Lecture Notes in Physics

The Lecture Notes in Physics

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

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

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

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

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

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

R. Dvorak F. Freistetter J. Kurths (Eds.)

Chaos and Stability in Planetary Systems

Springer

Editors

Rudolf Dvorak
Florian Freistetter

Institute of Astronomy
University of Vienna
Türkenschanzstr. 17
1180 Vienna
Austria
E-mail: dvorak@astro.univie.ac.at
freistetter@astro.univie.ac.at

Jürgen Kurths

University of Potsdam
Institute of Physics
Am Neuen Palais, Geb. 19
14415 Potsdam
Germany
E-mail: jkurths@agnld.uni-potsdam.de

R. Dvorak, F. Freistetter, J. Kurths *Chaos and Stability in Planetary Systems*,
Lect. Notes Phys. 683 (Springer, Berlin Heidelberg 2005), DOI 10.1007/b94975

Library of Congress Control Number: 2005930446

ISSN 0075-8450
ISBN-10 3-540-28208-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28208-2 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Printed in The Netherlands

Typesetting: by the author using a Springer LaTeX macro package

Printed on acid-free paper SPIN: 10978337 54/TechBooks 5 4 3 2 1 0

Preface

This book is the outcome of the 3rd HISP (Helmholtz Institute for Supercomputational Physics) summer school in Potsdam, which was dedicated to "Chaos and Stability in Planetary Systems". The idea was to give a group of some 40, carefully selected students from different nations a basic formation in the following topics: Hamiltonian dynamics (I), Celestial Mechanics (II), the structure of extrasolar planetary systems (III), cosmogonie and the formation of planets (IV) and last but not least in handling supercomputers (V). Lucky enough we had world known experts for achieving the goal and teaching these different topics which are substantially connected. All these lectures are written down in this volume, which should give students and colleagues interested in this topic the basis to understand the complexity of extrasolar planetary system, their formation, their detection and dynamical structure. The idea of the book was not to replace existing books in Hamiltonian Mechanics or in Celestial Mechanics, but it should be able to give the necessary basic information for interested readers.

In the friendly atmosphere of the lectures and especially of the afternoon exercises with the computers we all had the impression of having reached the goal of the school to initiate the interest in the subject. Young people from different countries all over the globe, from Brazil to Russia, from Marocco to Finland, learned together, wrote computer codes and – last but not least – built new friendships. We, as editors, also responsible for the success of the summer school of 2003 in Potsdam, hope that this book will be able to transport the spirit of these weeks of work to the reader.

We gratefully acknowledge the financial support of the Ministerium für Wissenschaft und Forschung des Landes Brandenburg and the hospitality of the University of Potsdam which was an important precondition for the school to take place. Special thanks go to Carmen Romano, Marco Thiel and Werner von Bloh, who are a good deal responsible for the scientific and also social success of the school.

Vienna
April 2005

Rudolf Dvorak
Florian Freistetter
Jürgen Kurths

Contents

Part II Extrasolar Planetary Systems

List of Contributors

Cristian Beaugé
Universidad Nacional de Córdoba
Observatorio Astronómico
Laprida 854 – Barrio Observatorio
Argentina
beauge@oac.uncor.edu

Nelson Callegari Jr.
Universidade Federal de São Carlos
Departamento de Matemática
13565-905 São Carlos, Brasil
calleg@dm.ufscar.br

Rudolf Dvorak
University of Vienna
Institute of Astronomy
Türkenschanzstr. 17
A – 1180 Vienna, Austria
dvorak@astro.univie.ac.at

Sylvio Ferraz-Mello
Universidade de São Paulo
Instituto de Astronomia
Geofísica e Ciências Atmosféricas
CEP 05508-900 São Paulo, Brasil
sylvio@astro.iag.usp.br

Florian Freistetter
Astrophysikalisches Institut
Universitätsternwarte Jena
Schillergässchen 2-3
D-07745 Jena, Germany
florian@astro.uni-jena.de

Tatiana Michtchenko
Universidade de São Paulo
Instituto de Astronomia
Geofísica e Ciências Atmosféricas
CEP 05508-900 São Paulo, Brasil
tatiana@astro.iag.usp.br

Tom Quinn
Astronomy Department
University of Washington
Box 351580,
Seattle WA 98195-1580, USA
trq@astro.washington.edu

Harry Varvoglis
Section of Astrophysics Astronomy
& Mechanics
Department of Physics
Aristotle University of Thessaloniki
541 24 Thessaloniki, Greece
varvogli@physics.auth.gr

Part I

Introduction to Orbital Dynamics and Chaotic Systems

Orbit Dynamics, Stability and Chaos in Planetary Systems

Rudolf Dvorak[1] and Florian Freistetter[2]

[1] University of Vienna, Institute for Astronomy, Türkenschanzstr. 17, A-1180 Wien, Austria
(dvorak@astro.univie.ac.at)
[2] Astrophysikalisches Institut, Universitätssternwarte Jena, Schillergässchen 2-3, D-07745, Jena, Germany
(florian@astro.uni-jena.de)

1 Prolog: The Golden Mean

Let us start with a problem of dynamical biology, which was posed about 800 years ago by Fibonacci[1]:

> "A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year, if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?"

Fig. 1 shows how the rabbits reproduce:

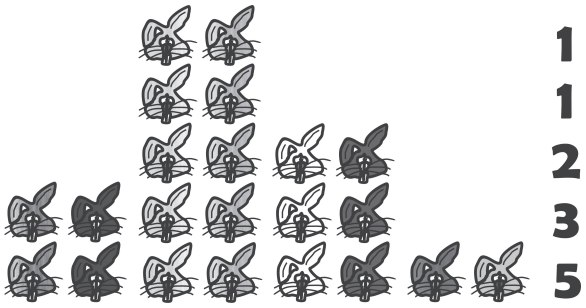

Fig. 1. The Fibonacci Tree

How can we interprete this diagram?

- After the first month there is still only 1 pair.
- After the second month the female gave birth to a new pair; consequently we have 2 pairs of rabbits.
- After the third month the original female has born a second pair and there are 3 pairs.

[1] Leonardo Pisano Fibonacci (1170 – 1250)

R. Dvorak and F. Freistetter: *Orbit Dynamics, Stability and Chaos in Planetary Systems*, Lect. Notes Phys. **683**, 3–140 (2005)
www.springerlink.com

- After the fourth month the original female has born another new pair but the female, born two months ago, has her first pair also, making 5 pairs.

Thus the number of pairs are 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, ... after one, two, three, ... months and the answer is that after 12 month the number of rabbit pairs populating the surrounded area is 144! But how can we compute these Fibonacci numbers ? The answer is easy because the number of pairs for a certain month is the sum of pairs of the two preceding months.

If we take the ratio of two successive numbers in Fibonacci's series, we will find the following series of numbers: $1/1 = 1$; $2/1 = 2$; $3/2 = 1.5$; $5/3 = 1.666...$; $8/5 = 1.6$; $13/8 = 1.625$; $21/13 = 1.61538...$

When we plot the sequence of the ratios on a graph (Fig. 2) we can see how fast this value converges.

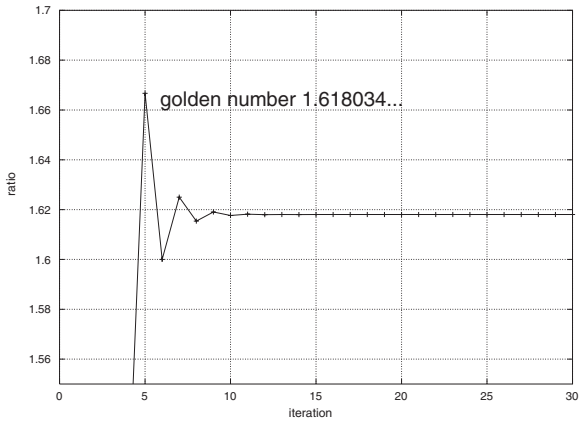

Fig. 2. The convergence of the mapping towards the golden number

The ratio settles to a particular value, namely the golden ratio or the **golden number** Φ. It has a value of approximately 1.618034...[2] This number can be constructed as follows (see Fig. 3).

It is the ratio between two sides a and b in the rectangle in Fig. 3 with the m property that

$$\frac{a}{b} = \frac{a+b}{a} \tag{1}$$

Setting $a = 1$ the ratio, which is the golden number Φ (the "most" irrational number) can be constructed by making use of Fig. 3:

$$\overline{MB} = \frac{\sqrt{5}}{2} = \overline{MF} \text{ and } \overline{CF} = \Phi \tag{2}$$

[2] Φ with 20000 digits can be found at `http://goldennumber.net/phi20000.htm`

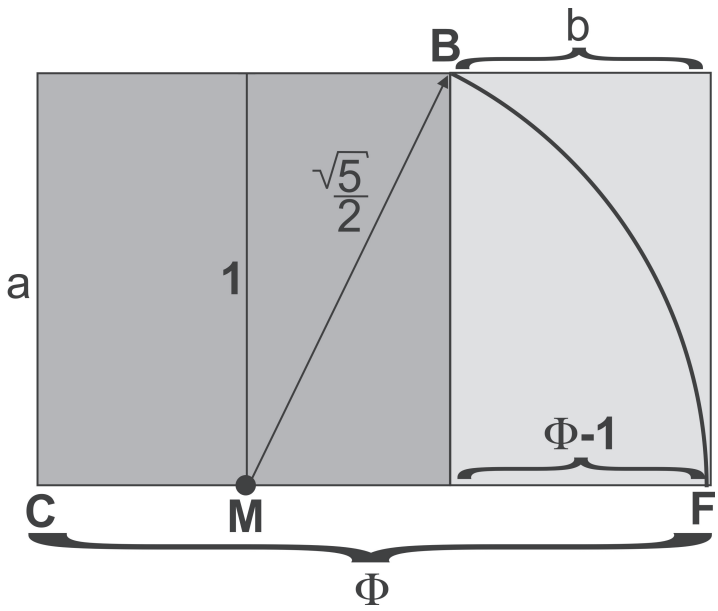

Fig. 3. The construction of the golden mean

and the golden number Φ which fulfills relation (1) is therefore $\frac{1+\sqrt{5}}{2}$.

This ratio, or proportion, determined by Φ was already known to the Greeks as the **Golden Section** and to Renaissance artists as the **Divine Proportion** and it is also called the **Golden Ratio** and the **Golden Mean**. Φ appears in mathematics, physics, biology, architecture, ... and seems to be an extraordinary number of unique importance for the construction of our world.[3] As we will see in the last section, the number Φ also plays an essential role in dynamical systems.

2 Mappings

2.1 Basic Definitions

When one wants to investigate the properties of a dynamical systems, mappings are an easy way to start with: first of all they are, in general, very simple (compared to a full dynamical system); nevertheless, they hold the same amount of information and complexity. Second, they are easy to compute – thus the general properties of dynamical systems can be best investigated by the help of mappings. This section will give an introduction to mappings, will give some examples and explain some of the basic characteristics of dynamical systems.

[3] compare the respective homepages e.g. `http://goldennumber.net/neophite.htm`

Mappings are mathematical objects with the following simple form

$$a_{n+1} = F(n, a_n) \tag{3}$$

with n a natural number where a_n is a sequence of real (or complex) numbers and $F(n, a_n)$ is a function which determines how to compute the next number a_{n+1}. A very simple example is the fore mentioned Fibonacci-sequence; there the mapping is defined as

$$a_{n+1} = a_n + a_{n-1} \tag{4}$$

with $a_0 = 0$ and $a_1 = 1$

Definition 1. *A mapping F is called autonomous when it does not depend explicitly on the index n, while in the opposite case it is called non-autonomous.*

Definition 2. *A mapping is defined as of order M if*

$$a_{n+M} = F(n, a_n, a_{n+1}, ...a_{n+M-1}). \tag{5}$$

In this sense the Fibonacci sequence is of order M=2 and is the result of an autonomous mapping.

A general solution of a mapping exists, when we can express any a_n by a mathematical formula which just computes any a_n from a_0. E.g. for the mapping

$$a_{n+1} = a_n^2 \tag{6}$$

the solution is

$$a_n = F(n, a_0) = a_0^{2^n} \tag{7}$$

which can be verified easily

$$a_0^{2^{n+1}} = [a_0^{2^n}]^2 = a_0^{2(2^n)} \tag{8}$$

2.2 Linear Mappings

In the following two examples of a non-autonomous linear mapping (the terms a_n appear only linearly) are shown:

$$a_{n+1} = F(n)a_n + G(n) \tag{9}$$

$$a_{n+2} = F_1(n)a_{n+1} + F_2(n)a_n + G(n) \tag{10}$$

In the mapping I (9) only the n-th element is used to compute the (n+1)th element – thus it is of order 1, whereas in the mapping II (10) the (n+2)th element is computed from the (n+1)th element and the n-th element – so it has the order 2 (similar to the Fibonacci sequence).

Remark 1. The theory of ordinary differential equations is very similar to the theory of mappings.

Theorem 1. *If a_n is a solution of the mappings I and II with the function G(n)=0 then also the function $C \cdot a_n$ is a solution if C is an arbitrary number ($\in \mathbb{C}, \mathbb{R}$) (no proof).*

Theorem 2. *If b_n and c_n are linearly independent solutions of the mapping II with G(n)=0 ($c_n \neq b_n$ and C arbitrary) then the function $C_1 \cdot b_n + C_2 \cdot c_n$ is also a solution (no proof).*

A special form of a linear mapping is given when we have the following

$$a_{n+2} = \alpha a_{n+1} + \beta a_n \tag{11}$$

where α and β are constants. Making the ansatz for a particular solution

$$a_n = \lambda^n \text{ and } \lambda = const \tag{12}$$

and inserting it into the mapping and solving for λ^n one gets

$$\lambda^n[\lambda^2 - \alpha\lambda - \beta] = 0 \tag{13}$$

where ($\lambda \neq 0$) one gets the **characteristic equation**[4]:

$$\lambda^2 - \alpha\lambda - \beta = 0 \tag{14}$$

Using the respective solutions λ_1 and λ_2 the general solution for the n-th term reads

$$a_n = C_1\lambda_1^n + C_2\lambda_2^n \tag{15}$$

where the constants are determined from the "initial conditions" a_0 and a_1:

$$a_0 = C_1 + C_2 \tag{16}$$

$$a_1 = C_1\lambda_1 + C_2\lambda_2 \tag{17}$$

For the Fibonacci sequence with $\alpha = \beta = 1$ the real solutions for the characteristic equation are

$$\lambda_{1,2} = \frac{1 \pm \sqrt{5}}{2} \tag{18}$$

which leads for $a_0 = a_1 = 1$ to the irrational solution

$$a_n = \frac{1}{\sqrt{5}} \left(\left(\frac{\sqrt{5}+1}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right) \tag{19}$$

for any a_n.[5] If, for example, $n = 7$ we find $a_7 = 21$.

[4] which is the same in the theory of ordinary linear differential equations of 2nd order.

[5] Note the interesting fact that a linear combination of powers of irrational numbers has an integer as solution.

As next example we will treat two coupled linear maps of the form

$$x_{n+1} = m_1 x_n + m_2 y_n \tag{20}$$

$$y_{n+1} = m_3 x_n + m_4 y_n \tag{21}$$

We write this 2-dimensional mapping in a form of a vector

$$\boldsymbol{X}_{n+1} = M\boldsymbol{X}_n \tag{22}$$

where M is the 2×2 matrix

$$M = \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}$$

where $m_1, \ldots, m_4$ are real constants which do not depend on n.

The mapping can be reduced to two independent homogeneous mappings of second order by replacing n by $n+1$ in equation (20) and using equation (21) to express y_{n+1}:

$$x_{n+2} = m_1 x_{n+1} + m_2 y_{n+1} \tag{23}$$

Using

$$y_n = \frac{1}{m_2}(x_{n+1} - m_1 x_n) \tag{24}$$

the result is

$$x_{n+2} = (m_1 + m_4)x_{n+1} - (m_1 m_4 - m_2 m_3)x_n \tag{25}$$

In fact we can express this equation with the aid of the matrix M for x_{n+2} but also for y_{n+2}

$$x_{n+2} = \mathtt{Tr}M x_{n+1} - \det M x_n \tag{26}$$

$$y_{n+2} = \mathtt{Tr}M y_{n+1} - \det M y_n \tag{27}$$

We have to solve the characteristic equation defined previously for $\alpha = \mathtt{Tr}M$ and $\beta = -\det M$ where the stability of the system depends on two conditions:

$$|\mathtt{Tr}M| \geq 2\sqrt{|\det M|} \tag{28}$$

$$|\mathtt{Tr}M| < 2\sqrt{|\det M|} \tag{29}$$

If the first condition is fulfilled, then all the solutions are unbounded (which is the hyperbolic case); if the second condition is fulfilled all solutions are bounded (which is the elliptic case).

Of special interest for the problems in celestial mechanics is the case with $\det M = 1$ which gives an area preserving mapping.

Theorem 3. *A mapping M is called* area preserving *if* $\det M = 1$. *This means that a closed curve of initial conditions $IC1$ around an area A is always mapped into another closed curve $IC2$ which encloses exactly the same area A.*

Proof. We can prove this statement just by applying the transformation M to a parallelogram with the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with the components (a_1,a_2) and (b_1,b_2), which is depicted in Fig. 4:

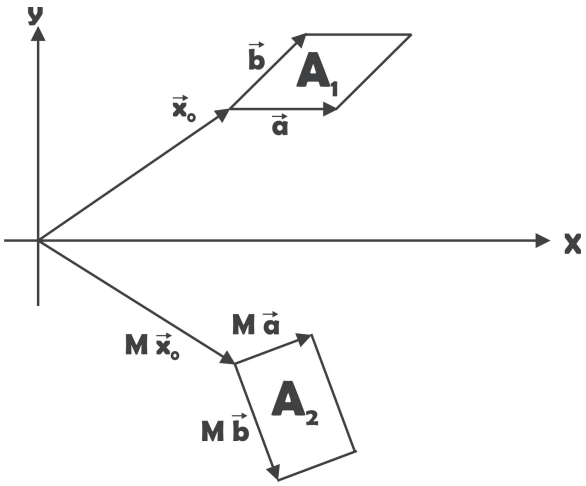

Fig. 4. Area preserving mapping

$$A_1 = |a_1 b_2 - a_2 b_1| \tag{30}$$

The transformation can be applied directly to the vectors where we can see that

$$A_2 = |(Ma_1)(Mb_2) - (Ma_2)(Mb_1)| \tag{31}$$

which is after multiplication of the different terms

$$A_2 = |\det M| A_1 \tag{32}$$

and because of the assumption that $\det M = 1$ it follows that

$$A_2 = A_1 \tag{33}$$

□

Fixed Points and Their Stability

Definition 3. *A mapping of any form may have a periodic solution with a period T when $\boldsymbol{x}_{n+T} = \boldsymbol{x}_n$. We relate the order k to a fixed point x_0 when the mapping F applied k times gives the same value of x_0:*

$$x_0 = \underbrace{F(F(F(F(\ldots F(x_0)\ldots))))}_{\text{k-times}} \tag{34}$$

A simple example is the mapping

$$x_{n+1} = 2x_n - x_n^2 \tag{35}$$

which has a 1^{st} order fixed point which can be computed by

$$x = 2x - x^2 \tag{36}$$

with the solutions $x = 0$ and $x = 1$. The 2^{nd} order fixed point can be found by

$$x_{n+2} = 2x_{n+1} - x_{n+1}^2 \tag{37}$$

which leads to

$$x = 2(2x - x^2) - (2x - x^2)^2 \tag{38}$$

The solution of the resulting 3rd order algebraic equation leads to three 2^{nd} order fixed points:

$$x_1 = 1 \tag{39}$$

$$x_{2,3} = 3/2 \pm i\sqrt{3}/2 \tag{40}$$

where, with $i = \sqrt{-1}$, two solutions are complex.

A mapping with a fixed point of order k will reproduce this point after applying the mapping k times. But what happens when we start close to this fixed point?

Definition 4. *A fixed point is called* **linearly stable** *when a point sufficiently close to the fixed point remains close to it for any number of iterations.*

Definition 5. *A fixed point is called* **linearly unstable** *when a point sufficiently close to the fixed point leaves the neighborhood after a certain number of iterations.*

To derive a criterion for linear stability we study the point $(x_F + \lambda_n)$ which lies infinitesimally close to a fixed point x_F of the mapping

$$x_{n+1} = F(x_n) \tag{41}$$

We thus substitute $x_n = x_F + \lambda_n$:

$$x_F + \lambda_{n+1} = F(x_F + \lambda_n) \tag{42}$$

and make a Taylor expansion with respect to λ_n

$$x_F + \lambda_{n+1} = x_F + \lambda_n \frac{dF}{d\lambda_n} + O(\lambda_n^2) \tag{43}$$

which leads to the so-called linearized mapping ($'$ indicates derivation according to λ)

$$\lambda_{n+1} = F'(x_F)\lambda_n + O(\lambda_n^2) \tag{44}$$

which has the solution

$$\lambda_n = \lambda_0[F'(x_F)]^n \tag{45}$$

Remark 2. A fixed point of the mapping (41) is linearly stable iff $|F'(x_F)| \leq 1$, otherwise ($|F'(x_F)| > 1$) it is linearly unstable.

2.3 The Logistic Map

As an example we want to deal with the well known **logistic map**, which describes in a simple way the size of a population of species depending on a parameter r:

$$x_{n+1} = F(x) = rx_n(1 - x_n) \tag{46}$$

If we examine a hunter-prey population, then the number of prey will decrease if the number of hunters increases (in a closed system). After some time, because of the lack of food, the number of hunters will decrease and thus the number of prey will increase again (the same mechanism can also be found e.g. in financial trading or electronics). Equation (46) describes the mentioned behavior mathematically.

Using the methods described in the last section, we now can compute the first order fixed points easily:

$$x = rx(1 - x) \tag{47}$$

from which follows that we have two 2 fixed points, namely $x_1 = 0$ and $x_2 = 1 - 1/r$. We can study the stability of these points by writing the derivative with respect to x according to equation (45):

$$\frac{dF}{dx} = r - 2xr \tag{48}$$

Substituting the two solutions for x_1 and x_2 gives

$$F'(x_1) = r \tag{49}$$

$$F'(x_2) = 2 - r \tag{50}$$

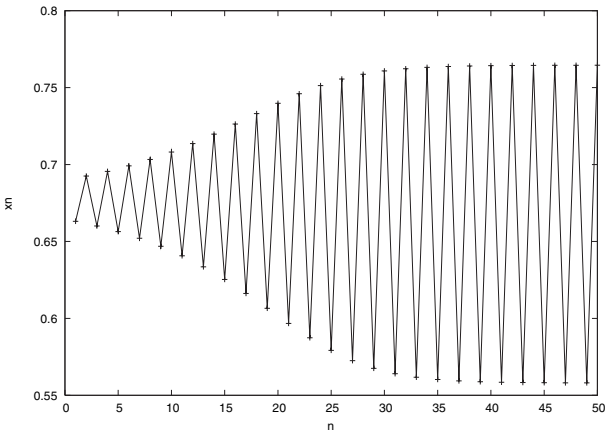

Fig. 5. Values of x_n for $r > 3$: the mapping converges to two fixed points

The behavior of this mapping is such it converges for special values of the parameter r to different points, which seem to attract the sequence of points.[6] We now investigate this mapping according to the value of r.

- If $r < 1$ then there are two period 1 fixed points: 0 (linearly stable), and $1 - \frac{1}{r} < 1$ which is linearly unstable (because of $F'(x_2) = 2 - r > 1$). Thus the value of x_n converges relatively fast to 0.
- For $r = 1$ there exists only one fixed point of period one : $x = 0$.
- If $1 < r < 3$ the value of x_n converges to a fixed point which depends on the parameter r.
- If $r > 3$ the dynamics become more complicated. For $r > 0$ both period 1 fixed points are linearly unstable. There are now four fixed points of period 2, two of them stable and two unstable – thus the mapping will converge to two fixed points (*period doubling* – see Fig. 5).
- For $3 < r < 3.5699...$ the period doubling continues and the mapping will converge to $4, 16, \ldots$ fixed points – see Fig. 6.
- If $r > 3.5699...$ the period doubling stops and the value of x_n now jumps chaotically between 0 and 1 (Fig. 7) – the mapping has become chaotic!

Fig. 8 shows the *Feigenbaum diagram*. The values of x_n are plotted against the value of r and one can see the behavior described above more clearly: for small values of r there exists only one solution; then, for a certain value

[6] "An attractor is a set of states (points in the phase space), invariant under the dynamics, towards which neighboring states in a given basin of attraction asymptotically approach in the course of dynamic evolution. An attractor is defined as the smallest unit which cannot be itself decomposed into two or more attractors with distinct basins of attraction. This restriction is necessary since a dynamical system may have multiple attractors, each with its own basin of attraction" (definition from `mathworld.wolfram.com`).

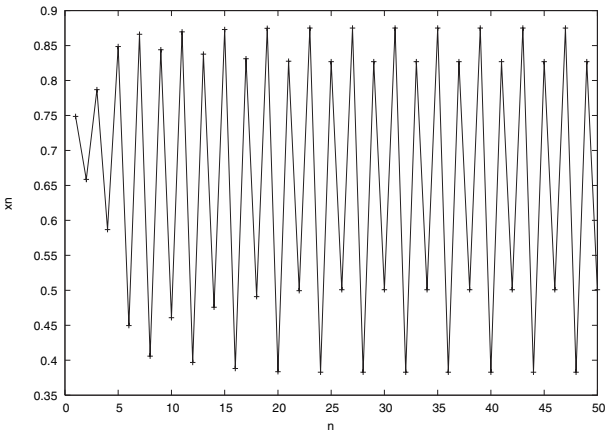

Fig. 6. Values of x_n for $r > 3$: the mapping converges to 4 fixed points

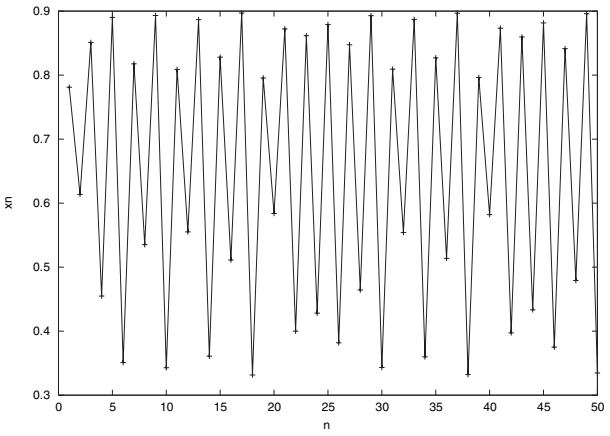

Fig. 7. Values of x_n for the chaotic case: x_n jumps chaotically between 0 and 1

r_1, the period doubles and there are two solutions, for r_2 there are four solutions,...After $r = r_\infty = 3.5699...$ the mapping becomes chaotic.

The values of r_i can be related by

$$\delta = \frac{r_i - r_{i-1}}{r_{i+1} - r_i} \tag{51}$$

which, for $i \to \infty$ has a value of $\delta = 4.6692...$, which is a universal constant – the *Feigenbaum constant*. For certain windows for larger parameter values new attractor cycles with different periodicity appear – thus the logistic map is *self-similar*. Fig. 9 shows a zoom inside the chaotic region where the "original structure" reappears.

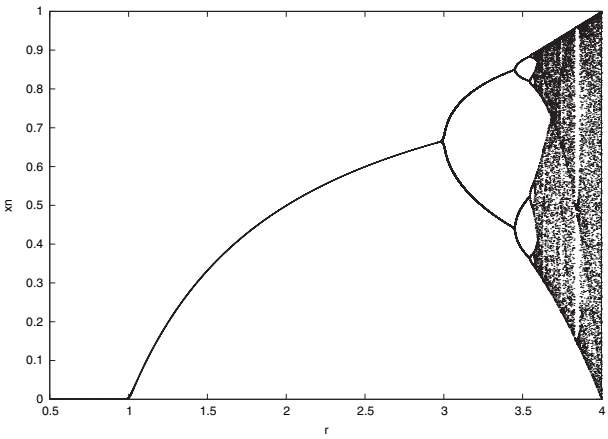

Fig. 8. The Feigenbaum diagram for the logistic equation

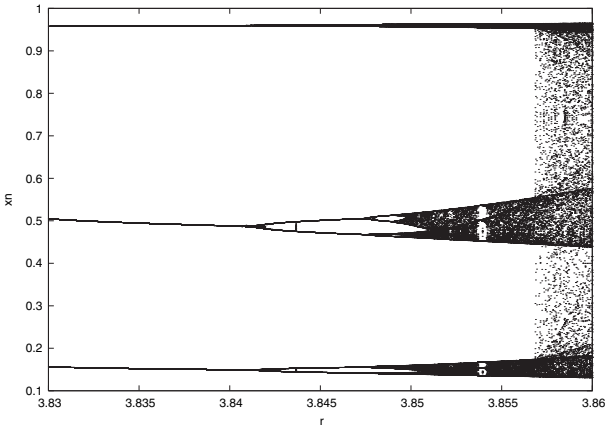

Fig. 9. Magnification of the chaotic region in the Feigenbaum diagram (Fig. 8)

2.4 Standard Mapping

Introduction

The origin of the nowadays well known and widely used *Standard Mapping* lies in the field of particle physics[7]: to examine a possible cosmic ray acceleration mechanism in which charged particles accelerate by collisions with moving magnetic field structures, Fermi [21] created an analog problem, where a ball is bouncing between a fixed and an oscillating wall. If, every time the ball impacts, the phase of the oscillation is chosen at random, the particle will be (in the average) accelerated. The question was now, if the ball would be

[7] The derivation of the equations of the Standard Mapping follows [42].

also accelerated, when the wall oscillation is a periodic function of time. This problem was investigated by Ulam [72] who found that the particle motion appeared to be stochastic, but did not increase its energy (on the average). In other publications (i.e. [44]) these results are explained. They demonstrated that in the case of smooth forcing functions, the phase plane shows three distinct regions with increasing ball velocity: a low-velocity region were all fixed points of period 1 are unstable and thus leading to stochastic motion; an intermediate velocity region in which islands of stability around elliptic fixed points are embedded in a stochastic sea and a high velocity region in which bands of stochastic motion are separated from each other by regular orbits. Because this problem of particle acceleration can be approximated by simple mappings, it became a well-suited case to study the parameter regions of phase space and the corresponding KAM surfaces.

In the following we want to show, how the equations of the *Standard Mapping* can be derived.

Ulam Mapping

The exact Ulam Mapping for the motion of a bouncing ball between a fixed and an oscillating wall (see Fig. 10), where the wall velocity is defined by a sawtooth function, is given by the following set of exact difference equations:

$$u_{n+1} = \pm u_n \left(\psi_n - \frac{1}{2}\right) \tag{52}$$

$$\psi_{n+1} = \frac{1}{2} - 2u_{n+1} + \left[\left(\frac{1}{2} - 2u_{n+1}\right)^2 + 4\phi_n u_{n+1}\right]^{\frac{1}{2}} \quad \left(u_{n+1} > \frac{1}{4}\psi_n\right) \tag{53}$$

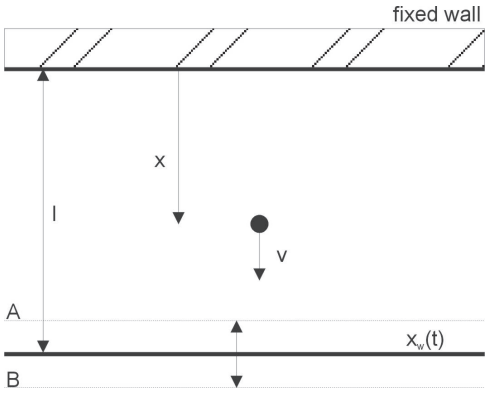

Fig. 10. Ulam version of the Fermi acceleration model in which a particle bounces between an oscillating and a fixed wall

$$\psi_{n+1} = 1 - \psi_n + 4u_{n1} \quad \left(u_{n+1} \leq \frac{1}{4}\psi_n\right) \tag{54}$$

$$\phi_n = \psi_n + \frac{\psi_n\,(1-\psi_n) + \frac{1}{4}l}{4u_{n+1}} \qquad \mathrm{mod}(1) \tag{55}$$

where $2a$ is the peak amplitude of the wall oscillation, l is the distance between the walls, u_n is the velocity of the wall (normalized to V, where $\frac{1}{4}V$ is the amplitude of the wall velocity), n is the number of collisions with the wall and ψ_n is the phase of the moving wall at the time of the collision.

These equations are exact, but not area-preserving. To obtain an area-preserving map, the difference equations can be written in terms of collision with the fixed instead of the moving wall. $\frac{\bar{u}}{2\omega a}$ is defined as the normalized velocity, θ_n is the phase of the moving wall at the nth collision with the fixed wall; the motion of the wall shall be given as $x = F(\psi)$ where F is an even periodic function of phase $\psi = \omega t$ and a period of 2π and $F_{\max} = F_{\min} = 1$. The equations of motion now can be written as

$$\bar{u}_{n+1} = \bar{u}_n - F'(\psi) \tag{56}$$

$$\theta_{n+1} = \psi_c + \frac{\left[\pi M + \frac{1}{2}F(\psi_c)\right]}{\bar{u}_{n+1}} \tag{57}$$

$$\psi_c = \theta_n + \frac{\left[\pi M + \frac{1}{2}F(\psi_c)\right]}{\bar{u}_n} \tag{58}$$

where ψ_c is the phase at the next collision with the moving wall after the nth collision with the fixed wall; $M = \frac{l}{2\pi a}$ and F' is the velocity impulse given to the ball. Using the extended phase space, given by $(v, x, -E, t)$ (where $E = \bar{u}^2$ is the energy), one can construct an area-preserving mapping $(-E, \theta)$ by choosing a surface of section (SOS)[8] with $x = 0$.

