW-20140806) and Airbus Defence & Space. The author would like to thank Mr Lorenzo Ricciardi for the contribution to the development and implementation of the algorithms used for the generation of the numerical results in this chapter.

References

1. Coverstone-Carroll, V., Hartmann, J.W., Mason, W.J.: Optimal multi-objective low-thrust spacecraft trajectories. Comput. Methods Appl. Mech. Eng. **186**(2), 387–402 (2000)
2. Yalçın Kaya, C., Maurer, H.: A numerical method for nonconvex multi-objective optimal control problems. Comput. Optim. Appl. **57**(3), 685–702 (2014)
3. Englander, J.A., Vavrina, M.A., Ghosh, A.R.: Multi-objective hybrid optimal control for multiple-flyby low-thrust mission design. In: 25th AAS/AIAA Space Flight Mechanics Meeting, 11–15 January (2015)
4. Ober-Blöbaum, S., Ringkamp, M., zum Felde, G.: Solving multiobjective optimal control problems in space mission design using discrete mechanics and reference point techniques. In: IEEE 51st Annual Conference on Decision and Control (CDC), 2012, pp. 5711–5716. IEEE, Piscataway (2012)
5. Chankong, Y.Y., Haimes, V.: Multiobjective Decision Making. Dover Publications, Inc., Mineola (2008)
6. Hillermeier, C.: Nonlinear Multiobjective Optimization. International Series of Numerical Mathematics. Birkhäuser, Basel (2001). https://doi.org/10.1007/978-3-0348-8280-4
7. Pascoletti, A., Serafini, P.: Scalarizing vector optimization problems. J. Optim. Theory Appl. **42**, 499–524 (1984)
8. Eichfelder, G.: Adaptive Scalarization Methods in Multiobjective Optimization. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-79159-1
9. Zuiani, F., Vasile, M.: Multi agent collaborative search based on tchebycheff decomposition. Comput. Optim. Appl. **56**(1), 189–208 (2013)
10. Zhang, Q., Li, H.: Moea/d: a multi-objective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comput. **11**(6), 712–731 (2007)
11. Shapiro, S.: Lagrange and mayer problems in optimal control. Automatica **3**(3), 219–230 (1966)
12. Vasile, M.: Finite elements in time: a direct transcription method for optimal control problems. In: AIAA/AAS Astrodynamics Specialist Conference, Guidance, Navigation, and Control and Co-located Conferences, Toronto, 2–5 August 2010
13. Vasile, M., Finzi, A.E.: Direct lunar descent optimisation by finite elements in time approach. Int. J. Mech. Control **1**(1) (2000)
14. Hodges, D.H., Bless, R.R.: Weak hamiltonian finite element method for optimal control problems. J. Guid. Control Dyn. **14**(1), 148–156 (1991)
15. Bottasso, C.L., Ragazzi, A.: Finite element and runge-kutta methods for boundary-value and optimal control problems. J. Guid. Control Dyn. **23**(4), 749–751 (2000)
16. Vasile, M., Bernelli-Zazzera, F.: Optimizing low-thrust and gravity assist maneuvers to design interplanetary trajectories. J. Astronaut. Sci. **51**(1), 13–35 (2003)
17. Vasile, M., Bernelli-Zazzera, F.: Targeting a heliocentric orbit combining low-thrust propulsion and gravity assist manoeuvres. Oper. Res. Space Air **79**, 203–229 (2003)
18. Ricciardi, L.A., Vasile, M.: Direct transcription of optimal control problems with finite elements on bernstein basis. AIAA J. Guid. Control Dyn. **42**(2), 229–243 (2019)
19. Zuiani, F., Kawakatsu, Y., Vasile, M.: Multi-objective optimisation of many-revolution, low-thrust orbit raising for destiny mission. Adv. Astronaut. Sci. **148**, 783–802. In: Proceedings of the 23rd AAS/AIAA Space Flight Mechanics Conference, January (2013)

20. Ricciardi, L.A., Vasile, M.: Improved archiving and search strategies for multi agent collaborative search. In: Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences, pp. 435–455. Springer, Cham (2018)
21. Ricciardi, L.A., Vasile, M., Maddock, C.: Global solution of multi-objective optimal control problems with multi agent collaborative search and direct finite elements transcription. In: IEEE Congress on Evolutionary Computation (CEC), 2016, pp. 869–876. IEEE, Piscataway (2016)
22. Ricciardi, L.A., Vasile, M., Toso, F., Maddock, C.A.: Multi-objective optimal control of the ascent trajectories of launch vehicles. In: AIAA/AAS Astrodynamics Specialist Conference, pp. 5669 (2016)
23. Vasile, M., Ricciardi, L.: A direct memetic approach to the solution of multi-objective optimal control problems. In: IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1–8. IEEE, Piscataway (2016)
24. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)
25. Bryson, A.E.: Applied Optimal Control: Optimization, Estimation and Control. CRC Press, Boca Raton (1975)
26. Van Veldhuizen, D.A.: Multiobjective evolutionary algorithms: classifications, analyses, and new innovations. Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB, Ohio (1999)

Practical Uncertainty Quantification in Orbital Mechanics

Massimiliano Vasile

Abstract The chapter provides an overview of methods to quantify uncertainty in orbital mechanics. It also provides an initial classification of these methods with particular attention to whether the quantification method requires a knowledge of the system model or not. For some methods the chapter provides applications examples and numerical comparisons on selected test cases.

Keywords Uncertainty quantification · Orbital mechanics · Uncertainty propagation

1 Introduction

Although orbital mechanics is fundamentally based on deterministic models, the position, velocity and attitude of a space object can only be known with some degree of uncertainty. Model uncertainty and uncertainty in measurements and observations concur to transform a seemingly deterministic problem into a stochastic one.

This chapter provides an overview of methods for the quantification of uncertainty in orbital mechanics with some considerations on their practical applicability to different scenarios.

The best known form of uncertainty quantification in orbital mechanics falls probably under what is commonly known as orbit determination [1, 2]. In fact, the problem dates back to Gauss [3] and is fundamental in astronomy. Classical techniques include batch and sequential filters [2] where the latter can be used to estimate model parameters and implement navigation and control loops in complex nonlinear dynamic environments [4, 5].

One key issue is the linearity of dynamics. In fact, the general objective is to achieve a good estimation of the expected state of a space object at a given time. The quantification of uncertainty associated to the expected state is the probability

M. Vasile (✉)
University of Strathclyde, Glasgow, UK
e-mail: massimiliano.vasile@strath.ac.uk

G. Baù et al. (eds.), *Satellite Dynamics and Space Missions*,
Springer INdAM Series 34, https://doi.org/10.1007/978-3-030-20633-8_7

associated to a variation of the expected state. Given a generic system of differential equations:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{p})\\ \mathbf{x}(t_0) &= \mathbf{x_0}\end{aligned} \tag{1}$$

with $\mathbf{x}$ the state vector and $\mathbf{p}$ a vector of model parameters, the question is whether the Encke's model:

$$\begin{aligned}\delta\dot{\mathbf{x}} &= \mathbf{J}(\mathbf{x}_0, \mathbf{p}) + \varepsilon\\ \delta\mathbf{x}(t_0) &= \delta\mathbf{x_0}\end{aligned} \tag{2}$$

is representative of the evolution of the variated state $\mathbf{x}_0 + \delta\mathbf{x}$, where ε is some uncertainty on the dynamics $f(\mathbf{x}, \mathbf{p})$, $\mathbf{J}$ is the Jacobian matrix of the vector function f, $\mathbf{x}_0$ is the expected state and $\delta\mathbf{x}$ is an uncertainty on the initial state.

In recent times a number of authors focused on developing methods to better capture nonlinearities in the case in which model (2) is not giving satisfactory results [6–10]. This area of research is generally concerned with the propagation of uncertainty and relies on standard probability theory and strong assumptions on the underlying probability distributions.

An important area of application, that has attracted more and more attention in recent times, is collision avoidance. The problem, in this case, is to have an accurate long term prediction of a possible collision in order to plan and implement one or more collision avoidance manoeuvres. Given the cost of the implementation of a collision avoidance manoeuvre, an accurate and reliable prediction is paramount. In view of an increase of the traffic in orbit, this problem becomes of fundamental importance and traditional approaches derived from orbit determination might not be sufficient. The key difficulty comes from two main issues: the effect of nonlinearities over long term predictions, or large uncertainty, and the intrinsic epistemic nature of the underlying uncertainty. The former issue has been widely investigated while the latter is still a somewhat open problem. In fact, the general approach is to treat the uncertainty as aleatory with a consequence dilution of the probability of a collision as the knowledge reduces. The same applies to rare but high risk events, such as the impact of an asteroid with Earth, for which epistemic uncertainty affects both the knowledge of the state of the system and the dynamical model that governs its motion.

To be noted that the quantification of uncertainty includes the uncertainty on the implementation of collision avoidance manoeuvre or any manoeuvre in general. The level of complexity, in this case, is increased by the uncertainty coming from system design aspects that are not directly dependent on the dynamics of space objects but have an impact on the prediction of their future state.

This chapter is structured as follows. In the first section we introduce a general formulation for the orbital dynamics of individual space objects. This formulation incorporates the system aspects via a model parameter vector $\mathbf{p}$. The section includes a brief discussion on formulations that look at the overall density distribution over the whole space occupied by space objects orbiting the Earth.

The second section will classify uncertainty and quantification methods. The main focus is not the technique to model uncertainty but the technique to handle and propagate uncertainty in orbital dynamics. The following sections will expand on each of the classes of techniques presenting the major approaches that can be found in the literature. Some illustrative examples will accompany each of the sections. The last two sections will explore some techniques to capture model uncertainty and to define an appropriate uncertainty model.

2 Problem Formulation

The general problem is to quantify the probability that a space object is at a given position with a given velocity at a given time, conditional to the uncertainty associated to its initial state, model parameters and the dynamics itself. The state of a space object at a given time is, therefore, called the *quantity of interest* in the remainder of this chapter.

If the interest is in the dynamics of a single object with state $\mathbf{x}(t)$ at time t, one can start from the following Cauchy problem:

$$\begin{cases} \dot{\mathbf{x}} & = f(\mathbf{x}, \mathbf{p}, \gamma(\mathbf{x}, \mathbf{p}))\eta(\mathbf{x}, \mathbf{p}) + \nu(\mathbf{x}, \mathbf{p}) \\ \mathbf{x}(t_0) & = \mathbf{x_0} \end{cases}, \tag{3}$$

where $\mathbf{p} \in \Upsilon \subseteq \mathbb{R}^q$ is a vector of model parameters and the initial conditions have value $\mathbf{x_0} \in \Sigma_0 \subseteq \mathbb{R}^c$. The uncertainty space of model parameters and initial conditions is defined as $\Omega = \Upsilon \cup \Sigma_0 \subset \mathbb{R}^d$. The three functions η, γ and ν indicate a multiplicative, a composition and an additive uncertainty function respectively. In this chapter the three functions η, γ and ν are not random processes unlike what appears in stochastic differential equations where ν is generally modelled as a Weiner process. We will consider, instead, that η, γ and ν belong to some normed functional space and are Lipschitz continuous.

In this framework, the quantification of uncertainty requires two different operations: one is the reconstruction of the uncertainty functions η, γ and ν, that we will call model uncertainty, and the propagation of the uncertainty set Ω.

When the interest is to calculate the density of objects ρ in a control volume, the problem can be formulated as:

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{v}) = \sum_k \Phi_k + n^+ + n^-, \tag{4}$$

where $\mathbf{v}$ is the velocity field, Φ_k the external force field and n^+ and n^- two processes that add and remove objects from the control volume. The continuity equation (4) was introduced for the treatment of debris fields in 1993 by Smirnov et al. [11] and extended by Nazarenko to include the dependency on orbital elements and probability distribution in 1997 [12]. A parallel development was proposed by [13] and later on [14]. It is interesting to mention that a similar approach using Jeans equation is used to study galactic dynamics [15].

Formulation (4) requires a careful interpretation. In fact, orbiting objects, even excluding collisions and active manoeuvres, behave more like a rarefied gas than a continuous fluid. Thus, conceptually, if Eq. (4) is understood as actual mass density, it predicts a non-zero density even when no objects are present.

2.1 Quantity of Interest, Uncertainty and Expectation

In the two formulations presented in the previous section the quantity of interest is different in nature. For problem (3) the quantity of interest is the state of the object at a given time, $\mathbf{x}(t)$, while in (4), the quantity of interest is the density of objects at a given time, $\rho(t)$.

If one takes formulation (3) the quantification of uncertainty can be expressed in general terms as:

$$P(\mathbf{x}(t)|\Omega) = \int_{\Omega} (\mathbf{x}(t) \in \Psi)\phi(\xi)d\xi, \tag{5}$$

where ξ is the uncertain vector, with distribution ϕ, and Ψ is a target set. This quantification does not introduce any assumption on the probability distribution or on the spatial distribution of all possible states at a given time. This approach is directly applicable to the calculation of collisions and conjunctions or to problems of rendezvous, docking, landing and flyby. On the other hand the calculation of integral (5) is not a trivial matter, especially in high dimensions.