The mapping (52)–(55) can be simplified by allowing the wall to add momentum to the ball according to its velocity (without a change in the position of the wall). For this problem, the ball velocity and phase just before the nth impact can serve as proper canonical variables. The equations (in normalized form) are given by

$$u_{n+1} = \left|u_n + \psi_n - \frac{1}{2}\right| \tag{59}$$

$$\psi_{n+1} = \psi_n + \frac{M}{u_{n+1}} \qquad \mathrm{mod}1$$

where we used the following notation: $M = \frac{l}{16a}$, $\frac{M}{u} = \frac{2l}{vT}$ is the normalized transit time, $T = 32a$ is the period of the wall oscillation and $v = uV$ is the

[8] for details how to construct a surface of section see Sect. 2.1 in the chapter on "Regular and Chaotic Motion in Hamiltonian Systems".

particle velocity.[9] With a sinusoidal momentum transfer – a nonlinear force function – equations (59) can be written as

$$u_{n+1} = \mid u_n + \sin\psi_n \mid \qquad (60)$$
$$\psi_{n+1} = \psi_n + \frac{2\pi M}{u_{n+1}}$$

where the phase of the wall oscillation now extends over 2π instead of unity.

From the Fermi Map to the Standard Mapping

To explore the transition between orbits bounded by KAM surfaces[10] and orbits that can move chaotically, now a *stochasticity parameter* K is introduced.

The Standard Mapping is obtained from (60) by linearization in action space near a given fixed point of period 1 located at

$$\frac{2\pi M}{u_1} = 2\pi m \qquad (61)$$

where m is an integer. Using $u_n = u_1 + \Delta u_n$ and shifting the angle $\theta_n = \psi_n - \pi$ $(-\pi < \theta_n \leq \pi)$ the equations now can be written as

$$I_{n+1} = I_n + K\sin\theta_n \qquad (62)$$
$$\theta_{n+1} = \theta_n + I_{n+1}$$

where

$$I_n = \frac{-2\pi M \Delta u_n}{u_1^2} \qquad (63)$$

is now the new action and[11]

$$K = \frac{2\pi M}{u_1^2} \qquad (64)$$

is the stochasticity parameter, now related to the old action u_1.

The period 1 fixed points of the Standard Mapping can be easily obtained by requiring that the phase (mod 2π) and the action are stationary:

$$I_1 = 2\pi m \qquad \mathrm{m} \in \mathrm{Z} \qquad (65)$$
$$\theta_1 = 0, \pi$$

[9] These simplified equations can be obtained as an approximation to the exact set for $\frac{l}{a} >> 1$ and $u >> 1$.

[10] KAM surfaces (or KAM tori) are the last invariant curves that separate chaotic from regular motion. If the perturbation increases these KAM tori break up and become so the called *Cantori*.

[11] for details on *action-angle variables* see see Sect. 3.2 in the chapter on "Regular and Chaotic Motion in Hamiltonian Systems".

Thus, for every integer m there are two fixed points. We now expand the mapping around a fixed point to obtain the linearized equations:

$$\boldsymbol{x} = \boldsymbol{x}_0 + \Delta\boldsymbol{x}. \tag{66}$$

With $\boldsymbol{x}^T = (I, \theta)$, and, by keeping only the linear terms, this gives

$$\Delta\boldsymbol{x}_{n+1} = \mathsf{A}\Delta\boldsymbol{x}_n \tag{67}$$

where A is a transformation matrix independent of $\boldsymbol{x}$ and reads

$$\mathsf{A} = \mathsf{M}(\boldsymbol{x_0}) \cdot \mathsf{M}(\boldsymbol{x_1}) \tag{68}$$

where $\mathsf{M}(\boldsymbol{x})$ is the Jacobian matrix

$$\mathsf{M}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial I_1}{\partial I_0} & \frac{\partial I_1}{\partial \theta_0} \\ \frac{\partial \theta_1}{\partial I_0} & \frac{\partial \theta_1}{\partial \theta_0} \end{pmatrix}$$

For the mapping (62) this leads to

$$\mathsf{A} = \begin{pmatrix} 1 & \pm K \\ 1 & 1 \pm K \end{pmatrix}$$

for $\theta_1 = 0$ and $\theta_1 = \pi$. Note that det A=1 is required for area preserving mappings. The stability condition depends on the trace of A (see equations (28)–(29)):

$$\mid 2 \pm K \mid < 2 \tag{69}$$

It is evident that the fixed point at $\theta = 0$ will be always unstable whereas the fixed point at $\theta = \pi$ will be stable for $K < 4$ and changes with larger K from an elliptic fixed point to an hyperbolic one.

Numerical Experiments

The behavior of the period-1 fixed points explained above can be seen in Figs. 11 and 12. It can be also seen, how the chaotic region in the phase space is increasing with K. For $K = 0.5$ (Fig. 11a) one can see the primary period 1 and 2 orbits very clearly; only local stochasticity near the seperatrices[12] occurs. In Fig. 11b the last KAM curve between the period 1 and 2 islands has been destroyed – now the chaos is global; only *islands of stability* remain (for $K > 0.9716...$).

These islands decrease in size if K is increased: for $K > 6$ no islands of stability can be seen in the phase space.

[12] "A phase curve (i.e. an invariant manifold) which meets a hyperbolic fixed point (i.e. an intersection of a stable and an unstable invariant manifold) or connects the unstable and stable manifolds of a pair of hyperbolic or parabolic fixed points. A separatrix marks a boundary between phase curves with different properties" (definition from `mathworld.wolfram.com`).

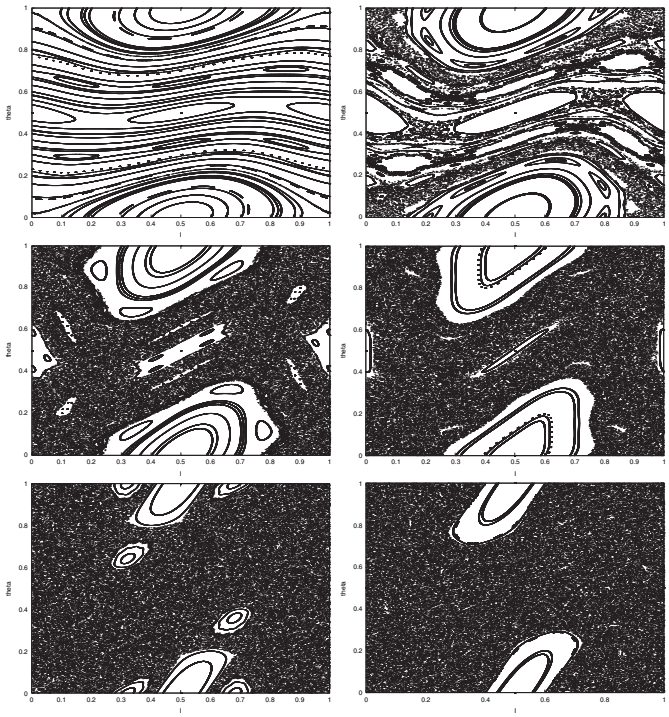

Fig. 11. Phase space for the Standard Mapping for K = **(a)** 0.5; **(b)** 1.0; **(c)** 1.5; **(d)** 2.0; **(e)** 2.5; **(f)** 3.0 *(from top left to bottom right)*

This does not mean, that the chaos is complete for $K > 6$ – if one magnifies sections of the phase space, one can still find areas of regular motion. If one divides the phase space of the Standard Mapping in a large number of squares and counts the ones not entered by a chaotic orbit, one can obtain a measure of the percentage of regular area. Fig. 13 shows, how the regular area decreases and increases as a function of the stochasticity parameter K.

It can be seen that, also for very large values of K, there are still regions in the phase space, where regular motion is possible. Fig. 14 shows, that the "creation and decay" of the periodic orbits always happens in the same way and repeats itself with a period of 2π. For a detailed description of this behavior see [6], [8] and [10].

2.5 Stickiness

An important feature in chaotic dynamics is the so called *stickiness*. This behavior can be seen in Fig. 15 which shows the logarithm of the escape time (that is the time a chaotic orbit needs to leave the vicinity of an island of stability) versus the initial x position. The edge of a large region of regular

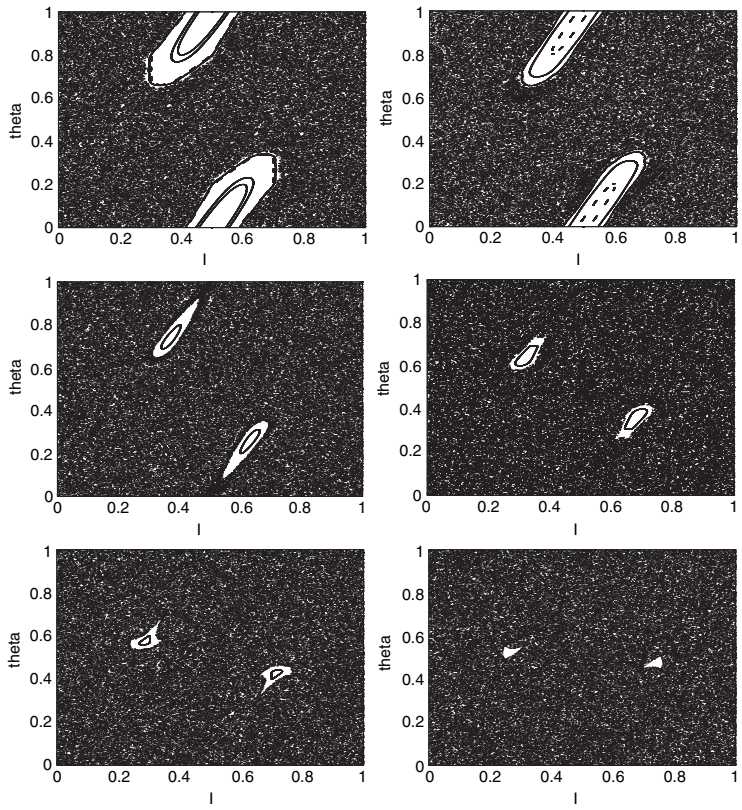

Fig. 12. Phase space for the Standard Mapping for K = **(a)** 3.5; **(b)** 4.0; **(c)** 4.5; **(d)** 5.0; **(e)** 5.5; **(f)** 6.0 *(from top left to bottom right)*

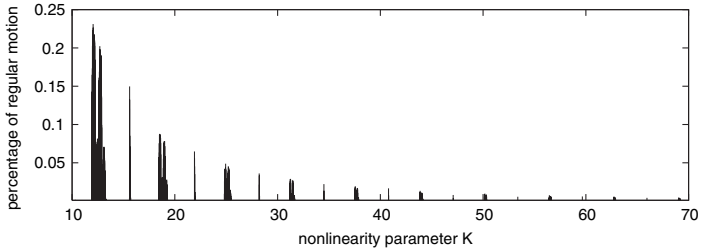

Fig. 13. Percentage of regular area in the phase space of the Standard Mapping in dependence of K

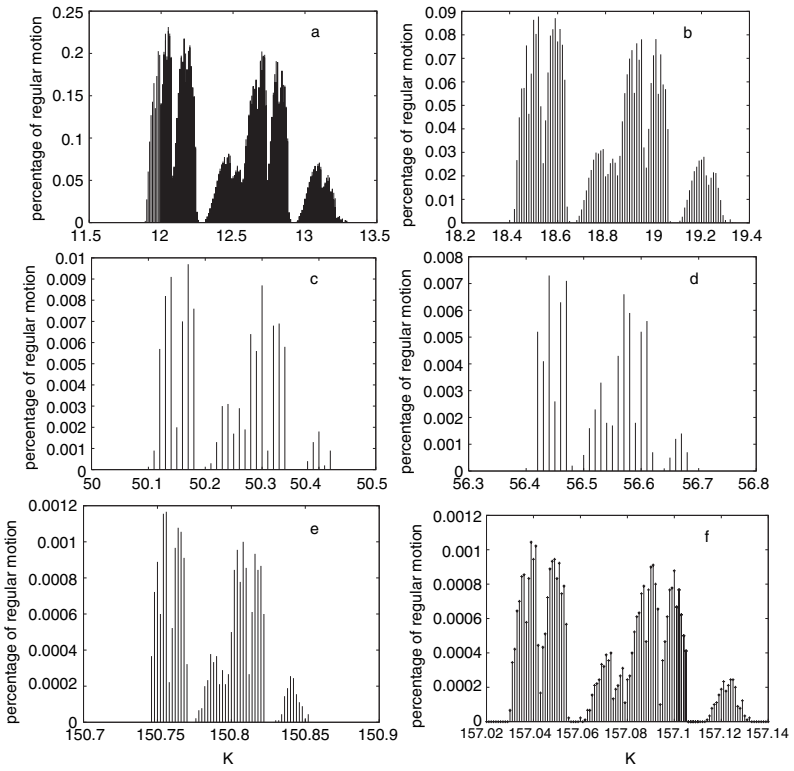

Fig. 14. Development of the regular areas for K close to **(a)** 12; **(b)** 18; **(c)** 50; **(d)** 56; **(e)** 150 and **(f)** 157. The pictures show the percentage of regular area versus the stochasticity parameter K

motion can be seen on the right side of the figure; also, in the middle, there is a smaller region of stability, where the orbit stays near the island all the time. In between, the trajectory leaves the vicinity of the regular orbit more or less fast. But it is clearly visible, that there is a chaotic region (left to the large island of stability) where, although the orbit is chaotic, the escape time is considerably higher than for the others. Thus, the orbit "sticks" to the island of stability. This behavior is due to the complicated structure of the Cantori (the remanents of the KAM curves that broke up when the stochasticity parameter was increased) which "trap" the trajectory near the islands of stability. Some sticky regions can be seen in Fig. 16 and 17. For details on the behavior of sticky orbits see [8].

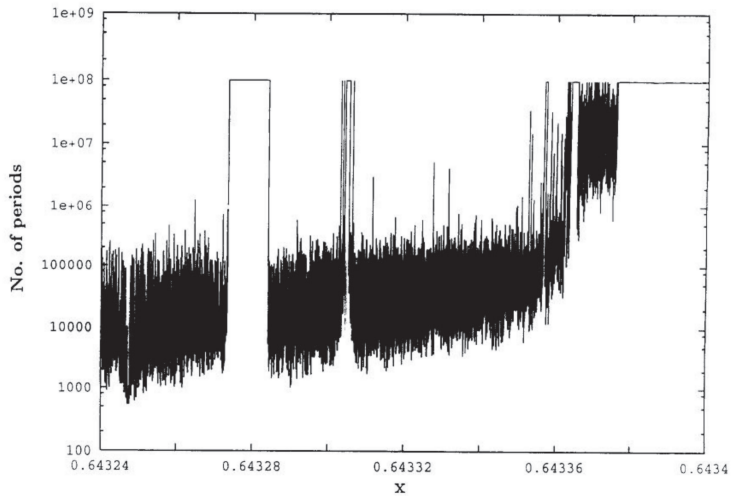

Fig. 15. Stickiness in the Standard Mapping: the logarithm of the escape time vs. the initial value *(from [8])*

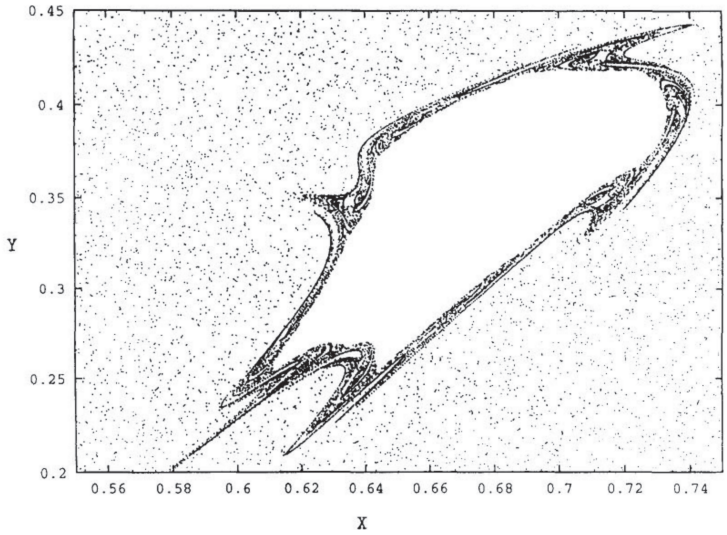

Fig. 16. Sticky region for the Standard Mapping (K=5) *(from [8])*

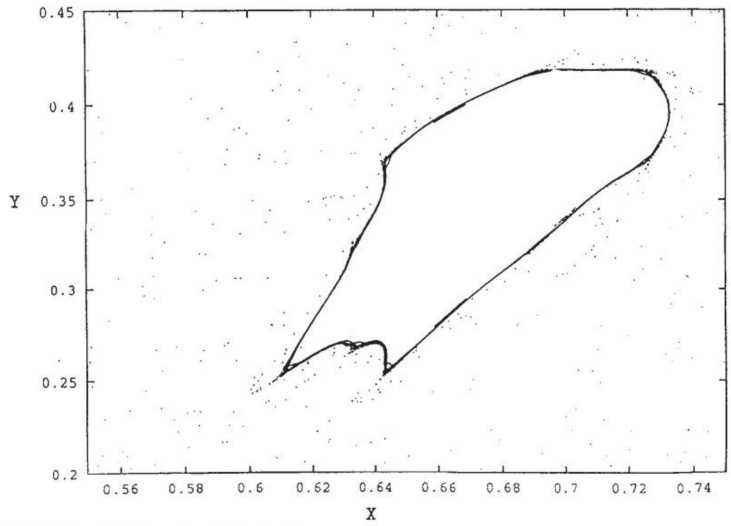

Fig. 17. Sticky regions for the Standard Mapping (K=5) *(from [8])*

3 Hamiltonian Systems

3.1 Introduction

In this section we will introduce two of the most important tools when dealing with dynamical systems: the *Lagrange equations* and the *Hamilton equations*. We will start with the Lagrange equations of first kind stemming from Newtonian axioms; out of this, the Lagrange equations of second kind will be derived. These equations can be used to obtain easily the equations of motions of any dynamical system if its kinetic and potential energy are known (This method will be applied in Sect. 5 to obtain the equations of motion for the restricted three body problem).

Then the powerful Hamilton formalism and the Hamilton equations will be introduced and explained. Throughout this book one will find many examples on how this formulation can help to investigate dynamical systems. We will show how canonical transformations can be used to simplify the system – appendix B gives an application of this method to the famous Hénon Heiles system which can be derived from the Toda-Lattice that will be discussed at the end of this section.

This derivation of the Lagrange and Hamilton equations closely follows the very good presentation of the topic in [22].

3.2 Lagrange Equations of the First Kind

Although the Newtonian axioms are very important, for certain problems it is not always possible to apply them directly. If we investigate a plain pendulum with a length l, moving on a circular path, because of the restriction due to the length it follows (in rectangular coordinates) that

$$x^2 + y^2 - l^2 = 0 \tag{70}$$

Thus, the thread induces a constraint force $\boldsymbol{C}$ and the second Newtonian axiom reads

$$m\ddot{\boldsymbol{r}} = \boldsymbol{F} + \boldsymbol{C} \tag{71}$$

where $\boldsymbol{r} = (x, y, z)$.

Because the force $\boldsymbol{C}$ in general depends on the actual motion it is not possible to obtain its value directly. To determine the constraint forces, we first write:

$$h_1(\boldsymbol{r}, t) = 0 \tag{72}$$

$$h_2(\boldsymbol{r}, t) = 0 \tag{73}$$

Equations (72)–(73) give the constraints. For the pendulum we would have

$$h_1(\boldsymbol{r}, t) = z = 0 \tag{74}$$

and

$$h_2(\boldsymbol{r}, t) = x^2 + y^2 - l^2 = 0 \tag{75}$$

Of course there can be only two constraint forces for one particle; three of them would determine the values of x, y and z and thus make any motion impossible. For more than one particle, equations (72)–(73) can be written as

$$h_i(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_N}, t) = 0 \tag{76}$$

with $i = 1, 2, \ldots, R$ and $R \leq 3N - 1$.

The constraints determine the *direction* of the constraint force $\boldsymbol{C}$. Using this property, equation (71) can be solved. A constraint restricts the motion of a particle on a surface – but not inside the surface. Thus, the constraint force has only a component orthogonal to the surface:

$$h(\boldsymbol{r}, t) = 0 \rightarrow C \mid\mid \nabla h(\boldsymbol{r}, t) \tag{77}$$

where ∇ is the Nabla operator. Now we can make the following ansatz for the constraint force $\boldsymbol{C}$:

$$\boldsymbol{C}(\boldsymbol{r}, t) = \alpha(t) \cdot \nabla h(\boldsymbol{r}, t) \tag{78}$$

where $\alpha(t)$ is a unknown function. Inserting (78) in (71) gives now the Lagrange equations of first kind:

$$m\ddot{\boldsymbol{r}} = \boldsymbol{F} + \alpha\,(t) \cdot \nabla h\,(\boldsymbol{r}, t) \tag{79}$$

and

$$h\,(\boldsymbol{r}, t) = 0 \tag{80}$$

We now have four equations and four unknown functions $(x(t), y(t), z(t), \alpha(t))$ – thus it is possible to solve the problem.

Equation (79) can be written in a different form, if one considers that the constraints determine a curve. The constraint force now has to be orthogonal to this curve. For $h_i(\boldsymbol{r}, t) = 0$, the $\nabla h_i(\boldsymbol{r}, t)$ are independent basis vectors. Thus, any force that has to be orthogonal to a surface determined by $h_i(\boldsymbol{r}, t) = 0$, $i = 1, \ldots, R$ can be written as

$$\boldsymbol{C}\,(\boldsymbol{r}, t) = \alpha_1(t)\nabla h_1(\boldsymbol{r}, t) + \ldots + \alpha_R(t)\nabla h_R(\boldsymbol{r}, t) \tag{81}$$

For N particles and $3N - 1$ constraints of the form (76) we can write (79) as

$$m_n\ddot{x}_n = F_n + \sum_{i=1}^{R} \frac{\partial h_i\,(x_1, \ldots, x_{3N}, t)}{\partial x_n} \tag{82}$$

and

$$h_i\,(x_1, \ldots, x_{3N}, t) = 0 \tag{83}$$

with $n = 1, 2, \ldots, 3N$ and $i = 1, 2, \ldots, R$.[13] These are now $3N + R$ equations for $3N + R$ unknown functions and therefore the equations can be solved. Equations (82) and (83) are called **Lagrange equations of the first kind**.

3.3 Lagrange Equations of Second Kind

The Lagrange equations of first kind (82)–(83) were defined only for cartesian coordinates. For R constraints, only

$$f = 3N - R \tag{84}$$

of the 3N cartesian coordinates are independent (f is called the number of *degrees of freedom*). We now select f *generalized coordinates*:

$$q_1, q_2, \ldots, q_n \tag{85}$$

The q_i have to be chosen in a way that the position of all particles is determined:

$$x_n = x_n(q_1, q_2, \ldots, q_f, t) \tag{86}$$

with $n = 1, 2, \ldots, 3N$. Additionally, all constraints have to be fulfilled for all values of the q_i:

[13] The first particle has the coordinates (x_1, x_2, x_3), the second one (x_4, x_5, x_6); the last one is given by $(x_{3N-2}, x_{3N-1}, x_{3N})$.

$$h_i\left(x_1\left(q_1,\ldots,q_f,t\right),\ldots,x_{3N}\left(q_1,\ldots,q_f,t\right),t\right)=0 \tag{87}$$

Using the generalized coordinates, we can now eliminate the constraint forces. The constraints in (87) do not depend on q_i because, for every value of q_i, $h_i = 0$. That means that the total derivative of h_i has to vanish:

$$\frac{dh_i}{dq_k}=0 \tag{88}$$

and thus

$$\sum_{n=1}^{3N}\frac{\partial h_i}{\partial x_n}\frac{\partial x_n}{\partial q_k}=0 \tag{89}$$

with $k = 1, 2, \ldots, f$. Multiplying (82) with $\frac{\partial x_n}{\partial q_k}$ and building the summation over n gives then

$$\sum_{n=1}^{3N} m_n\ddot{x}_n\frac{\partial x_n}{\partial q_k}=\sum_{n=1}^{3N}\boldsymbol{F_n}\frac{\partial x_n}{\partial q_k} \tag{90}$$

with $n = 1, 2, \ldots, 3N$.[14] Equation (90) now does not depend on the constraints or constraint forces.

We now want to simplify (90) furthermore and use the following notation:

$$x=(x_1,x_2,\ldots,x_{3N}) \tag{91}$$
$$q=(q_1,q_2,\ldots,q_f) \tag{92}$$
$$\dot{x}=(\dot{x}_1,\dot{x}_2,\ldots,\dot{x}_{3N}) \tag{93}$$
$$\dot{q}=(\dot{q}_1,\dot{q}_2,\ldots,\dot{q}_f) \tag{94}$$

Differentiation of (86) with respect to the time gives

$$\dot{x}_n=\frac{d}{dt}x_n(q,t)=\sum_{k=1}^{f}\frac{\partial x_n(q,t)}{\partial q_k}\dot{q}_k+\frac{\partial x_n(q,t)}{\partial t}=\dot{x}_n(q,\dot{q},t) \tag{95}$$

Thus it follows for $\dot{x}_n$

$$\frac{\partial\dot{x}_n(q,\dot{q},t)}{\partial\dot{q}_k}=\frac{\partial x_n(q,t)}{\partial q_k} \tag{96}$$

The kinetic energy (in cartesian coordinates) can be written as

$$T=T(\dot{x})=\sum_{n=1}^{3N}\frac{m_n}{2}\dot{x}_n^2 \tag{97}$$

If $x_n(q,t)$ does not explicitly depend on the time, equation (95) reduces to

$$\dot{x}_n=\sum_{k=1}^{f}\frac{\partial x_n(q)}{\partial q_k}\dot{q}_k \tag{98}$$

[14] Note that x_n stands for $x_n(q_1, q_2, \ldots, q_f, t)$.

Together with (97) it follows that

$$\frac{\partial T(q,\dot{q},t)}{\partial q_k} = \sum_{n=1}^{3N} m_n \dot{x}_n \frac{\partial \dot{x}_n}{\partial q_k} \tag{99}$$

and because of (96)

$$\frac{\partial T(q,\dot{q},t)}{\partial \dot{q}_k} = \sum_{n=1}^{3N} m_n \dot{x}_n \frac{\partial x_n}{\partial q_k} \tag{100}$$

Differentiation of (100) with respect to time gives[15]

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_k} = \sum_{n=1}^{3N} m_n \ddot{x}_n \frac{\partial x_n}{\partial q_k} + \sum_{n=1}^{3N} m_n \dot{x}_n \frac{\partial \dot{x}_n}{\partial q_k} \tag{101}$$

We now define the *generalized forces* Q_k:

$$Q_k = \sum_{n=1}^{3N} F_n \frac{\partial x_n}{\partial q_k} \tag{102}$$

Combining (101), (99) and (102) and inserting in the equations of motion (90) gives:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} = Q_k \tag{103}$$

with $k = 1, 2, \ldots, f$.

We now restrict the equations to forces F_n that originate from a potential U:

$$F_n = -\frac{\partial U(x)}{\partial x_n} \tag{104}$$

The generalized forces can now be written as derivatives of U

$$Q_k = \sum_{n=1}^{3N} F_n \frac{\partial x_n}{\partial q_k} = -\sum_{n=1}^{3N} \frac{\partial U(x)}{\partial x_n}\frac{\partial x_n}{\partial q_k} = -\frac{\partial U(q,t)}{\partial q_k} \tag{105}$$

where $U(q,t) = U(x_1(q,t), \ldots, x_n(q,t))$. Because $\frac{\partial U}{\partial \dot{q}_k} = 0$, (103) reads

$$\frac{d}{dt}\frac{\partial (T-U)}{\partial \dot{q}_k} = \frac{\partial (T-U)}{\partial q_k} \tag{106}$$

We now define the *Lagrange Function* L

$$L(q,\dot{q},t) = T(q,\dot{q},t) - U(q,t) \tag{107}$$

[15] Here we use that $\frac{d}{dt}\frac{\partial x_n}{\partial q_k} = \frac{\partial \dot{x}_n}{\partial q_k}$ which can be shown easily.

which is the difference between the kinetic and the potential energy. The final form of the **Lagrange equation of second kind** is now

$$\frac{d}{dt}\frac{\partial L(q,\dot{q},t)}{\partial \dot{q}_k} - \frac{\partial L(q,\dot{q},t)}{\partial q_k} = 0 \tag{108}$$

with $k = 1, 2, \ldots, f$.

Equation (108) is widely used in mechanics because it depends only on L – a scalar function – which can be easily found for most systems and thus the equations of motions can be obtained easily. In Sect. (5) we will give some examples on how to use equation (108).

The term $\frac{\partial L}{\partial \dot{q}_k}$ is also called the *generalized impulse* p_k:

$$\frac{\partial L(q,\dot{q},t)}{\partial \dot{q}_k} = p_k \tag{109}$$

3.4 The Hamilton Function

The Lagrangian notation uses the generalized coordinates and velocities q_i and $\dot{q}_i$. Now we want to use the generalized momenta p_i instead of $\dot{q}_i$. From equation (109) we get

$$\dot{q}_k = \dot{q}_k(q,p,t) \tag{110}$$

We now define the **Hamilton function**[16]:

$$H(q,p,t) = \sum_{i=1}^{f} \dot{q}_i\left(q,p,t\right) p_i - L\left(q,\dot{q}\left(q,p,t\right),t\right) \tag{111}$$

We can use the Lagrange equation (108) to obtain the partial derivatives of H:

$$\begin{aligned}
\frac{\partial H}{\partial q_k} &= \sum_{i=1}^{f} \frac{\partial \dot{q}_i}{\partial q_k} p_i - \frac{\partial L}{\partial q_k} - \sum_{i=1}^{f} \frac{\partial L}{\partial \dot{q}_i}\frac{\partial \dot{q}_i}{\partial q_k} \\
&= -\frac{\partial L}{\partial q_k} \\
&= -\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) \\
&= -\dot{p}_k
\end{aligned} \tag{112}$$

[16] Note that the Hamilton function is obtained from the Lagrange function by a Legendre transformation.

$$\frac{\partial H}{\partial p_k} = \sum_{i=1}^{f} \frac{\partial \dot{q}_i}{\partial p_k} p_i + \dot{q}_k - \sum_{i=1}^{f} \frac{\partial L}{\partial \dot{q}_i} \frac{\partial \dot{q}_i}{\partial p_k}$$
$$= \dot{q}_k \tag{113}$$

$$\frac{\partial H}{\partial t} = \sum_{i=1}^{f} \frac{\partial \dot{q}_i}{\partial t} p_i - \sum_{i=1}^{f} \frac{\partial L}{\partial \dot{q}_i} \frac{\partial \dot{q}_i}{\partial t} - \frac{\partial L}{\partial t}$$
$$= -\frac{\partial L}{\partial t} \tag{114}$$

Equations (112) and (113) are the **canonical** or **Hamiltonian equations**:

$$\dot{p}_k = -\frac{\partial H(q,p,t)}{\partial q_k} \tag{115}$$

$$\dot{q}_k = \frac{\partial H(q,p,t)}{\partial p_k} \tag{116}$$

These equations follow directly from the Lagrangian equations and thus they are equivalent to the equations of motion given by (108).

3.5 Canonical Transformations

Canonical transformations are a very important tool when one is dealing with Hamiltonian systems. These transformations are used to simplify the Hamiltonian and thus make it possible to find a solution of the Hamiltonian equations.[17]

The so-called *Hamiltonian principle* states that the variation of the action vanishes. The action S[18] is defined by

$$S(q) = \int_{t_1}^{t_2} L(q, \dot{q}, t) dt \tag{117}$$

One can replace the Lagrangian equations (108) with the condition that the variation of S has to vanish:

$$\delta S(q) = S(q + \delta q) - S(q) = 0 \tag{118}$$

where δq is an infinitesimal change of q.

Replacing L with H (equation (111)) gives

[17] An example for a canonical transformation can be found in the appendix, where we introduce the method of Lie transformations.

[18] For more information on the so-called *action–angle variables* see Sect. 3.2 in the chapter on "Regular and Chaotic Motion in Hamiltonian Systems".

$$\delta S(q) = \delta \int_{t_1}^{t_2} \left(\sum_{i=1}^{f} p_i \dot{q}_i - H(q,p,t) \right) dt = 0 \tag{119}$$

We assume now that

$$\begin{aligned} \delta q(t_1) &= 0 \\ \delta q(t_2) &= 0 \\ \delta p(t_1) &= 0 \\ \delta p(t_2) &= 0 \end{aligned} \tag{120}$$

and get[19]

$$\delta S[q,p] = \int_{t_1}^{t_2} \sum_{i=1}^{f} \left(\left(\dot{q}_i - \frac{\partial H}{\partial p_i} \right) \delta p_i - \left(\dot{p}_i + \frac{\partial H}{\partial q_i} \right) \delta q_i \right) dt \tag{121}$$

Because of the Hamiltonian equations (115) and (116) it follows that

$$\delta S[q,p] = 0. \tag{122}$$

H is determined up to the total time derivative of an arbitrary function $F(q,p,t)$ which can be added to (119) without affecting the canonical equations (115) and (116). The transformation

$$\sum_{i=1}^{f} p_i \dot{q}_i - H \rightarrow \sum_{i=1}^{f} p_i \dot{q}_i - H + \frac{d}{dt} F(q,p,t) \tag{123}$$

can now be used to change the form of the Hamiltonian.

We investigate the transformation from the old coordinates q_i and p_i to some new coordinates Q_k and P_k:

$$Q_k = Q_k(q_i, \ldots, q_f, p_i, \ldots, p_f, t) = Q_k(q,p,t) \tag{124}$$

$$P_k = P_k(q_i, \ldots, q_f, p_i, \ldots, p_f, t) = P_k(q,p,t) \tag{125}$$

Definition 6. *A transformation is called* canonical *if it does not change the form of the Hamiltonian equations:*

$$\dot{Q}_k = \frac{\partial H'}{\partial P_k} \tag{126}$$

$$\dot{P}_k = -\frac{\partial H'}{\partial Q_k} \tag{127}$$

where H' is now the transformed Hamiltonian.

[19] Note that $p_i \delta \dot{q}_i$ can be replaced with $\dot{p}_i \delta q_i$.