In a classical framework, where one is interested only in the expected state $\hat{\mathbf{x}}$ and associated covariance $Cov(\mathbf{x})$, under suitable hypothesis, one can calculate the expected state as a weighted average of a set of samples $\tilde{\mathbf{x}}_i$:

$$\hat{\mathbf{x}} = \sum_i w_i \tilde{\mathbf{x}}_i \tag{6}$$

with covariance:

$$Cov(\mathbf{x}) = \sum_i w_i [\tilde{\mathbf{x}}_i - \hat{\mathbf{x}}][\tilde{\mathbf{x}}_i - \hat{\mathbf{x}}]^T. \tag{7}$$

This approach is computationally far less complex than the calculation of (5) because it does not require the propagation of Ω and to calculate the inclusion $(\mathbf{x}(t) \in \Psi)$. On the other hand, it captures only the first two statistical moments of the distribution of the possible states of the system at a given time. It is important to keep in mind that in this framework the uncertainty is represented by $Cov(\mathbf{x})$.

2.1.1 Upper and Lower Expectations

When the uncertainty on the input quantities is epistemic the probability ϕ can belong to a family of parametric distributions or to a set of unknown distributions.

Consider the case in which one can reasonably assume that the uncertainty can be quantified with a family of beta distributions with unknown parameters α and β (any other parametric or non-parametric distribution would equally work). Equation (5) then translates into two equations defining the upper and lower probability associated to $\hat{\Omega}$:

$$P_l = \min_{\alpha,\beta} \int_{\hat{\Omega}} \phi(\xi)\, d\xi\ , \qquad P_u = \max_{\alpha,\beta} \int_{\hat{\Omega}} \phi(\xi)\, d\xi\ , \tag{8}$$

where ϕ is the product of probability $\phi = \prod_{j=1}^{d} \phi_j$, where each marginal density mass ϕ_j is a beta distribution function with parameters α_j, β_j. Here $\hat{\Omega}$ is the subset of Ω defined as:

$$\hat{\Omega} = \{\xi | (\mathbf{x}(\xi, t) \in \Psi)\} \tag{9}$$

As it will be shown later in this chapter the same idea can be extended to a generic set of distributions if one is representing ϕ in a suitable form so that:

$$P(\mathbf{c}) = \int_{\hat{\Omega}} \phi(\xi, \mathbf{c})\, d\xi\ , \tag{10}$$

is a probability function of a vector $\mathbf{c}$ of free parameters.

3 Classification and Definitions

Each method for uncertainty quantification is composed of three elements, broadly speaking: an uncertainty model, a propagation technique, and an inference process. This chapter will only consider the first two elements as the inference process, which defines how to make decisions on the results of the quantification, is a much broader topic that requires a dedicated discussion.

The uncertainty model defines the uncertainty that needs to be quantified, for example whether an uncertain quantity is normally distributed or not. In classical Probability Theory one starts from the definition of a probability space, a mathematical triplet (Ω, Φ, P) where Ω is a sample space, or set of outcomes, Φ is the collection of all the possible events $\Phi \subseteq 2^{\Omega}$ and P is the probability associated to each event such that $P : F \rightarrow [0, 1]$. When assigning a probability distribution is not possible, alternative models are considered. They fall under the broader group of imprecise probability theories. In this case P is a multivalued mapping and the single probability splits into an upper $\bar{P}$ and a lower $\underline{P}$ probability. Different theories exist and each one provides a different model to define the uncertain quantities and the probability associated to the quantity of interest [16, 17].

Once an uncertainty model is defined the second element is the propagation of the uncertainty to compute the quantity of interest. Given the specific problem (3) the propagation method maps the uncertainty in $\mathbf{x}_0$, $\mathbf{p}$, γ, η and ν at time t_0 into the uncertainty in $\mathbf{x}$ at time t.

The main difficulty in the propagation of uncertainty is to achieve a balance between accuracy and computational cost. The accuracy is in the representation of the quantity of interest and its probability at any time in the interval $[t_0, t]$. Before classifying the methods for uncertainty propagation, it is useful to classify the types of uncertainty that are normally considered in uncertainty quantification.

- **Aleatory** uncertainties are non-reducible uncertainties that depend on the very nature of the phenomenon under investigation. They can generally be captured by well defined probability distributions as one can apply a frequentist approach. E.g. measurement errors.
- **Epistemic** uncertainties are reducible uncertainties and are due to a lack of knowledge. Generally they cannot be quantified with a well defined probability distribution and a more subjectivist approach is required. Two classes: a lack of knowledge on the distribution of the stochastic variables or a lack of knowledge of the model used to represent the phenomenon under investigation.
- **Structural** (or model) uncertainty is a form of epistemic uncertainty on our ability to correctly model natural phenomena, systems or processes. If we accept that the only exact model of Nature is Nature itself, we also need to accept that every mathematical model is incomplete. One can then use an incomplete (and often much simpler and tractable) model and account for the missing components through some model uncertainty.
- **Experimental** uncertainty is aleatory. It is probably the easiest to understand and model, if enough data are available on the exact repeatability of measurements.
- **Geometric** uncertainty is a form of aleatory uncertainty on the exact repeatability of the manufacturing of parts and systems.
- **Parameter** uncertainty can be either aleatory or epistemic and refers to the variability of model parameters and boundary conditions.
- **Numerical** (or algorithmic) uncertainty, also known as numerical errors, refers to different types of uncertainty related to each particular numerical scheme, and to the machine precision (including clock drifts).
- **Human** uncertainty is difficult to capture as it has both aleatory and epistemic elements and is dependent on our conscious and unconscious decisions and reactions. It includes the possible variability of goals and requirements due to human decisions.

All the source of uncertainty listed above are applicable to orbital mechanics, including geometric uncertainty if one considers that the uncertainty in the execution of a manoeuvre depends on the manufacturing of the actuators. We can now consider the following classes of uncertainty propagation methods:

- **Intrusive methods**. These are methods that require accessing problem (3) to propagate the uncertainty and obtain a representation of states and probability at time t.
- **Non-intrusive methods**. These are methods that do not need any access to (3) but build a surrogate model based on a set of samples. In both cases, uncertainty can be directly propagated through (3) without any transformation. To be noted that if η and ν are not explicitly available, intrusive methods cannot be applied.
- **Direct vs Indirect methods**. Problem (3) describes a stochastic process when $\mathbf{x}$, $\mathbf{p}$ are stochastic quantities. Under appropriate assumptions on the nature of the uncertainty, generally described as a Weiner process, one can translate problem (3) into a stochastic differential equation, in the Itô form [18], and integrate forward in time. If the evolution of the distribution is of interest one would need to solve the Fokker–Plank equation. The integration of the Fokker–Plank equation poses remarkable challenges and some of the approaches in this paper were developed to overcome this challenges without resorting to a Monte Carlo simulation.

  When nonlinearities are small the propagation of the covariance with the state transition matrix is sufficient to give a correct value of the first two statistical moments and under the assumption of Gaussian a priori, they provide also a good representation of the a posteriori distribution. If nonlinearities are relevant first order approaches fail to correctly capture the distribution of the quantity of interest but also the first two moments result affected by a significant error. In the context of orbital mechanics some authors proposed to transform problem (3) with a different parameterisation [19] (Keplerian elements, averaged Keplerian elements, equinoctial elements) propagate uncertainty with a linear method in this new parameterisation and then transform back to the original set of parameters, position and velocity, where the measurements are typically acquired. In this chapter we will call this class of approaches *indirect*.

The accuracy in the representation of the quantity of interest and its probability depends on the ability of the propagation method to propagate nonlinearities but also to consider generic distributions, sets of distributions, fuzzy sets, belief functions, rough sets, etc. For this reason we can distinguish between methods that provide first a representation of the uncertain set and then of the probability distribution and methods that directly provide a representation of the probability distribution.

Table 1 provides a taxonomical classification of a number of methods that will be presented later on in this chapter: MC = Monte Carlo, STM = State Transition Matrix, STT = State Transition Tensor, GM = Gauss Mixture Models, UT = Unscented Transformation, PCE = Polynomials Chaos Expansions, PA = Polynomial Algebra, IA = Interval Arithmetic, TPE = Chebyshev Polynomial Expansions, HDMR = High Dimensional Model Representation, PPE = Positive Polynomial Expansions. A Yes in the table means that particular method has that particular property or can be used in that context. For example, MC is not intrusive, can be used to directly propagate uncertainty within any transformation of the problem or coordinates, provides the distribution of the states but cannot be used to directly calculate upper and lower expectations or belief functions.

Table 1 Taxonomy of uncertainty propagation techniques in orbital mechanics

Method	MC	STM	STT	GM	UT	PCE	PA	IA	TPE	HDMR	PPE
Intrusive	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No
Non-intrusive	Yes	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Direct	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Indirect	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes
State representation	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Probability representation	No	No	No	Yes	No	Yes	No	No	No	No	Yes
Imprecise Probability representation	No	No	No	Yes	No	No	No	No	No	No	Yes

4 Non-Intrusive Techniques

Non intrusive techniques are sampling based methods that work with generic models in the form of black-box codes. They have little requirements on the coding of the models or on their regularity. This advantage is interesting when a set is propagated through a complex system that cannot be expressed in a simple analytical form. Non-intrusive methods offer the additional advantage that can correct for epistemic model uncertainty identifying missing components from the assimilation of experimental data and measurements. The following partial list of methods will be considered in this chapter:

- **Monte Carlo (MC)** [20]: the most straight forward approach is to randomly sample the uncertainty region according to the probability distribution of its parameters, integrate the dynamical system for each of the sample point, to obtain the corresponding final state, and estimate the expectation of the final region of uncertainty. Despite the easiness of the methodology, it is the one with the highest computational cost. At comparable accuracy, the number of samples required by this technique is generally far greater than other sample-based method.

  Monte Carlo Simulations date back as far as Enrico Fermi's study on neutron diffusion, and can be used to derive statistical information via simulation of random samples or to compute multi-dimensional integrals. In uncertainty quantification MCS are used in both ways. The method starts from a probability distribution over the uncertainty space from which samples are drawn. Deterministic simulations are then run for all the samples to derive a quantification of the uncertainty in the output of the simulations.

  Under the hypotheses of the Central Limit Theorem, the expected value of a random variable X belongs with probability ε to the interval

$$E(X) \in \left[\bar{X}_n - \frac{c\bar{\sigma}_n}{\sqrt{n}}, \bar{X}_n + \frac{c\bar{\sigma}_n}{\sqrt{n}}\right] \tag{11}$$

where $\bar{X}_n = \frac{1}{n}\sum_i^n X_i$ and $\bar{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ and the probability ϵ is computed over the interval $[-c, c]$:

$$\varepsilon = \frac{1}{2\pi}\int_{-c}^{c} e^{-\frac{x^2}{2}} dx. \tag{12}$$

From these simple expressions one can deduce that for the mean to converge with confidence 0.95, the number of samples needs to be:

$$\left|E(X) - \bar{X}_n\right| \leq 1.96\frac{\sigma}{\sqrt{n}}. \tag{13}$$

The convergence rate of MCS to the correct mean value is therefore proportional to $1/\sqrt{n}$. To be noted that the convergence of the mean does not provide any information on the convergence of the distribution or an exact bound on the error.

- **Non-intrusive Polynomial Chaos Expansion (PCE)** [10]:
PCEs are popular in Computational Fluid Dynamics and have found recent applications also in astrodynamics [21, 22]. PCEs employ a set of orthogonal polynomial functions to approximate the functional form between the system response and each of the inputs [23–26]. The main advantage of this method is the ability to deal with nonlinear and non-Gaussian propagation of the uncertainty without any assumption on an a posteriori Gaussian distribution. PCEs allow one to use different polynomial kernels depending on the input distribution.
The chaos expansion for a component l of a the state vector $\boldsymbol{x}$ takes the form:

$$x_l = a_0 B_0 + \sum_{i_1=1}^{\infty} a_{i_1} B_1(\chi_{i_1}) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2} B_2(\chi_{i_1}, \chi_{i_2})$$
$$+ \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3} B_3(\chi_{i_1}, \chi_{i_2}, \chi_{i_3}) + \ldots \tag{14}$$

where χ are random inputs and B_i is a generic multivariate polynomial. This expression can be simplified by replacing the order-based indexing with a term-based indexing:

$$x_l = \sum_{j=0}^{\infty} \alpha_{lj} \Psi_j(\chi) \tag{15}$$

where there is a one-to-one correspondence between $a_{i_1 i_2 i_3}$ and α_{lj}, and between $B_n(\chi_{i_1}, \chi_{i_2}, \ldots, \chi_{i_v})$ and $\Psi_j(\chi)$. Each of the $\Psi_j(\chi)$ is a multivariate polynomials which involve products of the one-dimensional polynomials. In practice, one truncates the infinite expansion at a finite number of random variables and a finite expansion order, p:

$$x_l \cong \sum_{j=0}^{p} \alpha_{lj} \Psi_j(\chi) \tag{16}$$

Using Hermite polynomials, a multivariate polynomial $B(\chi)$ of order n is defined from:

$$B_n(\chi_{i_1}, \chi_{i_2}, \ldots, \chi_{i_v}) = e^{\frac{1}{2}\chi^T\chi}(-1)^n \frac{\partial^n}{\chi_{i_1}, \ldots, \chi_{i_v}} e^{-\frac{1}{2}\chi^T\chi} \tag{17}$$

which can be shown to be a product of one-dimensional Hermite polynomials involving a multi-index m_i^j:

$$B_n(\chi_{i_1}, \chi_{i_2}, \ldots, \chi_{i_v}) = \Psi_j(\chi) = \prod_{i=1}^{n} \psi_{m_i^j}(\chi_i) \tag{18}$$

For a multivariate polynomial the number of coefficients of the expansions for each uncertain variable is given by $\frac{(i_v+n)!}{i_v!n!}$ which shows that the expansions tend to increase quite rapidly with the number of variables and order. The coefficients of the expansion (15) are here calculated via spectral projection [27]. This approach projects the response $\mathbf{x}$ against each basis function using inner products and employs the polynomial orthogonality properties to extract each coefficient. Each coefficient in Eq. (16) is calculated as:

$$\alpha_{lj} = \frac{\langle x_l, \Psi_j \rangle}{\langle \Psi_j^2 \rangle} = \frac{1}{\langle \Psi_j^2 \rangle} \int_{\Omega} x \Psi_j \rho(\chi) d\chi \tag{19}$$

where the inner product involves a multi-dimensional integral over the support of the weighting function $\rho(\chi)$. Analytical expressions of the mean and covariance matrix are then available as:

$$\begin{aligned} \mu_x &= E[\mathbf{x}] \cong \textstyle\sum_{j=0}^{p} \alpha_j E[\Psi_j] = \alpha_0 \\ P_x &= E[(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T] \cong \textstyle\sum_{j=1}^{p} \alpha_j (\alpha_j)^T E[\Psi_j^2] \end{aligned}, \tag{20}$$

μ_G and P_G are the exact moments of the expansion, which converge to moments of the true response function; the vector α_j represents the j-th column of the matrix α of components α_{lj}. The computation of the multi-dimensional integral can be done using a MCS with low-discrepancy sequences or a quadrature formula using Gauss points and weights. The latter, however, requires a full tensor product and a number of points that increases exponentially with the number of dimensions. A more attractive choice, is based on sparse grids generated using Smolyak's algorithm [28]. Smolyak's approach provides a general tool for constructing efficient algorithms able to solve multivariate problems with orders of magnitude reduction in the number of support nodes while giving the same level of approximation as the usual tensor product. In this framework, the work of Genz and Keister [29] introduced fully symmetric interpolatory integration rules for Smolyak sparse grid of Gauss–Hermite nodes.

The multi-dimensional integral in Eq. (19) can be approximated as the sum of discrete number of terms:

$$\int_{\Omega} x_l \Psi_j \rho(\chi) d\chi \cong \sum_{i=1}^{ngrid} x_l(\chi_i)\Psi_j(\chi_i) w(\chi_i) \tag{21}$$

The set of points χ_i and weights $w(\chi_i)$ are defined by the Gauss–Hermite cubature rule in Genz and Keister [29]. These rules are optimal for the solution of multidimensional integrals over infinite regions with a Gaussian weight function. In the work of Genz and Keister [29], it is shown that a Gaussian integral for a polynomial of order n can be calculated perfectly using a grid of level $l = 2n + 1$. Figure 1 shows a normalised sparse grids, with different levels of accuracy, for 3 uncertain parameters χ_1, χ_2 and χ_3 using Hermite polynomials as bases.

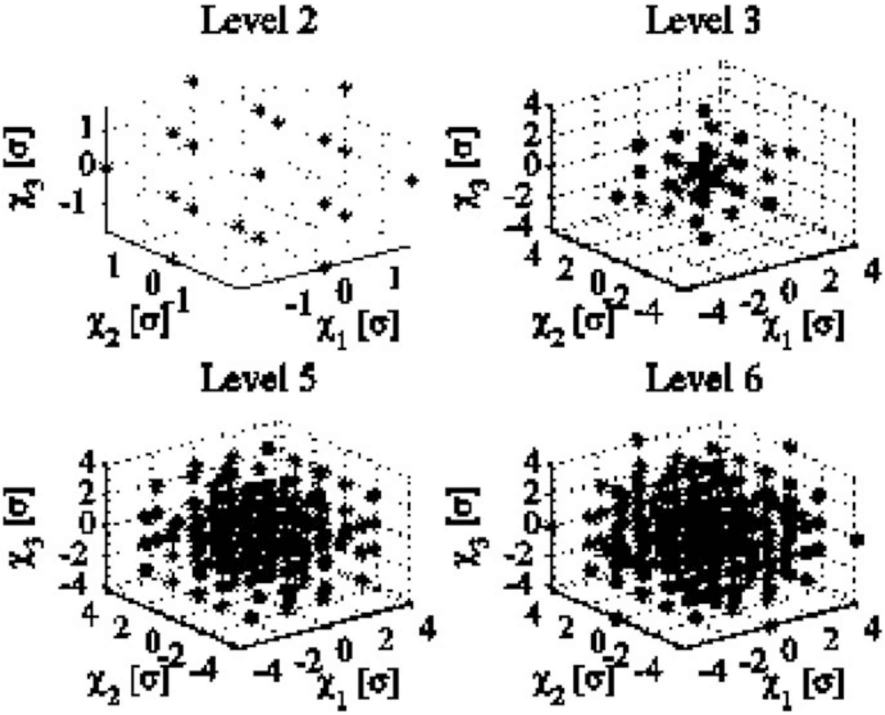


Fig. 1 Smoliak grid for Hermite polynomials with different levels of accuracy

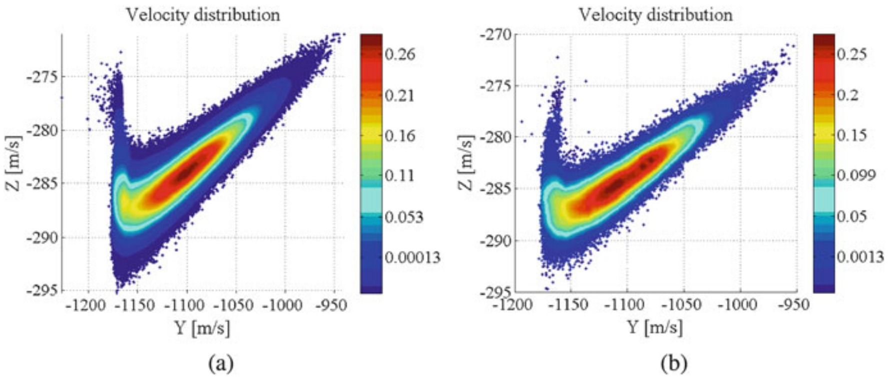


Fig. 2 Velocity distribution for a Libration Point to Moon trajectory: (**a**) Monte Carlo Simulation with 1e6 samples, (**b**) PCE of degree 6 with 28,000 sample

As one can see when the level of required accuracy is increased the grid is populated by a higher number of samples, which cover also a broader portion of the uncertainty space. For example, the max and min values of the uncertain parameters are respectively 1.7σ and -1.7σ for a level 2 grid, while they are 4σ and -4σ for a level 6 grid. This gives the possibility to better cover the uncertainty space, generating samples with low associated probability. Figure 2 shows the distribution of the velocity vector of a spacecraft along a trajectory from a Libration Point Orbit (in the Earth-Moon system) to the Moon. Figure 2a presents the result of an MC simulation with one million samples while Fig. 2b is the result of a PCE of degree 6 that uses only 28,000 samples (see [22] for more details).

- **Chebyshev Polynomial Expansions (TPE)** [30]. Univariate Chebyshev polynomials are an orthogonal basis over the space $\mathscr{C}^{\infty}[-1, 1]$ and the truncate Chebyshev series are close to the best uniform polynomial approximation for a given continuous function [31, 32]. When TPE are used as interpolating polynomials, samples are taken at points defined by a structured grid. The most popular sampling methods use Smolyak sparse grids [28], where the number of samples grows polynomially with the degree d, instead of exponentially. The number of elements to be included is controlled by a parameter l, called level of approximation, which has the same role as the order of expansion in the Taylor series. In this work, sparse grids are generated using extrema of unidimensional Chebyshev polynomials as described in Judd et al. [33].

  The reduced number of points allows one to reduce the number of terms in the Chebyshev polynomial basis, and so the number of unknown coefficients. The basis functions are chosen from all the polynomials up to degree d in n variables according to the level of approximation. Some products of higher order terms are not included under the assumption that their contribution is negligible.

Using the same notation as in the PCE section, we want to find the linear combination of multivariate Chebyshev polynomials of level of approximation l (and maximum degree 2^l) in n variables:

$$\hat{X}(\mathbf{x}) = \sum_{\alpha \in \mathscr{H}^{n,l}} c_\alpha \mathscr{T}_\alpha(\mathbf{x}) \,, \tag{22}$$

where

$$\mathscr{H}^{n,l} = \{\alpha \in \mathbb{N}^n : \alpha \text{ satisfies the Smolyak rule at level } l\} \,.$$

The unknown coefficients are computed via a Lagrange interpolation at the Chebyshev nodes given by the sparse grid of level l. Sparse grids have been introduced by Sergey Smolyak [28] and allow to represent, integrate and interpolate functions on multidimensional hypercubes. A complete polynomial basis of maximum degree 4 in 10 unknown variables consists of 1001 elements, while the corresponding sparse basis contains only 221 elements. We follow the construction of disjoint sparse grid presented in Judd et al. [33], that use the extrema of Chebyshev polynomials (also known as Chebyshev–Gauss Lobatto points or Clenshaw–Curtis points).

Let n be the number of uncertain variables and $l \in \mathbb{N}^+$ be the level of approximation of the sparse grid. The complete polynomial basis is given by

$$\mathscr{B} = \{\mathscr{T}_{\alpha_1}, \mathscr{T}_{\alpha_2} \, \dots, \mathscr{T}_{\alpha_s}\} \,, \qquad s \in \mathbb{N}^+ \,,$$

where $\alpha_i = (\alpha_{i_1}, \dots, \alpha_{i_n})$ denotes the multi-index array corresponding to the i-th multidimensional Chebyshev polynomial

$$\mathscr{T}_{\alpha_i} = \prod_{j=1}^{n} T_{\alpha_{i_j}} \,,$$

chosen in the space of all polynomial of degree at most 2^l in n variables such that

$$\alpha_i \in \mathscr{H}^{n,l} = \{\alpha \in \mathbb{N}^n \ : \ \alpha \text{ satisfies the Smolyak rule at level } l\} \,,$$

and $T_{\alpha_{i_j}}$ is the univariate Chebysehv polynomial corresponding to the variable of index j. For example, for $n = 2$ and $l = 1$ the Smolyak rule gives

$$\mathscr{H}^{2,1} = \{(0,0), (1,0), (0,1), (2,0), (0,2)\} \,,$$

and the corresponding Chebyshev polynomial basis is

$$\mathscr{T}_{(0,0)} = 1 \,, \ \mathscr{T}_{(1,0)} = x \,, \ \mathscr{T}_{(1,0)} = y \,, \ \mathscr{T}_{(2,0)} = 2x^2 - 1 \,, \ \mathscr{T}_{(0,2)} = 2y^2 - 1 \,.$$

Note the absence of the cross term $\mathscr{T}_{(1,1)} = xy$ from the basis. The response function can be approximated with the finite series

$$\hat{Y}(X_0) = \sum_{\alpha \in \mathscr{H}^{n,l}} c_\alpha \, \mathscr{T}_\alpha(X_0) \,, \tag{23}$$

where each c_α is the unknown coefficient with respect to the element $\mathscr{T}_\alpha$, and X_0 are the initial uncertainty variables and belong to an hypercube.
The unknown coefficients can be computed by inverting the linear system

$$HC = Y \,, \tag{24}$$

with

$$H = \begin{bmatrix} T_{\alpha_1}(x_1) & \dots & T_{\alpha_s}(x_1) \\ \vdots & \ddots & \vdots \\ T_{\alpha_1}(x_s) & \dots & T_{\alpha_s}(x_s) \end{bmatrix}, \qquad C = \begin{bmatrix} c_{\alpha_1} \\ \vdots \\ c_{\alpha_s} \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_s \end{bmatrix}, \tag{25}$$

where $x_1, \dots, x_s$ are the Chebyshev nodes in the sparse grid and the components of Y are the true values of the dynamical systems in these points. The system (24) cannot be inverted if the matrix H has not full rank. In most of the cases, this is guaranteed by choosing the Chebyshev nodes.

- **High Dimensional Model Representation (HDMR)** [34]. HDMR decomposes a generic quantity of interest, function of a generic parameter vector **b**, in a sum of functions of the components of **b**. If the quantity of interest is the solution of problem (3) at a given instant of time t, the corresponding HDMR decomposition can be written as:

$$\mathbf{x}_t(\mathbf{b}) = \mathbf{x}_0 + \sum_{i=1}^{d} c_i \alpha_i(b_i) + \sum_{1 \le i_1, i_2 \le n} c_{i_1,i_2} \alpha_{i_1,i_2}(b_{i_1}, b_{i_2}) + \\ + \sum_{1 \le i_1, i_2, \dots, i_d \le n} c_{i_1,i_2,\dots,i_d} \alpha_{i_1,i_2,\dots,i_d}(b_{i_1}, b_{i_2}, \dots, b_{i_d}),$$

where the $\mathbf{x}_0$ represents the mean value of the propagated states and the terms $\alpha_{i_j,i_k}(b_{i_j}, b_{i_k})$ represent the cooperative effects of the i_j and i_k input variables on the output. If the components of **b** are weakly coupled, this decomposition allows one to build an approximated representation by sampling only some slices of the space to which **b** belongs. Furthermore, it provides information on the influence of each component of **b** and their interactions in a similar fashion to an analysis of variance [35].