Applying the Hamiltonian principle we get

$$\delta \int_{t_1}^{t_2} \left(\sum_{i=1}^{f} p_i \dot{q}_i - H(q,p,t) \right) dt = \delta \int_{t_1}^{t_2} \left(\sum_{i=1}^{f} P_i \dot{Q}_i - H'(Q,P,t) \right) dt \quad (128)$$

From (123) it follows that

$$\sum_{i=1}^{f} p_i \dot{q}_i - H(q,p,t) = \sum_{i=1}^{f} P_i \dot{Q}_i - H'(Q,P,t) + \frac{d}{dt} F(q,p,Q,P,t) \quad (129)$$

The function $F(q,p,Q,P,t)$ is called the *generating function* of the transformation. In the following we will describe the process of canonical transformation with a generating function of the type $F(q,Q,t)$.[20] The additional term in (129) now becomes

$$\frac{dF(q,Q,t)}{dt} = \sum_{i=1}^{f} \frac{\partial F}{\partial q_i} \dot{q}_i + \sum_{i=1}^{f} \frac{\partial F}{\partial Q_i} \dot{Q}_i + \frac{\partial F}{\partial t} \quad (130)$$

With (129) we derive

$$p_k = \frac{\partial F(q,Q,t)}{\partial q_k} \quad (131)$$

$$P_k = -\frac{\partial F(q,Q,t)}{\partial Q_k} \quad (132)$$

$$H'(Q,P,T) = H(q,p,t) + \frac{\partial F(q,Q,t)}{\partial t} \quad (133)$$

Equation (133) shows the great advantage of a canonical transformation: one can choose an arbitrary function $F(q,Q,t)$ in such a way, that the new Hamiltonian H' becomes simpler. Because the Hamiltonian equations are not affected by the transformation, it is also simpler to solve the equations of motion!

3.6 Example: The Toda Lattice

As an example for an integrable Hamiltonian with more than one degree of freedom, we present the three-particle *Toda Lattice*.[21] It is given by

$$H = \frac{1}{2}\left(p_1^2 + p_2^2 + p_3^2\right) + e^{-(\phi_1 - \phi_3)} + e^{-(\phi_2 - \phi_1)} + e^{-(\phi_3 - \phi_2)} - 3 \quad (134)$$

This Hamiltonian describes the motion of three particles on a ring with exponentially decreasing repulsive forces between them (see Fig. 18).

[20] The process is similar for the other types of generating functions $(F(q,P,t), F(Q,p,t), F(p,P,t))$.

[21] For a detailed investigation of the Toda Lattice see also Sect. 3.4 in the chapter "Regular and Chaotic Motion in Hamiltonian Systems".

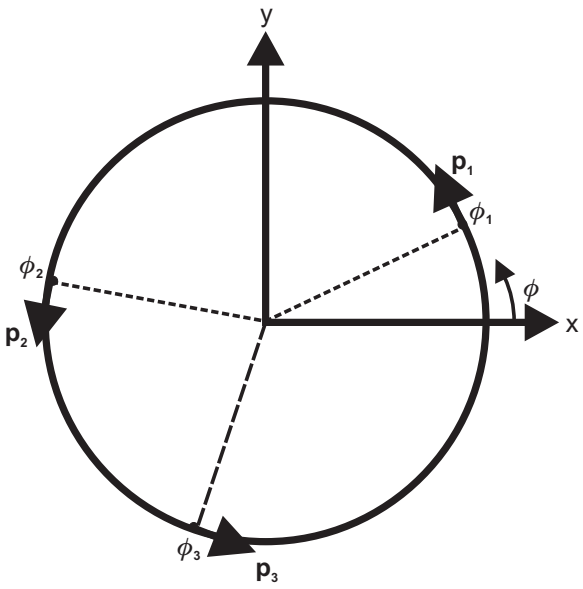

Fig. 18. The three particle Toda lattice

Of course the energy of this system has to be conserved; additionally the total momentum is constant:

$$P_3 = p_1 + p_2 + p_3 = \text{const.} \tag{135}$$

This has to be true because if we apply a rigid rotation $\phi_i \rightarrow \phi_i + \phi_0$, the Hamiltonian is not affected. This can be seen, when the Hamiltonian is transformed to the new momenta $P_1 = p_1$, $P_2 = p_2$ and P_3 given as above. When using the generating function

$$F = P_1\phi_1 + P_2\phi_2 + (P_3 - P_1 - P_2)\,\phi_3 \tag{136}$$

one obtains

$$H' = \frac{1}{2}\left(P_1^2 + P_2^2 + (P_3 - P_1 - P_2)^2\right) + e^{-\phi_1} + e^{-(\phi_2 - \phi_1)} + e^{\phi_2} - 3 \tag{137}$$

Here the ϕ's are canonical with the P's and H' is independent of ϕ_3 – thus P_3 is constant. We now choose a rotating coordinate system in which the total momentum equals zero which allows us to set $P_3 = 0$. In addition to the two known isolating integrals (energy, momentum) there seems no further obvious constant.

By applying the following transformation, we obtain a Hamiltonian in a form of a particle moving in a 2-D potential well.

$$F' = \left(4\sqrt{3}\right)^{-1}\left[\left(p'_x - \sqrt{3}p'_y\right)\phi_1 + \left(p'_x + \sqrt{3}p'_y\right)\phi_2\right] \tag{138}$$

and

$$p'_x = 8\sqrt{3}p_x \qquad x' = x \qquad p'_y = 8\sqrt{3}p_y \qquad y' = y \qquad \bar{H} = H'/\sqrt{3} \tag{139}$$

This gives the so-called *Toda Hamiltonian*:

$$\bar{H} = \frac{1}{2}\left(p_x^2 + p_y^2\right) + \frac{1}{24}\left[e^{2y+2\sqrt{3}x} + e^{2y-2\sqrt{3}x} + e^{-4y}\right] - \frac{1}{8} \tag{140}$$

Figure 19 shows the potential curves; they vary smoothly and show a ternary symmetry.

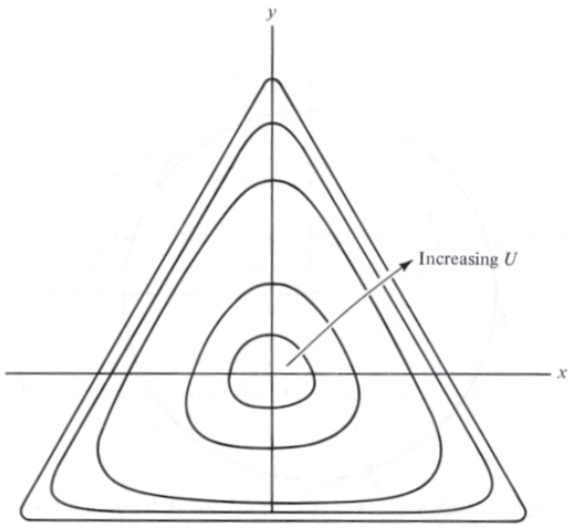

Fig. 19. The potential curves for the Toda Hamiltonian; shown are the lines for constant U *(taken from [42])*

We can expand equation (140) to 3rd order and obtain the now **non-integrable Hénon and Heiles Hamiltonian**:

$$\bar{H}' = \frac{1}{2}\left(p_x^2 + p_y^2 + x^2 + y^2\right) + x^2 y - \frac{1}{3}y^3 \tag{141}$$

3.7 Hénon Heiles System

We will now describe shortly the former derived system which played an outstanding role in the exploration of nonlinear dynamics: it was the first numerical proof for the existence of chaotic motion in a dynamical system [29].

The motion of a star in the gravitational potential $U_{\rm g}(R,z)$ of a galaxy corresponds to a trajectory in the 6-dimensional phase-space described by the polar coordinates (R,Θ,z,$\dot{R}$,$\dot{\Theta}$,$\dot{z}$). In a 6-d phase-space there exist five independent conservative integrals of motion: $I_{\rm j} = C_{\rm j}$ with $j = 1,\ldots,5$ where the $C_{\rm j}$ are constants. Each integral can be *isolating* or *non-isolating*. The hypersurface corresponding to a non-isolating integral consists of an infinity of sheets which fill the phase-space very densely; thus $I_{\rm j} = C_{\rm j}$ does not give any useful (for physical applications) information. Therefore the non-isolating integrals are ignored and only the isolating integrals are called "integrals'. For a very detailed explanation of the integrals of motion in galactic dynamics see [5].

In this case, two isolating integrals are known:

$$I_1 = U_{\rm g}(R,z) + \frac{1}{2}\left(\dot{R}^2 + \dot{R}\dot{\Theta}^2 + \dot{z}^2\right) \tag{142}$$

$$I_2 = R^2\dot{\Theta}^2 \tag{143}$$

corresponding to the total energy and the angular momentum per unit mass of the star around the z-axis. It can be shown, that I_4 and I_5 are non-isolating in general. But there are no informations about the nature of the third integral.

The properties of any potential third integral can be investigated numerically [5]. For that purpose, one introduces the potential

$$U(R,z) = U_{\rm g}(R,z) + \frac{C_2}{2R^2} \tag{144}$$

where C_2 is the constant value of the angular momentum. This problem is equivalent to the problem of the motion of a particle in a plane in an arbitrary potential U. If one substitutes x and y for R and z, the phase-space $(x,y,\dot{x},\dot{y})$ has four dimensions and there exist three independent conservative integrals. One of them is the known isolating integral

$$I_1 = U(x,y) + \frac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) = E \tag{145}$$

which corresponds to the total energy of the star divided by its mass. The third integral is generally non-isolating – but nothing is known about the second integral.

We now investigate the fore mentioned potential by means of a **surface of section**: because of the existence of the energy integral E, one of the four coordinates can be substituted. Thus, the trajectory of a star for a given energy can be investigated in the three dimensional phase-space $(x,y,\dot{y})$. Then we consider the successive intersections of the trajectory with the plane $x = 0$ (in the upward direction) – this plane $(y,\dot{y})$ is called the surface of section (SOS). If there is no other isolating integral, the trajectory will fill a bounded volume and is called *ergodic*; the successive points on the SOS

will fill an area. But if there exists a second isolating integral, the trajectory will lie on a surface; on the SOS the successive points will lie on an invariant curve.

The potential for the Hénon Heiles system reads

$$U(x,y) = \frac{1}{2}\left(x^2 + y^2 + 2x^2y - \frac{2}{3}y^3\right) \tag{146}$$

which corresponds to the potential of the third order Toda-Hamiltonian (equation 141).

Figures 20a to f show the SOS for different values of the energy. If the energy is very low, the whole phase-space consists of closed curves – so it seems that an additional isolating integral exists. But if the energy is increased, chaotic regions appear and the regular areas shrink. Finally, for very large values of the energy, the whole phase space seems to be chaotic.

So it is impossible to answer the question, if there exists a third integral of motion definitely. If the energy is small, the third integral seems to exist always. If the energy is increased to values larger than a certain critical energy we find an infinite number of regions in phase-space (seperated from each other) where a third integral exists; if we increase the energy further it seems that there is no third integral. For a detailed discussion we refer to the extended literature on this nonlinear dynamical system (for a review see [5]).

4 The Two-Body Problem

4.1 Historical Remarks

For the ancient Babylonian, Greek and Egyptian astronomers the major interest was to predict the positions of the Sun, the Moon and the planets on the celestial sphere, because the main tasks for them were to provide an exact calendar and the precise determination of the times of the eclipses in advance. This was accomplished with long and difficult observations to detect the different periods of the motion of the celestial bodies on the sky. The believed structure of the universe for more than one and a half thousand years was written down in the *Almagest* of Ptolemäus (100–160): the Earth is in the center, the sphere of the stars is rotating around it, and the planets, the Sun and the Moon are separately moving on perfect circles on other spheres.[22]

From the point of view of understanding the nature of physics an important step forward was the change of the "center" of the world from the Earth to the Sun, namely the change from the geocentric to the heliocentric system by Nikolaus Copernicus (1473–1543). This time coincides also with the

[22] A poetic description of the structure of the universe as it was known in the middle ages can be found in the "divina comedia" by Dante (1265–1321).

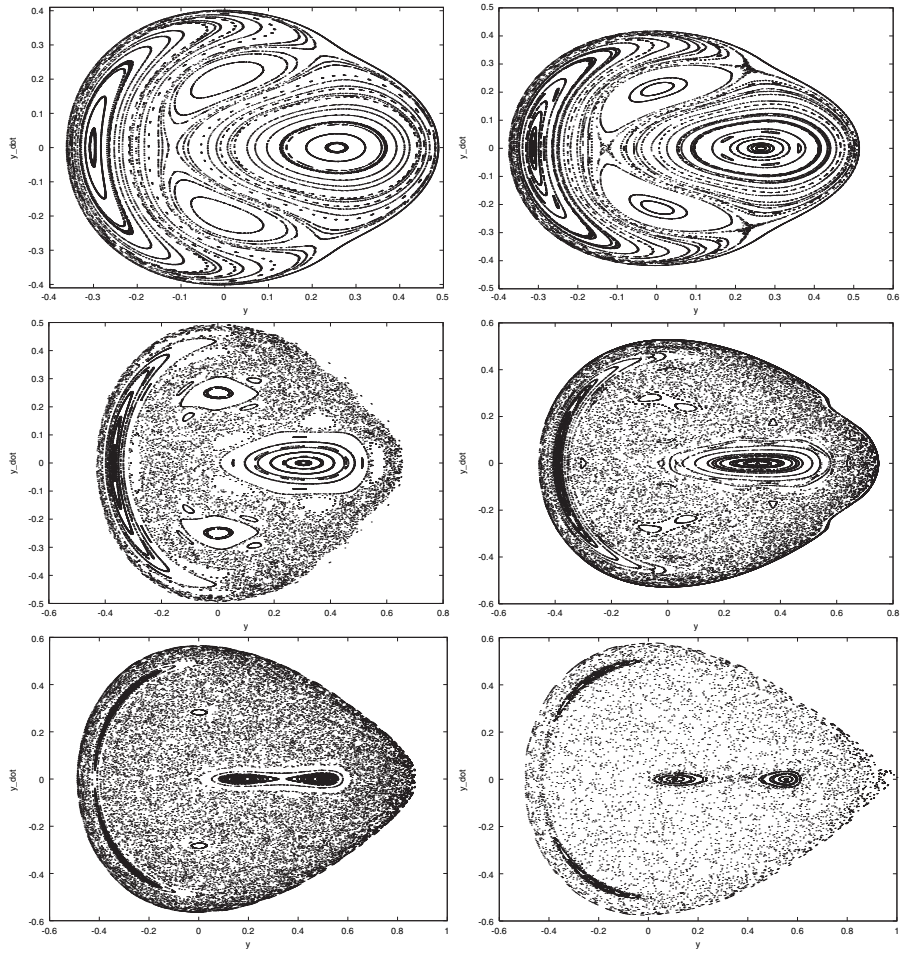

Fig. 20. SOS of the Hénon-Heiles system for an energy E = **(a)** 0.08333; **(b)** 0.09; **(c)** 0.125; **(d)** 0.14; **(e)** 0.16 and **(f)** 0.1675

discovery of America, which is seen as the end of the "dark" middle ages. Another finding can be regarded also as an essential change in the human mind, namely that the planets do not fulfill the perfect "divine" circular motion. It was Johannes Kepler (1571–1630) who found that the motion of Mars is an ellipse, with the Sun in its focus. He did computations as successor of Tycho Brahe (1546–1601) an imperial mathematician on the court of Rudolf II in Prag. Two facts were favorable for Kepler's discovery: first Tycho was a very careful observer and provided the best observations ever done up to his time, second Mars has the most eccentric orbit besides Mercury and therefore the deviations from the circle are the most prominent ones. In his *Astronomia Nova*, published in 1609, he formulated his laws of the motion of the planets,

Fig. 21. Johannes Kepler *(Picture courtesy of Sternwarte Kremsmünster, Upper Austria)*

which are the basis of all future work in Celestial Mechanics. It should be mentioned that Kepler was quite close to the idea that the cause of motion is a force acting between two celestial bodies, but he did not find the law behind!

In this short introductionary section we should also mention the great Italian, who was always in war with the christian authorities in Rome, namely Galileo Galilei (1564–1642). He was the one who discovered, with the aid of the telescope, that Jupiter can be regarded as being the center of a small planetary system with its four satellites. With this discovery the Copernican heliocentric system was finally accepted (with the exception of the Roman Catholic church).[23]

The final step forward to understanding the motion of the planets, the Sun and the Moon was the discovery of the universal law of gravitation by Isaac Newton (1642–1717), who stands at the beginning of the epoch of

[23] The Vatican waited until 1992 to accept the work of Galileo officially!

deterministic physics, where everything can be determined when we know the laws governing it.[24]

Why is the knowledge of the two-body motion of such importance? There are different reasons for that:

1. the motion of a single planet around the Sun (the two-body problem) is the only astrodynamical problem where we have a complete and general solution (although it is not as simple as it looks like) besides very special cases of the three-body problem not exactly realized in nature.
2. for many problems in the dynamics of celestial bodies it is a very good first approximation.
3. it is the starting point of analytical theories from which we can develop then the solution to higher orders with respect to small parameters involved like the eccentricity, the inclination of the orbit and the small masses of the other planets, which perturb the elliptic motion of a planet around the Sun (see Sect. 7)

4.2 The First and Second Law of Kepler

The motion in a two-body system is governed by the three Kepler Laws. This section will give a kinematic description of the first and second law, that were found by Kepler by the means of observation.[25]

Because the motion of a planet takes place on an ellipse (Fig. 22), the **first law of Kepler** can be written as

$$r = \frac{p}{1 + e \cos v} \tag{147}$$

where r is the distance between the planet and the central star, e the eccentricity of the ellipse and v the true anomaly. From Fig. 22, the following properties for the configuration parameters can be derived easily:

$$\begin{aligned} a &= \frac{p}{1 - e^2} \\ b &= \frac{p}{\sqrt{1 - e^2}} \\ p &= \frac{b^2}{a} \\ e &= \sqrt{1 - \frac{b^2}{a^2}} \end{aligned} \tag{148}$$

The **second law of Kepler** can be derived by investigating Fig. 23 when one takes into account the motion of a planet along its orbit:

[24] This belief ended in the 20$^{\text{th}}$ century with the discovery of quantum physics and also the finding of chaotic behavior of orbits in dynamical systems, which makes long time predictions impossible.

[25] In this chapter we follow mainly the extensive book by Stumpf [63]

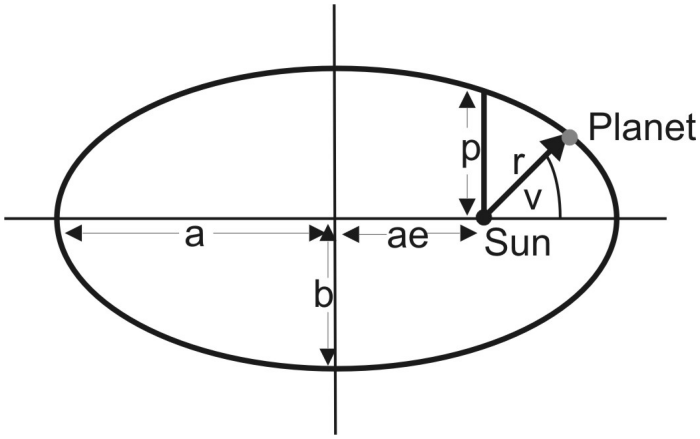

Fig. 22. Configuration of the two-body problem

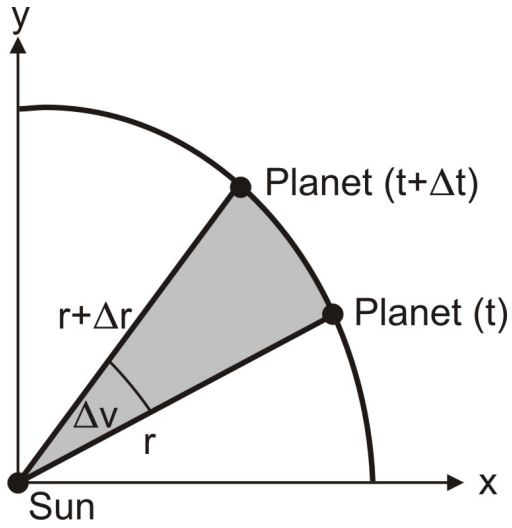

Fig. 23. Motion of a planet along the ellipse

The area covered by the connecting line between the planet and the central body is given by:

$$F_\Delta = \frac{1}{2} r \left(r + \Delta r\right) \sin \Delta v \tag{149}$$

If Δv and Δr are small, this can be written as

$$F_\Delta \sim \frac{1}{2} r^2 \Delta v \tag{150}$$

Kepler discovered that the value of $r^2 \Delta v$ is constant for equal time intervals:

$$r^2 \Delta v \sim c \Delta t \tag{151}$$

or, using differential notation

$$r^2 \frac{dv}{dt} = c \tag{152}$$

We can now derive the value of the constant by introducing rectangular coordinates:

$$\begin{aligned} x &= r \cos v \\ y &= r \sin v \\ r &= \sqrt{x^2 + y^2} \\ \tan v &= \frac{y}{x} \end{aligned} \tag{153}$$

By implicit differentiation we can express the derivate of v:

$$\frac{\dot{v}}{\cos^2 v} = \frac{\dot{v} r^2}{x^2} = \frac{\dot{y}x - y\dot{x}}{x^2} \tag{154}$$

and thus

$$r^2 \frac{dv}{dt} = r^2 \dot{v} = \dot{y}x - y\dot{x} = c \tag{155}$$

We can get a different representation of the constant by starting with the following formula for the area in Fig. 23:

$$\Delta f = \frac{ab\pi}{N} \tag{156}$$

where $ab\pi$ is the area of the ellipse and N the number of sections. By introducing the orbital period U one obtains with (151)

$$\frac{ab\pi}{N} = \frac{1}{2} c \Delta t = \frac{1}{2} c \frac{U}{N} \tag{157}$$

and thus

$$c = \frac{2\pi ab}{U} = nab = \frac{nap}{\sqrt{1 - e^2}} \tag{158}$$

where

$$n = \frac{2\pi}{U} \tag{159}$$

is the mean motion of the body which is related to the mean anomaly M (see Fig. 24) by

$$M = nt \tag{160}$$

Equation (158) can be transformed by using (148):

$$c = nab = n\sqrt{pa^3} = na^2 \sqrt{1 - e^2} = \frac{nap}{\sqrt{1 - e^2}} \tag{161}$$

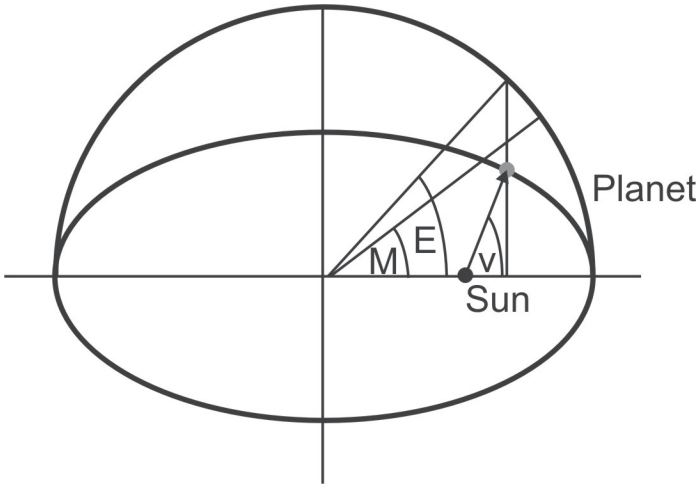

Fig. 24. Relation between true, mean and eccentric anomaly. Note that the "fictitious angle" $M = nt$ cannot be found in an exact way by geometrical construction

To obtain the true anomaly from the mean anomaly, we first have to define the eccentric anomaly E (see Fig. 24). We define $E = 0$ if the planet is in perihelion position ($r = a(1 - e)$) and $E = \pi$ if the planet is in aphelion position ($r = a(1 + e)$). This leads to the following (indirect) definition of E:

$$r = a\,(1 - e\cos E) \tag{162}$$

We start by transforming the first law of Kepler (147) and use the relations (148) and (162) to obtain an expression for $x = r\cos v$:

$$r + er\cos v = p \tag{163}$$

$$re\cos v = p - r = a\left(1 - e^2\right) - a\,(1 - e\cos E) \tag{164}$$

$$r\cos v = a\,(cosE - e) \tag{165}$$

The expression for $y = r\sin v$ can be obtained by

$$r^2 - r^2\cos^2 v = r^2\sin^2 v = a^2\left(\left(1 + e^2\cos^2 E\right) - \left(\cos^2 E + e^2\right)\right) \tag{166}$$

This simplifies to

$$r^2\sin^2 v = a^2\left(1 - \cos^2 E\right)\left(1 - e^2\right) = a^2\sin^2 E\left(1 - e^2\right) \tag{167}$$

and thus

$$r\sin v = a\sqrt{1 - e^2}\sin E \tag{168}$$

Now making use of (165) and (162) one finds that

$$\cos v = \frac{\cos E - e}{1 - e\cos E} \tag{169}$$

and find a relation between E and v

$$1 - \cos v = \frac{(1+e)(1-\cos E)}{1 - e\cos E}$$
$$1 + \cos v = \frac{(1-e)(1+\cos E)}{1 - e\cos E} \tag{170}$$

which simplifies via the formula

$$\tan\frac{\alpha}{2} = \frac{1-\cos\alpha}{1+\cos\alpha} \tag{171}$$

to

$$\tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2} \tag{172}$$

The next step consists in finding a relation between eccentric anomaly E and mean anomaly M. From (147) the derivate of r can be obtained:

$$dr = \frac{pe\sin v}{(1+e\cos v)^2}dv = \frac{r^2 e}{p}\sin v dv \tag{173}$$

Because of the second law of Kepler (151)

$$r^2 dv = cdt = \frac{nap}{\sqrt{1-e^2}}dt \tag{174}$$

it follows that

$$dr = \frac{na}{\sqrt{1-e^2}}e\sin v dt \tag{175}$$

Equation (162) gives for dr

$$dr = ae\sin E dE \tag{176}$$

Combining (168), (175) and (176) gives

$$dE = \frac{na}{r}dt \tag{177}$$

or

$$(1 - e\cos E)\, dE = ndt \tag{178}$$

and, after integrating, one obtains

$$E - e\sin E = n\,(t - t_0) \tag{179}$$

and because of (160)

$$E - e\sin E = M \tag{180}$$

which is the well-known Kepler equation that cannot be solved in closed form.[26]

[26] There exist literally hundreds of different ways to solve it numerically and even today new algorithms are found that make the evaluation more efficient.

4.3 From Kepler to Newton

By using the properties above, it is now possible to derive the *acting forces* between the two bodies. For this purpose, the force is divided in a radial part $\boldsymbol{r}$ and a tangential part $\boldsymbol{t}$ (the configuration can be seen in Fig. 25) which are perpendicular to each other and can be described by

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{i}\cos\phi + \boldsymbol{j}\sin\phi \\ \boldsymbol{t} &= -\boldsymbol{i}\sin\phi + \boldsymbol{j}\cos\phi \end{aligned} \tag{181}$$

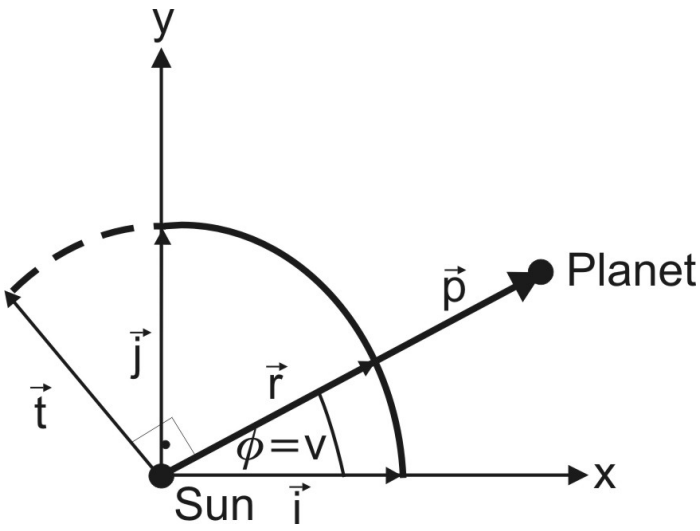

Fig. 25. Forces acting between the two bodies

According to Fig. 25 the acting force is proportional to $\ddot{\boldsymbol{p}}$. Because of

$$\boldsymbol{p} = r\boldsymbol{r} \tag{182}$$

(where $\boldsymbol{r}$ is the unit vector in the r-direction) we have

$$\ddot{\boldsymbol{p}} = \ddot{r}\boldsymbol{r} + 2\dot{r}\dot{\boldsymbol{r}} + r\ddot{\boldsymbol{r}} \tag{183}$$

From (181) it follows that

$$\begin{aligned} \dot{\boldsymbol{r}} &= \left(-\boldsymbol{i}\sin\phi + \boldsymbol{j}\cos\phi\right)\dot{\phi} \\ &= \dot{\phi}\boldsymbol{t} \\ \ddot{\boldsymbol{r}} &= \left(-\boldsymbol{i}\sin\phi + \boldsymbol{j}\cos\phi\right)\ddot{\phi} - \left(\boldsymbol{i}\cos\phi + \boldsymbol{j}\sin\phi\right)\dot{\phi}^2 \\ &= \ddot{\phi}\boldsymbol{t} - \dot{\phi}^2\boldsymbol{r} \end{aligned} \tag{184}$$

This leads to

$$\ddot{\boldsymbol{p}} = \left(\ddot{r} - r\dot{\phi}^2\right)\boldsymbol{r} + \left(2\dot{r}\dot{\phi} + \ddot{\phi}r\right)\boldsymbol{t} \tag{185}$$

where now the vector $\ddot{\boldsymbol{p}}$ has been decomposed in its radial and tangential components $\boldsymbol{r}$ and $\boldsymbol{t}$. We can now use the first and second law of Kepler to obtain expressions for $\dot{r}$,$\ddot{r}$,$\dot{\phi} = \dot{v}$ and $\ddot{\phi} = \ddot{v}$:

$$\dot{v} = \frac{c}{r^2} \tag{186}$$

$$\dot{r} = \frac{pe\sin v}{(1 + e\cos v)^2}\dot{v} = c\frac{e}{p}\sin v \tag{187}$$

$$\ddot{r} = \frac{c^2 e}{pr^2}\cos v \tag{188}$$

$$\ddot{v} = -\frac{2}{r^3}c\dot{r} \tag{189}$$

Equations (186)–(189) can be used together with (185) to obtain an expression for the tangential part $\boldsymbol{t}$ (please note that v and ϕ are identical):

$$2\dot{r}\dot{\phi} + \ddot{\phi}r = 0 \tag{190}$$

thus **there are no tangential forces acting in the two body problem**.

The acceleration in the radial direction can be derived by substituting (147) in (185):

$$\ddot{r} - r\dot{\phi}^2 = \frac{c^2}{r^2}\left(\frac{e}{p}\cos v - \frac{1}{r}\right) = -\frac{c^2}{pr^2} \tag{191}$$

It can be seen that the acting acceleration is *negative* and thus *attracting*; it is also proportional to $\frac{1}{r^2}$. This is the famous *law of gravitation* that was established by Isaac Newton one hundred years after the discoveries of Kepler which reads

$$\ddot{\boldsymbol{p}} = -\frac{c^2}{pr^2}\boldsymbol{r} = -\frac{c^2}{pr^3}\boldsymbol{p} \tag{192}$$

where p is the parameter of the ellipse defined above (148).

The fact, that the motion in the two body system is only governed by a central force, leads directly to the second law of Kepler: because of (190)

$$\frac{2\dot{r}}{r} + \frac{\ddot{\phi}}{\dot{\phi}} = 0 \tag{193}$$

By integration one gets

$$2\log r + \log\dot{\phi} = \log c \tag{194}$$

or

$$r^2\dot{\phi} = c \tag{195}$$

which is identical to (151). As a consequence the second law of Kepler is valid for all central force fields.

4.4 From Newton to Kepler

By including also the masses involved, we can write the differential equation describing the motion in the two body problem by starting with the second law of motion of Newton:

> "The acceleration of an object as produced by a net force is directly proportional to the magnitude of the net force, in the same direction as the net force, and inversely proportional to the mass of the object"

$$\boldsymbol{F} = m\ddot{\boldsymbol{p}} = -k^2 \frac{Mm}{r^2}\boldsymbol{r} \tag{196}$$

The gravitational force on M (the Sun) by m (the planet) and vice versa on m by M is therefore given by (where $\boldsymbol{r}$ is the unit vector between the Sun and the planet - see Fig. 26)

$$\begin{aligned} \boldsymbol{F_s} &= M\ddot{\boldsymbol{q}}_s = -k^2 \frac{Mm}{r^2}\boldsymbol{r} \\ \boldsymbol{F_p} &= m\ddot{\boldsymbol{q}}_p = k^2 \frac{Mm}{r^2}\boldsymbol{r} \end{aligned} \tag{197}$$

which reads with $\boldsymbol{p} = \boldsymbol{q}_p - \boldsymbol{q}_s$ because of Newton's third law of motion

> "For every action, there is an equal and opposite reaction."

$$\ddot{\boldsymbol{p}} = -k^2 (M + m) \frac{\boldsymbol{p}}{r^3} \tag{198}$$

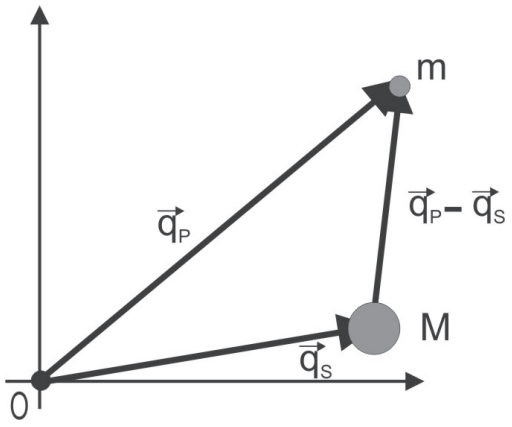

Fig. 26. Configuration of the Sun and planet in an inertial frame

where the constant k still has to be determined. This is a three dimensional differential equation of second order – thus one needs 6 constants of motion to solve the equation.

For an easier treatment of the equations, we first set $k^2\,(M+m) = \kappa^2$. Also, in the following, squared brackets $[.,.]$ will denote vectorial multiplications.