- **Unscented Transformation (UT)** [36]: The Unscented Transformation works on the underlying hypothesis that one can well approximate the posteriori covariance by propagating a limited set of optimally chosen samples, called sigma points. The set of sigma points are defined as then given as:

$$\chi_l = \begin{cases} \mathbf{x}_k \\ \mathbf{x}_k + \left(\sqrt{(n+k_f)\mathbf{P}_k}\right)_l \\ \mathbf{x}_k - \left(\sqrt{(n+k_f)\mathbf{P}_k}\right)_l \end{cases}, \tag{26}$$

where P_k is the covariance matrix, χ_l is a matrix consisting of $(2n+1)$ vectors, with $k_f = \alpha_f^2(n+\lambda_f) - n$, k_f is a scaling parameter, constant α_f determines the extension of these vectors around $\mathbf{x}_k$. The sigma points are transformed or propagated through the nonlinear function, the so-called Unscented Transformation, to give:

$$\chi_{l,k+1} = f(t, \chi_{l,k}), \tag{27}$$

From the collection of the propagated sigma points one then derives mean and covariance at stage $k+1$ and the process is iterated till time t.
Although the UT allows one to fully integrate the nonlinear system, it still capture only the first two moments and introduces a strong assumption on the symmetry of the prior distribution.

- **Gaussian Mixture (GM)** [37]. It is assumed that the probability density function of the uncertain parameters $\mathbf{b}$ is given by the weighted sum of M component Gaussian densities

$$p(\mathbf{b}) = \sum_{i=1}^{M} \omega_i g(\mathbf{b}|\mu_i, \Sigma_i).$$

Then also the probability density function of the uncertain parameters at a given time can be approximated as a Gaussian mixture. The weights associated with each Gaussian elements are computed so that the Fokker–Planck–Kolmogorov equation (FPKE) residual error is minimized (this equation, for continuous-time dynamic systems, gives the exact evolution of the states pdf). This minimisation problem is convex and hence has a unique solution. The main limitation of the method are the assumptions on the initial distribution.

Guassian Mixtures can be revisited as Kriging models [38] if one attempts an interpolation of the samples. In this case the weighted sum of Gaussian kernels is used to represent the shape of the propagated states. In the next section the use of Kriging will be compared to other non-intrusive representations, namely PCE, Chebyshev interpolation and HDMR.

4.1 Comparative Example

In this section some of the non-intrusive methods are compared on four different scenarios, please refer to Tardioli et al. [30] for further details:

1. Low-Earth orbit with 6 uncertain parameters (LEO6): the components of position and velocity at the initial states.
2. Low-Earth orbit with 10 uncertain parameters (LEO10): the components of position and velocity at the initial states, plus two uncertain model parameters.
3. Highly elliptical orbit with 6 uncertain parameters (HEO6): the components of position and velocity at the initial states.
4. Highly elliptical orbit with 10 uncertain parameters (HEO10): the components of position and velocity at the initial states, plus two uncertain model parameters.

The goal is to compare accuracy and computational cost, where the computational cost is measured using the number of sample points.

4.1.1 Dynamical Model

To compare the approximation provided by the four methods we use a dynamical model containing the main perturbations acting on a satellite of negligible mass orbiting in low-Earth orbit. The main gravitational perturbation is due to the non-spherical shape of the Earth: the most relevant effect is due to the J_2 coefficient in the development of the Earth's potential in spherical harmonics. Among the non-gravitational perturbations there are the solar radiation pressure (SRP) and the atmospheric drag.

In an equatorial reference frame, the dynamical equations can be written as

$$\begin{aligned} \dot{\mathbf{r}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= \mathbf{F}_{J_2} + \mathbf{F}_{SRP} + \mathbf{F}_{drag}\,, \end{aligned} \tag{28}$$

where $\mathbf{r}, \mathbf{v}$ are the position and velocity vectors, $\mathbf{r}_0 = \mathbf{r}(t_0), \mathbf{v}_0 = \mathbf{v}(t_0)$ are the initial conditions at the initial time t_0, and (see, e.g., Milani et al. [39], Sharaf and Selim [40])

$$\mathbf{F}_{J_2} = -\frac{\mu}{r^3} r + 3\,\frac{\mu J_2 R_e^2}{2}\,\frac{\mathbf{r}}{r^5}\left(\mathbf{r} + 2\,z - \frac{5z^2}{r^2}\right), \tag{29}$$

$$\mathbf{F}_{SRP} = \frac{\phi_\odot}{c}\,C_R\,\frac{A}{m}\,\hat{\mathbf{S}}\,, \tag{30}$$

$$\mathbf{F}_{drag} = -\frac{1}{2}\,C_D\,\frac{A}{m}\,\rho\,v^2\hat{\mathbf{v}}\,, \tag{31}$$

where μ is the gravitational parameter, R_e is the mean Earth's equatorial radius, (x, y, z) and r are, respectively, the components and the modulus of $\mathbf{r}$, $\phi_\odot$ is the solar radiation flux, c is the velocity of light, C_R is the reflectivity coefficient, A/m is the area-to-mass ratio, $\hat{\mathbf{S}}$ is the direction of the Sun, C_D is the drag coefficient, and ρ is the density of the air atmosphere given by the NRLSISE-00 athmospheric model [41].

4.1.2 Uncertainty Space

The uncertainty space is assumed to be a hypercube. The uncertainty variables are the components of the position and velocity vectors $\mathbf{r}, \mathbf{v}$ and/or four dynamical parameters A/m, C_R, C_D and $F_{10.7}$. The last one represents the daily solar flux for previous days, and it is varied here to model the uncertainty on the air density. The bounds for the dynamical parameters are reported in Table 2. As initial conditions for the state vector, a LEO and HEO orbit have been chosen from the TLE orbit catalog available from the space-track website [42]. The values are reported in Table 3. The uncertainty bounds are set in the Cartesian coordinate space and are assumed to be $10^{-5} \cdot r_0$ and $10^{-5} \cdot v_0$, where r_0 and v_0 are the magnitude of the initial position and velocity vector expressed in km and km/s, respectively.

The propagation time span is set to $40\,P$, where P is the period of the unperturbed orbit. It is to about 4 days for the LEO orbit and 60 days for the HEO orbit. All simulations have been implemented in MATLAB and run on an Intel i7 3.40 GHz.

4.1.3 Experimental Set Up

In this comparative test, PCE were built using Legendre bases and both PCEs and Kriging use a random Latin Hypercube sampling scheme to collect samples. Chebyshev interpolation (spelled the French way, Tchebycheff, in the figures) and the HDMR, called UQ-HDMR, use sparse grids with Clenshaw–Curtis points, instead. By its nature, UQ-HDMR uses a different numbers of samples for each term in the expansion to economise on the total number of samples.

Table 2 Uncertainty bounds for the dynamical parameters

	A/m	C_R	C_D	$F_{10.7}$
Lower bound	0.001	1.0	1.5	100
Upper bound	0.1	2.0	3.0	200

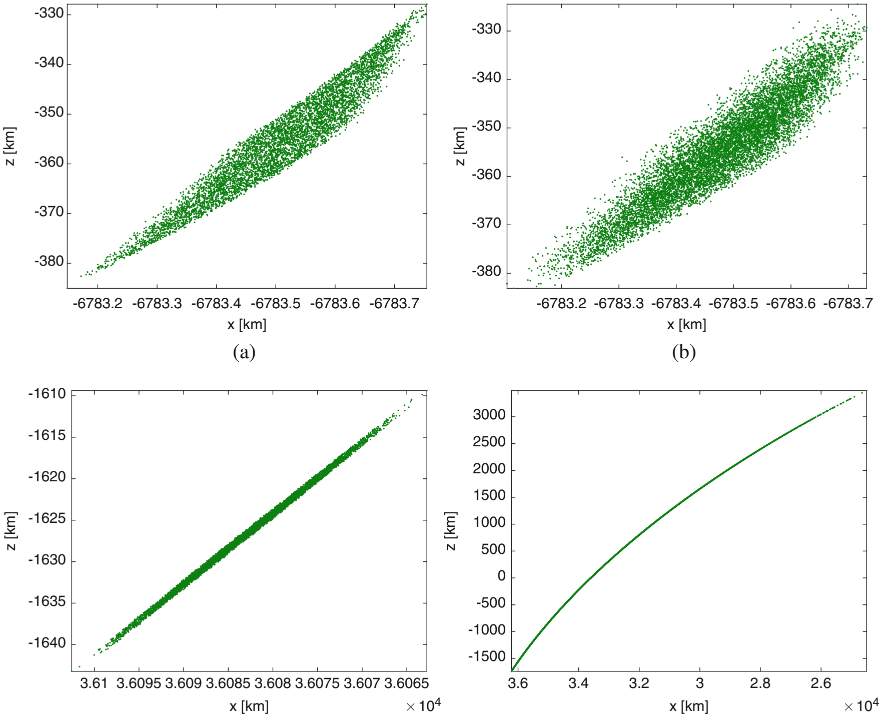
Table 3 Keplerian orbital elements of the LEO and HEO orbit as of May 26, 2015

ID	a [km]	e	i [deg]	Ω [deg]	ω [deg]	ℓ [deg]
40,650	7006.96	0.0008315	98.1533	165.9974	100.2845	259.5405
40,618	24204.56	0.7278988	25.4766	31.5897	179.4183	182.5857

The accuracy of the polynomial computed with each one of the four non-intrusive methods is evaluated at $M = 1000 \times n$ points, where n is the number of uncertainty variables. The M samples are once again generated with a Latin Hypercube sampling scheme. The result is then compared with the true state given by the forward propagation of the dynamics. The estimation of the error between the approximation $\hat{X}$ and the true value X is given by the root mean square error

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{X}_j(\mathbf{x}_i) - X_j(\mathbf{x}_i))^2}\,, \qquad j = 1, \ldots, 6\,, \tag{32}$$

where $\mathbf{x}_i$ represents a single sample vector. Figure 3 shows the uncertainty regions in the 3D space for each scenarios. The effect of the dynamical parameters is to enlarge the uncertainty region for the LEO orbit and stretch it along the trajectory for the HEO. As a result, the dependence of the final state with respect to the initial conditions is highly non-linear (Fig. 3).



(a)

(b)

(c)

(d)

Fig. 3 Uncertainty region of the final state. (**a**) Scenario 1: LEO6. (**b**) Scenario 2: LEO10. (**c**) Scenario 3: HEO6. (**d**) Scenario 4: HEO10

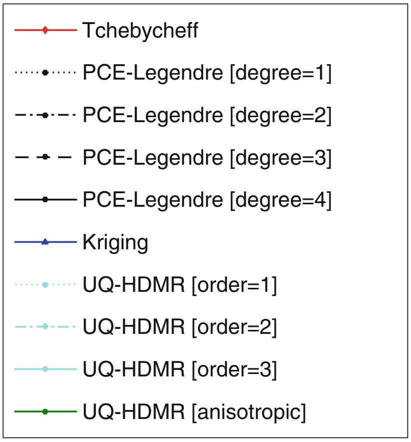


Fig. 4 Legend of the Figs. 5, 6, 7 and 8

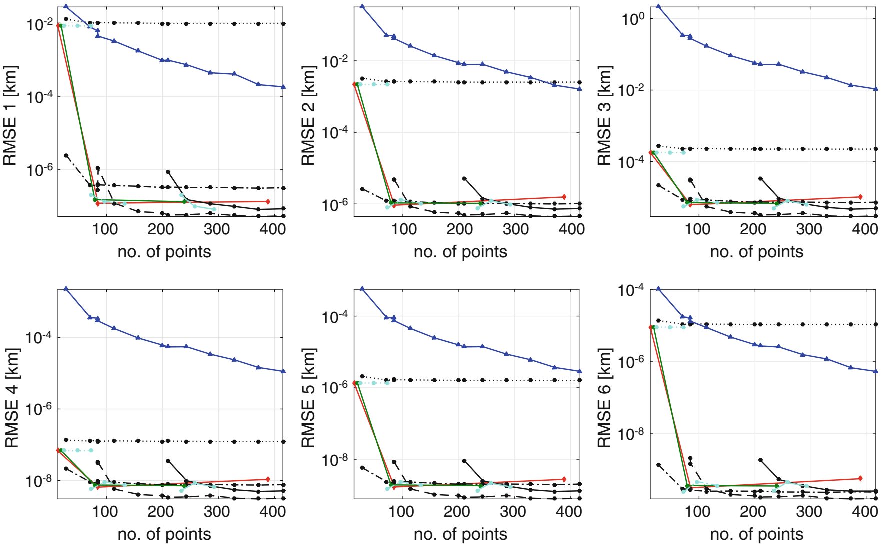


Fig. 5 The RMSE as a function of the number of the sample points for scenario 1 using 6000 test points

The legend in Fig. 4 applies to all the figures. The convergence of the polynomial approximation is presented in Figs. 5, 6, 7 and 8. The estimation of the accuracy is given by the RMSE of each component of the final state vector, computed with the Monte Carlo outcomes, as a function of the number of samples used to build the polynomial approximation. In all the examples, Kriging exhibits the slowest convergence, i.e. for a fixed number of samples it has the highest value for the RMSE. The uncertainty region of scenario 1 (LEO6) can be approximated with

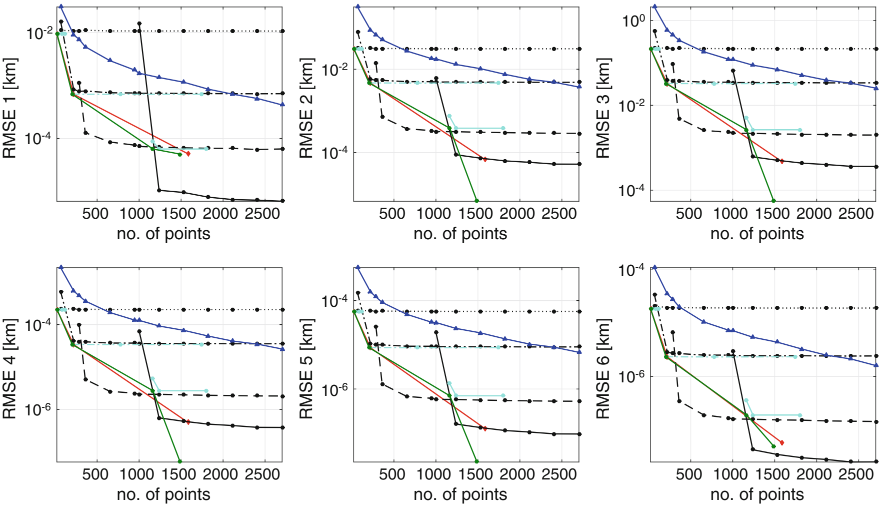


Fig. 6 The same as Fig. 5 applied to scenario 2 with 10,000 test points

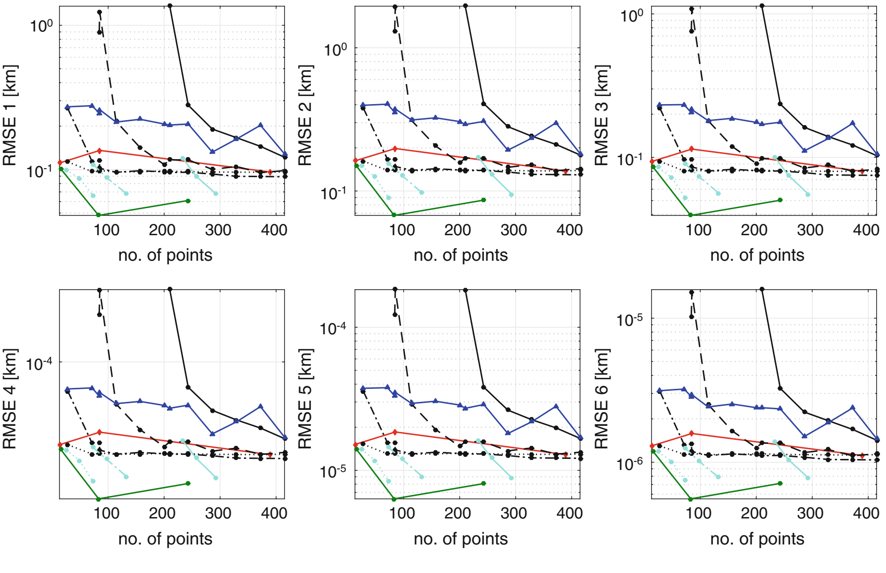


Fig. 7 The same as Fig. 5 applied to the HEO orbit in scenario 3 with 6,000 test points

less than 100 sample points and a maximum RMSE of 10^{-5} km by all methods with the exception of Kriging. The best accuracy is achieved with a PCE-Legendre of degree 3. Chebyshev and UQ-HDMR show equal performance (see Fig. 5). When additional four dynamical parameters become uncertain (scenario 2), the resulting functional dependency between quantity of interest and input parameters becomes

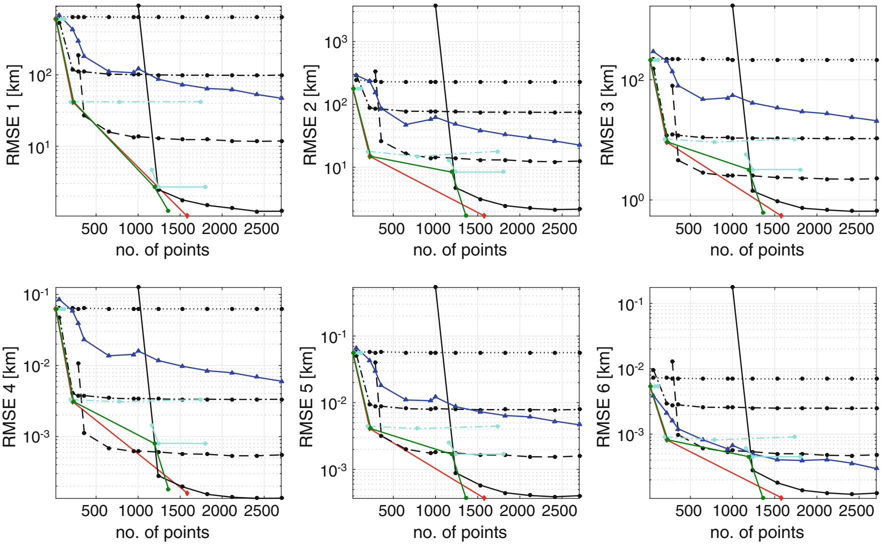


Fig. 8 The same as Fig. 6 applied to the HEO orbit in scenario 4 with 10,000 test points

highly non-linear. In order to obtain an accuracy of 10^{-3} km, a PCE-Legendre of degree equal to 4 or a Chebyshev sparse basis of level 3 need to be used (see Fig. 6).

The results for scenario 3 (HEO6) are shown in Fig. 7. A PCE-Legendre of degree 1 dominates higher order PCE-Legendre and all the Chebyshev approximations and Kriging. However, the best approximation is given by the UQ-HDMR. Figure 8 presents the analysis for scenario 4 (HEO10). As for scenario 2, non-linearities are rather important and Chebyshev and UQ-HDMR show comparable results.

Finally, Tables 4, 5, 6 and 7 report the number of samples to achieve and accuracy $\max(\text{RMSE}) < 4D \cdot 10^{-4}$, where $\max(\text{RMSE})$ is the maximum error across all components of the final state vector and D is the diameter of the projection of the uncertainty region of the final state on the (x, z)-plane.

Table 4 Summary of the comparison for scenario 1 with a reference accuracy of 0.229 km

Method	No. of sample points	max (RMSE)
Chebyshev	13	0.00833
PCE-Legendre	28	0.01262
UQ-HDMR	19	0.00833
Kriging	144	0.16753

Table 5 Summary of the comparison for scenario 2 with a reference accuracy of 0.235 km

Method	No. of sample points	max (RMSE)
Chebyshev	21	0.21433
PCE-Legendre	66	0.22553
UQ-HDMR	23	0.21437
Kriging	652	0.18307

Table 6 Summary of the comparison for scenario 3 with a reference accuracy of 0.209 km

Method	No. of sample points	max (RMSE)
Chebyshev	13	0.16224
PCE-Legendre	28	0.16147
UQ-HDMR	15	0.15068
Kriging	286	0.19265

Table 7 Summary of the comparison for scenario 4 with a reference accuracy of 51.317 km

Method	No. of sample points	max (RMSE)
Chebyshev	221	41.970
PCE-Legendre	359	26.658
UQ-HDMR	233	41.976
Kriging	2703	41.044

4.2 Representation with Positive Polynomials

Positive polynomials, like Bernstein polynomials for example, have been used to represent generic distributions. Bernstein polynomials in particular can approximate any generic distribution on a finite supports and represent exactly Beta distributions. Their use in orbital mechanics was recently introduced by the author [43] to calculate upper and lower expectations of the quantity of interest by solving a simple linear optimisation programme with a single linear constraint. In the general case the integrals in Eq. (8) can calculated numerically via multidimensional quadrature formula. As an example we can replace the calculation of the exact integrals with an approximation using Halton low discrepancy sequence to generate M sample points (called quasi-Monte Carlo points) in the domain $\mathfrak{U}_0$ and then re-write the integrals in the form:

$$\int_{\Omega} \phi(\boldsymbol{\xi})\, d\xi \approx \frac{1}{M} \sum_{k=1}^{M} I_{\Omega}(\xi_k)\, \phi(\xi_k) \tag{33}$$

where the samples ξ_k are taken from the low discrepancy sequence. Similarly, we can approximate the integrals in Eq. (8):

$$\min_{\alpha,\beta} \sum_{k=1}^{M} I_{\Omega}(\xi_k) \prod_{j} \phi_j(\xi_k)\,, \qquad \max_{\alpha,\beta} \sum_{k=1}^{M} I_{\Omega}(\xi_k) \prod_{j} \phi_j(\xi_k)\,. \tag{34}$$

subject to the constraint:

$$\frac{1}{M} \sum_{k=1}^{M} \phi(\xi_k) = 1. \tag{35}$$

If the family of distributions is unknown or does not contain only one particular type, one can use an a representation with an expansion in positive polynomials to approximate the extrema of $[\phi]$ and obtain the upper and lower expectation on Ω as solutions of a linear problem. In this chapter, in particular, we propose the use of Bernstein polynomials [44, 45]. The family of probability distributions to which the uncertain variable ξ_j belongs can be expressed as

$$[\phi_{c_j}] = \left\{ \sum_{i=1}^{n} c_i^{(j)} \, B_i(\tau_j(\xi_j)) \right\}, \tag{36}$$

where $B_i : [0, 1] \mapsto [0, 1]$ is the i^{th}-univariate Bernstein polynomials of dimension n and τ_j is the change of coordinate from the uncertain interval $[\xi_j]$ to $[0, 1]$.

Under the independence and non-correlation assumption among the variables, the joint probability distribution is the product of the marginal masses and it is contained in the p-box $[\phi_{\tilde{c}}] = \prod_{j=1}^{d} [\phi_{c_j}]$ which can be re-written as

$$[\phi_c] = \left\{ \sum_{\kappa \in \mathscr{K}} c_\kappa \, \mathscr{B}_\kappa(\tau(\xi)) \right\}, \tag{37}$$

with $\mathscr{K} = \{\kappa = (k_1, \ldots, k_d) \in \mathbb{N}^d : 0 \leq k_j \leq n, \forall j\}$, $\mathscr{B}_\kappa$ is a multivariate Bernstein polynomial, $\tau = \prod_{j=1}^{d} \tau_j$, and c is the unknown coefficient vector. Then, the upper and lower expectation are the solutions of the two linear optimization problems:

$$E_l(\Omega) = \min_{c \in \mathscr{C}} \int_\Omega \phi_c(\xi) \, \mathrm{d}\xi \, , \qquad E_u(\Omega) = \max_{c \in \mathscr{C}} \int_\Omega \phi_c(\xi) \, \mathrm{d}\xi \, , \tag{38}$$

The set $\mathscr{C} \in \mathbb{R}^M$ can be assumed to be an hyper-cube, for example, $\mathscr{C} = [0, M]^M$. In discrete form programmes (38) translate into:

$$E_l(\Omega) = \min_{c \in \mathscr{C}} \sum_{s=1}^{M} I_\Omega(\xi_s) \sum_{\kappa \in \mathscr{K}} c_\kappa \, \mathscr{B}_\kappa(\tau(\xi_s)), \tag{39}$$

and

$$E_u(\Omega) = \max_{c \in \mathscr{C}} \sum_{s=1}^{M} I_\Omega(\xi_s) \sum_{\kappa \in \mathscr{K}} c_\kappa \, \mathscr{B}_\kappa(\tau(\xi_s)) \, . \tag{40}$$

subject to the linear constraint:

$$\frac{1}{M} \sum_{s=1}^{M} \sum_{\kappa \in \mathscr{K}} c_\kappa \, \mathscr{B}_\kappa(\tau(\xi_s)) = 1. \tag{41}$$

This technique is very efficient in low dimension but with Bernstein polynomials the number of coefficients increases exponentially with the number of uncertain parameters and can quickly lead to a very large constrained linear programming problem. An alternative is to solve the nonlinear problem:

$$E_u(\Omega) = \max_{c \in \mathscr{C}} \sum_{s=1}^{M} I_\Omega(\xi_s) \prod_j \sum_i c_i B_i(\tau(\xi_s)), \tag{42}$$

subject to the linear constraint:

$$\frac{1}{M} \sum_{s=1}^{M} \prod_j \sum_i c_i \; B_i(\tau(\xi_s)) = 1. \tag{43}$$

In this case the number of coefficients grows linearly with the number of dimensions and the optimisation problem remains tractable even for a large number of uncertain parameters.

5 Intrusive Techniques

Intrusive techniques cannot treat computer codes as a black box. They require full access to the mathematical model and computer code computing the quantity of interest and introduce a modification of the code and model. The goal of intrusive techniques is still to provide a surrogate representation of the variation of the quantity of interest as a function of the uncertainty in model, parameters and boundary conditions. Most existing methods are used to propagate the uncertainty space Ω.

The main advantage of intrusive methods lays in the better control of the truncation error versus the complexity of the polynomial expansion. They also automatically provide a polynomial representation of the propagated set at every propagation step.

- **State Transition Matrix and Tensors (STM/STT)** The State Transition Matrix is the most traditional approach and requires the expansion of the dynamics only to the first order (see system (2)). For this reason the STM cannot properly capture nonlinearities of higher order. In order to overcome this limitation the use of high order State Transition Tensors were proposed in 2006 by Park and Scheeres [6] and Tapley et al. [46] .