Multiplying (198) with $\boldsymbol{p}$ gives

$$[\boldsymbol{p}, \ddot{\boldsymbol{p}}] = -\frac{\kappa^2}{r^3}\,[\boldsymbol{p}, \boldsymbol{p}] = 0 \tag{199}$$

Thus we introduce the constant angular momentum vector $\boldsymbol{g}$ which is perpendicular to the plane of motion given by

$$\boldsymbol{g} = [\boldsymbol{p}, \dot{\boldsymbol{p}}] \tag{200}$$

and $\mid \boldsymbol{g} \mid = c$.

The three components of this vector deliver the first three constants of motion. An additional constant can be found by

$$[\ddot{\boldsymbol{p}}, \boldsymbol{g}] = -\kappa^2 \frac{[\boldsymbol{p}, \boldsymbol{g}]}{r^3} = -\kappa^2 \frac{[\boldsymbol{p}, [\boldsymbol{p}, \dot{\boldsymbol{p}}]]}{r^3} \tag{201}$$

By solving the multiple vectorial multiplication one obtains

$$[\ddot{\boldsymbol{p}}, \boldsymbol{g}] = -\frac{\kappa^2}{r^3}\,((\dot{\boldsymbol{p}}\boldsymbol{p})\,\boldsymbol{p} - (\boldsymbol{p}\boldsymbol{p})\,\dot{\boldsymbol{p}}) \tag{202}$$

Substituting $\dot{\boldsymbol{p}}\boldsymbol{p}$ with $\dot{r}r$ and $\boldsymbol{p}\boldsymbol{p}$ with r^2 gives

$$[\ddot{\boldsymbol{p}}, \boldsymbol{g}] = -\frac{\kappa^2}{r^3}\left((\dot{r}r)\,\boldsymbol{p} - r^2\dot{\boldsymbol{p}}\right) = \frac{\kappa^2}{r^2}\,(r\dot{\boldsymbol{p}} - \dot{r}\boldsymbol{p}) \tag{203}$$

and thus

$$[\ddot{\boldsymbol{p}}, \boldsymbol{g}] = \kappa^2 \frac{d}{dt}\left(\frac{\boldsymbol{p}}{r}\right) \tag{204}$$

Because of

$$\frac{d}{dt}\,[\dot{\boldsymbol{p}}, \boldsymbol{g}] = [\ddot{\boldsymbol{p}}, \boldsymbol{g}] + [\dot{\boldsymbol{p}}, \dot{\boldsymbol{g}}] \tag{205}$$

with (204) one obtains by integrating

$$[\dot{\boldsymbol{p}}, \boldsymbol{g}] = \frac{\kappa^2}{r}\boldsymbol{p} + \boldsymbol{f} \tag{206}$$

The constant vector $\boldsymbol{f}$ stemming from the integration is the so-called Laplace vector, which is perpendicular to $\boldsymbol{g}$ and has $\mid \boldsymbol{f} \mid = d = const$. Because $[\boldsymbol{g}, \boldsymbol{f}] = 0$, this gives 2 more constants of motion.

Now we can write

$$(\boldsymbol{p}, [\dot{\boldsymbol{p}}, \boldsymbol{g}]) = \frac{\kappa}{r}(\boldsymbol{pp}) + (\boldsymbol{pf}) \tag{207}$$

Because of

$$(\boldsymbol{p}\,[\dot{\boldsymbol{p}}, \boldsymbol{g}]) = (\boldsymbol{g}\,[\dot{\boldsymbol{p}}, \boldsymbol{p}]) = (\boldsymbol{gg}) = c^2 \tag{208}$$

and $(\boldsymbol{pf}) = rd\cos(\phi - \phi_0)$ it follows that

$$c^2 = r\left(\kappa^2 + d\cos(\phi - \phi_0)\right) \tag{209}$$

and

$$r = \frac{c^2}{(\kappa^2 + d\cos(\phi - \phi_0))} \tag{210}$$

This can also be written as

$$r = \frac{\frac{c^2}{\kappa^2}}{\left(1 + \frac{d}{\kappa^2}\cos(\phi - \phi_0)\right)} \tag{211}$$

which is identical to the first law of Kepler (147) with

$$p = \frac{c^2}{\kappa^2} \qquad\qquad e = \frac{d}{\kappa^2} \tag{212}$$

Substituting into the second law of Kepler ($r^2 d\phi = cdt$) gives

$$\frac{p^2}{(1 + e\cos(\phi - \phi_0))^2} d\phi = cdt \tag{213}$$

and, after integrating

$$p^2 \int_{\phi_0}^{\phi} \frac{d\phi}{(1 + e\cos(\phi - \phi_0))^2} = c(t - t_0) \tag{214}$$

t_0, the perihelion time (the moment, when planet is in its perihelion), is now the sixth and last constant of motion.

According to (212), (198) can also be written as

$$\ddot{\boldsymbol{p}} = -\kappa^2 \frac{\boldsymbol{r}}{r^2} = -\frac{c^2}{r^3 p}\boldsymbol{p} \tag{215}$$

with

$$\kappa^2 = \frac{c^2}{p} \rightarrow c = \kappa\sqrt{p} \tag{216}$$

From (158) and (148) it follows that

$$c^2 = n^2 a^3 p \tag{217}$$

and because of (159) we get

$$\kappa^2 = n^2 a^3 = \frac{4\pi^2}{U^2} a^3 \tag{218}$$

which gives the third law of Kepler:

$$\frac{a_i^3}{U_i^2} = \frac{\kappa^2}{4\pi^2} \tag{219}$$

where the index i stands now for the i-th body.

Comparing equation (215) with (198) shows that the constant κ is in fact the quantity introduced above

$$\kappa^2 = k^2 (M + m) \tag{220}$$

but the actual value of k is still unknown and can only be determined by astronomical observations or experimental measurements. The value[27] of k is identical with the *Gaussian constant* G determined as

$$G = 6.672 \cdot 10^{-11} \mathrm{m}^3 \mathrm{kg}^{-1} \mathrm{s}^{-2} \tag{221}$$

4.5 Motion in the Space and Computation of Ephemerides

Until now, we did not take into account, that the motion of a planet takes place in the three dimensional space, with three degrees of freedom. It was not necessary, because one can reduce it to a motion on a plane, via an adequate change of variables. In three dimension we need 6 orbital elements, which are the following three action-like variables:

1. semimajor axis **a**
2. eccentricity of the orbit **e**
3. inclination **i** (of the orbital plane, with respect to an inertial system)

and the three angle-like variables

1. the longitude of the ascending node **Ω**
2. the argument of the perihelion ω
3. the true anomaly **v**.

These elements are plotted in Figs. 27 and 22.

We already discussed in detail the three Kepler laws, which govern the motion of a planet. What we are interested in is now the position of a planet on the celestial sphere. To compute from the given orbital elements for a given time t the right ascension α and declination δ of a planet (asteroid) we need to compute step by step:

1. the eccentric anomaly E from the mean anomaly M of the celestial body for the instant of time we want to know via the Kepler equation

$$E - e \sin E = n(t - T) = M(t_0) + n(t - t_0) \tag{222}$$

[27] if measured in years, AU and Sun's mass, $k^2 = 0.01720209895$

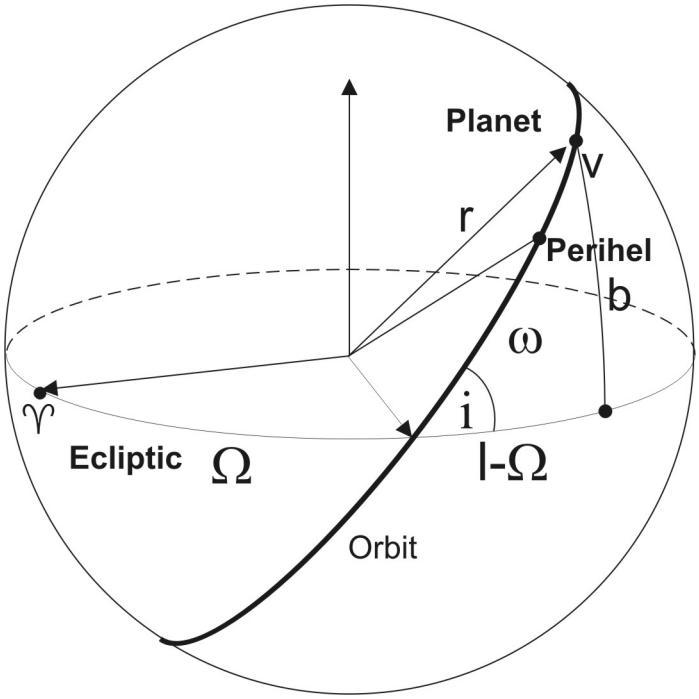

Fig. 27. The orbital elements

2. the distance r, from the Sun is

$$r = a(1 - e\cos E) \tag{223}$$

3. the true anomaly v giving the position in the orbit

$$\tan\frac{v}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2} \tag{224}$$

4. the so-called argument of the latitude $u = \omega + v$
5. the heliocentric polar coordinates r, l, b (see Fig. 27)

$$\begin{aligned}\cos b\cos l &= \cos u\cos\Omega - \sin u\sin\Omega\cos i\\ \cos b\sin l &= \cos u\sin\Omega - \sin u\cos\Omega\cos i\\ \sin b &= \sin u\sin i\end{aligned} \tag{225}$$

6. the rectangular heliocentric coordinates x, y and z

$$\begin{aligned}x &= r(\cos u\cos\Omega - \sin u\sin\Omega\cos i)\\ y &= r(\cos u\sin\Omega - \sin u\cos\Omega\cos i)\\ z &= r\sin u\sin i\end{aligned} \tag{226}$$

7. which we need to transform into the equatorial heliocentric coordinates, taking into account the variable obliquity of the ecliptic $\varepsilon \sim 23$ degrees[28]

$$\begin{aligned} \bar{x} &= x \\ \bar{y} &= y\cos\varepsilon - z\sin\varepsilon \\ \bar{z} &= z\cos\varepsilon - y\sin\varepsilon \end{aligned} \tag{227}$$

8. the geocentric equatorial coordinates making use of the published coordinates of the Sun $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ (from the Nautical Almanac)

$$\begin{aligned} \bar{\xi} &= \bar{x} + \bar{X} \\ \bar{\eta} &= \bar{y} + \bar{Y} \\ \bar{\zeta} &= \bar{z} + \bar{Z} \end{aligned} \tag{228}$$

9. via a final transformation the right ascension α and declination δ of the planet

$$\begin{aligned} \tan\alpha &= \left(\frac{\bar{\eta}}{\bar{\xi}}\right) \\ \tan\delta &= \left(\frac{\bar{\zeta}}{\sqrt{\bar{\xi}^2 + \bar{\eta}^2}}\right) \end{aligned} \tag{229}$$

The whole procedures sketched here are nowadays written in computer programs, which take into account all – in the foregoing schema neglected – effects like aberration etc. In this short section we only wanted to demonstrate how one can, in principle, compute ephemerides of a body when the six Keplerian elements are known. For a very detailed and profound description we refer to the Nautical Almanac, which is published every year by the US Naval observatory.

5 The Restricted Three Body Problem

The history of the restricted problem (by which we mean the circular restricted three body problem = R3BP) dates back to Leonhard Euler,[29] who worked on a Lunar theory. His main contribution to the R3BP was the introduction of a synodic coordinate system, where the two massive bodies have fixed positions. He also solved a special case of the R3BP – the so-called two fixed center problem, where two fixed centers of force act on a third one.[30]

[28] The actual value for a given date has to be computed from the respective formula given in the Nautical Almanac.

[29] Leonhard Euler (1707 – 1783)

[30] Later in this chapter we will see, that this problem can be regarded as a generalization of the MacMillan problem, where a massless body moves up and down in between two equally massive bodies.

Fig. 28. Leonhard Euler (1707 – 1783) *(picture taken from MacTutor History of Mathematics archive, University of St Andrews Scotland)*

After the two-body problem was investigated (and solved analytically) in the last section, we will now demonstrate how, adding a massless third body, adds an enormous amount of complexity to the dynamics of the system. The most complete study on the R3BP was published by Szebehely [67], where one finds also an extensive literature. In the following demonstration we follow closely to the very well book by [64].

The restricted three body problem is defined as follows:

- two bodies, named primaries, with the masses $m \neq 0$ and $\mu \neq 0$ move on circular orbits,
- a third, massless body $m_3 = 0$ moves in the same plane as the primaries.

The restricted three body problem can serve as a good dynamical model for the investigation of many different types of motion in the Solar system and in other planetary systems:

- the motion of an asteroid, with the Sun and Jupiter as primaries,
- the motion of a satellite with the Sun and the planet as primaries,
- the motion of a comet with the Sun and Jupiter (which is – due to its large mass – the principle perturbing body) as primaries,
- the motion of terrestrial planets in extrasolar planetary systems with the star and a large Jupiter-like planet as primaries (or motions in double stars),
- ...

For a detailed investigation of the R3BP we choose a rotating coordinate system (η, ξ) (centered at the Sun) where the primaries are fixed (see Fig. 29).

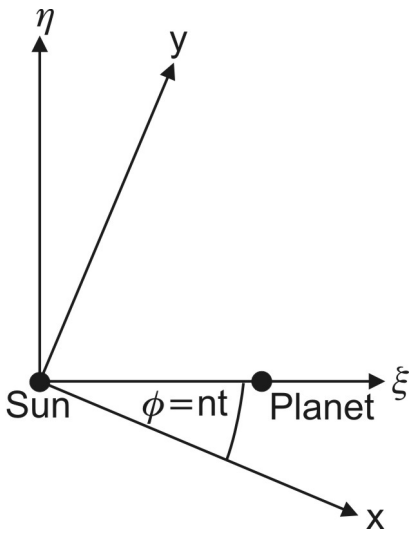

Fig. 29. Rotating coordinate system for the R3BP

The equations of motions can be obtained by the *Lagrange equation* (108):

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}_\mathrm{j}} - \frac{\partial L}{\partial x_\mathrm{j}} = 0 \tag{230}$$

where the function $L = T + U$ is the sum of potential energy U and kinetic energy T which (if we assume that the factor $\kappa^2 = -k^2(m+\mu) = 1$) are given by:

$$T = \frac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) \tag{231}$$

$$U = \frac{1}{r} \tag{232}$$

where (according to Fig. 29)

$$
\begin{aligned}
x &= \xi\cos\phi - \eta\sin\phi \\
y &= \xi\sin\phi + \eta\cos\phi \\
\phi &= nt \\
r &= \sqrt{\xi^2+\eta^2}
\end{aligned}
\tag{233}
$$

with n the mean motion of the primaries and r the distance to the Sun. By differentiation of x and y with respect to time t

$$
\begin{aligned}
\dot{x} &= \cos\phi\left(\dot{\xi} - \eta n\right) - \sin\phi\left(\dot{\eta} + \xi n\right) \\
\dot{y} &= \sin\phi\left(\dot{\xi} - \eta n\right) + \cos\phi\left(\dot{\eta} + \xi n\right)
\end{aligned}
\tag{234}
$$

one obtains the kinetic energy T in the rotating system:

$$
T = \frac{1}{2}\left(\dot{\xi}^2 + \dot{\eta}^2\right) + n\left(\xi\dot{\eta} - \eta\dot{\xi}\right) + \frac{n^2}{2}\left(\xi^2+\eta^2\right) \tag{235}
$$

To get the equations of motion from (230) we now have to differentiate (235) respectively (232):

$$
\begin{aligned}
\frac{\partial U}{\partial \xi} &= -\frac{\xi}{r^3} \\
\frac{\partial U}{\partial \dot{\xi}} &= 0 \\
\frac{\partial U}{\partial \eta} &= -\frac{\eta}{r^3} \\
\frac{\partial U}{\partial \dot{\eta}} &= 0
\end{aligned}
\tag{236}
$$

$$
\begin{aligned}
\frac{\partial T}{\partial \xi} &= n\left(\dot{\eta} + n\xi\right) \\
\frac{\partial T}{\partial \dot{\xi}} &= \dot{\xi} - n\eta \\
\frac{\partial T}{\partial \eta} &= n\left(-\dot{\xi} + n\eta\right) \\
\frac{\partial T}{\partial \dot{\eta}} &= \dot{\eta} + n\xi
\end{aligned}
\tag{237}
$$

Substituting (236) and (237) in (230) gives

$$
\begin{aligned}
\frac{d}{dt}\left(\dot{\xi} - n\eta\right) &= n\left(\dot{\eta} + n\xi\right) - \frac{\xi}{r^3} \\
\frac{d}{dt}\left(\dot{\eta} + n\xi\right) &= n\left(-\dot{\xi} + n\eta\right) - \frac{\eta}{r^3}
\end{aligned}
\tag{238}
$$

which leads, with the introduction of the potential Ω,

$$\Omega = \frac{1}{r} + \frac{1}{2}n^2r^2 \tag{239}$$

to the equations of motion of the R3BP:

$$\ddot{\xi} - 2n\dot{\eta} = \frac{\partial \Omega}{\partial \xi} \tag{240}$$

$$\ddot{\eta} + 2n\dot{\xi} = \frac{\partial \Omega}{\partial \eta} \tag{241}$$

5.1 Lagrange Points

We will now investigate the potential (239) in more detail, but we use now a rotating coordinate system with the barycentre of the masses m and μ as origin and the angular velocity $\phi = nt$ (Fig. 30). First, we write it in a way, where the gravitational and rotational part are separated:

$$\Omega = \Omega_{\text{grav}} + \Omega_{\text{rot}} = \left(\frac{m}{r} + \frac{\mu}{\rho}\right) k^2 + \frac{n^2R^2}{2} \tag{242}$$

From Fig. 30 we obtain the following relations:

$$\begin{aligned} R^2 &= \xi^2 + \eta^2 \\ r^2 &= (\xi + \mu)^2 + \eta^2 \\ \rho^2 &= (m - \xi)^2 + \eta^2 \end{aligned} \tag{243}$$

In addition we choose the units such that $k = 1$, $n = 1$ and $m + \mu = 1$; as a consequence (242) can be written as

$$\Omega = \frac{m}{2}\left(r^2 + \frac{2}{r}\right) + \frac{\mu}{2}\left(\rho^2 + \frac{2}{\rho}\right) - \frac{m\mu}{2} \tag{244}$$

The maxima and minima of the potential can be found by setting the derivatives equal to zero:

$$\Omega_\xi = \Omega_\eta = 0 \tag{245}$$

and therefore

$$\begin{aligned} \Omega_{\text{r}} r_\xi + \Omega_\rho \rho_\xi &= 0 \\ \Omega_{\text{r}} r_\eta + \Omega_\rho \rho_\eta &= 0 \end{aligned} \tag{246}$$

where the subscripts indicate derivations. Differentiating the potential gives:

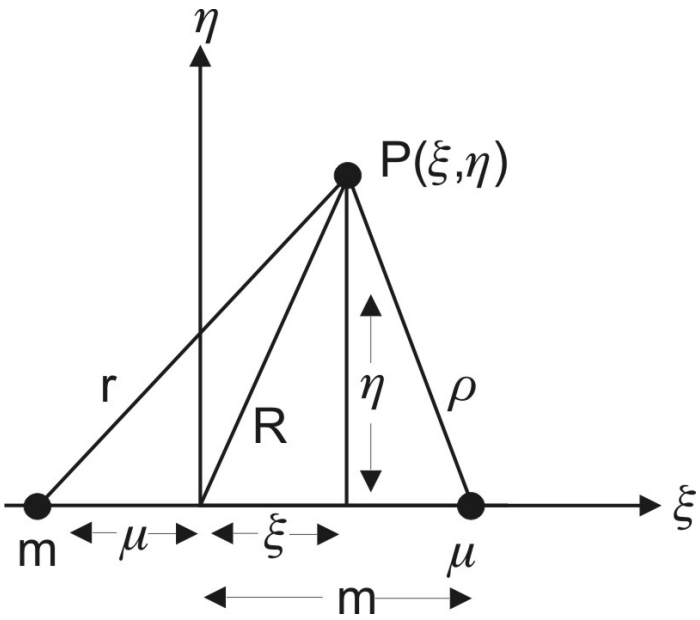

Fig. 30. Configuration of the bodies in the rotating frame ξ, η

$$\begin{aligned}
\Omega_{\mathrm{r}} &= m\left(r - \frac{1}{r^2}\right) \\
\Omega_{\rho} &= \mu\left(\rho - \frac{1}{\rho^2}\right) \\
r_{\xi} &= \frac{\xi + \mu}{r} \\
r_{\eta} &= \frac{\eta}{r} \\
\rho_{\xi} &= \frac{\xi - m}{\rho} \\
\rho_{\eta} &= \frac{\eta}{\rho}
\end{aligned} \tag{247}$$

This leads to the conditions that have to be fulfilled at the extreme values:

$$\begin{aligned}
m\left(r - \frac{1}{r^2}\right)\frac{\xi + \mu}{r} + \mu\left(\rho - \frac{1}{\rho^2}\right)\frac{\xi - m}{\rho} &= 0 \\
m\left(r - \frac{1}{r^2}\right)\frac{\eta}{r} + \mu\left(\rho - \frac{1}{\rho^2}\right)\frac{\eta}{\rho} &= 0
\end{aligned} \tag{248}$$

A trivial solution can be found for $r = \rho = 1$ which corresponds (see Fig. 30) to two positions where the three bodies form an equilateral triangle.

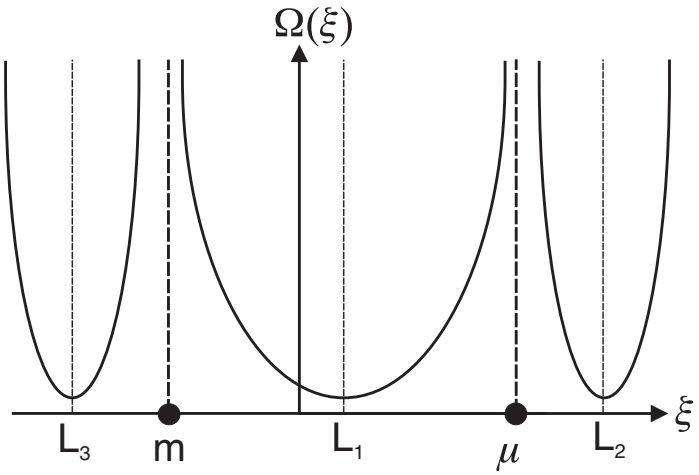

Fig. 31. Location of the extreme values on the ξ-axis

To find the other solutions, it can be seen from (246) that the following equation has to be fulfilled:

$$\frac{\xi+\mu}{r}\frac{\eta}{\rho} - \frac{\xi-m}{\rho}\frac{\eta}{r} = \frac{\eta}{r\rho} = 0 \tag{249}$$

which is true for $\eta = 0$ – therefore the remaining extreme values will be placed on the ξ-axis. Because of $\Omega(m) = \infty$ and $\Omega(\mu) = \infty$, there are three more extreme values situated in three intervals (see also Fig. 31):

1. $-\mu < \xi < m$
2. $m < \xi < \infty$
3. $-\infty < \xi < -\mu$

Their location on the ξ-axis can be found in the following way (we show it for L_1, the interval 1 $-\mu < \xi < m$). Because of

$$\begin{aligned} r &= \xi + \mu \\ \rho &= 1 - \xi - \mu \end{aligned} \tag{250}$$

and

$$\begin{aligned} r_\xi &= 1 \\ \rho_\xi &= -1 \end{aligned} \tag{251}$$

$$\tag{252}$$

it results from (246) that the following condition has to be satisfied:

$$\Omega_\mathrm{r} - \Omega_\rho = 0 \tag{253}$$

or

$$m\left(r-\frac{1}{r^2}\right)-\mu\left(\rho-\frac{1}{\rho^2}\right)=0 \tag{254}$$

Substituting

$$\lambda=\frac{\rho}{r} \tag{255}$$

leads to

$$m\left(\frac{1}{1+\lambda}-\frac{(1+\lambda)^2}{1}\right)-\mu\left(\frac{\lambda}{1+\lambda}-\frac{(1+\lambda)^2}{\lambda^2}\right)=0 \tag{256}$$

or

$$\left(-\lambda^2+(1+\lambda)^3\lambda^2\right)=\frac{\mu}{m}\left(-\lambda^3+(1+\lambda)^3\right) \tag{257}$$

and after simplifying

$$\lambda^3=\frac{\mu}{3}\left(\frac{1+3\lambda+3\lambda^2}{1+\lambda+\frac{1}{3}\lambda^2}\right) \tag{258}$$

we get the approximate solution

$$\lambda\sim\sqrt[3]{\frac{\mu}{3}}=\nu \tag{259}$$

Expanding (258) into a power series around ν gives for L_1

$$\lambda=\nu+\frac{2}{3}\nu^2+\frac{2}{9}\nu^3-\frac{32}{81}\nu^4+\ldots \tag{260}$$

Inserting the values of m and μ for the Sun-Jupiter system, this yields a value for $\lambda=\frac{\rho}{r}\sim\frac{1}{15}$.

The same procedure can be performed for the other intervals. The respective positions for the other four Lagrangian equilibrium points are the following:

- L_2: $\frac{\rho}{r}=\nu+\frac{1}{3}\nu^2-\frac{1}{9}\nu^3-\frac{31}{81}\nu^4\ldots$
- L_3: $\frac{\rho}{r}=1+\frac{7}{12}\mu-\frac{35}{144}\mu^2+\frac{3227}{20736}\mu^3\ldots$
- L_4: $\rho=r=1$
- L_5: $\rho=r=1$

In Fig. 32 we show in a three-dimensional plot the value of the potential function Ω: The location of the primaries is well visible through the two tubes where the potential field is going to infinity (marked with "M1" and "M2"). In between the two tubes we see the Lagrange point L_1; we also plotted the other two collinear Lagrange points L_2 (right side of the smaller tube) and L_3 (left side if the large tube). The two equilateral equilibrium points are inside the black banana shaped regions. We explain the meaning of these two regions in the next section.

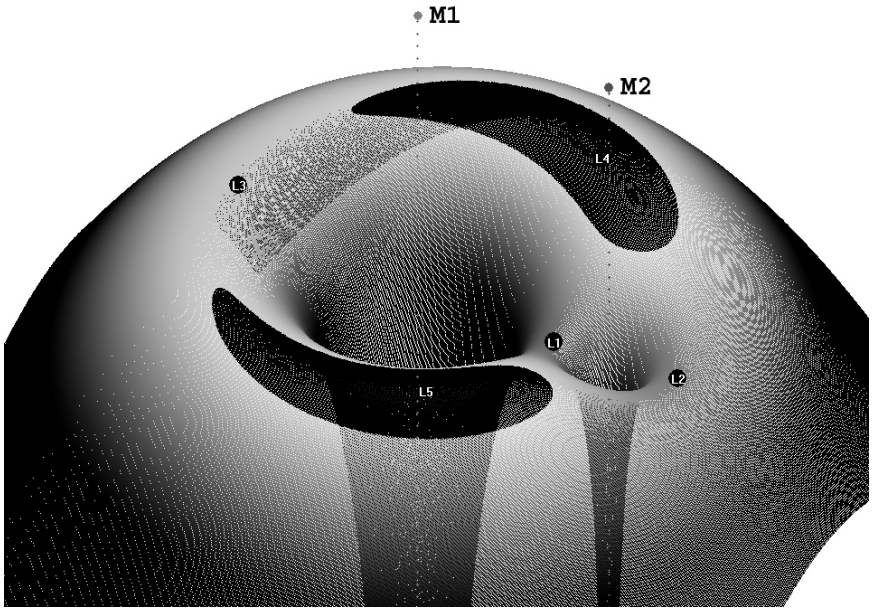

Fig. 32. The potential function Ω in the rotating frame

5.2 The Jacobi-Constant and the Zero-Velocity Curves

The problem of finding the motion of the third massless body m_3 is to find solutions $\xi(t)$ and $\eta(t)$ of the equations of motion (240-241): Fortunately it is possible to find an integral of motion, which leads to significant constraints for the motion of m_3. We multiply (240) with $\dot{\xi}$, and (241) with $\dot{\eta}$ and add them which leads to a cancellation of the terms with $\dot{\eta}\dot{\xi}$:

$$\ddot{\xi}\dot{\xi} + \ddot{\eta}\dot{\eta} = \frac{\partial \Omega}{\partial \xi}\dot{\xi} + \frac{\partial \Omega}{\partial \eta}\dot{\eta}. \tag{261}$$

This expression can be integrated because the right hand side is the total derivative with respect to the time, $\frac{d\Omega}{dt}$, and thus leads to

$$2(\dot{\xi}^2 + \dot{\eta}^2) \equiv 2v^2 = 2\Omega - C \tag{262}$$

where $C(\xi, \eta, \dot{\xi}, \dot{\eta})$ is the Jacobian constant.[31] With the aid of C the dimension of the system may be reduced by one order. This constant is an integral of motion which can be written

$$C = n^2(\xi^2 + \eta^2) + 2k^2\left(\frac{m}{r} + \frac{\mu}{\rho}\right) - \dot{\xi}^2 - \dot{\eta}^2 \tag{263}$$

[31] after Carl Gustav Jacobi (1804–1851).

We mention that C can also be expressed in a non-rotating coordinate system (x, y) by[32]

$$C = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - k^2 \left(\frac{m}{r} + \frac{\mu}{\rho} \right) - n(x\dot{y} - y\dot{x}) \tag{264}$$

We will see how useful the Jacobian constant is for a qualitative analysis of the motion of m_3: it is evident that the square of velocity in the rotating system has to be positive. This means that in the (ξ, η) coordinate system – according to the quantity of C – there are so-called *forbidden regions*, where the velocity would be imaginary. Thus one defines the zero-velocity curves; also called Hill's zero-velocity curves[33]

$$2U = C \text{ for } v^2 \equiv \dot{\xi}^2 + \dot{\eta}^2 = 0. \tag{265}$$

One should mention that this is *not* the energy integral, which does not exist in the restricted problem, nor exists the angular momentum integral, because m_3 is regarded as a massless body. In a certain sense the Jacobian integral replaces the energy-integral.

An interesting property of the system, which holds for any system of 2-degrees of freedom, is the following: if additional to the Jacobian constant C one could derive a second integral, then the problem would be integrable. Let us denote it by $D(\xi, \eta, \dot{\xi}, \dot{\eta}) = const$. We can then express the velocities by

$$\begin{aligned} \frac{d\xi}{dt} &= f(\xi, \eta, C, D) \\ \frac{d\eta}{dt} &= g(\xi, \eta, C, D). \end{aligned} \tag{266}$$

By division one gets only one equation, namely

$$\frac{d\eta}{d\xi} = \frac{g}{f} = h(\xi, \eta, C, D) \tag{267}$$

which could be solved with the aid of a factor of the form $M(\xi, \eta)$ such that

$$0 = fMd\eta - gMd\xi \equiv dF \tag{268}$$

is a complete differential expression. This yields $F(\xi, \eta) = const$ which is an algebraic equation of the orbit we are looking for. Additionally with the aid of (268) one could eliminate from (266) the coordinates

$$\dot{\xi} = \alpha(\xi), \ \dot{\eta} = \beta(\eta) \tag{269}$$

and from this one easily via a quadrature could find the final solution $\xi = \xi(t)$ and $\eta = \eta(t)$. Because no additional integral of motion is known for the R3BP

[32] note that we expicitely write 'n' although it was set to 1 in the preceding chapter

[33] after George William Hill (1838-1914), who also developed a Lunar theory on the basis of the R3BP.

this way of solution cannot be used. One should also mention that – even with two integrals of motion in a two degrees of freedom system – it is sometimes impossible to solve the former mentioned equation (268) because the factor M is rather difficult – if not impossible – to find. In addition when one can find M quite often one cannot solve it in elementary function (compare Sect. 3.4 in the chapter "Regular and Chaotic Motion in Hamiltonian Systems" of [42]).

We now continue to use the zero-velocity curves for a quantitative analysis of the motion of a satellite, an asteroid and a comet in the gravitational field of the Sun and a massive planet on a circular orbit, neglecting the gravitational force of the other planets. Jupiter has a mass of 1/1047.355 of the Sun's mass – thus the ratio is of the order of 0.001, but in order to explain the essential of the zero-velocity-curves we use in Figs. 33, 34 and 35 a different mass ratio between the two primaries ($\mu = 0.1$).

- **Large C:** We start to describe the motion of an asteroid on an eccentric orbit with a small semimajor axis (see Fig. 33): from the position of the third mass we can compute the potential and from (262) we compute the constant C for $v^2 = 0$ which defines the zero-velocity curves. For this motion they correspond to two separated closed curves around the Sun (left region) and around the planet (right region) and the asteroid can never leave these areas. But also the motion of a satellite of Jupiter is limited by the zero-velocity curves and can never escape from this region (right region). On the other hand a massless body could move outside the outer circles-like curve and could never penetrate the inner system. We thus have a forbidden region of motion (dark grey) separating different possible regions of motion. In this sense we can speak of an absolute stability found for $t \to \infty$, which is very rare for problems in astrodynamical problems.
- **Moderate C:** Next we check the stability of a massless asteroid with a large eccentricity in the main belt. Again the position defines the potential and for $v^2 = 0$ we find a value for C which defines the zero-velocity curves (see Fig. 34). We can see that the regions of motion around the two primaries opened and are now connected. This means that the third body can move freely around both primaries, but is still captured inside and may not escape into the outer regions.
- **Small C:** With even a higher eccentricity the third body may now escape through the large gap, which opens around the Lagrange point L_2 (Fig. 35). A massless body may stay around the primaries for quite a long time, but then it can escape to the outer region, which is now connected to the inner one. The forbidden region is shrinking more and more towards two points in the rotating coordinate system $(\xi - \eta)$, a fact that is quite well visible in the three dimensional representation of Ω (Fig. 32) (the very dark region around the two Lagrange points L_4 and L_5). Finally

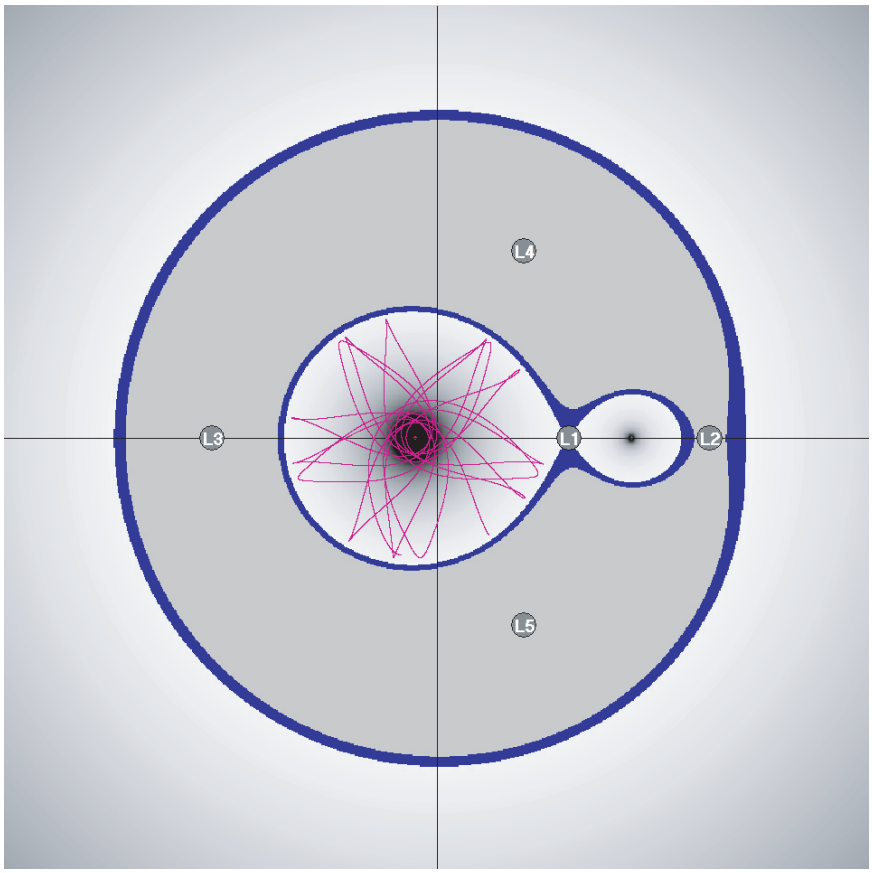

Fig. 33. Zero-velocity curves in a rotating frame for an asteroid close to the Sun *(large black spot)* or a satellite around a planet *(small black point)*. The grey region can never be penetrated by a third body moving either close to the planet, close to the Sun or far outside both primaries. The small circles show the so-called Lagrange points of equilibrium *(for an explanation see in the text)*

the zero-velocity curves degenerate to the equilateral Lagrange points and disappear completely for small values of C.

5.3 Stability of the Lagrange Points

To investigate the stability of the Lagrange points, we start by examining the motion of a body that moves very close to L_4 (see Fig. 38).

The coordinates ξ and η for the body close to the Lagrange point are given by

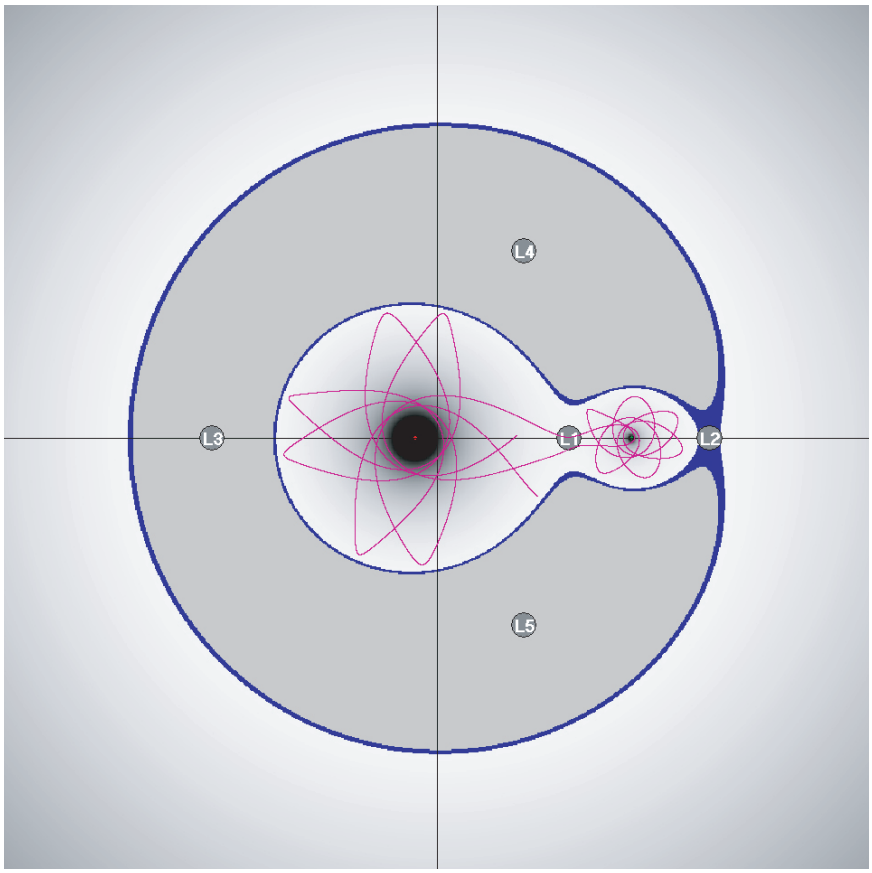

Fig. 34. Zero-velocity curve in a rotating frame for a body *(asteroid)* which can move in a region around both primaries *(caption like in Fig. 33)*

$$\begin{aligned} \xi &= a + x \\ \eta &= b + y \end{aligned} \tag{270}$$

where the position of L_4 is fixed by $a = 0.5 - \mu$ and $b = \sqrt{3}/2$ (see Fig. 36).

This leads to the following equations of motion for the deviation from the Lagrange point

$$\begin{aligned} \ddot{x} - 2\dot{y} &= \frac{\partial \Omega}{\partial x} \\ \ddot{y} + 2\dot{x} &= \frac{\partial \Omega}{\partial y}. \end{aligned} \tag{271}$$

With a Taylor expansion of the potential Ω close to the Lagrange point up to 2^{nd} order

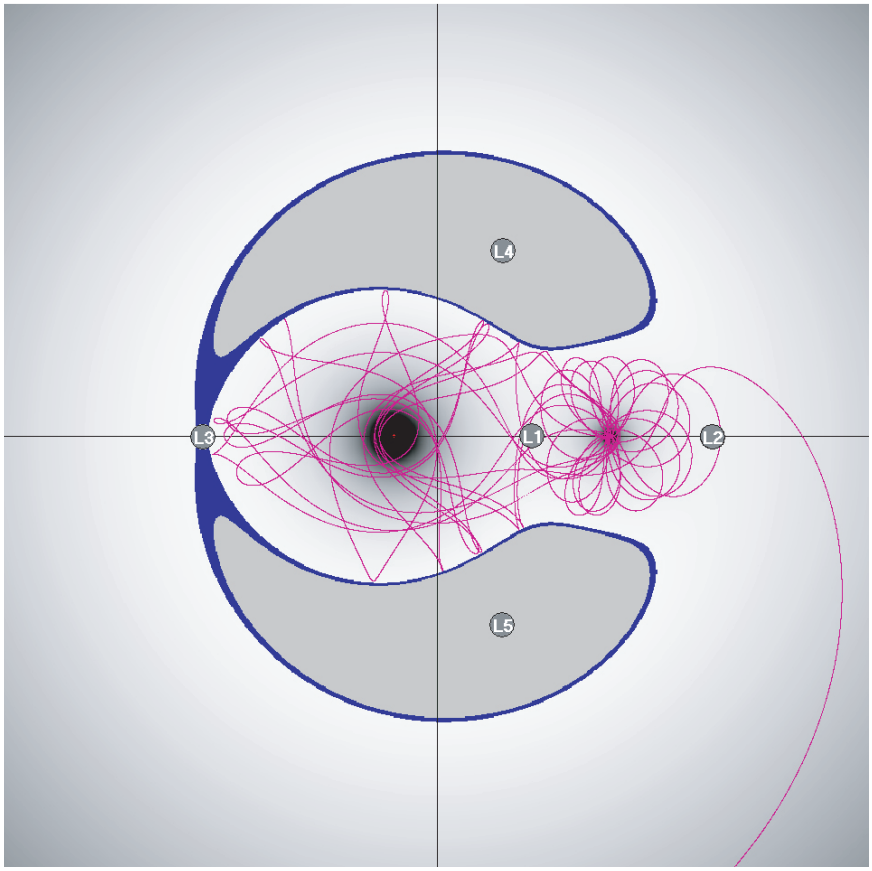

Fig. 35. Zero-Velocity Curve in a rotating frame for a body *(comet)* which can escape from the region around the Sun and the planet *(caption like in Fig. 33)*

$$\Omega(\xi,\eta) = \Omega(a+x, b+y) \tag{272}$$

we derive the equations of motion for the asteroid close to L_4

$$\begin{aligned} \ddot{x} - 2\dot{y} &= x\Omega_{\mathrm{aa}} + y\Omega_{\mathrm{ab}} \\ \ddot{y} + 2\dot{x} &= x\Omega_{\mathrm{ab}} + y\Omega_{\mathrm{bb}} \end{aligned} \tag{273}$$

where the Ω_{aa}, Ω_{bb} and Ω_{ab} are the numerical values of the partial derivatives at the Lagrange points (for $x = y = 0$).

To obtain a solution, we make the ansatz

$$\begin{aligned} x &= Ae^{\lambda \mathrm{t}} \\ y &= Be^{\lambda \mathrm{t}} \end{aligned} \tag{274}$$

Inserting (274) into (273) gives

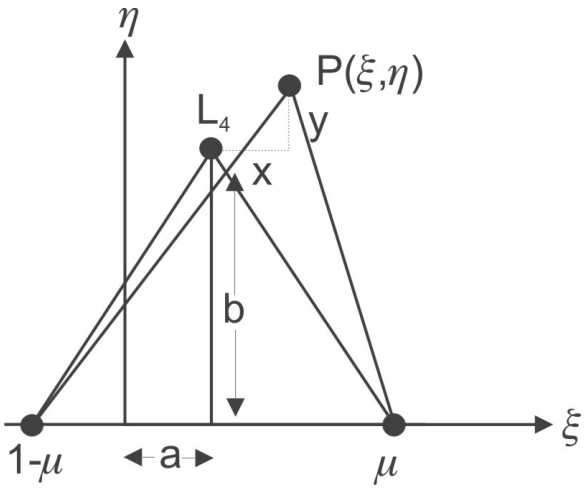

Fig. 36. Configuration and parameters of a body close to the Lagrange point L_4

$$\begin{aligned} \lambda^2 A e^{\lambda t} - 2\lambda B e^{\lambda t} &= A e^{\lambda t}\Omega_{aa} + B e^{\lambda t}\Omega_{ab} \\ \lambda^2 B e^{\lambda t} + 2\lambda A e^{\lambda t} &= A e^{\lambda t}\Omega_{ab} + B e^{\lambda t}\Omega_{bb} \end{aligned} \tag{275}$$

from which we derive

$$\begin{aligned} A\left(\lambda^2 - \Omega_{aa}\right) - B\left(2\lambda + \Omega_{ab}\right) &= 0 \\ A\left(2\lambda - \Omega_{ab}\right) + B\left(\lambda^2 - \Omega_{bb}\right) &= 0 \end{aligned} \tag{276}$$

The characteristic equation for λ is therefore

$$\lambda^4 - \lambda^2\left(\Omega_{aa} + \Omega_{ab} - 4\right) + \Omega_{aa}\Omega_{bb} - \Omega_{ab}^2 = 0 \tag{277}$$

After performing the derivations and inserting the values for L_4 one obtains:

$$\Omega_{aa} = \frac{3}{4} \qquad \Omega_{ab} = \frac{3\sqrt{3}}{4}(1 - 2\mu) \qquad \Omega_{bb} = \frac{9}{4} \tag{278}$$

This gives the following solution for λ:

$$\lambda^2 = -\frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 27\mu(1 - \mu)} \tag{279}$$

from which the condition for a real solution for λ is found:

$$\mu < \frac{1}{25} \tag{280}$$

This means that L_4 (and also L_5) are stable only, if the mass ratio of the two primaries is smaller than $\frac{1}{25}$! This is the case in our Solar system for all

planets and also the Earth-Moon system. For the Sun and Jupiter the ratio is $\sim \frac{1}{1000}$ and indeed we can observe a large group of asteroids ("Trojans") that move around L_4 and L_5.

For the general solution of (273) we can make the ansatz

$$x = \sum_{i=1}^{4} A_i e^{\lambda_i t} \tag{281}$$

$$y = \sum_{i=1}^{4} B_i e^{\lambda_i t} \tag{282}$$

where the A_i can be expressed through the B_i. One can replace the λ_is with

$$\lambda_1 = j\nu_1, \lambda_2 = -j\nu_1 \tag{283}$$

$$\lambda_3 = j\nu_2, \lambda_4 = -j\nu_2 \tag{284}$$

with $j = \sqrt{-1}$, which now leads to

$$x = A_1 e^{j\nu_1 t} + A_2 e^{-j\nu_1 t} + A_3 e^{j\nu_2 t} + A_4 e^{-j\nu_2 t} \tag{285}$$

$$y = B_1 e^{j\nu_1 t} + B_2 e^{-j\nu_1 t} + B_3 e^{j\nu_2 t} + B_4 e^{-j\nu_2 t} \tag{286}$$

Furthermore by replacing $e^{\pm j\nu t}$ (making use of the *de Moivre formula*) one can see – after some algebraic work[34] – that the solutions for x and y consists of two purely sinusoidal functions with the periods

$$p_1 = \frac{2\pi}{\nu_1};\ p_2 = \frac{2\pi}{\nu_2} \tag{287}$$

where the frequencies ν_1 and ν_2 are independent of the initial conditions. They determine the amplitudes and the phases of the libration, because

$$\nu^2 = -\lambda^2 = \frac{1}{2}\left(1 \pm \sqrt{1-\chi}\right) \text{ with } \chi = 27\mu(1-\mu) \tag{288}$$

only depends on the masses but not on the initial conditions. With this equation the two periods as function of the masses (respectively the mass ratio μ) can be developed as follows:

$$\begin{aligned} \nu_1 &= 1 - \left(\frac{1}{8}\chi + \frac{5}{128}\chi^2 + \frac{21}{1024}\chi^3 + \cdots\right) \\ \nu_2 &= \frac{1}{2}\sqrt{\chi}\left(1 + \frac{1}{8}\chi + \frac{7}{128}\chi^2 + \frac{31}{1024}\chi^3 + \cdots\right) \end{aligned} \tag{289}$$

Taking into account the property that $\nu_2^2 = 1 - \nu_1^2$ we can explicitely find the two frequencies (expressed in the correct units) :

[34] e.g. [65], 127ff

$$\nu_1 = 1 - \frac{27}{8}\mu - \frac{3213}{126}\mu^2 - \frac{22717827}{65600}\mu^3 + \cdots \tag{290}$$

For Jupiter Trojans the two periods are a short one very close to Jupiter's own period $p_1 = 11.9$ years and a longer one with $p_2 = 147.42$ years. We can see a similar behavior in Figs. 37-39. Fig. 37 shows a "banana" shaped orbit around the Lagrange point L_4 where one clearly sees, how the motion is dominated by two frequencies; also the shape becomes more elongated if the angle ϕ[35] is increased. For even larger ϕ, the orbit encloses L_4 *and* L_5 – and it is called "horseshoe" orbit (Fig. 39).

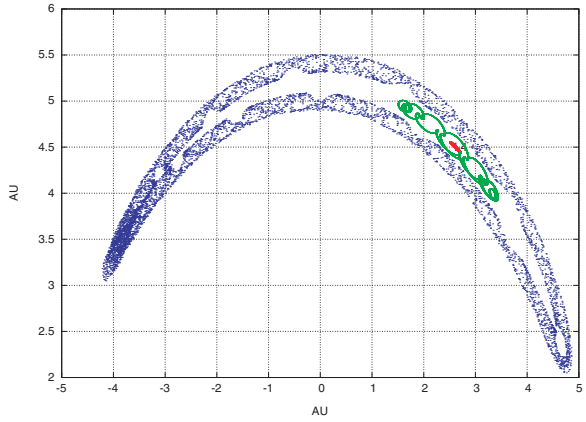

Fig. 37. Orbits around the Lagrange point L_4 in the R3BP: the large "banana" shaped orbit is for the initial conditions $\phi = 140^o$ and $e = 0.02$; the smaller orbit with several loops is for $\phi = 50^o$ and $e = 0.01$; the most innermost orbit only visible as line through L_4 is for $\phi = 60^o$ and $e = 0.001$. (Note that the eccentricity is the one of the Trojan!)

Finally one should say that even in extrasolar planets one can expect stable motion around the Lagrange points L_4 and L_5, which is the subject of several recent papers (e.g. [14, 39]). The period of libration for such systems can be easily determined using the formula given above (290) with the appropriate value of μ.

5.4 On the Stability Regions of the Trojan Asteroids

Until the first discovery of an asteroid, which moves always in the vicinity of L_4 by Max Wolf in Heidelberg in 1906 (the Asteroid (588) Achilles), the R3BP with stable orbits around the equilateral Lagrange points seemed to be

[35] ϕ is the angle between the line connecting the Sun and the planet and the planet and the point where the massless body has its closest approach to the planet.

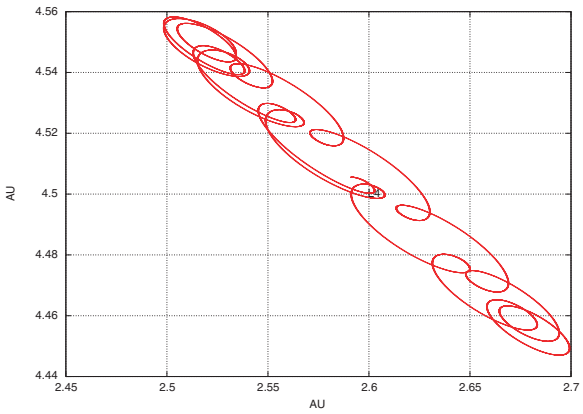

Fig. 38. Magnification of Fig. 37: the orbit around L_4 for $\phi = 60^o$ and $e = 0.001$

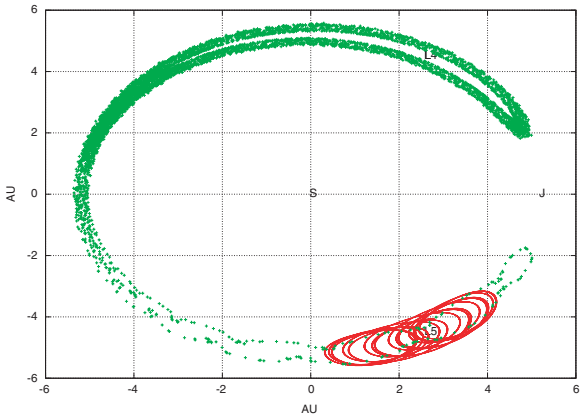

Fig. 39. Orbits around the Lagrange point L_4 in the R3BP:"Horseshoe" orbit around L_4 and L_5 for $\phi = 150^o$ and $e = 0.02$ and an orbit around L_5 for $\phi = 40^o$ and $e = 0.1$

only of theoretical interest. Ever since the search for bodies librating around the Lagrange points L_4 and L_5 of the Sun-Jupiter system was quite successful and nowadays we know of about 1000 Trojan asteroids[36] around the leading point L_4, but only 600 around the trailing L_5. This difference between the two populations is still an open question of Solar System dynamics. It should be mentioned that the stability of the equilibrium points is retained also for an elliptic orbit of the planet and for orbits outside the plane of motion of the primaries. In many papers (e.g. [17],[18]) the motion of the Trojans was

[36] These asteroids are called "Trojans" since all of them are named after heroes of the Trojan war.

studied also with analytical methods in the model of the spatial elliptic restricted three-body problem and even in a more realistic model where one takes into account Saturn's perturbations on Jupiter. Using numerical methods Milani [48], [49] was able to show that some of the real Trojans are on chaotic orbits. Also possible escapes from the Trojan cloud were discussed by different groups (e.g. [54],[70],[71]) in connection with chaotic orbits.

In Fig. 40 we show a histogram of the Trojans with respect to the inclinations and another one with respect to the eccentricities (Fig. 41). Here the differences in the population for L_4 and L_5 Trojans mentioned before are well visible.

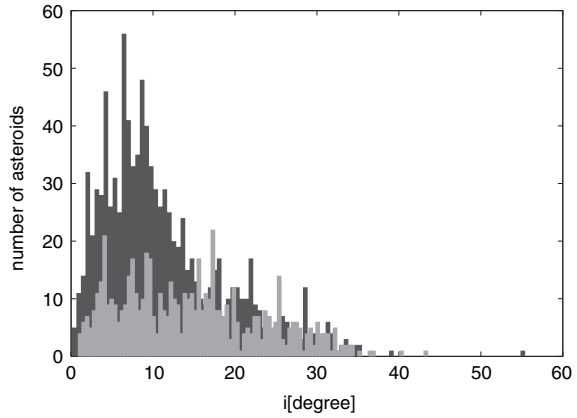

Fig. 40. Distribution of all real L_4 *(darkgrey)* and L_5 *(lightgrey)* Trojans with respect to their inclination

One can see that the mean value of the inclinations is slightly above zero degrees – this means that the Trojans with small inclined orbits with respect to Jupiter's plane of orbit seem to be more stable. The eccentricities of the orbits show a maximum around $e = 0.05$, which is caused by the variation of the eccentricity of Jupiter's orbit ($0.025 < e_{\rm Jup} < 0.06$).

To check the largeness of the stability region around L_4 and L_5 of Jupiter, numerical test computations were undertaken for time scales up to 10^8 years. The respective results are summarized in Fig. 42, where one can see how large these regions are. For a fine grid of initial conditions (semimajor axis versus initial eccentricity) of fictitious Trojans, we performed long-term numerical integrations. We used the entropy K_2[37] as measure of the chaoticity of an orbit. We started with the R3BP and did also computations in the realistic model of the outer Solar System, where we included the planets from Jupiter to Neptune (we show the results for the more realistic model including also

[37] for a definition see [2]

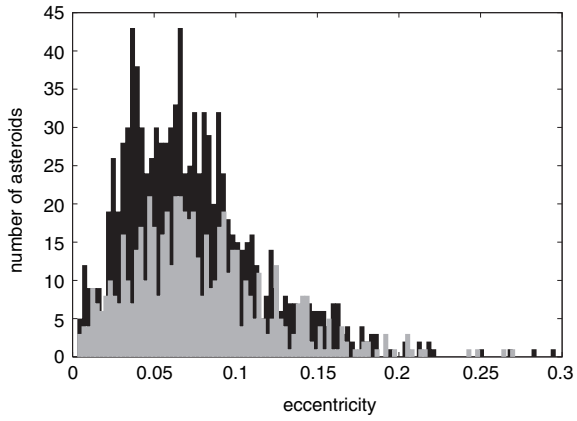

Fig. 41. Distribution of all real L_4 *(darkgrey)* and L_5 *(lightgrey)* Trojans with respect to their eccentricity

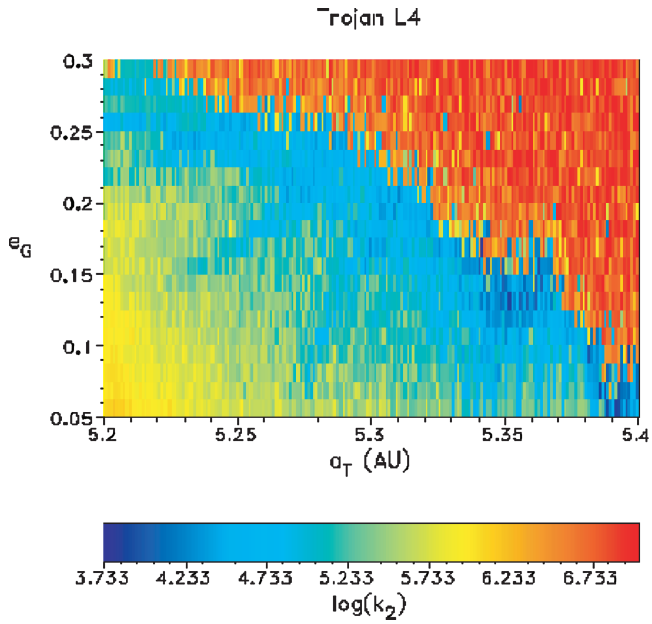

Fig. 42. Stability diagram around the equilateral Lagrange points L_4 showing the semi-major axis of the fictitious Trojan versus its initial eccentricity. Initial conditions in the red regions lead to unstable and in the yellow ones to stable orbits (after von Bloh, 2005, unpublished). The color in the diagram gives the value of K_2 – the entropy, a chaos indicator *(for details see [2])*

Saturn). Recently [62, 16] also analytical methods were used for an estimation of the "effective time of stability" where it is shown (still in the framework of the restricted three body problem) that most of the actually observed Trojans are "effectively" stable for up to 10^9 years.

6 The Sitnikov Problem

There are good reasons to include a description of the Sitnikov problem in this introductionary course on Celestial Mechanics:

> "We argue that it provides a very clear example for introducing students to non-integrability and chaos" (see Hevia [30])

In fact this dynamical system it is quite often cited as a perfect example for the chaoticity of a dynamical system. The Sitnikov problem can be regarded as a special case of the restricted three body problem: the configuration consists of two equally massive primary bodies on Keplerian orbits. A third, massless body moves perpendicular to the orbital plane of the primaries through the center of mass (see Fig. 43).

To describe the extreme sensitivity with respect to the initial conditions (which is one main property of a chaotic system) we cite J. Moser [50]:

> "*... we consider a solution of* $z(t)$ *with infinitely many zeroes* $t_k (k = 0, \pm 1, \pm 2, \cdots)$ *which are ordered according to size,* $t_k < t_{k+1}, z(t_k) = 0$. *Then we introduce the integers*
>
> $$s_k = \left[\frac{t_{k+1} - t_k}{2\pi} \right] \tag{291}$$
>
> *which measures the number of complete revolutions of the primaries between two zeroes of z(t). This way we can associate to every such solution a double infinite sequence of integers. The main result can be expressed as the converse statement:*
>
> Theorem. Given a sufficiently small eccentricity $\varepsilon > 0$ there exists an integer $m = m(\varepsilon)$ such that any sequence s with $s_k \geq m$ corresponds to a solution ..."

6.1 Circular Case

The case, when the primaries move on circular orbits was already known to Leonard Euler as a special case of the two fixed center problem, when the two masses involved are equal ($m_1 = m_2$) and the third massless body m_3

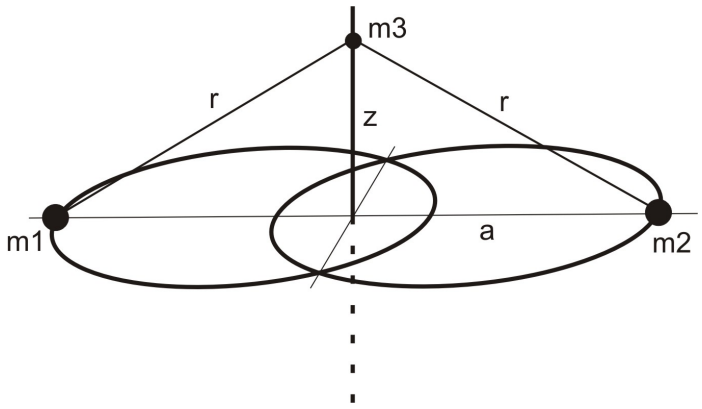

Fig. 43. Configuration of the Sitnikov problem

moves in the same plane, but perpendicular to the line of connection of the two masses through the barycentre. He solved the two fixed center problem generally for any motion of the massless body with the aid of elliptic functions (see Sect. 3.4 in the chapter "Regular and Chaotic Motion in Hamiltonian Systems"and also [74],[75]). In 1907 G. Pavanini [53] expressed the solutions of the circular Sitnikov problem by means of Weierstrass elliptic functions. MacMillan [45] gave a solution in terms of elliptic integrals in Legendre's normal form and was able to find a solution in form of a Fourier series expansion where the coefficients are power series depending on the perturbation parameter z (sufficiently small).

With $m_1 = m_2 = m$, $r^2 = a^2 + z^2$ and $a =$ constant, the total energy h of the systems reads

$$h = \frac{1}{2}\dot{z}^2 - \frac{2m}{r} \tag{292}$$

from which one finds the equation of motion by differentiating with respect to the time

$$\ddot{z} = -\frac{2mz}{r^3}. \tag{293}$$

From the energy relation it follows furthermore

$$\frac{dz}{dt} = \sqrt{2h + \frac{4m}{r}} \tag{294}$$

which can be separated and prepared for integration

$$dt = \pm \frac{dz}{\sqrt{2h + \frac{4m}{r}}}. \tag{295}$$

Setting $2m = 1$ and $a = 1$ leads to the equation of motion from (295)

$$\ddot{z} = -\frac{z}{\left(\sqrt{1+z^2}\right)^3}. \tag{296}$$

We find a restriction for the square of the velocity from the energy relation

$$\frac{1}{2}\dot{z}^2 = h + \frac{z}{\sqrt{1+z^2}} = h + \frac{1}{r} \geq h + 1 \tag{297}$$

and by integration of equation (295)

$$t - t_0 = \tau = \frac{1}{\sqrt{2}} \int_{z_0}^{z} \frac{dz}{\sqrt{h + \frac{1}{\sqrt{1+z^2}}}} \tag{298}$$

We set $t = t_0$, which is the time when the massless body crosses the barycentre. For every motion $\dot{z} \geq 0$ and consequently condition (297) is fulfilled and we find a restriction for bounded motions, namely $h \geq -1$. It follows that there exists a point of return $z_{\max} = \pm\gamma$ for the massless body for $\dot{z} = 0$

$$h = -\frac{1}{\sqrt{1+\gamma^2}}. \tag{299}$$

That means we have bounded motion only for negative h. With the respective energy $-1 \leq h < 0$ the motion is confined between

$$z_{\max} = \pm\gamma = \pm\sqrt{\frac{1}{h^2} - 1} \tag{300}$$

For $h = -1$ we have $\gamma = 0$ and the body is motionless in the barycentre. For $h = 0$, which is the parabolic case, the reversion point is shifted to infinity; when $h > 0$ we have the hyperbolic case, where the velocity is always nonzero.

For $-1 \leq h < 0$ the motion is always symmetric to $z = 0$ and reaches its maximum velocity in the barycentre

$$\dot{z}_{\max} = \pm\sqrt{2(h+1)} = \pm 2k; \text{ with } 0 \leq k = \sqrt{\frac{1+h}{2}} < \frac{1}{\sqrt{2}} \tag{301}$$

This purely periodic motion can be represented with the aid of a Fourier series

$$z = a_1 \sin \nu\tau + a_3 \sin 3\nu\tau + a_5 \sin 5\nu\tau + \ldots \tag{302}$$

where $\nu = \frac{2\pi}{\sqrt{p}}$ and p is the period of the oscillation

$$\frac{p}{4} = \frac{1}{\sqrt{2}} \int_0^{\gamma} \frac{dz}{\sqrt{h + \frac{1}{\sqrt{1+z^2}}}} \tag{303}$$

The solution leads to an elliptic integral of the 3rd kind. After rather lengthy transformations[38] finally one ends up with a time series for z:

[38] for details see Stumpff [64], p.75 ff

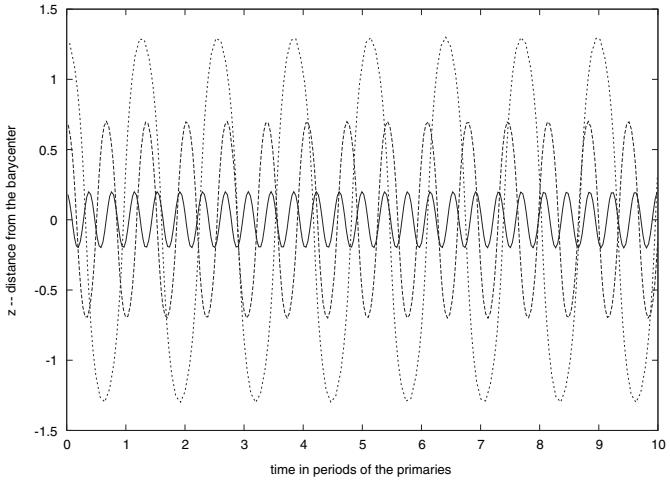

Fig. 44. Oscillations of m_3 for $z_{\rm ini} = 0.2, 0.7, 1.3$ for 10 revolutions of the primary bodies

$$z = \gamma \left[\sin \nu\tau + \frac{3}{64}\mu(\sin \nu\tau + \sin 3\nu\tau) + \right. \tag{304}$$

$$\left. + \frac{1}{4096}\mu^2 \left(79 \sin \nu\tau + 108 \sin 3\nu\tau + 29 \sin 5\nu\tau\right) + \ldots \right] \tag{305}$$

with $\gamma = \sqrt{\frac{\mu}{1-\mu}}$.

6.2 Numerical Results

Almost no dynamical system is integrable and therefore a way of deriving an orbit is to use numerical techniques to solve the equations of motion. The problem there is that we just can follow one orbit for one set of initial conditions. An appropriate way to study the dynamics of – low dimensional – systems is to construct surfaces of section (=SOS) with the aid of extensive numerical investigations, where we compute orbits for various initial conditions. As example we show in Fig. 44 the solutions for 10 revolutions of the primaries for the MacMillan problem for three different initial conditions, where we set the velocity $\dot{z} = 0$. It is obvious that the frequency diminishes with the amplitudes and converges to a final value for infinitesimal oscillations; the respective numerical value will be given later. In Fig. 46 (upper left plot) the SOS of the MacMillan problem is plotted for initial values of $0.1 < z_{\rm ini} < 2.$ $(e = 0)$, where one can see that all invariant curves are closed: this means, that all motions are quasiperiodic ones on a KAM torus.

6.3 Elliptic Case

In the Sitnikov problem ($e \neq 0$) the situation is different: we can observe the structure of phase space which we know from nonlinear dynamical systems with periodic orbits, quasiperiodic orbits and chaotic ones.