  This section briefly reviews the approach proposed by Park and Scheeres [6] to propagate uncertainty in dynamical systems and highlights some key properties through a simple example. The method expands the variation $\delta x(t)$ of the states at time t with respect to a reference point $\phi(t, x_0; t_0)$ in Taylor series of some initial deviation δx_0. The s-th order expansion can be expressed using the Einstein summation convention:

$$\delta x^i(t) = \sum_p^s \frac{1}{p!} \phi_{(t,t_0)}^{i,\gamma_1 \ldots \gamma_p} \delta x_0^{\gamma_1} \cdots \delta x_0^{\gamma_p} \tag{44}$$

where $\gamma_1 .. \gamma_p \in \{1, \ldots., n\}$denotes the γ_i component of the state vector corresponding to the s-th derivative, n is the number of components of the state vector and:

$$\phi_{(t,t_0)}^{i,\gamma_1 \ldots \gamma_p}(t; x_0; t_0) = \left. \frac{\partial^p \phi_{(t,t_0)}^i(t; \xi_0; t_0)}{\partial \xi_0^{\gamma_1} \cdots \partial \xi_0^{\gamma_p}} \right|_{\xi_0^{\gamma_j} = \mathrm{x}_0^{\gamma_j}} . \tag{45}$$

In this way, a generic trajectory x, whose initial conditions are defined with respect to the reference trajectory as $x_0 + \delta x_0$, will evolve as follows:

$$x^i(t) = x_0^i(t) + \sum_p^s \frac{1}{p!} \phi_{(t,t_0)}^{i,\gamma_1 \ldots \gamma_p} \delta x_0^{\gamma_1} \cdots \delta x_0^{\gamma_p} . \tag{46}$$

The partials of the flow in Eq. (44) form the so called global State Transition Tensors, which map the initial deviations δx_0 at time t_0 to the deviation $\delta x(t)$ at time t. For $s = 1$, the STTs reduces to the simple state transition matrix. The partials in Eq. (45) can be computed by numerical integration of a set of ordinary differential equations (see [6]). An example of these differential equations up to the third order follows:

$$\dot{\phi}^{i,a} = f^{i,\alpha} \phi^{\alpha,a} \tag{47}$$

$$\dot{\phi}^{i,ab} = f^{i,\alpha} \phi^{\alpha,ab} + f^{i,\alpha\beta} \phi^{\alpha,a} \phi^{\beta,b} \tag{48}$$

$$\dot{\phi}^{i,abc} = f^{i,\alpha} \phi^{\alpha,abc} + f^{i,\alpha\beta} \left(\phi^{\alpha,a} \phi^{\beta,bc} + \phi^{\alpha,ab} \phi^{\beta,c} + \phi^{\alpha,ac} \phi^{\beta,b} \right) + f^{i,\alpha\beta\delta} \phi^{\alpha,a} \phi^{\beta,b} \phi^{\delta,c} \tag{49}$$

where $\alpha, \beta, \in \delta \{1, \ldots., n\}$and $a, b, c = \{1, \ldots, n\}$are the indexes for the first, second and third order derivative. $f^{i,\gamma_1 \cdots \gamma_p}$ are the partials of the dynamics and are computed as follows:

$$f^{i,\gamma_1 \ldots \gamma_p} = \left. \frac{\partial^p f^i(t; \xi_0; t_0)}{\partial \xi_0^{\gamma_1} \cdots \partial \xi_0^{\gamma_p}} \right|_{\xi_0^{\gamma_j} = \mathrm{x}_0^{\gamma_j}} \tag{50}$$

Note that the partial derivatives in Eqs. (45) and (50) are calculated with respect to the nominal trajectory $\phi(t, x_0; t_0)$ (also equivalent to $x_0(t)$ of Eq. (46)).

If the partials in Eq. (45) are obtained by numerical integration, the calculation of the STTs requires the forward propagation of $\sum_{q=1}^{s+1} 6^q$ differential equations starting with initial values $\phi^{i,a}_{(t_0,t_0)} = 1$, if $i = a$, and zero otherwise. When the order is $s = 3$, the 1554 equations need to be integrated simultaneously. Moreover, the computational time and complexity are increased by the numerical evaluations of the analytical partials of the dynamics. In Vetrisano and Vasile [22], the partials in Eq. (45) were computed analytically using the symbolic manipulator in the MATLAB$^{\text{RM}}$ Symbolic Toolbox. As an example, the third order STTs integration, along a 5 day period, considering only Earth, Moon, Sun and light pressure, required approximately 8 h using a Windows 7 OS 3.16 GHz Intel$^{(R)}$Core$^{(TM)}$2 Duo CPU. To be noted that the coupled integration of thousands of equations could introduce numerical errors when integrated over a long period of time. For this reason, it is good practice to consider the nominal trajectory and to integrate the STTs over short periods of time, say 1 day to reduce possible numerical errors [22]. The intermediate STTs are called local STTs. While the global STTs map the deviation at the initial time t_0 to the deviation at time t_{k+1}, the local STTs map the deviation at time $t_k$to the deviation at time t_{k+1}.

Once the state transition tensors are available for some time interval $[t_k, t_{k+1}]$, the mean and covariance matrix of the relative dynamics at t_k can be mapped analytically to t_{k+1} as a function of the probability distribution at t_k. In the remainder of this paper we will make use of the mean and covariance to compare different methods, therefore, here we briefly summarise the procedure proposed in [6]. From t_k to t_{k+1} the propagated mean and covariance can be computed as:

$$m^i_{k+1} = \phi^i(t_{k+1}; m_k) + \delta m^i_{k+1} = \phi^i(t_{k+1}; m_k) + \sum_{p=1}^{s} \frac{1}{p!} \phi^{i.\gamma_1\cdots\gamma_p}_{(t_{k+1},t_k)} E[\delta x_k^{\gamma_1} \cdots \delta x_k^{\gamma_p}] \tag{51}$$

$$\begin{aligned} P^{ij}_{k+1} &= E[(\delta x^i_{k+1} - \delta m^i_{k+1})(\delta x^j_{k+1} - \delta m^j_{k+1})] = \\ &\textstyle\sum_{p=1}^{s} \sum_{q=1}^{s} \frac{1}{p!q!} \phi^{i.\gamma_1\cdots\gamma_p}_{(t_{k+1},t_k)} \phi^{i.\varsigma_1\cdots\varsigma_q}_{(t_{k+1},t_k)} E[\delta x_k^{\gamma_1} \cdots \delta x_k^{\gamma_p} \delta x_k^{\varsigma_1} \cdots \delta x_k^{\varsigma_q}] - \delta m^i_{k+1} \delta m^j_{k+1} \end{aligned} \tag{52}$$

where $\{\gamma_i, \varsigma_j\} \in \{1, \ldots, n\}$ are the indexes for the different order derivative. If one sticks to the hypothesis of an initial Gaussian distribution, the joint characteristic function for a Gaussian random vector can be defined as [6]

$$\vartheta(u) = E[e^{ju^T x}] = \exp(ju^T m - \frac{1}{2} u^T P u) \tag{53}$$

where$j = \sqrt{-1}$ and the expected higher moments can be computed using:

$$E[x^{\gamma_1} x^{\gamma_2} \cdots x^{\gamma_p}] = j^{-p} \left. \frac{\partial^p \vartheta(u)}{\partial u^{\gamma_1} \partial u^{\gamma_2} \cdots \partial u^{\gamma_p}} \right|_{u=0} \tag{54}$$

- **Intrusive Polynomial Chaos Expansion (PCE)** [25] Intrusive PCEs are probably the first version of chaos expansions and date back to the work of Ganhem in 1988, Ganhem and Spanos 1991 end Xiu and Karniadakis in 2002 that extended the expansion from Hermit polynomials to the general Askey scheme. The idea behind intrusive PCEs is the same as the one of non-intrusive but the chaos expansion of the input parameters is introduced in the governing equation (3) and the quantity of interest is obtained by integration one differential equation per coefficient of the expansion.
- **Interval Arithmetic (IA)** If the interest is to propagate sets of values an option is to propagate intervals. In this case, it is possible to define an Algebra on the space of intervals [47] $\mathscr{I} := \{[a, b],\ a \leq b,\ a, b \in \mathbb{R}\}$, such that

  $$C := A \otimes B = \{a \oplus b \,|\, a \in A, b \in B\} \in \mathscr{I}$$

  where $A, B \in \mathscr{I}$, $\oplus \in \{+, -, \cdot, /\}$ and $\otimes$ is then the corresponding operation in the algebra of intervals. If the propagation is performed in the algebra of intervals it is possible to compute validated solutions of ODE systems. The tight enclosure of the solution of the dynamics at a certain instant of time, is computed taking into account truncation errors due to floating point implementation, errors due to approximation integration scheme, and parameters uncertainties. However this approach could lead to overestimation, depending on the problem and the method used. An hybridization of Polynomial Algebra (see next section) and IA techniques has given promising results in mitigating the overestimation problem. These are known as Taylor Models [48].

5.1 Polynomial Algebra (PA)

The idea is to redefine states and parameters as polynomials of the uncertain quantities and all algebraic operations between real numbers as algebraic operations among polynomial functions.

The function space $\mathscr{P}_{n,d}(\alpha) =< \alpha_{\mathbf{I}}(\mathbf{b}) >$ where $\mathbf{b} \in \Omega \subset \mathbb{R}^d$, $\mathbf{I} = (i_1, \ldots, i_d) \in \mathbb{N}^d_+$ and $|\mathbf{I}| = \sum_{j=1}^{d} i_j \leq n$, is the space of polynomials in the α basis up to degree n in d variables [49]. This space can be equipped with a set of elementary arithmetic operations, generating an algebra on the space of polynomials such that, given two elements $A(\mathbf{b}),\ B(\mathbf{b}) \in \mathscr{P}_{n,d}(\alpha)$ approximating any two real multivariate functions $f_A(\mathbf{b})$ and $f_B(\mathbf{b})$, it stands that

$$f_A(\mathbf{b}) \oplus f_B(\mathbf{b}) \sim A(\mathbf{b}) \otimes B(\mathbf{b})\,, \tag{55}$$

where $\oplus \in \{+, -, \cdot, /\}$ and $\otimes$ is the corresponding operation in $\mathscr{P}_{n,d}(\alpha_i)$. This allows one to define the algebra $(\mathscr{P}_{n,d}(\alpha_i), \otimes)$, of dimension $\dim(\mathscr{P}_{n,d}(\alpha_i), \otimes) = \mathscr{N}_{d,n} = \binom{n+d}{d}$, the elements of which belong to the polynomial ring in d

indeterminates $\mathbb{R}[\mathbf{b}]$ and have degree up to n. Each element $P(\mathbf{b})$ of the algebra, is uniquely identified by the set of its coefficients $\mathbf{p} \in \mathbb{R}^{\mathscr{N}_{d,n}}$ such that

$$P(\mathbf{b}) = \sum_{\mathbf{I},|\mathbf{I}|\leq n} p_{\mathbf{I}}\alpha_{\mathbf{I}}(\mathbf{b}) \ . \tag{56}$$

In the same way as for arithmetic operations, it is possible to define a composition rule in the polynomial algebra and hence the counterpart, in the algebra, of the elementary functions $\{\sin(y), \cos(y), \exp(y), \log(y), \ldots\}$. Differentiation and integration operators can also be defined. By defining the initial conditions and model parameters of the dynamics as element of the algebra and by applying any integration scheme with operations defined in the algebra, at each integration step is available the polynomial representation of the state flow. The main advantage of the method is in the control of the trade-off between computational complexity and representation accuracy at each step of integration. Furthermore, sampling and propagation are decoupled, therefore, irregular regions can be propagated with a single integration, provided that a polynomial expression is available. It has been shown that the polynomial algebra approach presents overall good performance and scalability (with respect to the size of the algebra) compared to its non-intrusive counterpart. On the other hand, being an intrusive method, it cannot treat the dynamics as black box. Its implementation requires operator overloading for all the algebraic operations and elementary functions defining the dynamics, making it more difficult to implement than a non-intrusive method. There are currently two different polynomial representations that have been successfully used in Orbital Mechanics: Taylor polynomials (known in the literature as Differential Algebra [50] or Jet Transports [51]) and Chebyshev polynomials.

- **Taylor Algebra**. In Taylor Algebra (TA) all quantities are expanded in Taylor series and all algebraic operations are defined among Taylor polynomials. One advantage of this approach is that the truncated product of two Taylor polynomials is again a Taylor polynomial. This means that one can control the number of terms in the expansion retaining all properties of Taylor polynomials. The product and other algebraic operations requires only to apply the operator among Taylor bases, leading to fast execution of most computation when propagating the uncertainty set.

  Taylor Algebra, as for STM and STT, however, provide only a local model that is centred into a reference point. This means that the approximation error is not globally minimised over a region but tends to increase as one departs from the central point. Taylor series have also other undesirable properties, for example they can converge to a function that is not the one they are trying to represent.
- **Chebyshev Polynomials and Generalised Algebra**. The use of Chebyshev polynomials instead of Taylor expansions provides a better global accuracy because of the min-max properties of Chebyshev polynomial approximation. It also allows one to develop an approximation of the uncertainty region without any particular central point of expansion. This is particularly useful when the

interest is the quantification of the probability of inclusion in a particular set but with no assumption on the nature of the distribution.

On the other hand an algebra on the space of Chebyshev polynomials is not as straightforward because, for example, the product of two Chebyshev bases is not a Chebyshev basis. The current implementation of an algebra using Chebyshev polynomials transforms all polynomial expansions into monomials and then defines an algebra on the space of the monomials [52]. This approach has proven to be very effective and allows one to generalise polynomial algebra to any type of polynomials. The computational overhead of this transformation is limited and one can partially preserve the property of the original polynomial expansion even after the transformation to monomials.

5.1.1 Example: Orbit Re-Entry Under Uncertainty

Figure 9 shows an example of propagation of the uncertainty in the initial conditions of a satellite orbiting the Earth. The figure compares a full Monte Carlo simulation against a single integration of the equations of motion with a Taylor-based algebra (i.e algebra based on an expansion in Taylor polynomials) or with a Chebyshev-based algebra (i.e algebra based on an expansion in Chebyshev polynomials). The figure shows that Taylor diverges as one departs from the estimated state of the

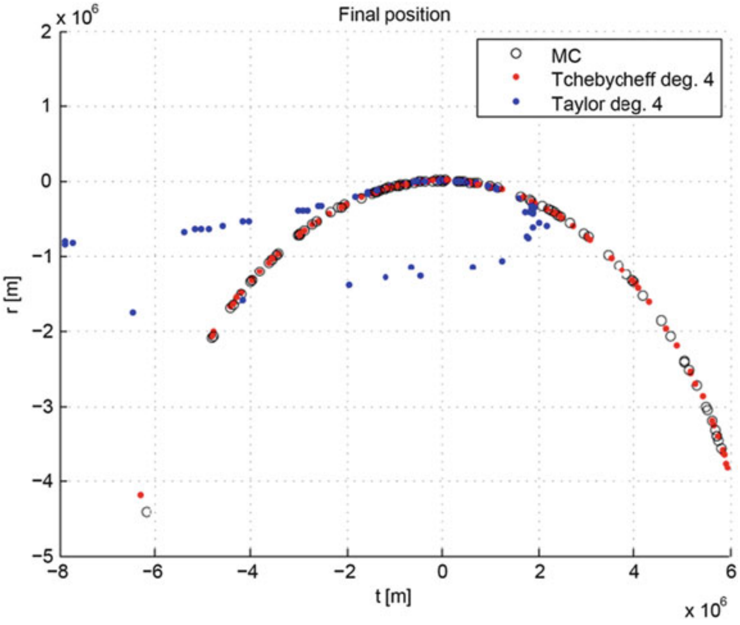


Fig. 9 Propagated uncertainty with Taylor and Chebyshev polynomial algebra

satellite. On the contrary, in this case, Chebyshev offers a more stable global representation of the uncertainty space (see [52] for more details).

6 Handling of Model Uncertainty

Model uncertainty is difficult to quantify because it requires capturing missing parts of the model itself. The nature of this uncertainty is epistemic. In this section two ideas are presented that can be used to capture model uncertainty.

With reference to what is commonly done in precise orbit determination one way to capture the missing components of the dynamics is to introduce so called empirical accelerations in the form of a time dependent polynomial or other time dependent functional forms. This form of data assimilation is generally effective but does not provide the actual dependency of the missing parts of the dynamics on the state of the system.

6.1 Reconstruction via Polynomial Expansions

Consider problem (3) with only the additive term:

$$\begin{cases} \dot{\mathbf{x}} & = f(\mathbf{x}, \mathbf{p}) + \nu(\mathbf{x}, \mathbf{b}) \\ \mathbf{x}(t_0) & = \mathbf{x_0} \end{cases}, \tag{57}$$

and assume that the function ν can be expanded in some form of polynomial series of the state vector $\mathbf{x}$:

$$\begin{cases} \dot{\mathbf{x}} & = f(\mathbf{x}, \mathbf{p}) + \sum_i c_i(\mathbf{b}) Q_i(\mathbf{x}) \\ \mathbf{x}(t_0) & = \mathbf{x_0} \end{cases}, \tag{58}$$

with unknown coefficients c_i. The idea is now to determine the value of the coefficients by matching the observations of the state vector $\bar{\mathbf{x}}(t_j)$ at given times t_j with the result of the propagation of model (58) at the same time instants. We can then solve the following optimisation problem:

$$\begin{array}{l} \min\limits_{c \in \mathscr{C}} J(\mathbf{x}, c) \\ s.t. \\ \mathbf{x}(t_j) \in \Sigma \quad j = 0, \ldots, N_o \end{array}, \tag{59}$$

where Σ is an arbitrary set, N_o is the total number of observations and $\mathscr{C}$ is the space of the coefficients c_i. The main advantage of this formulation is that no statistical moments are required and no exact distribution needs to be known a priori. Note that the initial conditions $\mathbf{x}(t_0)$ are treated as an observed state. The second example is an orbital motion with unknown drag component. The gravity component of the model is fully known but the observations show an additional component that is not modelled. The real dynamics is assumed to be governed by the following system of differential equations in polar coordinates:

$$\begin{aligned}
\dot{v}_r &= -\tfrac{\mu}{r^2} + \tfrac{v_t^2}{r} - \tfrac{1}{2}\rho C_d v v_r \\
\dot{v}_t &= -\tfrac{v_t v_r}{r} - \tfrac{1}{2}\rho C_d v v_t \\
\dot{r} &= v_r \\
\dot{\theta} &= \tfrac{v_t}{r}
\end{aligned} \quad . \tag{60}$$

We assume a unitary area to mass ratio, and a constant density ρ such that the product of the density times the drag coefficient C_d is $\rho C_d = 10^{-6}\text{kg/m}^3$. Furthermore, we assume that the expected trajectory, given the known dynamic components, is a circular orbit with $v_r(t = 0) = v_{r_0} = 0$ and $v_t(t = 0) = v_{t_0}$. The orbital period, without drag, is $T = 2\pi\sqrt{r^3/\mu}$. If one expands the modulus of the velocity v in Taylor series up to the first order, the differential equations with the drag term can be approximated as:

$$\begin{aligned}
\dot{v}_r &= -\tfrac{\mu}{r^2} + \tfrac{v_t^2}{r} - \tfrac{1}{2}\rho C_d v_t v_r \\
\dot{v}_t &= -\tfrac{v_t v_r}{r} - \tfrac{1}{2}\rho C_d v_t^2 \\
\dot{r} &= v_r \\
\dot{\theta} &= \tfrac{v_t}{r}
\end{aligned} \tag{61}$$

In order to capture the unmodelled component of the dynamics, we assume the following expansion with terms up to order 2 in velocity and position:

$$\begin{aligned}
\dot{v}_r &= -\tfrac{\mu}{r^2} + \tfrac{v_t^2}{r} + c_1 + c_3 r + c_5 r^2 + \\
&\qquad c_7 r\theta + c_9 v_r + c_{11} v_r^2 + c_{13} v_r v_t \\
\dot{v}_t &= -\tfrac{v_t v_r}{r} + c_2 + c_4\theta + c_6\theta^2 + \\
&\qquad c_8 r\theta + c_{10} v_t + c_{12} v_t^2 + c_{14} v_r v_t \\
\dot{r} &= v_r \\
\dot{\theta} &= \tfrac{v_t}{r}
\end{aligned} \tag{62}$$

If the linear effects in Eq. (61) are dominant over a given time span Δt, then the prediction given by Eq. (62) should be of the form:

$$\begin{aligned}
\dot{v}_r &= -\tfrac{\mu}{r^2} + \tfrac{v_t^2}{r} + c_{13} v_r v_t \\
\dot{v}_t &= -\tfrac{v_t v_r}{r} + c_{12} v_t^2 \\
\dot{r} &= v_r \\
\dot{\theta} &= \tfrac{v_t}{r}
\end{aligned} \tag{63}$$

We can now introduce observations at time $t = T$ and $t = T/2$, for a total of 8 constraint equations and 14 parameters, and solve problem (59) with cost function $J = c^T c$. This cost function implies that we look for the minimum energy solution under the assumption that this solution corresponds to a minimum noise state. An alternative, not presented here, is to use a maximum Entropy principle and maximise, for example, the Shannon entropy function of the coefficients **c**.

If measurements are affected by an error, problem (59) needs to be solved under some assumptions on the initial conditions. The assumption in this chapter is that the initial conditions are distributed uniformity over a given confidence interval. The size of the confidence interval for the measurements is 10^{-4} of the measured value; accordingly the confidence interval on the initial conditions is set to the same value.

The parameters c estimated by solving problem (59) are represented in Fig. 10 together with their associated confidence intervals. As one can see, the expected value is close to the true solution. One thing that has to be taken into consideration is that the dynamics that are simulated and measured are the true dynamics, not the linearised equations. Therefore, some components that are not in the linear model might be different from zero.

The other interesting result is that some components are nearly zero for every initial condition while other components, c_4 for example, have a wide variability. This result suggests that some components are irrelevant as they do not contradict the observations no matter which initial conditions are taken, while others substantially affect the evolution of the trajectory. Starting from this first iteration, one can then update the confidence intervals on the parameters c and eventually converge

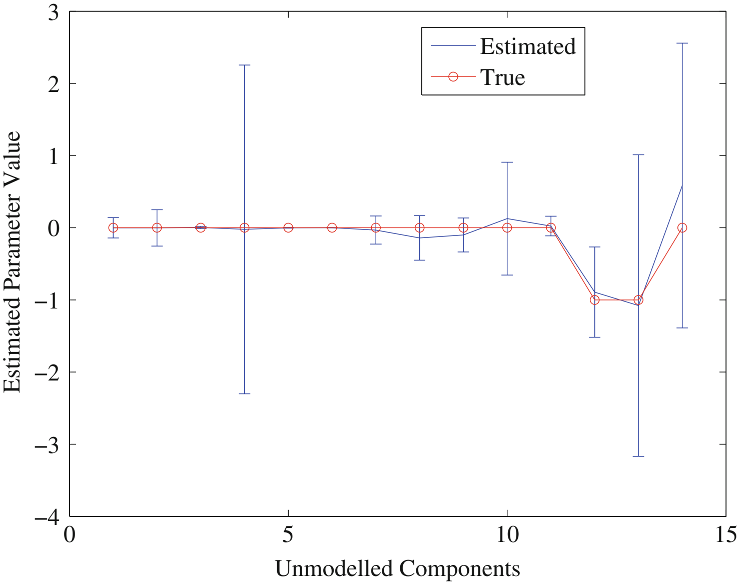


Fig. 10 Example of reconstructed gravity-drag dynamics with confidence intervals

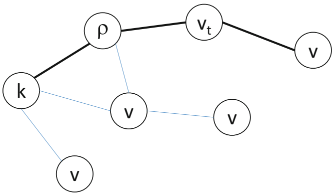


Fig. 11 Example of decision tree for the simple re-entry case with unknown drag

to the correct missing components. Indeed since the uncertain function is based on a truncated series some components of the expansion might absorb the truncation error.

6.2 Reconstruction via Symbolic Regression

The idea of symbolic regression is to use a decision tree, or tree in which each node is a decision, to construct an analytical formula, of the some unknown variables, that once evaluated produces outputs that match the values in a dataset.

The regression process starts from a database of symbols that includes some basic algebraic operations ($+, -, *$) and a number of elementary functions of the variables of interest. The tree is then progressively grown by adding new symbols to different branches. Each branch represents a partial formula. The regression process then evaluates a branch (for example adds the sequence of symbols to a differential equation and then integrates the equation) and compares the result to a dataset using a suitable metric, typically an Euclidean distance. Figure 11 is an example of a simple decision tree for the reconstruction of the drag component in (60). In this case the only algebraic operator is the product and the symbols to be identified are only functions of the state variables. The path with thick black lines is the right solution. What the algorithm does is to evaluate a branch, every time a node is added to it, by propagating the dynamics and comparing the result of the propagation at a given time with some reference measurements. Typically Genetic Programming [53] is used to generate and evaluate the tree, although in recent times the author experimented with Ant Colony Optimisation [54] as well, with good results.

Recent work on the use of symbolic regression to capture the relationship between re-entry time and system uncertainty can be found in Minisci et al. [55].

7 Evidence-Based Quantification

An interesting aspect of modelling uncertainty in orbital mechanics is that the nature of this uncertainty is epistemic more often than not. This realisation has important consequences on the significance of the quantification of the uncertainty, for example, in the prediction of a collision or impact.

One can re-state the hypothesis that a conjunction or a collision occurs in the form of an inclusion statement [56]:

$$A = \{\boldsymbol{\xi}(t)|\mathbf{x}(t) \in \Phi\}, \tag{64}$$

where Φ defines either a region around a target within which a collision avoidance is triggered or its complement.

The interest is now to calculate the degree of belief associated to statement (64). If one uses belief functions [57] this can be computed as follows:

$$Bel(A) = \sum_{\theta_i \in A} m(\theta_i), \tag{65}$$

where θ_i is a piece of evidence supporting A and $m(\theta_i)$ is a belief mass associated to θ_i. Or alternatively one can calculate the degree of belief associated to the complement of A:

$$Pl(\neg A) = \sum_{\theta_i \cap A \neq \emptyset} m(\theta_i). \tag{66}$$

The use of belief functions is not the only choice of course. One can use rough sets, fuzzy sets or other methods that allow the treatment of partial knowledge.