As example we show the orbital behavior of the third body for the same initial conditions ($z = 0.05$, the primaries in *periastron*) but for different eccentricities of the primary bodies (Fig. 45):

1. The regular periodic motion with one specific amplitude is well visible on the top graph for $e = 0$.
2. For $e = 0.3$ the difference is already evident (small differences in the amplitudes and the periods).
3. A big change is visible in the next graph for $e = 0.6$ with small and large amplitudes and periods, although the motion still seems to be quasiperiodic (we can verify from Fig. 47 that it is very close to be approximated by a high order perturbation theory, see next subsection).
4. Finally with the eccentricity $e = 0.9$ (bottom graph) the system is in a mode, where the consecutive intervals of time of the crossing of the barycentre are completely different. Whenever the primaries are in the *periastron* (multiples of 2π) there is a strong interaction with m_1 and m_2 because they are in this position very close to m_3 ($a = 0.1$, compared to the value $a > 1$ in the MacMillan problem). In fact after some ten

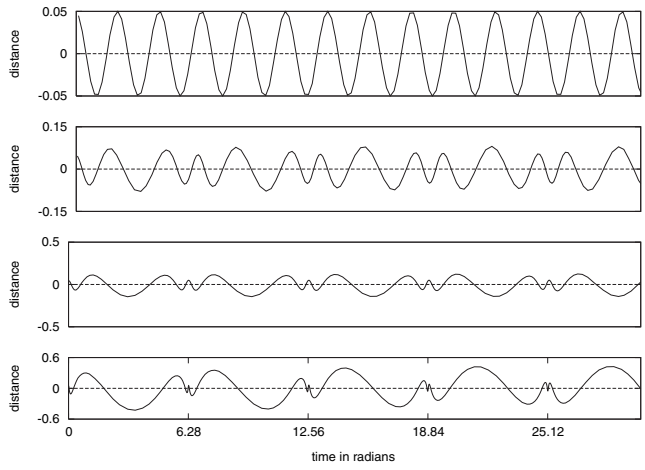

Fig. 45. Distance z of the massless body from the barycentre versus the time in the Sitnikov problem for e = 0, 0.3, 0.6, and 0.9 *(from top to bottom)* for a few revolutions of the primaries *(in radians)*

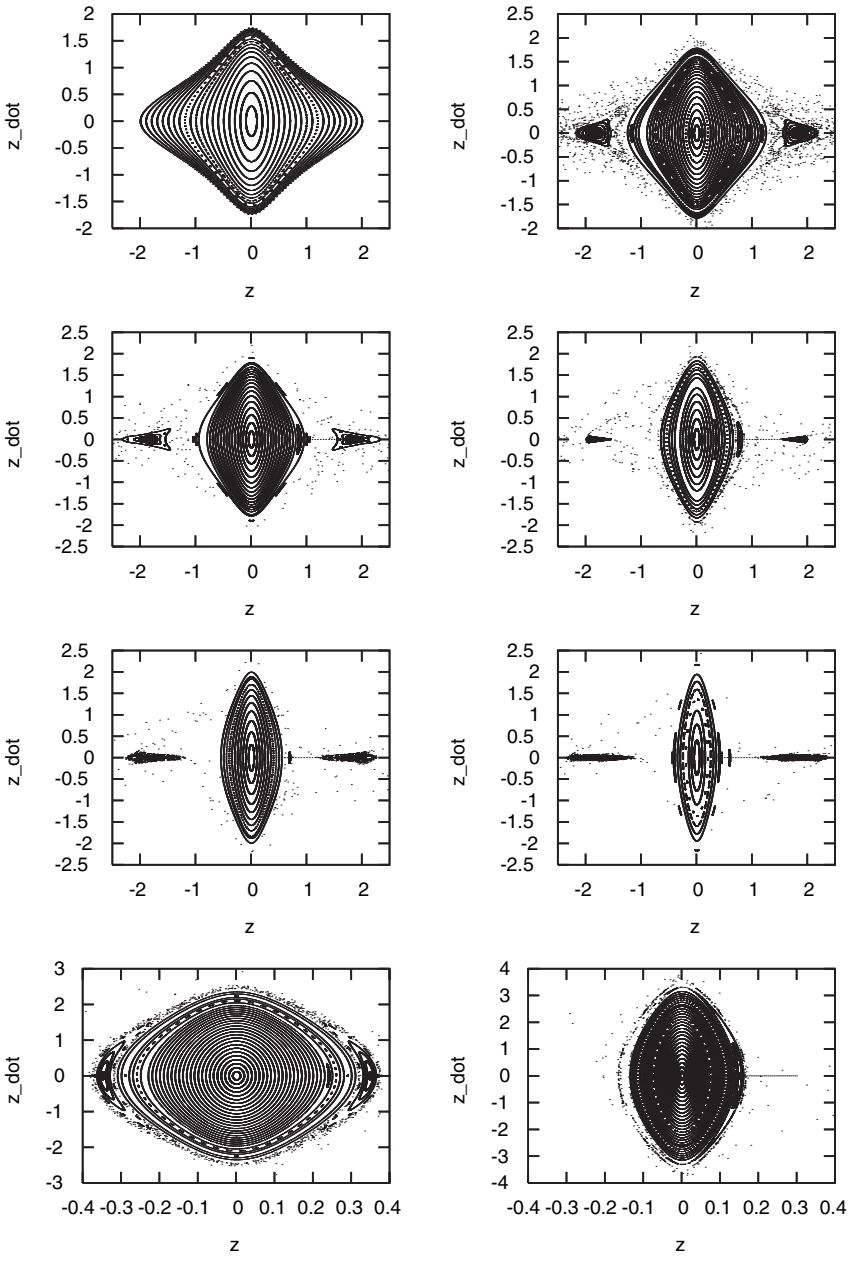

Fig. 46. Surface of Section ($z - \dot{z}$) for different eccentricities $e = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8$ *(detailed description in the text)*

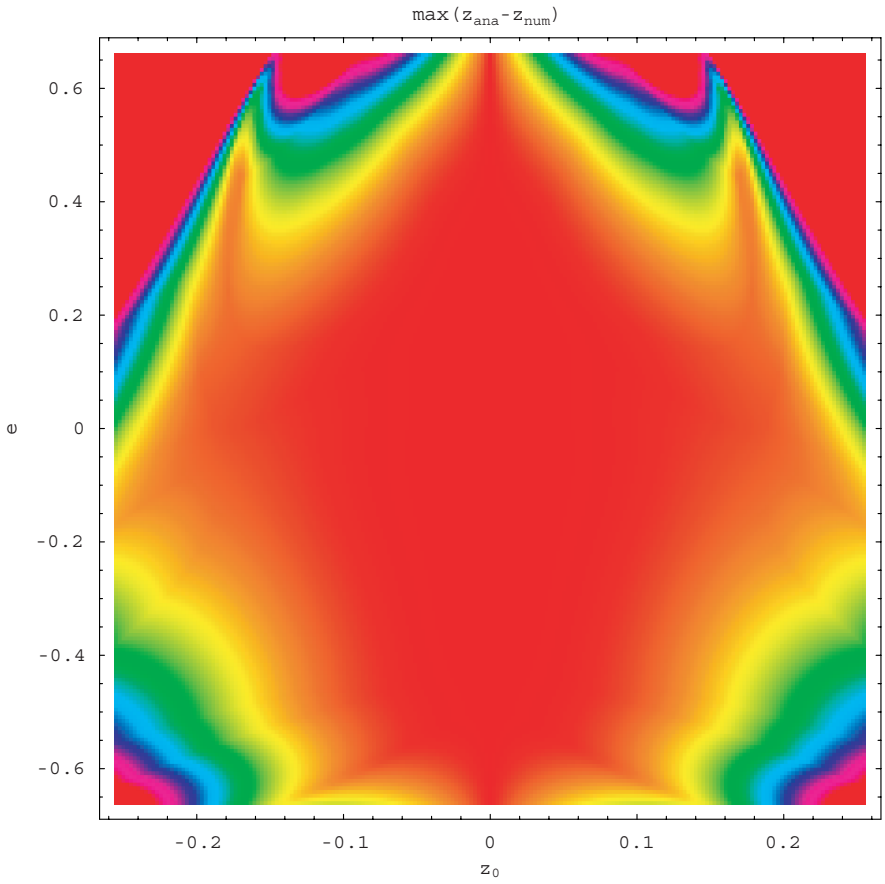

Fig. 47. Comparison of the analytical solution with the numerical one for 5 periods of the primaries. In this $e - z(0)$ diagram in the central red region the errors stays within 10^{-7}, in the yellow, green and light blue region the error grows from 10^{-7} to 10^{-2} and reaches very fast 50 percent in the violet corners, after Lhotka [41]

more revolutions of the primaries the third body escapes! The other three motions described above are stable.

The structure of the phase space for different eccentricities is shown in Fig. 46. Any of these orbits in the SOS starts with different initial conditions for z with $\dot{z} = 0$ and $a = a_{\min}$ (= periastron); one can imagine that we "drop" the third mass from this position. Points on this subspace of the three dimensional phase space are plotted for the next crossing of the orbit with the SOS when the primaries are again in their perihelion position. It means that we have chosen the positional angle of the primaries being zero

as initial condition and as defining the surface of section. Let us describe in Fig. 46 the different SOS one by one:

- $e = 0$ (upper left graph): As stated above all initial conditions lead to closed curves, which we can understand as sections of KAM-tori with a plane.
- $e = 0.1$ (upper right graph): We already can see the structure known from nonlinear dynamics: outside the "main island" of stable orbits, where all motions are bounded. We see two islands of invariant curves around a stable point, which correspond to the 2:1 resonance (where the primaries make two complete revolution whereas the third mass finishes exactly one oscillation – two crossings of the barycentre). Inside the "main island" one can also see two other islands, which correspond to different resonant motions (in Table 1 we give initial conditions for some specific resonant orbits).
- $e = 0.2$: The "main island" shrinks towards the point of equilibrium in the center and the 2:1 island is more and more isolated and shrinks also in size. The triangle like structure has a vertex directly on the x-axis opposite to the main land and is caused by a 3:1 resonant orbit.
- $e = 0.3$, $e = 0.4$ and $e = 0.5$: The size of the island starts to grow again after going through a minimum and change of the orientation on the SOS: the triangle like structure of the island now has a vertex on the abscissa on the side of the "main island"[39]
- $e = 0.6$ and $e = 0.8$: The size of the "main island" is shrinking more and more while from the point $(z = 0, \dot{z} = 0)$ new periodic orbits emerge when the eccentricities of the primaries become larger. These fixed points and the surrounding secondary islands are shifted outwards with larger and larger values of the eccentricity until they disappear in the large chaotic see.

What we describe here is the well known behavior of dynamical systems which was already described in the discussion of the Standard Mapping. We emphasize that the structure of the phase space is such, that in between any periodic orbits there is an infinity of invariant curves corresponding to quasiperiodic orbits. This foliation into KAM-tori shows the close connection of dynamical systems and number theory which will be discussed in the very last section of this part. A detailed discussion can be found e.g. in [16].

Nevertheless there exist periodic orbits which correspond to points in the SOS. They are dense with respect to the initial conditions (like the rational numbers). Table 1 shows the initial condition z_{ini} for some of them:

[39] this well known effect in non-integrable Hamiltonian systems is described in details in [73]

Table 1. Initial conditions for the distance z for selected periodic orbits in the Sitnikov problem; the primaries are in periastron position and $\dot{z} = 0$; n is the mean motion

$n_{\rm prim} : n_{\rm m3}$	$e = 0$	$e = 0.2$	$e = 0.4$	$e = 0.6$	$e = 0.8$
1:1	1.043698	0.872719	0.691427	0.496293	0.278653
1:2	1.848460	1.836787	1.816155	1.786429	1.747073
1:3	2.495394	2.924811	3.024252	3.048240	3.040044
1:4	3.064713	3.071739	3.071158	3.062020	3.042885
1:5	3.584274	3.137770	3.090335	3.067454	3.043343

6.4 Analytical Results for the Elliptic Case

In the elliptic case ($e > 0$) the distance of the primaries $2a$ is not constant, a fact that changes the equations of motion into a form which is explicitely time dependant

$$\ddot{z} = -\frac{2m}{\left(\sqrt{z^2 + a(t)^2}\right)^3} \tag{306}$$

Using the known development for the distance a

$$a(t) = \frac{1}{2}\left(1 - e\cos(t)\right) + O(e^2) \tag{307}$$

we can develop (306) and set $m = 1$

$$\ddot{z} = -\frac{8z}{\left(\sqrt{4z^2+1}\right)^3} - e\frac{24z}{\left(\sqrt{4z^2+1}\right)^5}\cos(t) \tag{308}$$

For the case $e = 0$ we see, that for $z << a$ the equation of motion reduces to a simple harmonic oscillator

$$\ddot{z} + 8z = 0 \tag{309}$$

with the solution

$$z(t) = z(0)\cos\sqrt{8}t \tag{310}$$

which is the aforementioned limiting period $p = 2\sqrt{2}$ for small amplitudes.

For treating the low energy case Wodnar [76] introduced a transformation of the independent variable time to the angle $\zeta = \frac{z}{a(\varphi)}$ which is the tangent of the angle between the barycentre and one primaries' position seen from the mass m_3. The equation of motion in the dependant variable ζ is

$$\zeta'' + \frac{e\cos\varphi + (0.25 + \zeta^2)^{-1.5}}{1 + e\cos\varphi}\zeta = 0 \tag{311}$$

where the $'$ stands for the derivation with respect to the true anomaly φ. Following Hagel [26] the expression $\Xi(\zeta) = (0.25 + \zeta^2)^{-1.5}$ can be developed into a polynomial differential equation of the form

$$\zeta'' + g_1(\varphi)\zeta + \sum_{k=2}^{M} g_k(\zeta)\zeta^k = 0 \tag{312}$$

with an appropriately chosen M to ensure accurate results for ζ. To achieve this goal one can expand the expression $\Xi(\zeta) = (0.25+\zeta^2)^{(-1.5)}$ into a Taylor series or, which turns out to be more efficient, into Chebycheff polynomials which yield

$$\Xi(\zeta) = 8 - 47\zeta^2 + 203\zeta^4 - 616\zeta^6 + 1168\zeta^8 + O(\zeta^{10}) \tag{313}$$

which is valid in the range $-0.8 < \zeta < 0.8$. Thus one is lead to the following equation which describes motions in the vicinity of the barycentre and can be used for further investigations. From the linearized equations

$$\zeta'' + g_1(\varpi)\zeta = 0 \tag{314}$$

one can derive exact analytical solutions for the frequencies and the amplitudes. This can be used for a subsequent perturbation theory which gives quite good solutions up to large eccentricities of the primaries.

$$\zeta'' + \frac{8 + e\cos\varphi}{1 + e\cos\varphi}\zeta + \sum_{k=2}^{n} \frac{A_k}{1 + e\cos\varphi}\zeta = 0 \tag{315}$$

The linear part is a differential equation of Hill's type and the solution can be discussed using the Floquet theory. For very small oscillations but large eccentricities Hagel and Trenkler [27] succeeded with the aid of quasi integrals to confirm a result found by Alfaro and Chiralt [1] that the fixed point for $z = \dot{z} = 0$ is unstable for very small intervals of the eccentricity (e.g. close to $0.544469\cdots$, between $0.855862\cdots < e < 0.855863\cdots$, etc. Based on this work Lhotka [41] treated the non-integrable case ($e > 0$) of small amplitudes and used then Floquet's theory like Hagel. Using computer algebra (*Mathematica*) he was able to expand the transfermatrix R (monodromy matrix) up to the 17^{th} order in the eccentricity of the primaries. After the reduction of the problem to a harmonic oscillator, an automatic application of the method of Poincaré-Lindstedt was implemented. Thus it was possible to find a solution consisting of 24 time dependant frequencies and amplitudes:

$$z(t) = \sum_{k=1}^{24} a_k(t)\cos(\Psi_k(\sigma(t))) + \sum_{k=1}^{24} a_k(t)\sin(\Psi_k(\sigma(t))) \tag{316}$$

where the respective functions $\Psi(\sigma(t))$ can be computed according to the initial conditions with a *Mathematica* program.[40] The range of validity for z and the eccentricity e is shown in Fig. 47, where the red central region marks a very good precision even after 10 revolutions of the primaries. It is

[40] Available at `http://www.astro.univie.ac.at/~adg`

quite remarkable that with this approach Lhotka succeeded to have very good solutions even for relatively large values of the "small quantities" eccentricity and/or initial distance from the barycentre.

6.5 The Sitnikov Sequences

One motivation to discuss the main properties of the Sitnikov problem was already mentioned in the introduction: whatever sequence of integer numbers s_k– corresponding to revolutions of the primaries – one can imagine, there is always one orbit which corresponds to these successive crossings of the barycentre (with the restriction $s_k \geq m$ already stated in the theorem).

To realize such a sequence numerically we can proceed as follows (see [9]): we choose a grid of initial conditions for the true anomaly φ of the primaries and for the velocity $\dot{z}$ and start our computations always from the position $z = 0$. We follow the orbit up to the first crossing of the massless body and ask whether this crossing was still during the first revolution of the primaries, or after the second, third, fourth etc. more revolutions. In such a way we can characterize an orbit by a sequence of numbers. We mark these numbers in an initial condition diagram for a great number of orbits, which now define different regions. Then we proceed to the next crossing of the barycentre, but work on a finer grid and count again the number of revolutions between the last crossing and this one. We have done this up to four crossings and found regions in the initial condition diagram (phase of the primaries versus the velocity of m_3) which correspond exactly to the sequences of numbers we were speaking above. We characterized each of this sequences with different colors in Fig. 48: e.g. the large violet region corresponds to a crossing already during the first revolution of the primaries; consequently the sequence would start with 1. The black region on the top of the plot corresponds to a large initial velocity and to a first crossing after 4 or more revolution of the primaries. In this figure one can see the complicated dynamical structure of the phase space of the Sitnikov problem and how a tiny difference in the initial conditions can change the orbit after only a few crossings of the barycentre.

In Fig. 49 we finally show an orbit which excellently demonstrates the sensitivity of the motion to a change of the initial conditions(which characterizes nonlinear dynamical systems and chaotic motions). The initial position of these three orbits just differ by 10^{-6} and lead to different behavior:

1. for $z(0) = 10^{-3}$ m_3 leaves after 2 primary revolutions along the negative z-axis (red line),
2. for $z(0) = 10^{-3} + 10^{-6}$ m_3 leaves the system in the other direction (green line);
3. for $z(0) = 10^{-3} - 10^{-6}$ m_3 the orbit stays bounded for long time but nevertheless escapes.

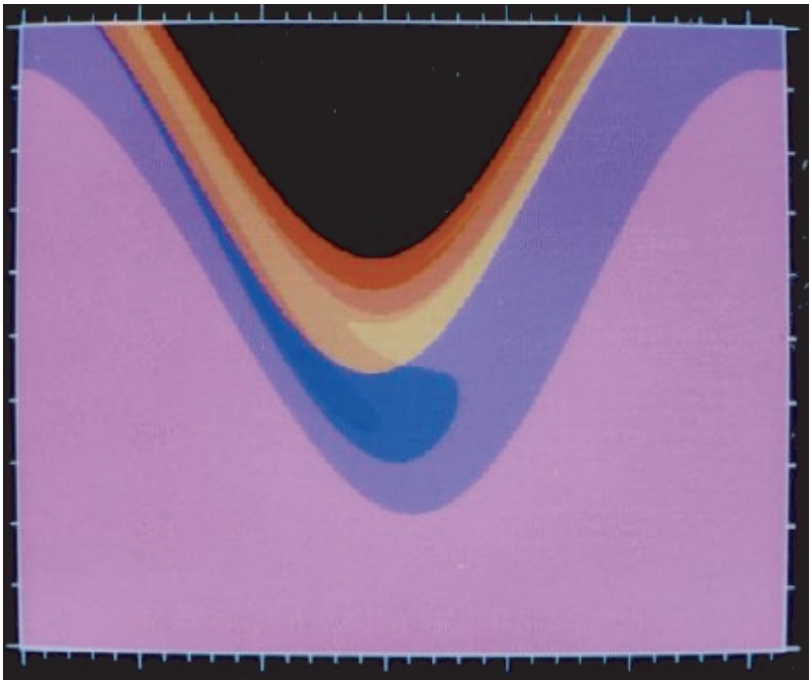

Fig. 48. Initial condition diagram for the eccentricity $e = 0.3$: x-axis is the phase of the primaries, y-axis is the velocity of the third body ($z(0) = 0$). The fine structure of the phase space is marked with different colors indicating initial conditions which lead to a specific Sitnikov sequence (after F. Vrabec (unpublished)) *(see the text for details)*

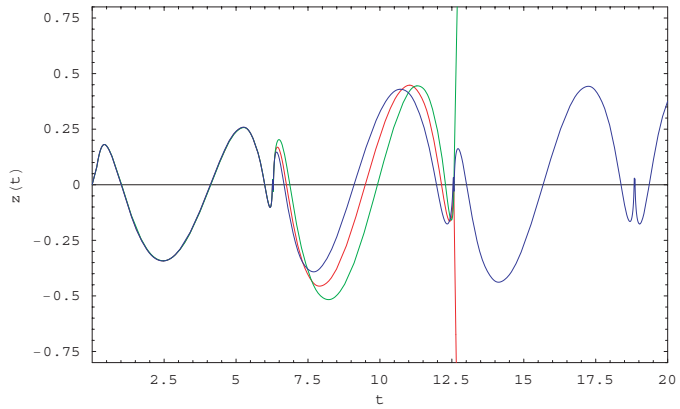

Fig. 49. For a large eccentricity of the primaries *(e=0.9)* we see the different behavior of three orbits initially only separated by 10^{-6} in position *(for details see text)*

7 Planetary Motion and Classical Perturbation Theory

7.1 Historical Remarks

A remarkable confirmation of Newton's law of gravitation was the prediction of the appearance of the comet Halley for 1758, which was done by Edmund Halley (1656–1742) who used several observations and assumed an elliptic orbit with a very large eccentricity for this comet. It was much more difficult for the next generation of astronomers to solve the n-body problem[41] – although the only force acting between them is the gravitation:

- Leonhard Euler (1707–1783) was the one who found the ten integrals for the n-body problem.
- Joseph-Louis Lagrange (1786–1813) solved the equations of motions for special cases of the general three-body problem (when they always are in the configuration of an equilateral triangle).
- Pierre Simon Laplace (1749–1827) discovered that up to the second order in the small masses the semi-major axis of the orbits of the planets are constant; this was an important indication of the stability of the planetary system. It was also Laplace who summarized the knowledge of the epoch in its famous books: *Mécanique céleste* and *Exposition du Système du Monde* where – as a consequence of Newton's law – he stated that

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at any given moment knew all of the forces that animate nature and the mutual positions of the beings that compose it, if this intellect were vast enough to submit the data to analysis, could condense into a single formula the movement of the greatest bodies of the universe and that of the lightest atom; for such an intellect nothing could be uncertain and the future just like the past would be present before its eyes."[42]

In 1781 William Herschel (1738–1822) discovered, after systematically searching the sky with his telescope, the planet Uranus. Because of systematically differences in the position of Uranus, Adams in England and Leverrier in France computed the orbit of an additional planet outside the orbit of Uranus. In fact in 1846 this planet – Neptune – was discovered in Berlin by Galle; this was a magnificent proof of the possibilities of using perturbation theory based only on Newton's law of gravitation. The outermost planet Pluto was discovered in 1931; it turned out in recent years that Pluto is just one large object out of many – called Plutinos –in the 3:2 mean motion resonance with Neptune. With the use of better and better telescopes

[41] with $n > 2$.
[42] quite often cited as "Laplace's demon"

Fig. 50. Pierre Simon Laplace (1749-1827) *(picture taken from MacTutor History of Mathematics archive, University of St Andrews Scotland)*

astronomers now discover many more small but also larger object far outside Pluto, which built the so-called Edgeworth-Kuiper belt.[43]

Finally we should mention that the spacing of the planets seemed to be quite regular for the planets known up to the middle of the eighteenth century; therefore Titius and Bode formulated a "law" from which one could compute the distance d of the known six planets (Mercury to Saturn) in units of the distance Earth-Sun:

$$d = 0.4 + 0.3(2^n) \text{ with } n = -\infty, 0, 1, 2, 4, 5 \tag{317}$$

[43] There is an estimated number of some ten thousands of objects larger than 100km; the largest known up to now are Quaoar and Sedna, discovered recently, with more than 1000 km in diameter.

At the distance corresponding to n=5 in fact Uranus was discovered ($d \sim 19$), for $n = 3$ in 1801 the first planetoid Ceres was observed ($d \sim 2.8$); both seemed to be a confirmation of the Titius-Bode law. But at least, when Neptune was discovered in a quite different distance from the Sun from the one given by this "law" (for $n = 7$ Neptune should have distance $d \sim 39$ but it has $d \sim 30$) it turned out to be just a kind of a rule to compute the distances for some planets and nothing more.[44]

In this section we will give an introduction into classical perturbation theory, then we will investigate mean motion resonances and their influence on the motion of the planets. Finally, at the end of this section, we explain the principles of the secular perturbation theory.

7.2 Classical Perturbation Theory of the Solar System

There exist quite good textbooks on this problem like the one by Roy [60] or the excellent one by Brouwer and Clemence [3].

Heliocentric Equations of Motion

Newton's law gives the force $\boldsymbol{F}$ acting between two bodies with masses m_i and m_j[45]

$$\boldsymbol{F}_{ij} = k^2 \frac{m_i m_j}{r_{ij}^2} \frac{\boldsymbol{p_i} - \boldsymbol{p_j}}{r_{ij}} \tag{318}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are the position vectors in an inertial frame (see Fig. 51); $r_{ij} = |\mathbf{p_i} - \mathbf{p_j}|$. To find out the force from n-1 bodies with masses $m_j; (j = 0, ..., n, j \neq i)$ acting on m_i, one has to add all the forces

$$\boldsymbol{F}_i = m_i \cdot \ddot{\boldsymbol{p}}_i = k^2 m_i \sum_{j=0, j\neq i}^{n} \frac{m_j}{r_{ij}^3} (\boldsymbol{p}_i - \boldsymbol{p}_j) \tag{319}$$

Adding all forces leads to the equation

$$\sum_{i=0}^{n} m_i \cdot \ddot{\boldsymbol{p}}_i = k^2 \sum_{i=0}^{n} \sum_{j=0, j\neq i}^{n} \frac{m_i m_j}{r_{ij}^3} (\boldsymbol{p}_i - \boldsymbol{p}_j) = 0 \tag{320}$$

which is identically zero because every vector $\boldsymbol{p}_i - \boldsymbol{p}_j$ is canceled by a vector in the opposite direction $\boldsymbol{p}_j - \boldsymbol{p}_i$. Double integration defines the barycentre of the dynamical system, which is moving linearly in space. This vector $\boldsymbol{s} = \boldsymbol{a}t + \boldsymbol{b}$ can be used to reduce the system because $\boldsymbol{a}$ and $\boldsymbol{b}$ are constants;

[44] Every now and then it is tried to find another law of the spacing of the planets. We refer to a systematic exploration in the book of Murray and Dermott [51] in a separate chapter (1.5).

[45] This derivation follows [65].

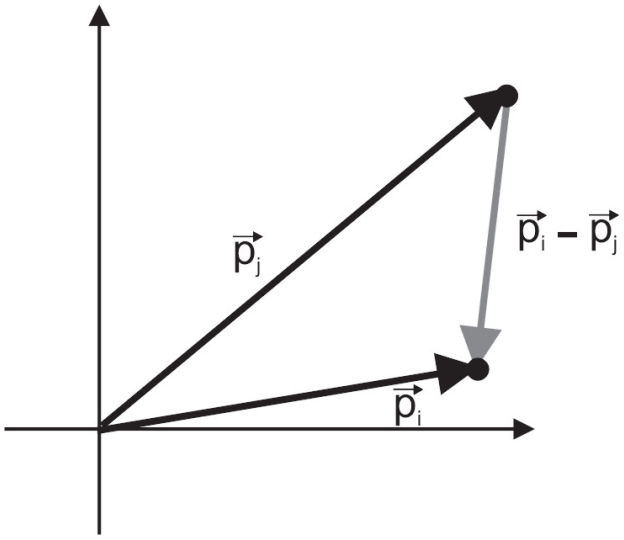

Fig. 51. Configuration in the inertial frame

the barycentre may serve as the new origin for the coordinate system. Besides these 6 constants of motions – the components of $\boldsymbol{a}$ and $\boldsymbol{b}$ – four more integrals of motion are the angular momentum integral (3 constants) and the energy integral $E_{kin} + E_{pot} = const.$ These 10 constants of motion are the classical integrals of the Newtonian n-body problem. In planetary theory, where one wants to describe the motion of the planets, we use the fact that more than 99.9 percent of the mass is in the Sun (m_0). Therefore it is useful to transform to a heliocentric relative coordinate system, where we denote the relative (heliocentric) vectors with $\boldsymbol{q}_i = \boldsymbol{p}_i - \boldsymbol{p}_0$.[46] To transform the equations of motion to a heliocentric system, we first separate in equation (320) the attraction of the Sun on the planet m_i from the other terms.

$$\ddot{\boldsymbol{p}}_i = k^2 \left[\frac{m_0}{r_{0i}^3}(\boldsymbol{p}_0 - \boldsymbol{p}_i) + \sum_{j=1, j\neq i}^{n} \frac{m_j}{r_{ij}^3}(\boldsymbol{p}_j - \boldsymbol{p}_i) \right] \tag{321}$$

and separate in the Sun's equation of motion the attracting force of the planet m_i on the Sun

$$\ddot{\boldsymbol{p}}_0 = k^2 \left[\frac{m_i}{r_{0i}^3}(\boldsymbol{p}_i - \boldsymbol{p}_0) + \sum_{j=1, j\neq i}^{n} \frac{m_j}{r_{j1}^3}(\boldsymbol{p}_j - \boldsymbol{p}_0) \right] \tag{322}$$

[46] which also is true for the velocities and the accelerations.

Subtracting now equation (322) from (321) and separating the vector $\boldsymbol{p}_i - \boldsymbol{p}_0$, we get the following – now heliocentric – equations of motion of the planet m_i

$$\ddot{\boldsymbol{q}}_i = k^2 \left[-\frac{m_0}{r_{0i}^3}(\boldsymbol{p}_i - \boldsymbol{p}_0) + \frac{m_i}{r_{0i}^3}(\boldsymbol{p}_i - \boldsymbol{p}_0) \right] + \\ + \sum_{j=1, j\neq i}^{n} \left[\frac{m_j}{r_{ji}^3}(\boldsymbol{p}_j - \boldsymbol{p}_i) - \frac{m_j}{r_{j0}^3}(\boldsymbol{p}_j - \boldsymbol{p}_0) \right] \tag{323}$$

which finally can be written as

$$\ddot{\boldsymbol{q}}_i = k^2 \left[-\frac{m_0 + m_i}{r_{0i}^3}\boldsymbol{q}_i + \sum_{j=1, j\neq i}^{n} m_j \left(\frac{\boldsymbol{q}_j - \boldsymbol{q}_i}{r_{ji}^3} - \frac{\boldsymbol{q}_j}{r_{j0}^3} \right) \right] \tag{324}$$

In this formulation it is evident why the orbits of the planets are – in the first approximation – ellipses, as we pointed out at the beginning: the first term describes an unperturbed two-body motion around the Sun, and the second term includes a factor that depends always on the other planets' masses, which are only in the order of 10^{-3} to 10^{-6} percent of the Sun's dominating mass. Consequently we can write for planetary motions the equations of motion in the form

$$\ddot{\boldsymbol{q}}_i = -k^2 \frac{m_0 + m_i}{r_{0i}^3}\boldsymbol{q}_i + \boldsymbol{P}_i \tag{325}$$

where the perturbing vector $\boldsymbol{P}_i$ reads

$$\boldsymbol{P}_i = k^2 \sum_{j=1, j\neq i}^{n} m_j \left(\frac{\boldsymbol{q}_j - \boldsymbol{q}_i}{r_{ji}^3} - \frac{\boldsymbol{q}_j}{r_{j0}^3} \right) \tag{326}$$

These n 2nd order differential equations are the equations of motion for all solar system bodies and apply to planets, asteroids and also comets. For the Moon it is better to use a coordinate system centered at the Earth; for satellites of the planets it is also better to work in a planetocentric system. In the perturbing vector $\boldsymbol{P}_i$ the appearance of the 3rd power of the distances $\mid \boldsymbol{q}_j - \boldsymbol{q}_i \mid = r_{ji}$ in the denominator may cause large accelerations, when these distances become small although the masses of the planets are small. In planetary theories this fact is not a problem, because planets move on well separated orbits. It is different for comets, which come quite often close to planets (especially Jupiter) who then changes their orbits significantly, so that comets, having initially a parabolic orbit, may end on an elliptic orbit with moderate eccentricity.[47] The equations of motion in this form are also not practical for "Near Earth Asteroids" like the Atens, Apollos and Amors on their high eccentric orbits, which lead them quite often close to a planet.

[47] These captures are responsible for the Jupiter family of comets.

One should mention the two different integration methods of solving the equations of motion (325):[48]

1. With the *numerical integration method*, the solution (only ONE) has to be calculated step by step (with regard to time), and can be found with any desired precision (but only for a limited time interval which depends on the problem; for comets and Near Earth Asteroids the time is in the order of years, for the planets in the order of ten thousands of years). The Nautical Almanac Service at the Naval observatory in Washington publishes ephemerides on the basis of numerical integrations provided by the JPL in Pasadena, which take into account even the perturbations of the major asteroids, make corrections due to the general relativity and incorporate precise astrometric observations.
2. The *general perturbation method* works with complicated series expansions including thousands of terms and computes the solutions for a whole bundle of initial conditions. The solution is such, that inserting the time in the series (also in form of a Fourier series) gives immediately the position in space (and on the sky). This method will be discussed in detail, because it gives deep insight in the nature of the dynamics of the planetary system!