The inclusion statement implies also that the common likelihood function that is normally used to relate the measurements to the propagated state in a classical Bayesian framework might need to be reconsidered because the hypothesis on normally distributed measurements might not apply in general.

8 Final Remarks

The chapter provided a broad overview of a number of methods for uncertainty quantification in orbital mechanics. An attempt was made to classify them according to the context in which they can be applied. Most of the chapter was dedicated to the propagation of uncertainty because that is the area where a lot of work has been developed in recent times. An attempt was made to include only methods that were not specialised or specific to a particular class of problems or were dependent on the characteristics of the problem. Thus, specific sampling and UQ techniques developed in orbit determination or impact monitoring, like the use of the line of variations, were not included. The interested reader is advised to read the relevant literature on the subject.

The choice of the propagation method is closely dependent on the uncertainty model and on the nature of the uncertainty to be propagated. In this respect, an open problem is the representation of epistemic uncertainty in measurements and physical model. The two sources of uncertainty result to be interdependent when one attempts to derive an improved model representation as the uncertainty in the measurements makes a number of physical models all equally possible. It is the opinion of the author that the problem with epistemic uncertainty is often overlooked and deserves more attention if one to handle complex systems, like large constellations, swarms of debris. high risk rare events and anomalies.

Most recent developments for uncertainty propagation in orbital mechanics have advantages and disadvantages. Non-intrusive methods are ideal when the dynamic model is not well known or is a black box. They are also the ideal solution when a mix a of experimental and simulated data are available. On the other hand intrusive methods provide a very interesting alternative if one can have access to the dynamic equations. Although Taylor algebra has been extensively used in practice, it is the opinion of the author, that a lot still needs to be done, both theoretically and algorithmically, on the use of intrusive methods for orbital mechanics.

Last but not least the chapter has shown that the use of some machine learning techniques, like symbolic or polynomial regressions, can be powerful tools to capture unknown dynamic components and reconstruct physical models from experimental data.

Acknowledgements The work in this chapter was partially supported by the FP7-PEOPLE-2012-ITN Stardust, grant agreement 317185 and by the H2020-MSCA-ITN-2016 UTOPIAE, grant agreement 722734. The author would like to acknowledge the work of all the PhD students who worked on the generation of the results in this chapter: Chiara Tardioli, Carlos Ortega, Martin Kubicek, Massimo Vetrisano.

References

1. Celletti, A., Pinzari, G.: Four classical methods for determining planetary elliptic elements: a comparison. Celest. Mech. Dyn. Astron. **93**, 1–52 (2005)
2. Milani, A., Gronchi, G.: Theory of Orbit Determination. Cambridge University Press, Cambridge (2010)
3. Gauss, C.F.: Theoria motus corporum coelestium in sectionibus conicis Solem ambientium (Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections). Reprinted by Dover Publications, 1963
4. Vetrisano, M., Vasile, M.: Autonomous navigation of a spacecraft formation in the proximity of an asteroid. Adv. Space Res. **57**(8), 1783–1804 (2016). Advances in Asteroid and Space Debris Science and Technology—Part 2
5. Vetrisano, M., Colombo, C., Vasile, M.: Asteroid rotation and orbit control via laser ablation. Adv. Space Res. **57**(8), 1762–1782 (2016). Advances in Asteroid and Space Debris Science and Technology—Part 2
6. Park, R.S., Scheeres, D.J.: Nonlinear mapping of Gaussian statistics: theory and applications to spacecraft trajectory design. J. Guid. Control. Dyn. **29**(6), 1367–1375 (2006)
7. Giza, D., Singla, P., Jah, M.: An approach for nonlinear uncertainty propagation: application to orbital mechanics. In: AIAA Guidance, Navigation, and Control Conference, pp. 1–19 (2009)
8. Armellin, R., Di Lizia, P., Bernelli-Zazzera, F., Berz, M.: Asteroid close encounters characterization using differential algebra: the case of Apophis. Celest. Mech. Dyn. Astron. **107**(4), 451–470 (2010)

9. De Mars, K.J., Jah, M.K.: Probabilistic initial orbit determination using gaussian mixture models. J. Guid. Control. Dyn. **36**(5), 1324–1335 (2013)
10. Jones, B.A., Doostan, A., Born, G.H.: Nonlinear propagation of orbit uncertainty using non-intrusive polynomial chaos. J. Guid. Control. Dyn. **36**, 430–444 (2013)
11. Smirnov, N.N., Dushin, V.R., Panfilov, I.I., Lebedev, V.V.: Space debris evolution mathematical modeling. In: Proceedings of the European Conference on Space Debris, ESA-SD-01, Darmstadt, pp. 309–316 (1993)
12. Nazarenko, A.: The development of the statistical theory of a satellite ensemble motion and its application to space debris modeling. In Kaldeich-Schuermann, B. (ed.) Second European Conference on Space Debris. ESA Special Publication, vol. 393, p. 233 (1997)
13. McInnes, C.R.: An analytical model for the catastrophic production of orbital debris. ESA J. **17**(4), 293–305 (1993)
14. Lewis, H.G., Letizia, F., Colombo, C.: Multidimensional extension of the continuity equation method for debris clouds evolution. Adv. Space Res. **57**(8), 1624–1640 (2016)
15. Fujiwara, T.: Integration of the collisionless Boltzmann equation for spherical stellar systems. Publ. Astron. Soc. Jpn. **35**(4), 547–558 (1983)
16. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. **38**, 325–339 (1967)
17. Walley, P.: Towards a unified theory of imprecise probability. Int. J. Approx. Reason. **24**(2), 125–148 (2000)
18. Itô, K.: Stochastic integral. Proc. Imp. Acad. **20**(8), 519–524 (1944)
19. Aristoff, J.M., Horwood, J.T., Singh, N., Poore, A.B.: Nonlinear uncertainty propagation in orbital elements and transformation to cartesian space without loss of realism. In: AIAA/AAS Astrodynamics Specialist Conference, pp. 1–14 (2014)
20. Sabol, C., Sukut, T., Hill, K., Alfriend, K.T., Wright, B., Li, Y., Schumacher, P.: Linearized orbit covariance generation and propagation analysis via simple Monte Carlo simulations. In: Paper AAS 10-134 presented at the AAS/AIAA Space Flight Mechanics Conference, pp. 14–17 (2010)
21. Jones, B.A., Doostan, A., Born, G.H.: Nonlinear propagation of orbit uncertainty using non-intrusive polynomial chaos. J. Guid. Control. Dyn. **36**(2), 430–444 (2013)
22. Vetrisano, M., Vasile, M.: Analysis of spacecraft disposal solutions from lpo to the moon with high order polynomial expansions. Adv. Space Res. **60**(1), 38–56 (2017)
23. Ghanem, R.G., Spanos, P.D.: Spectral stochastic finite element formulation for reliability analysis. J. Eng. Mech. **117**(10), 2351–2372 (1991)
24. Ghanem, R., Dham, S.: Stochastic finite element analysis for multiphase flow in heterogeneous porous media. Transp. Porous Media **32**(3), 239–262 (1998)
25. Xiu, D., Karniadakis, G.E.: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)
26. Ghanem, R.G., Doostan, A.: On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data. J. Comput. Phys. **217**(1), 63–81 (2006)
27. Eldred, M.S., Swiler, L.P., Tang, G.: Mixed aleatory-epistemic uncertainty quantification with stochastic expansions and optimization-based interval estimation. Reliab. Eng. Syst. Saf. **96**(9), 1092–1113 (2011)
28. Smolyak, S.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dofl. Akad. Nauk. **158**, 1042–1045 (1963)
29. Genz, A., Keister, B.D.: Fully symmetric interpolatory rules for multiple integrals over infinite regions with gaussian weight. J. Comput. Appl. Math. **71**(2), 299–309 (1996)
30. Tardioli, C., Kubicek, M., Vasile, M., Minisci, E., Riccardi, A.: Comparison of non-intrusive approaches to uncertainty propagation in orbital mechanics. In: AAS Astrodynamics Specialists Conference, Vail, Colorado (2015). AAS 15-544
31. Mason, J.C.: Near-best multivariate approximation by fourier series, tchebycheff series and tchebycheff interpolation. J. Approx. Theory **28**(4), 349–358 (1980)
32. Franklin, B.: Chebyshev Expansions. SIAM, Philadelphia (2007)

33. Judd, K.L., Maliar, L., Maliar, S., Valero, R.: Smolyak method for solving dynamic economic models: lagrange interpolation, anisotropic grid and adaptive domain. J. Econ. Dyn. Control. **44**(1–2), 92–123 (2014)
34. Kubicek, M., Minisci, E., Cisternino, M.: High dimensional sensitivity analysis using surrogate modeling and high dimensional model representation. Int. J. Uncertain. Quantif. **5**(5), 393–414 (2015)
35. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Math. Comput. Simul. **55**(1–3), 271–280 (2001). The Second IMACS Seminar on Monte Carlo Methods
36. Julier, S.J., Uhlmann, J.K.: A new extension of the kalman filter to nonlinear systems. In: Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition VI, pp. 182–193 (1997)
37. Terejanu, G., Singla, P., Singh, T., Scott, P.D. Singh, T., Scott, P., Terejanu, G., Singla, P.: Uncertainty propagation for nonlinear dynamic systems using gaussian mixture models. J. Guid. Control. Dyn. **31**(6), 1623–1633 (2008)
38. Kleijnen, J.P.C.: Kriging metamodeling in simulation: a review. Eur. J. Oper. Res. **192**(3), 707–716 (2009)
39. Milani, A., Nobili, A., Farinella, P.: Non-Gravitational Perturbations and Satellite Geodesy. Adam Hilger, Bristol (1987)
40. Sharaf, M.A., Selim, H.H.: Final state predictions for $j - 2$ gravity perturbed motion of the earth's artificial satellites using bispherical coordinates. NRIAG J. Astron. Geophys. **2**(1), 134–138 (2014)
41. Picone, J.M., Hedin, A.E., Drob, D.P., Aikin, A.C.: Nrlmsise-00 empirical model of the atmosphere: statistical comparisons and scientific issues. J. Geophys. Res. **107**(A12), 1468 (2002)
42. Space-track website. https://www.space-track.org
43. Vasile, M., Tardioli, C.: On the use of positive polynomials for the estimation of upper and lower expectations in orbital dynamics. In: Proceedings of Stardust Conference, ESA/ESTEC, 2016 (2016)
44. Ghosal, S.: Convergence rates for density estimation with bernstein polynomials. Ann. Stat. **10**(3), 1264–1280 (2001)
45. Zhao, Y., Ausín Olivera, M.C., Wiper, M.P.: Bayesian multivariate bernstein polynomial density estimation. In: UC3M Working papers. Statistics and Econometrics. Universidad Carlos III de Madrid. Departamento de Estadística, 2013
46. Tapley, B.D., Schutz, B.E. Born, G.H.: Statistical Orbit Determination. Elsevier, Amsterdam (2004)
47. Zazzera, F.B., Vasile, M., Massari, M., Di Lizia, P.: Assessing the accuracy of interval arithmetic estimates in space flight mechanics. Technical report, Politecnico of Milano and European Space Agency, Contract No.18851/05/NL/MV 2004. Ariadna 04/4105
48. Makino, K., Berz, M.: Taylor models and other validated functional inclusion methods. Int. J. Pure Appl. Math. 6, 239–316 (2003)
49. Riccardi, A., Tardioli, C., Vasile, M.: An intrusive approach to uncertainty propagation in orbital mechanics based on Tchebycheff polynomial algebra. In: AAS Astrodynamics Specialists Conference, Vail, Colorado (2015). AAS 15-544
50. Berz, M., Makino, K.: Verified integration of ODES and flows using differential algebraic methods on highorder Taylor models. Reliab. Comput. **4**, 361–369 (1998)
51. Jorba, A., Zou, M.: A software package for the numerical integration of ODEs by means of high-order Taylor methods. Exp. Math. **14**(1), 99–117 (2005)
52. Absil, C.O., Serra, R., Riccardi, A., Vasile, M.: De-orbiting and re-entry analysis with generalised intrusive polynomial expansions. In: 67th International Astronautical Congress. Proceedings of the International Astronautical Congress, Guadalajara (2016)
53. Riolo, R.L., Worzel, B. (eds): Genetic Programming Theory and Practice. Genetic Programming, vol. 6. Kluwer, Boston (2003). Series Editor - John Koza
54. Stützle, T.: Ant colony optimization. IEEE Comput. Intell. Mag. **1**, 28–39 (2004)

55. Minisci, E., Serra, R., Vasile, M., Riccardi, A., Grey, S., Lemmens, S.: Uncertainty treatment in the GOCE re-entry. In: 1st IAA Conference on Space Situational Awareness (ICSSA) Orlando, IAA-ICSSA-17-01-01, Orlando, FL, 2017
56. Tardioli, C., Vasile, M.: Collision and re-entry analysis under aleatory and epistemic uncertainty. In: Majji, M., Turner, J.D., Wawrzyniak, G.G., Cerven, W.T. (eds.) Advances in Astronautical Sciences, vol. 156, pp. 4205–4220. American Astronautical Society, San Diego (2016). This paper was originally presented at the AAS/AIAA Astrodynamics Specialist Conference held August 9–13, 2015, Vail, Colorado, USA, and was originally published in the American Astronautical Society (AAS) publication Astrodynamics, 2015. http://www.univelt.com
57. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)