The Lagrange Equations

The equations of motion of one planet perturbed by the other ones can also be written with the aid of a scalar, namely the *perturbing function* F_i

$$F_i = k^2 \sum_{j=1, j\neq i}^{n} m_j \left(\frac{1}{r_{ij}} - \frac{\boldsymbol{q}_i \boldsymbol{q}_j}{r_j^3} \right) \tag{327}$$

When we compute the gradient of F_i, we find

$$\nabla_i F_i = k^2 \sum_{j=1, j\neq i}^{n} m_j \left(\frac{\boldsymbol{q}_j - \boldsymbol{q}_i}{r_{ji}^3} - \frac{\boldsymbol{q}_j}{r_{j0}^3} \right) \tag{328}$$

which is identical to the perturbing acceleration $\boldsymbol{P}_i$ in equation (326). F_i consists of the direct part $\frac{1}{\rho}$ ($\rho = r_{ij} = |\boldsymbol{q}_j - \boldsymbol{q}_i|$) and the indirect part involving the inner product of the two heliocentric positions vectors $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$ of the planets. The direct part can be computed via the triangle consisting of $|\boldsymbol{q}_1| = r_1$, and $|\boldsymbol{q}_2| = r_2$ and ρ (see Fig. 52):

$$\rho^2 = r_1^2 - r_1 r_2 \cos\phi + r_2^2 \tag{329}$$

where we now restrict the calculations to one planet perturbed by another one; ϕ is the angle between the two position vectors of the planets involved:

[48] Examples of these methods are discussed in the appendix.

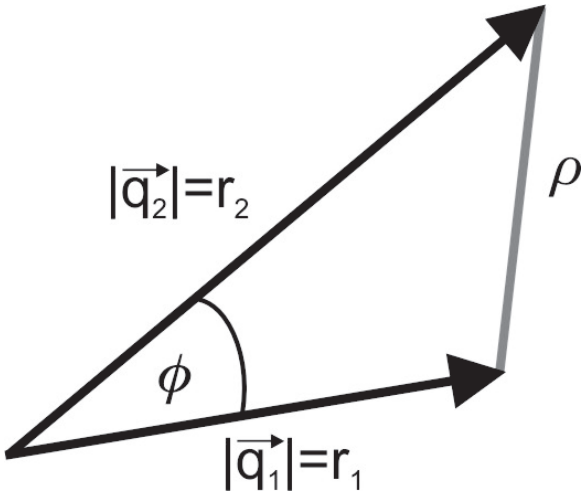

Fig. 52. Heliocentric position vectors

$$\frac{1}{\rho} = \frac{1}{r_2}\left[1 - 2\alpha\cos\phi + \alpha^2\right]^{\frac{1}{2}} \tag{330}$$

with $\alpha = r_1/r_2 < 1$ and thus $r_1 < r_2$, which means that we deal with the problem of an inner planet perturbed by an outer one.

We now can develop (330) with respect to the fact that $\alpha < 1$ into a power series of α and trigonometric functions of the angle ϕ. With the aid of the *Legendre polynomials*[49] the whole perturbing function for a planet with mass m_1 perturbed by an other planet with mass m_2 can be expressed as

$$F_{12} = \frac{k^2 m_2}{r_2}\sum_{n=2}^{\infty}\left(\frac{r_1}{r_2}\right)^n \wp_n(\cos\phi) \tag{331}$$

where $\wp_s$ stands for the Legendre polynomials. With some more algebra we end up with a Fourier series in time:

$$F_{12} = k^2 m_2 \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} C_{jk}\cdot\cos[(j\cdot n_1 + k\cdot n_2)t + D_{jk}] \tag{332}$$

where $C_{jk} = C_{jk}(a_1, e_1, i_1, a_2, e_2, i_2)$ are polynomial expressions in the small parameters inclination and eccentricity which can be derived from the above mentioned development into Legendre polynomials. Furthermore D_{jk} is a function of $\omega_1, \Omega_1, \omega_2$ and Ω_2.

The perturbations of other planets on a planet (m_1) may just be added leading to a perturbating function $F_1 = m_2F_{12} + m_3F_{13} + \ldots + m_nF_{1n}$ depending on the number of planets which are involved.

[49] see appendix D for details

As we have seen in a first approximation, the orbits of the planets are Kepler ellipses with constant elements, where the theory of the two-body-motion can be applied, which has been developed in Sect. 4 in its basic ideas. Under the attraction of the other planets, the constant orbital elements $\sigma_i^T = (a_i, e_i, i_i, \Omega_i, \omega_i, M_i)$[50] can be modeled with the aid of a set of 1st order differential equations – the *Lagrange equations*[51] where $F = \sum_{j=1}^{n} F_j$:

$$\begin{aligned}
\frac{da}{dt} &= \frac{2}{na}\frac{\partial F}{\partial M} \\
\frac{de}{dt} &= \frac{-\sqrt{1-e^2}}{na^2 e}\frac{\partial F}{\partial \omega} + \frac{1-e^2}{na^2 e}\frac{\partial F}{\partial M} \\
\frac{di}{dt} &= \frac{-1}{na^2\sqrt{1-e^2}\sin i}\frac{\partial F}{\partial \Omega} + \frac{\cos i}{na^2\sqrt{1-e^2}\sin i}\frac{\partial F}{\partial \omega} \\
\frac{d\Omega}{dt} &= \frac{1}{na^2\sqrt{1-e^2}\sin i}\frac{\partial F}{\partial i} \\
\frac{d\omega}{dt} &= \frac{\sqrt{1-e^2}}{na^2 e}\frac{\partial F}{\partial e} - \frac{\cos i}{na^2\sqrt{1-e^2}\sin i}\frac{\partial F}{\partial i} \\
\frac{dM}{dt} &= n - \frac{2}{na}\frac{\partial F}{\partial a} - \frac{1-e^2}{na^2 e}\frac{\partial F}{\partial e}
\end{aligned} \tag{333}$$

Perturbations in Delaunay Elements

To understand how mean motion resonances act in a planetary theory, we introduce the Delaunay elements (for $i = 1, \ldots, n$ planets)

$$\begin{aligned}
L_i &= \kappa_i\sqrt{a} \\
G_i &= L_i\sqrt{(1-e_i^2)} \\
H_i &= G_i\cos i_i \\
l_i &= M_i \\
g_i &= \omega_i \\
h_i &= \Omega_i
\end{aligned} \tag{334}$$

with $\kappa_i = k^2(m_i + m_0)$ where m_0 is the mass of the Sun. These canonical elements obey the canonical equations in the form of

$$\begin{aligned}
\frac{d\Gamma_i}{dt} &= \frac{\partial F_i}{\partial \gamma_i} \\
\frac{d\gamma_i}{dt} &= -\frac{\partial F_i}{\partial \Gamma_i}
\end{aligned} \tag{335}$$

[50] the ^{T} indicates transposition of the vector
[51] In this representation of the Lagrange equations the indices are omitted.

where the elements $\Gamma_i = (L_i, H_i, G_i)^T$ have as conjugate elements $\gamma_i = (l_i, h_i, g_i)^T$.

In analogy to the Hamilton formalism, we can use these conjugate elements to obtain a "semi-canonical" Hamilton function, which, in the case of two bodies ($i = 1, 2$) (a planet with mass m_1 perturbed by a planet with mass m_2) consists of an unperturbed part F_0 and the perturbation function F_{12}:

$$F = F_0 + F_{12} = \frac{\kappa_1^4}{2L_1^2} + m_2 \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{jk} \cdot \cos[(j \cdot n_1 + k \cdot n_2)t + D_{jk}] \quad (336)$$

The formulation of the perturbations shows that the mean anomaly $l = M$ is a special case, because we have to take into account F_0 in the second of equations (335), when we built the derivatives with respect to the conjugate Delaunay element L.

Using the equation for M

$$\frac{dM}{dt} = -\frac{\partial F_0}{\partial L} \quad (337)$$

just for the 2-body problem without the perturbation F_{12} and taking into account the definition of L gives for the right hand side

$$\frac{\kappa^4}{\kappa^3 a^{3/2}} = \kappa a^{-3/2} \equiv n \quad (338)$$

It is also clear that F_0 is the Hamiltonian of the 2-body problem which is the total energy H

$$H = \frac{v^2}{2} - \frac{\kappa^2}{r} \quad (339)$$

When we make use of the so-called velocity relation[52]

$$v^2 = \kappa^2 \left(\frac{2}{r} - \frac{1}{a} \right) \quad (340)$$

it follows that F_0 and H are in fact identical. For the 2-body problem one can see that the five Delaunay-elements are constant, but $M = nt$ is explicitely time dependant.

For now we concentrate on these other 5 Delaunay elements for which we can directly apply the integration after the derivation to get the perturbations. For any element – we now use the greek letter Γ for any of the five elements L, G, H, g and h and γ for their conjugate element – the form of the equations (335) is the following

[52] from which one can easily compute the velocity in the perihelion and aphelion position.

$$\frac{d\Gamma_i}{dt} = \pm m_i \frac{\partial F_i}{\partial \gamma_i} \tag{341}$$

Integrating we get the first order approximation

$$\Gamma_0 \pm \int_0^t \frac{\partial F_i}{\partial \gamma_i} dt = \Gamma_0 + \delta\Gamma \tag{342}$$

We show – as an example – the procedure to compute δL_1 where we need to build the derivative in formula (335) with respect to the conjugate Delaunay element l_1. Keeping in mind that the Delaunay element $l = M = nt$ and therefore $(j \cdot n_1 + k \cdot n_2)t = jl_1 + kl_2$ this expression reads

$$\delta L_1 = -m_2 \int_0^t \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{jk} \cdot j \sin[(j \cdot n_1 + k \cdot n_2)t + D_{jk}] dt \tag{343}$$

The integration leads to

$$\delta L_1 = m_2 \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{jk} \cdot j \frac{\cos[(j \cdot n_1 + k \cdot n_2)t + D_{jk}]}{jn_1 + kn_2} \tag{344}$$

For δL_2 the same small divisor $jn_1 + kn_2$ appears, a fact that will be discussed below. For G_i and H_i one has to build the derivative of $D_{j,k}$ with respect to the corresponding Delaunay element (e.g. for H_i with respect to h_i) and then to integrate:

$$\delta H_i = m_2 \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{jk} \cdot \frac{\partial D_{jk}}{\partial h_i} \frac{\cos[(j \cdot n_1 + k \cdot n_2)t + D_{jk}]}{jn_1 + kn_2} \tag{345}$$

For the elements g_i and h_i the perturbations read (we show it for h_i)

$$\delta h_i = -m_2 \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \frac{\partial C_{jk}}{\partial H_i} \cdot \frac{\cos[(j \cdot n_1 + k \cdot n_2)t + D_{jk}]}{jn_1 + kn_2} \tag{346}$$

We have to separate the term for $j = 0$ and $k = 0$ in the expansion of F (equation (336)) which is the term $C_{0,0} \cdot \cos D_{0,0}$; after integration it leads to a so-called secular term. This means that any Delaunay element – except L_i,[53] which we will treat later in detail – has a form

$$\Gamma = \Gamma_0 + \Gamma_1 t + \sum_{j,k \neq 0} \frac{E_{jk}}{j \cdot n_1 + k \cdot n_2} \cos[(j \cdot n_1 + k \cdot n_2)t + D_{j,k}] \tag{347}$$

We now will discuss the role of the small divisors and their connection with mean motion resonances. Whenever the ratio of the mean motions fulfills the condition of being closely commensurable

[53] in the semimajor axes no secular terms arise because for $j = 0$ and $k = 0$ the first order perturbation disappears.

$$\frac{n_1}{n_2} \approx \frac{k}{j} \tag{348}$$

this leads to a divisor which is close to zero and is called a *mean motion resonance*. Consequently the respective perturbations for such planets (or asteroids or comets) in a mean motion resonance are large, because the amplitudes $E_{j,k}$ are divided by a quantity close to zero. In fact in our Solar system we have the example of the giant planets Jupiter and Saturn where the mean motions are $n_{\rm Jup} = 0.^o08309/\text{day}$ and $n_{\rm Sat} = 0.^o03346/\text{day}$ which is close to the ratio 5 : 2. The perturbations acting between both planets have a small divisor

$$2n_{\rm Jup} - 5n_{\rm Sat} = 0.^o00112 \tag{349}$$

which leads to a period of $360^o/0.^o00112$ days corresponding to about 880 years. In fact this perturbation is rather large and causes a variation in the amplitude in Jupiter's longitude of almost $20^{'}$ and for Saturn $48^{'}$ with the above mentioned period.

Table 2. Mean motion resonances in the Solar System *(the * denotes exact mean motion resonances)*

System		Resonances		
Solar System	Jupiter–Saturn	5:2	Saturn–Uranus	3:1
	Uranus–Neptune	2:1	Neptune–Pluto*	3:2
Jupiter System	Io–Europa*	2:1	Europa–Ganymed*	2:1
Saturn System	Mimas–Tethys*	2:1	Enceladus–Dione*	2:1
	Dione–Rhea	5:3	Titan–Hyperion*	4:3
Uranus System	Miranda–Umbriel	3:1	Ariel–Umbriel	5:3
	Umbriel–Titania	2:1	Titania–Oberon	3:2

In our planetary system resonances occur quite often, which can be seen in Table 2 for the planets and for the satellite systems of the giant planets Jupiter, Saturn and Uranus:

Mean motion resonances are of special importance for the asteroids:

- The group of Trojan asteroid shows a 1:1 resonance with Jupiter with more than 1000 asteroids around the Lagrange point L_4 and about 600 around the Lagrange point L_5.
- The group of the Plutinos are asteroids in a 3:2 resonance with Neptune like Pluto itself (up to now we have evidence for some 100 of these bodies.[54])
- The role of mean motion resonances can be seen in the structure of the main belt asteroids between Mars and Jupiter (see Fig. 53) where one can see that for some resonances the number of asteroids is quite small; on

[54] these numbers significantly increase every year

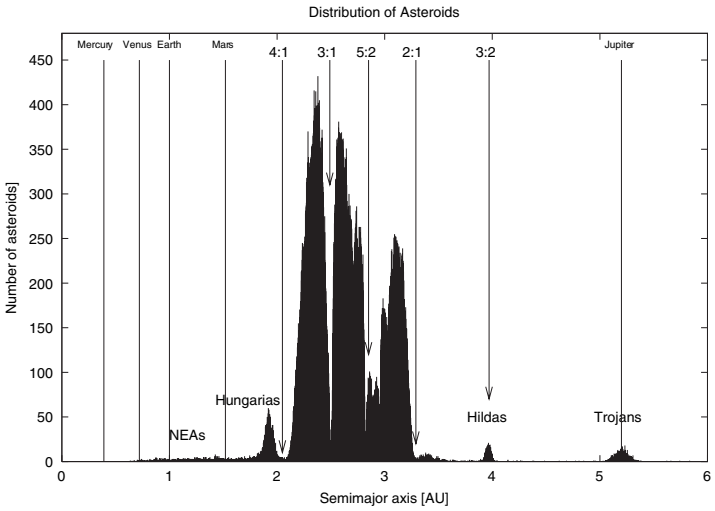

Fig. 53. The distribution of the asteroids in the inner Solar system

the other hand a resonance may protect an asteroid (e.g. visible in the 3:2 resonance).

First Order Perturbation for the Delaunay Element l

A separate treatment is necessary for the Delaunay element l, for which we have to take into account

$$\delta l_1 = \int_0^t \frac{\partial F_1}{\partial L_1} dt \tag{350}$$

Inserting the Fourier series for F_1 one has to evaluate the partial derivative with respect to the conjugate variable, which leads to an additional term for l because in the first term in equation (341) the conjugate element L_1 is present

$$-\frac{\partial F_0}{\partial L_1} = \frac{\kappa_1^4}{L_1^3} \tag{351}$$

such that we get the perturbations in the following form:

$$L_1 = L_1^{(0)} + m_2 \int_0^t \frac{\partial F_1}{\partial l} dt = L_1^{(0)} + \delta L_1 \tag{352}$$

We now find

$$\frac{dl_1}{dt} = -\left(\frac{\partial F_0}{\partial L_1} + m_2 \frac{\partial F_1}{\partial L_1}\right) = \kappa_1^4 \left(L_1^{(0)} + \delta L_1\right)^{-3} - m_2 \frac{\partial F_1}{\partial L_1} \tag{353}$$

Dividing the quantity $(L_1^{(0)} + \delta L_1)^{-3}$ by $L_1^{(0)}$ we develop the expression into a series and just keep the first term[55]

$$\frac{dl_1}{dt} = n_1^{(0)} \left(1 - 3\frac{\delta L_1}{L_1^{(0)}}\right) - m_2 \frac{\partial F_1}{\partial L_1} \tag{354}$$

where we also used the 3rd Kepler law: $n_1^{(0)} = a_1^{-3/2}\sqrt{\kappa_1^2}$. An integration with respect to time now leads to

$$l_1 = n_1^{(0)} t - m_2 \int_0^t \frac{\partial F_1}{\partial L_1} dt - 3\frac{n_1^{(0)}}{L_1^{(0)}} \int_0^t \delta L_1 dt \tag{355}$$

and furthermore

$$\begin{aligned} l_1 = &-m_2 \int_0^t \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \frac{\partial C_{j,k}}{\partial L_1} j \cos[(jn_1 + k \cdot n_2)t + D_{j,k}]dt - \\ &- 3\frac{n_1^{(0)}}{L_1^{(0)}} \int_0^t \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{j,k} j \frac{\cos[(j \cdot n_1 + kn_2)t + D_{j,k}]}{jn_1 + kn_2} dt + \\ &+ n_1^{(0)} t \end{aligned} \tag{356}$$

After integration this becomes

$$\begin{aligned} l_1 = &-m_2 \sum_{j=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \frac{\partial C_{j,k}}{\partial L_1} \frac{\sin[(jn_1 + k \cdot n_2)t + D_{j,k}}{(j \cdot n_1 + kn_2)} - m_2 \frac{\partial C_{00}}{\partial L_1} t \cos D_{00} - \\ &- 3\frac{n_1^{(0)}}{L_1^{(0)}} \sum_{j=-\infty}^{\infty} C_{j,k} j \frac{\sin[(jn_1 + kn_2)t + D_{j,k}]}{(jn_1 + kn_2)^2} + n_1^{(0)} t \end{aligned} \tag{357}$$

and finally we get

$$\begin{aligned} l_1 &= \left(n_1^{(0)} - m_2 \frac{\partial C_{00}}{\partial L_1} \cos D_{00}\right) t + \delta l_1 \\ &= \left(n_1^{(0)} + \delta n_1\right) t + \delta l_1 \end{aligned} \tag{358}$$

and a similar expression for

$$\begin{aligned} l_2 &= \left(n_2^{(0)} - m_2 \frac{\partial C_{00}}{\partial L_2} \cos D_{00}\right) t + \delta l_2 \\ &= \left(n_2^{(0)} + \delta n_2\right) t + \delta l_2 \end{aligned} \tag{359}$$

We can see that we have secular and periodic perturbations also for the mean anomalies. It is remarkable that the periodic perturbations – in contrary to

[55] we make use of the known formula $(1 + x)^{-3} = 1 - 3x + 6x^2 - \ldots$

the perturbations on the other Delaunay elements – have a denominator which causes very large amplitudes, due to the fact that we have to perform a second integration (which as a consequence leads to the square of the denominator in equation (357)).

A final statement should be made about the validity of the third law of Kepler: when we have δn_i and δa_i (the perturbations in mean motion and in semimajor axis) for a planet with mass m_i, then the relation

$$(n_i^{(0)} + \delta n_i)(a_i^{(0)} + \delta a_i)^{-3/2} = n_i^{(0)}(a_i^{(0)})^{-3/2} = \kappa_i \tag{360}$$

holds. We emphasize that this is not true for the mean values of a and n, which we derive from observations and which are published in the Nautical Almanac; for the herein published "mean" elements the 3rd Kepler law does not hold.

8 Secular Perturbation Theory

8.1 Introduction

Secular perturbation theory is a powerful tool for the investigation of planetary dynamics. Although the general N-body problem ($N \geq 3$) is non-integrable, one can find analytical solutions to particular problems by considering the secular terms of the disturbing functions. The derivation of the secular system and proper elements presented in the following was done according to [51] and follows closely this very clear presentation.

8.2 Two Planet Perturbations

At first, the motion of two planets (with masses m_1 and m_2 that orbit around a central mass m_0) under their mutual gravitational attraction will be investigated (with $m_1 << m_0$ and $m_2 << m_0$). The perturbations on the orbits of planet 1 and planet 2 can be described by the perturbation functions F_1 and F_2, which are functions of the osculating elements of the bodies. The perturbations on these orbital elements can be described by the Lagrangian planetary equations (333).

The secular perturbations can be obtained by isolating the terms in the disturbing functions that are independent from the mean longitude and that do not depend on the semimajor axis (it can be seen in Lagrange's equation for the element a that there is no influence on the secular evolution). Going to second order in the eccentricities and inclinations and first order in the masses gives then the general, averaged, secular direct part of the disturbing function:

$$F_{\rm D}^{\rm (sec)} = \frac{1}{8}\left[2\alpha_{12}D + \alpha_{12}^2 D^2\right] b_{\frac{1}{2}}^{(0)}\left(e_1^2+e_2^2\right) - \frac{1}{2}\alpha_{12}b_{\frac{3}{2}}^{(1)}\left(s_1^2+s_2^2\right)$$
$$+\frac{1}{4}\left[2-2\alpha_{12}D-\alpha_{12}^2D^2\right] b_{\frac{1}{2}}^{(1)} e_1e_2\cos\left(\overline{\omega}_1-\overline{\omega}_2\right)$$
$$+\alpha_{12}b_{\frac{3}{2}}^{(1)} s_1s_2\cos\left(\Omega_1-\Omega_2\right) \tag{361}$$

The subscripts 1 and 2 refer to the inner and the outer body; $\alpha_{12} = \frac{a_1}{a_2}$ with $a_1 < a_2$; $b_s^{(\rm j)}(\alpha)$ denotes the Laplace coefficients[56] and $s_j = \sin\frac{1}{2}i_j$.

F_1 and F_2 can now be derived from $F_{\rm D}^{\rm (sec)}$:

$$F_1 = \frac{Gm_2}{a_2}F_{\rm D}^{\rm (sec)} = \frac{Gm_2}{a_1}\alpha_{12}F_{\rm D}^{\rm (sec)} \tag{362}$$
$$F_2 = \frac{Gm_1}{a_1}\alpha_{12}F_{\rm D}^{\rm (sec)} = \frac{Gm_1}{a_2}F_{\rm D}^{\rm (sec)} \tag{363}$$

Using the relations between the Laplace coefficients and their derivatives and substituting $Gm_0 \approx n_1^2a_1^3 \approx n_2^2a_2^3$ one obtains:

$$F_1 = n_1^2a_1^2\frac{m_2}{m_0+m_1}\left[\frac{1}{8}\alpha_{12}^2b_{\frac{3}{2}}^{(1)}e_1^2 - \frac{1}{8}\alpha_{12}^2b_{\frac{3}{2}}^{(1)}i_1^2\right.$$
$$-\frac{1}{4}\alpha_{12}^2b_{\frac{3}{2}}^{(2)}e_1e_2\cos\left(\tilde{\omega}_1-\tilde{\omega}_2\right)$$
$$\left.+\frac{1}{4}\alpha_{12}^2b_{\frac{3}{2}}^{(1)}i_1i_2\cos\left(\Omega_1-\Omega_2\right)\right] \tag{364}$$

$$F_2 = n_2^2a_2^2\frac{m_1}{m_0+m_2}\left[\frac{1}{8}\alpha_{12}b_{\frac{3}{2}}^{(1)}e_2^2 - \frac{1}{8}\alpha_{12}b_{\frac{3}{2}}^{(1)}i_1^2\right.$$
$$-\frac{1}{4}\alpha_{12}b_{\frac{3}{2}}^{(2)}e_1e_2\cos\left(\tilde{\omega}_1-\tilde{\omega}_2\right)$$
$$\left.+\frac{1}{4}\alpha_{12}^2b_{\frac{3}{2}}^{(1)}i_1i_2\cos\left(\Omega_1-\Omega_2\right)\right] \tag{365}$$

We emphasize that these equations are only valid for small values of i_1 and i_2 when the approximations $s_1 = \sin\frac{1}{2}i_1 \approx \frac{1}{2}i_1$ and $s_2 = \sin\frac{1}{2}i_2 \approx \frac{1}{2}i_2$ are correct.

Both equations can be combined to

$$F_j = n_ja_j^2\left[\frac{1}{2}A_{jj}e_j^2 + A_{jk}e_1e_2\cos(\tilde{\omega}_1-\tilde{\omega}_2) + \frac{1}{2}B_{jj}I_j^2 + B_{jk}I_1I_2\cos(\Omega_1-\Omega_2)\right] \tag{366}$$

with $j = 1,2; k = 2,1 (j \neq k)$ and

[56] for details see appendix C.

$$A_{jj} = +n_j \frac{1}{4} \frac{m_k}{m_0 + m_j} \alpha_{12} \overline{\alpha}_{12} b^{(1)}_{\frac{3}{2}}(\alpha_{12}) \tag{367}$$

$$A_{jk} = -n_j \frac{1}{4} \frac{m_k}{m_0 + m_j} \alpha_{12} \overline{\alpha}_{12} b^{(2)}_{\frac{3}{2}}(\alpha_{12}) \tag{368}$$

$$B_{jj} = -n_j \frac{1}{4} \frac{m_k}{m_0 + m_j} \alpha_{12} \overline{\alpha}_{12} b^{(1)}_{\frac{3}{2}}(\alpha_{12}) \tag{369}$$

$$B_{jk} = +n_j \frac{1}{4} \frac{m_k}{m_0 + m_j} \alpha_{12} \overline{\alpha}_{12} b^{(2)}_{\frac{3}{2}}(\alpha_{12}) \tag{370}$$

with $\overline{\alpha}_{12} = \alpha_{12}$ if the perturbation is external ($j = 1$) and $\overline{\alpha}_{12} = 1$ if the perturbation is internal ($j = 2$).

The A_{ij} and B_{ij} can also be written as elements of two matrices $\mathbf{A}$ and $\mathbf{B}$ with elements that only depend on the masses and the semimajor axes of the two bodies.

If one takes only terms of the lowest order in e and i in the Lagrange equations (333), one obtains a simplified form of the Lagrange equations:

$$\dot{e}_j = -\frac{1}{n_j a_j^2 e_j} \frac{\partial F_j}{\partial \tilde{\omega}_j} \tag{371}$$

$$\dot{\tilde{\omega}}_j = +\frac{1}{n_j a_j^2 e_j} \frac{\partial F_j}{\partial e_j} \tag{372}$$

$$\dot{i}_j = -\frac{1}{n_j a_j^2 i_j} \frac{\partial F_j}{\partial \Omega_j} \tag{373}$$

$$\dot{\Omega}_j = +\frac{1}{n_j a_j^2 i_j} \frac{\partial F_j}{\partial i_j} \tag{374}$$

One can see that in equations (371)–(374) singularities can appear, if e or i become very small. Therefore a new set of variables will be used:

$$h_j = e_j \sin \tilde{\omega}_j \tag{375}$$

$$k_j = e_j \cos \tilde{\omega}_j \tag{376}$$

$$p_j = i_j \sin \Omega_j \tag{377}$$

$$q_j = i_j \cos \Omega_j \tag{378}$$

Equation (366) can now be written as

$$F_j = n_j a_j^2 \left[\frac{1}{2} A_{jj} \left(h_j^2 + k_j^2 \right) + A_{jk} \left(h_j h_k + k_j k_k \right) + \right.$$
$$\left. + \frac{1}{2} B_{jj} \left(p_j^2 + q_j^2 \right) + B_{jk} \left(p_j p_k + q_j q_k \right) \right] \tag{379}$$

The Lagrange equations in the new variables are given by

$$\dot{h}_j = +\frac{1}{n_j a_j^2}\frac{\partial F_j}{\partial k_j} \tag{380}$$

$$\dot{k}_j = -\frac{1}{n_j a_j^2}\frac{\partial F_j}{\partial h_j} \tag{381}$$

$$\dot{p}_j = +\frac{1}{n_j a_j^2}\frac{\partial F_j}{\partial q_j} \tag{382}$$

$$\dot{q}_j = -\frac{1}{n_j a_j^2}\frac{\partial F_j}{\partial p_j} \tag{383}$$

These differential equations for the non-singular elements can be written as

$$\dot{h}_1 = +A_{11}k_1 + A_{12}k_2 \tag{384}$$

$$\dot{h}_2 = +A_{21}k_1 + A_{22}k_2 \tag{385}$$

$$\dot{k}_1 = -A_{11}h_1 - A_{12}h_2 \tag{386}$$

$$\dot{k}_2 = -A_{21}h_1 - A_{22}h_2 \tag{387}$$

$$\dot{p}_1 = +B_{11}q_1 + B_{12}q_2 \tag{388}$$

$$\dot{p}_2 = +B_{21}q_1 + B_{22}q_2 \tag{389}$$

$$\dot{q}_1 = -B_{11}p_1 - B_{12}p_2 \tag{390}$$

$$\dot{q}_2 = -B_{21}p_1 - B_{22}p_2 \tag{391}$$

The advantage of this representation is that – at least for the first order – the development of h_j and k_j is decoupled from that of p_j and q_j. Because these equations are linear differential equations with constant coefficients, they can be solved by using the eigenvalues and eigenvectors of the matrices **A** and **B**:

$$h_j = \sum_{i=1}^{2} e_{ji} \sin(g_i t + \beta_i) \tag{392}$$

$$k_j = \sum_{i=1}^{2} e_{ji} \cos(g_i t + \beta_i) \tag{393}$$

$$p_j = \sum_{i=1}^{2} I_{ji} \sin(f_i t + \gamma_i) \tag{394}$$

$$q_j = \sum_{i=1}^{2} I_{ji} \cos(f_i t + \gamma_i) \tag{395}$$

where the frequencies g_i and f_i are the eigenvalues of the matrices **A** and **B** and e_{ji} and I_{ji} are the components of the corresponding eigenvectors. The phases β_i and γ_i (and also the amplitude of the eigenvectors) are determined by the initial conditions.

The solution described by equations (392)–(395) is known as the classical Laplace-Lagrange secular solution.

An important difference in the solutions for $\{h,k\}$ and $\{p,q\}$ is worth mentioning, namely that the characteristic equation for matrix $\mathbf{B}$ is

$$\begin{vmatrix} B_{11}-f & B_{12} \\ B_{21} & B_{22}-f \end{vmatrix} = 0$$

or

$$f\left(f-(B_{11}+B_{22})\right)=0 \tag{396}$$

Therefore one solution of the characteristic equation $f_1 = 0$ and consequently a degeneracy exists in the problem. This stems from the definition of the non-singular variables: $\{h,k\}$ involve the eccentricity and $\{p,q\}$ the inclination. Whereas an eccentric orbit causes an asymmetry (and thus a "natural" reference line), there is no sense in introducing a reference plane on a spherical or point-mass body and therefore it is only meaningful to talk about *mutual inclinations*.[57]

It is also an interesting fact that the solution (392)–(395) does not depend on the mean longitude M and therefore one is only able to predict the variations in eccentricity, inclination, pericenters and nodes; the positions of the planets cannot be determined via a secular theory.

8.3 Proper Elements

After establishing the secular theory of two bodies, we can now formulate the disturbing function for a third massless body (e.g. an asteroid perturbed by Jupiter) with the orbital elements $a, n, e, i, \tilde{\omega}$ and Ω:

$$F = na^2\left[\frac{1}{2}Ae^2+\frac{1}{2}Bi^2+\sum_{j=1}^{2}A_j e e_j\cos\left(\tilde{\omega}-\tilde{\omega_j}\right)+\sum_{j=1}^{2}B_j i i_j\cos\left(\Omega-\Omega_j\right)\right] \tag{397}$$

with

$$A = +n\frac{1}{4}\sum_{j=1}^{2}\frac{m_j}{m_0}\alpha_j\overline{\alpha_j}b_{\frac{3}{2}}^{(1)}(\alpha_j) \tag{398}$$

$$A_j = -n\frac{1}{4}\frac{m_j}{m_0}\alpha_j\overline{\alpha_j}b_{\frac{3}{2}}^{(1)}(\alpha_j) \tag{399}$$

$$B = -n\frac{1}{4}\sum_{j=1}^{2}\frac{m_j}{m_0}\alpha_j\overline{\alpha_j}b_{\frac{3}{2}}^{(1)}(\alpha_j) \tag{400}$$

$$B_j = +n\frac{1}{4}\frac{m_j}{m_0}\alpha_j\overline{\alpha_j}b_{\frac{3}{2}}^{(1)}(\alpha_j) \tag{401}$$

and

[57] By introducing non-spherical planet terms this degeneracy problem would be removed.

$$\alpha_j = \begin{cases} \frac{a_j}{a} \text{ if } a_j < a \\ \frac{a}{a_j} \text{ if } a_j > a \end{cases}$$

$$\overline{\alpha_j} = \begin{cases} 1 \text{ if } a_j < a \\ \frac{a}{a_j} \text{ if } a_j > a \end{cases}$$

When we use again the non-singular variables h, k, p and q the perturbing function reads

$$F = na^2 \left[\frac{1}{2} A \left(h^2 + k^2 \right) + \frac{1}{2} B \left(p^2 + q^2 \right) + \right.$$
$$\left. + \sum_{j=1}^{1} A_j \left(h h_j + k k_j \right) + \sum_{j=1}^{1} B_j \left(p p_j + q q_j \right) \right] \tag{402}$$

For the equations of motion, one obtains

$$\dot{h} = +\frac{1}{na^2} \frac{\partial F}{\partial k} \tag{403}$$

$$\dot{k} = -\frac{1}{na^2} \frac{\partial F}{\partial h} \tag{404}$$

$$\dot{p} = +\frac{1}{na^2} \frac{\partial F}{\partial q} \tag{405}$$

$$\dot{q} = -\frac{1}{na^2} \frac{\partial F}{\partial p} \tag{406}$$

Differentiation of (402) and inserting in (403)–(406) gives

$$\dot{h} = +Ak + \sum_{j=1}^{2} A_j k_j \tag{407}$$

$$\dot{k} = -Ah - \sum_{j=1}^{2} A_j h_j \tag{408}$$

$$\dot{p} = +Bq + \sum_{j=1}^{2} B_j q_j \tag{409}$$

$$\dot{q} = -Bp + \sum_{j=1}^{2} B_j p_j \tag{410}$$

The values for h_j, k_j, p_j and q_j can now be obtained from the solutions of the secular system (392)–(395), which gives

$$\dot{h} = +Ak + \sum_{j=1}^{2} A_j \sum_{j=1}^{2} e_{ji} \cos\left(g_i t + \beta_i \right) \tag{411}$$

$$\dot{k} = -Ah + \sum_{j=1}^{2} A_j \sum_{j=1}^{2} e_{ji} \sin(g_i t + \beta_i) \tag{412}$$

$$\dot{p} = +Bq + \sum_{j=1}^{2} B_j \sum_{j=1}^{2} i_{ji} \cos(f_i t + \gamma_i) \tag{413}$$

$$\dot{h} = -Bp + \sum_{j=1}^{2} B_j \sum_{j=1}^{2} i_{ji} \sin(f_i t + \gamma_i) \tag{414}$$

A solution can be obtained by building the second derivatives with respect to the time; see (411)–(414):

$$\ddot{h} = -A^2 h - \sum_{j=1}^{2} \nu_i (A + g_i) \sin(g_i t + \beta_i) \tag{415}$$

$$\ddot{k} = -A^2 k - \sum_{j=1}^{2} \nu_i (A + g_i) \cos(g_i t + \beta_i) \tag{416}$$

$$\ddot{p} = -B^2 p - \sum_{j=1}^{2} \mu_i (B + f_i) \sin(f_i t + \gamma_i) \tag{417}$$

$$\ddot{q} = -B^2 q - \sum_{j=1}^{2} \mu_i (B + f_i) \cos(f_i t + \gamma_i) \tag{418}$$

with

$$\nu_i = \sum_{j=1}^{2} A_j e_{ji} \tag{419}$$

$$\mu_i = \sum_{j=1}^{2} B_j i_{ji} \tag{420}$$

Equations (415)–(418) are a system of uncoupled differential equations with the solution

$$h = e_{\text{free}} \sin(At + \beta) + h_0(t) \tag{421}$$

$$k = e_{\text{free}} \cos(At + \beta) + k_0(t) \tag{422}$$

$$p = i_{\text{free}} \sin(Bt + \gamma) + p_0(t) \tag{423}$$

$$q = i_{\text{free}} \cos(Bt + \gamma) + q_0(t) \tag{424}$$

The constants $e_{\text{free}}, i_{\text{free}}, \beta$ and γ can be determined by the boundary conditions and the functions $h_0(t), k_0(t), p_0(t)$ and $q_0(t)$ are given by

$$h_0(t) = -\sum_{i=1}^{2} \frac{\nu_i}{A - g_i} \sin(g_i t + \beta_i) \tag{425}$$

$$k_0(t) = -\sum_{i=1}^{2} \frac{\nu_i}{A - g_i} \cos(g_i t + \beta_i) \tag{426}$$

$$p_0(t) = -\sum_{i=1}^{2} \frac{\mu_i}{B - f_i} \sin(f_i t + \gamma_i) \tag{427}$$

$$q_0(t) = -\sum_{i=1}^{2} \frac{\mu_i}{B - f_i} \cos(f_i t + \gamma_i) \tag{428}$$

It can be seen that $h_0(t), k_0(t), p_0(t)$ and $q_0(t)$ are only functions of the semi-major axes (which are constant) and do not depend on any other orbital elements of the test particle. Nevertheless they will vary with time because they also depend on the secular solution of the two perturbing bodies.

Introducing

$$e_{\text{forced}} = \sqrt{h_0^2 + k_0^2} \tag{429}$$

$$i_{\text{forced}} = \sqrt{p_0^2 + q_0^2} \tag{430}$$

allows a simple geometric interpretation of the solution (421)–(424).

Figures 54 and 55 show the connection between the free and forced elements: The solution for h and k defines a point in the (h, k) plane – a vector from the origin to this point has the length e and encloses an angle ϖ. This vector can also be constructed by the sum of two other vectors: one with a

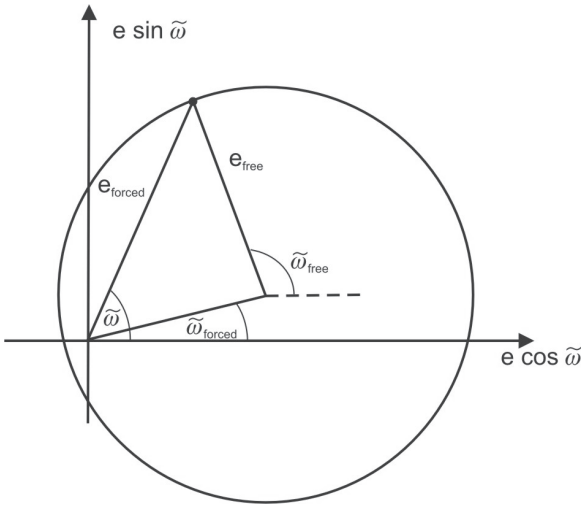

Fig. 54. Geometrical relationship between osculating, free and forced eccentricities and longitudes of pericentre ($e_{\text{free}} > e_{\text{forced}}$) *(after [51])*

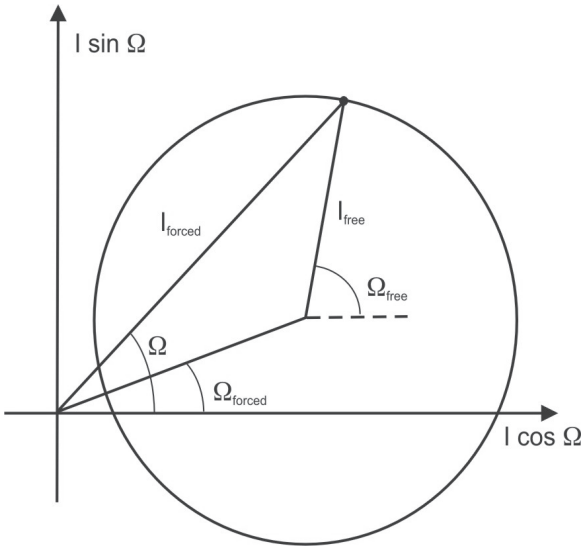

Fig. 55. Geometrical relationship between osculating, free and forced inclinations and longitudes of ascending nodes($i_{\text{free}} < i_{\text{forced}}$) *(after [51])*

length of e_{forced} and an angle $\tilde{\omega}_{\text{forced}}$ going from the origin to the point (h_0, k_0) and the second going from there to (h, k) with a length of e_{free} and an angle $\tilde{\omega}_{\text{free}} = At + \beta$. Thus in the plane h–k the motion of the particle resembles the motion of a circle centered at (h_0, k_0) at a constant rate A; the center-point itself moves according to the secular solution of the two perturbing bodies. The two quantities e_{forced} and $\tilde{\omega}_{\text{forced}}$ depend only on the semimajor axis of the test particle and the secular solutions; e_{free} and $\tilde{\omega}_{\text{free}}$ are the *fundamental orbital parameters* of the test particle. They are called *proper elements* of the particle's orbits.

A similar geometrical explanation exists for the relation between the osculating, free and forced inclinations and longitudes of ascending nodes.

9 The chaotic Motion of the Planets and the Stability of the Solar System

This section deals with the problem of the long-term stability of the Solar System. This question was posed in 1889 in a contest which was held as part of the 60^{th} birthday of King Oscar II, of Sweden and Norway. For this contest, mathematicians were invited to write an original paper where one question concerned celestial mechanics and the stability of a collection of orbiting bodies:

"Given a system of arbitrarily many mass points that attract each other according to Newton's laws, assuming that no two points ever collide, give the coordinates of the individual points for all time as the sum of a uniformly convergent series whose terms are made up of known functions."

Henri Poincaré (1854–1912, Fig. 56) wrote a prize paper,[58] where he showed that the perturbation techniques to represent the solution in form of power series of small parameters like the eccentricities, inclinations and/or the planets' masses are not convergent due to small divisors. As a consequence these solutions are not valid for infinite times.

Fig. 56. Henri Poincaré (1854-1912) *(picture taken from MacTutor History of Mathematics archive, University of St Andrews Scotland)*

Moreover Poincaré was well in advance his time with his deep insight in dynamical problems: in his famous work *Methodes nouvelles de la Mécanique céleste* he laid down the basic ideas of chaotic motion, something which was rediscovered and explored in detail only some fifty years later.

[58] Although he did not answer the question, because of his brilliant ideas the member of the jury were in favor to give him the prize.

It is evident that one could use a straight forward integration of the equations of motion to follow the orbit of a planet in the solar system. But even for the best integrators the accuracy is lost for chaotic systems, because for non-regular motions the nonlinear dependence of the initial conditions creates an inacurracy in the results due to the accumulation of the round off errors. Nevertheless the orbits are well determined in a qualitative way over long time scales. One may try to find a better solution – valid for long times with the aid of analytical perturbation theory. But we know from the preceding chapter that, due to the presence of small divisors in the series expansions of the solutions, the series do not converge!

This section will first present numerical integrations for the long time motion of the bodies in the Solar System and then will show results obtained from semi-analytical approaches.

9.1 Numerical Solutions

There exist many different papers on the long-term evolution of the planetary system (e.g. [40] and [32]). We report the results of an integration of the full equations of motion in a model consisting of the whole planetary system from Mercury to Neptune (planets as point-masses, Earth+Moon as one body, no relativistic effects) which demonstrate how small the variation for the orbital elements in fact are [13]. The integration of the differential equations of motion was done with the Lie-Integration Method [28], which is explained in detail in appendix A. The time interval covered was 10^8 years (from -500 Myr to +500 Myr). As examples we show in Fig. 57 the time development of the elements a, e and i for Mercury for 200 million years (from -100 million years to +100 million years). One can see the relative large variations in the elements eccentricity and inclination and the almost constant value for the semimajor axes. Additionally in Fig. 58 we plotted the respective elements from 300 Myrs to 500 Myrs. The plot of the semimajor axis shows the precision of the method used, because the first sign of the lack of precision of an integration is a secular drift of the semimajor axes of the most inner planet in a simulation of the motion of the planetary system.

In Fig. 59 we can see the strong coupling between the planets Earth and Venus, which have almost the same masses and therefore suffer from almost the same perturbations in the elements. We show – besides the time interval -2.5 to $+2.5$ million years around the present epoch – the time evolution of their eccentricities for the first 5 million years and the last 5 million years of our 200 million years integration. It is remarkable that no qualitative differences at all can be seen there!

Finally we show the eccentricities for the outer planets in Fig. 60: for Jupiter and Saturn the coupling due to the 5:2 resonance is well visible; for the outer planets the variation in eccentricity and inclinations (not shown here) is more regular than for the inner planets. In Table 3 we show also the

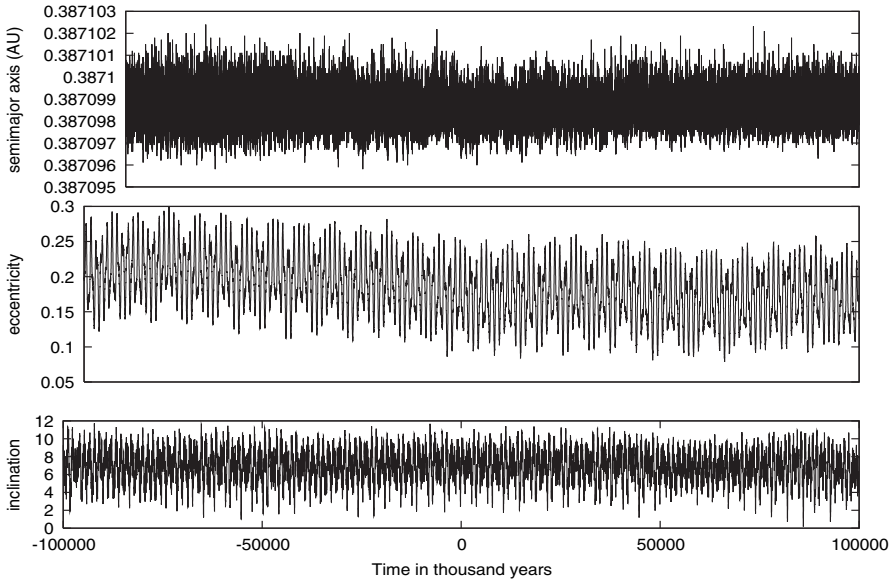

Fig. 57. Semimajor axis, eccentricity and inclination of Mercury from -100 Myr to 100 Myr

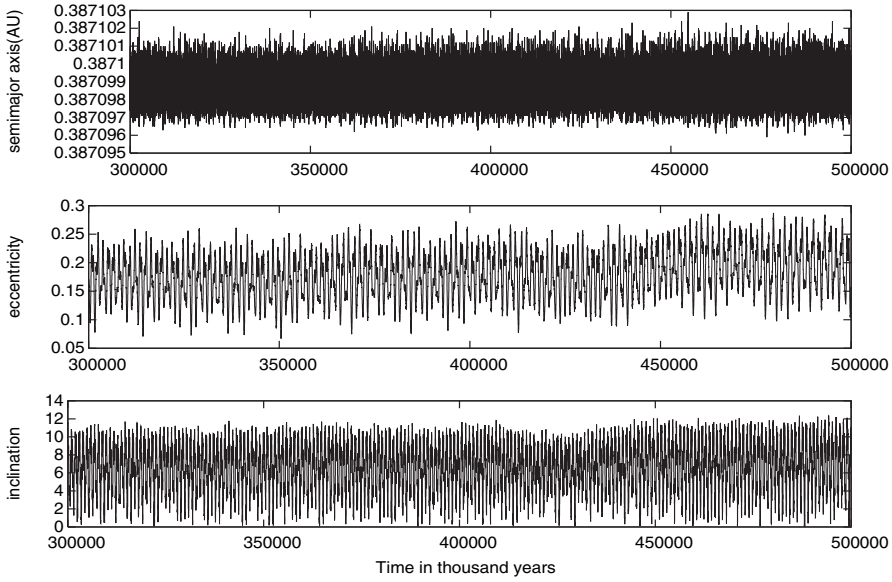

Fig. 58. Semi-major axis, eccentricity and inclination of Mercury from 300 Myr to 500 Mrs

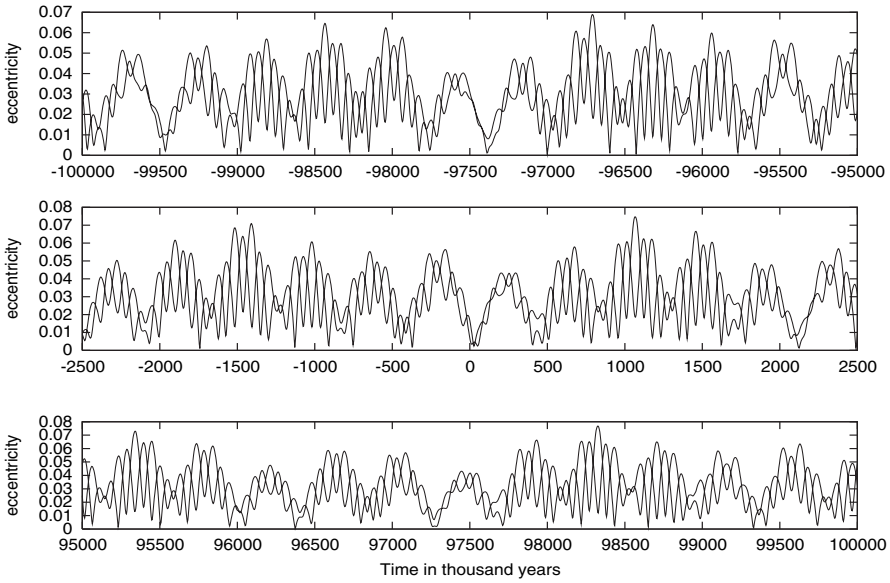

Fig. 59. Eccentricities of Earth and Venus

smallest (respectively largest) values of the three action like variables for all the planets.

Table 3. Maximum, minimum and actual values of the action like variable for 10^9 years

planet	a_{min}	a_{max}	a_{now}	e_{min}	e_{max}	e_{now}	i_{min}	i_{max}	i_{now}
Mercury	0.38710	0.38710	0.38710	0.07874	0.29988	0.20562	0.17600	11.72747	7.00468
Venus	0.72332	0.72336	0.72333	0.00002	0.07709	0.00680	0.00076	4.91515	3.39467
Earth	0.99997	1.00004	1.00000	0.00002	0.06753	0.01668	0.00075	4.49496	0.00065
Mars	1.52354	1.52386	1.52369	0.00008	0.13110	0.09342	0.00291	8.60320	1.84937
Jupiter	5.20122	5.20504	5.20187	0.02513	0.06191	0.04895	1.09172	2.06598	1.30377
Saturn	9.51281	9.59281	9.56872	0.00742	0.08959	0.05612	0.55867	2.60187	2.48618
Uranus	19.09807	19.33511	19.13833	0.00008	0.07835	0.04991	0.42170	2.73888	0.77187
Neptune	29.91013	30.32452	29.99772	0.00001	0.02317	0.00761	0.77977	2.38597	1.77179

9.2 A semi-analytical approach

By using a mixture of methods, namely representing the right hand sides of the secular equations (compare Sect. 8) via series expansions and a numerical integrations of these equations one can determine – again only for a limited but significant larger time span – the orbits of the planets qualitatively (without giving the positions on the sky).

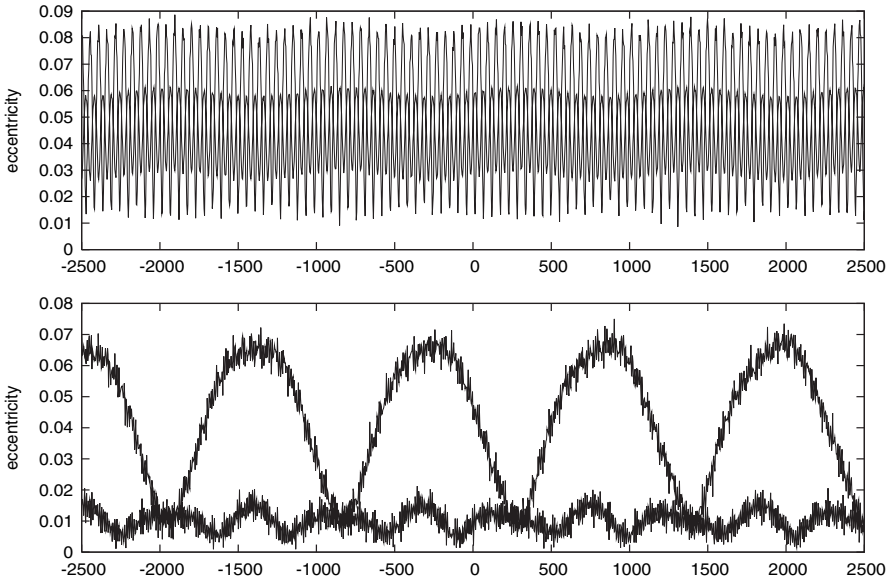

Fig. 60. Eccentricities of Jupiter and Saturn *(upper graph)* and Uranus and Neptune; (large (respectively smaller) amplitudes *lower graph)*.

Using this method of averaging over the fast angle, the mean motion of the planets,[59] J. Laskar [35, 37, 38] from the Bureau des longitudes in Paris[60] succeeded to find qualitative solutions of the motions of the planets up to Gigayears. His work consisted of two steps: first he developed with the aid of computer algebra the secular part of the differential equations of motion (see Sect. 8) up to the second order in the masses and to the order five with respect to the small inclinations and eccentricities of the orbits of all eight planets. These truncated secular equations consist of several hundred thousand terms. The next step was to integrate numerically these differential equations via an adequate numerical method with a rather big time step of 500 years. In a first run Laskar integrated the equation over 200 million years and determined the long periodic terms of the system, which are connected to the motions of the nodes and of the perihelions of the planets *which are the fundamental frequencies of the secular system*. There were two main secular resonances present among the planets, namely

- $\theta = 2(g_4 - g_3) - (s_4 - s_3)$ related to Mars and the Earth and
- $\sigma = (g_1 - g_5) - (s_1 - s_2)$ related to Mercury, Venus and Jupiter;

[59] which means to keep the semi-major axis of all planets constant.
[60] now as IMCCE part of the Observatoire de Paris.

The frequencies g_n and s_n involved refer to the secular motion of Ω_n and ω_n of each planet n. All these resonances and their linear combinations were found with the aid of a sophisticated frequency analysis [35] of the numerical solutions of the secular systems; the forementioned combinations of the frequencies turned out to have quite large amplitudes. During the evolution of the system the fundamental frequencies itself change slightly, a fact that leads to a characteristic change from libration to circulation several times during the time interval from $-200 \cdot 10^6$ years to $200 \cdot 10^6$ years. This is shown in Fig. 61.

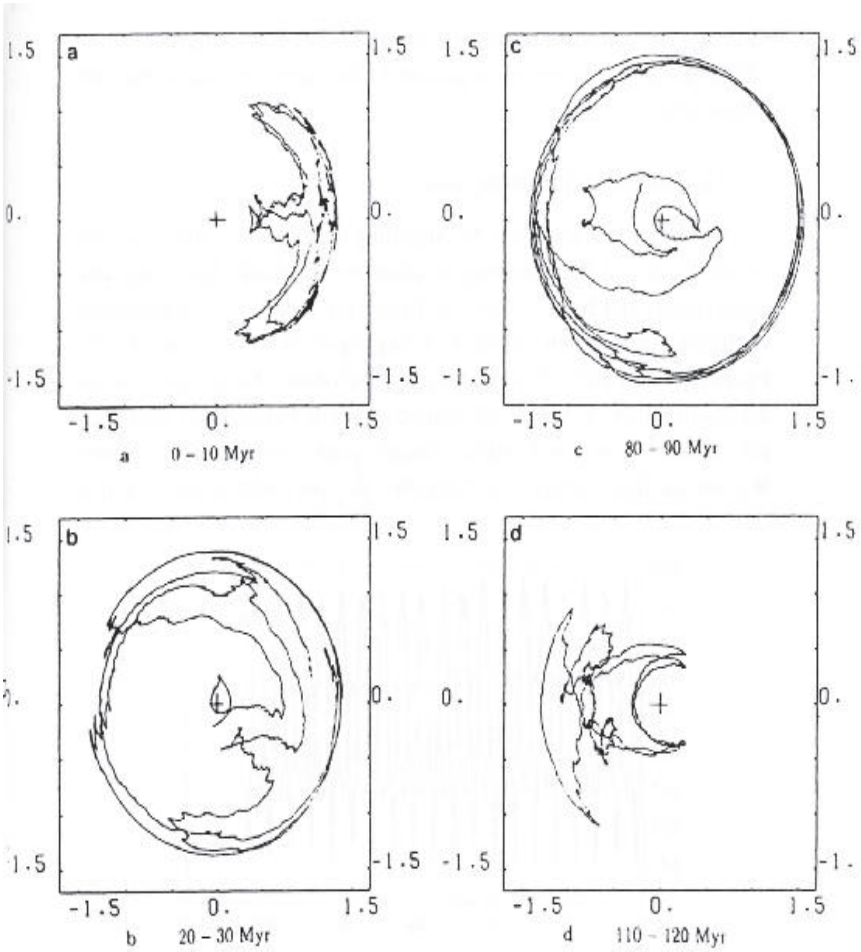

Fig. 61. Change of libration to circulation *(left graphs)* and from circulation to libration *(right graphs)* for the critical argument θ *(from Laskar [35])*

This qualitative change of the motion is a typical sign of chaotic behavior. Think of the perturbed pendulum having a motion close to the separatrix: the slightest change in the initial conditions may lead to a separatrix crossing – and the pendulum is no longer librating but circulates and vice versa. In a comparison with an integration of the full equations of motion the quantitative behavior of the argument θ was confirmed for the first 6 million years [36] A direct comparison for the whole time of 500 million years is under construction.

In a continuation of the computations Laskar extended the time scale up to several 10^9 years, which could only done by a further simplification of the right hand sides of the secular equations of motion. This was accomplished by dropping the small amplitudes which led to a significant smaller number of terms ($\sim$ 5000 terms). These simplified equations of motion could be integrated now much faster, although the system had still 15 independent degrees of freedom. Because of the further simplification one can doubt the significance of results which he found, which was also expressed by the author himself, who used a different argument: *"Indeed, because of the exponential divergence with a Lyapunov time of 5 Myr, after 100 Myr the computed solution will be very different from the real solution followed by the actual solar system"*.

His main idea was to integrate several different solutions with slightly different initial conditions up to times as long as some 10^9 years to find the diffusion time for the action-like variables inclination and eccentricity. Therefore, in his integrations, he used backward and forward slightly different conditions for the moments when the eccentricity of Mercury was large. It is evident that the crucial orbital element is the eccentricity, because even with a fixed value of the semi-major axes, as it has been done by averaging over the short periods, large eccentricities may lead to orbit crossings. Unlike that of the inner planets, the outer planets have almost no variations in their eccentricities. Fig. 62 gives the values for Mercury, Venus, Earth and Mars: The eccentricity of Mercury was large, and his computations led finally, after

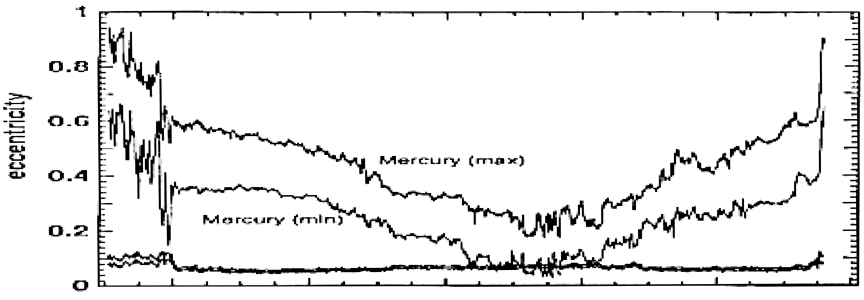

Fig. 62. Evolution of the eccentricities of Venus, Earth (lowest lines) and Mercury from −6.6 Gyrs to 3.5 Gyrs *(from Laskar [38])*

several 10^9 years, to a value close to 1. As a consequence, a close approach to Venus could happen because their orbits now cross. Thus, as a consequence, Mercury could escape from the Solar System! But, as also the author emphasized, *the solution computed here which led to an escape of Mercury was very carefully tailored by selecting at each step one solution among 4 or 5 equivalent ones*. And he finally concludes that *From the present computation, it can be thought that this probability is small, but not null, which is compatible with the present existence of Mercury.*[61]

9.3 Conclusions

After all what can be concluded for the long-term stability of our planetary system? Is there any final answer to the question of stability which astronomers are concerned with since the days of Lagrange? From both approaches we only can say for sure, that – also being in a state of chaos as shown by the computed Lyapunov exponent and also the qualitative change of some critical angles – within some 10^9 years of evolution – the overall qualitative behavior (but also the quantitative one) did not change significantly within the last Gigayears of evolution and will not change for the next Gigayears from the dynamical point of view.[62]

10 Terrestrial Planets in Extrasolar Planetary Systems

A whole part of this book is devoted to the description of extrasolar planetary systems (=ESP), their detection and the complicated dynamics involved when a system with two gas giants (=GG) in mean motion resonance is on high eccentric stable orbits (see the chapter in this book by Ferraz-Mello et al). We have no evidence up to now that terrestrial planets may exist in such systems but for our human spirit it seems quite normal that Earth-like planets may also exist in ESPs. The reason is simple: in our Solar System inside the orbits of Jupiter, Saturn, Uranus and Neptune four terrestrial planets are orbiting the Sun and many other terrestrial like planets are satellites of Jupiter and Saturn or are on orbits outside Pluto in the Kuiper-belt. The central question is whether such terrestrial planet in ESPs may move in the

[61] A critical remark should be placed here concerning the results presented in this paragraph: when, as it was done by Laskar, we truncate the order of the perturbations in the small parameters from 6 on, for eccentricities $e > 0.4$ the significant digits are limited to two. Thus the system cannot be any more an adequate description of the planetary system. The important point is that the very slow diffusion may lead to such large eccentricities as $e = 0.4$ of Mercury.

[62] The small masses of the asteroids, the shape of the planets, relativity and also galactic tides are too small to change this picture.

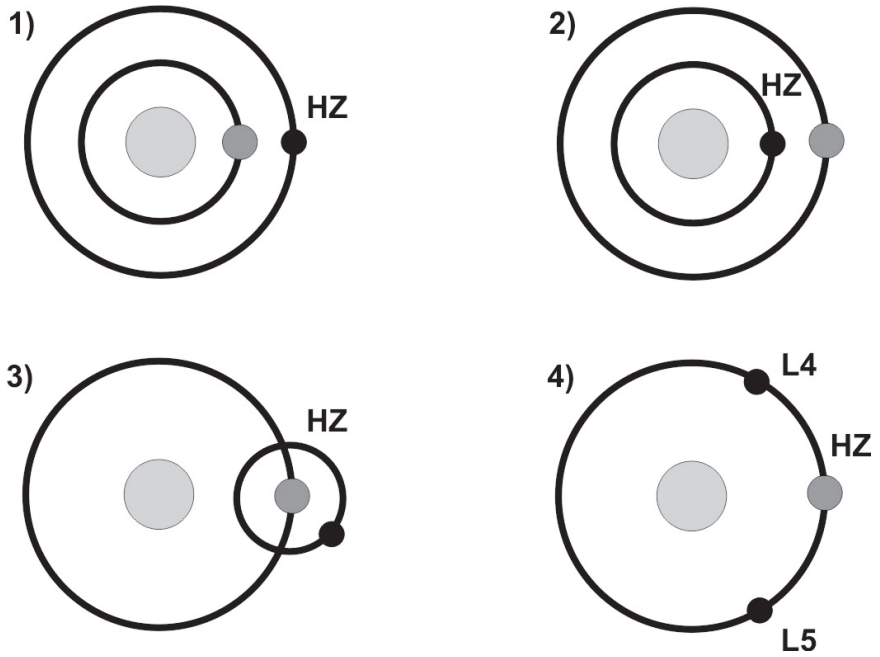

Fig. 63. Possible habitable regions for terrestrial planets in ESPs (C1 to C4): the full black circles show the terrestrial planet, the light grey full circles show the gas giant. *(for details of the four classes see in the text)*

so-called habitable zones[63]. From the orbital characteristics and the criterion to be in a habitable region one can distinguish 4 different classes (see Fig. 63):

1. C1: when the GG is very close to the star there could exist such stable orbits for time scales long enough to develop a biosphere.
2. C2: when this GG moves far away from the central star (like Jupiter) then stable low eccentric orbits for additional planets can exist.
3. C3: when the GG itself moves in the habitable region, a terrestrial like satellite (like e.g. Titan in the system of Saturn) could have a stable orbit.
4. C4: when the GG itself moves in the habitable region a Trojan like terrestrial planet may move on a stable orbit around the Lagrangian equilibrium points L_4 or L_5.

[63] We will not discuss the complex problem of habitable regions around a host star in detail, because it is still somewhat in contradiction; it depends not only on the dynamical parameters of the orbit of a planet, but also on the astrophysical parameters of the star like the spectral type and the age (e.g. [34]). A somewhat rough estimate for it is *where water could exist in liquid form on the surface of a planet*; for a more precise definition see [33].

Besides the extensive study of [47] concerning the stability of orbits of terrestrial planets in extrasolar systems there exist investigations for specific systems for the classes C1 and C2: e.g. [19], [20], [10], [11], [12],[2]. In recent investigations also the dynamical stability of possible terrestrial planets in the 1:1 resonance with the gas giant (class C4) was studied (e.g. [20]).

About the formation mechanism of planets in general from the protoplanetary disk we refer to the chapter by Tom Quinn (in this book). It seems that we have reasonable theories for the formation of terrestrial planets of C1, C2 and C3 types. From the cosmogonical point of view one can even imagine a possible formation of two planets in a 1:1 mean motion resonance [39] as result of an interaction with the protoplanetary disc.

10.1 Theoretical Considerations

For the C1, C2 and C3 types there exist many numerical studies concerning the stability of an additional massless regarded body in the three dimensional elliptic restricted problem via extensive numerical integrations (e.g. [58], [31], [55]). Quite often for this kind of stability studies the use of supercomputers is appropriate, because of the huge amount of CPU necessary to get good statistical results. Many thousands of orbits are computed simultaneously for a fine grid of initial conditions with such computers.

For the C4 group there exist a lot of analytical work concerning the stability of the Lagrangian points in the model of the elliptic restricted problem depending on the mass ratio of the primaries and the eccentricity of their orbits (e.g. [57], [4]). Additional work has been done even for cases when the third mass is not regarded as massless [46]. The results of a first order stability analysis in the framework of the general three-body-problem (loc.cit. p. 46ff) are presented there. With M the total mass and $m_1 \leq m_2 \leq m_3$ a mass parameter R was defined as $R = (m_2+m_3)/M+m_2.m_3/m_1^2+O(m_2^3.m_3/m_1^4)$. Using these results one can see[64] that in the case of a terrestrial like planet with a relatively small mass compared to the two primary bodies there is practically no difference in the stability of the equilibrium points When one takes into account the observed eccentricities of the orbits of the GG in the ESPs, furthermore the estimated (minimum) mass of the giant planet and a terrestrial planet comparable to the mass of our Earth it turned out that all planetary systems of the list given by Jean Schneider[65] have stable equilateral Lagrangian points. But the extension of the stable region around this equilibrium points cannot be determined with such an analysis.

The methods of establishing such zones for all 4 groups C1 to C4 are numerical ones; suitable methods are e.g. the Lie integration method (see appendix A) or classical Runge-Kutta and Bulirsch-Stoer methods [56] to determine the orbits via solving the equations of motion for an appropriate

[64] Fig. 13, p.49 in Marchals's book

[65] `http://www.obspm.fr/encycl/catalog.html`

grid of initial conditions. One uses different tests for the stability of an orbit like the Lyapunov exponents (see the Chapter of Harry Varvoglis – in this book) or similar methods ([23], [61]). In the following we will show – as examples – the results of such investigations for terrestrial planets (C2 and C4) in the habitable zones of Gl 777A and HD 23079.

10.2 Gl 777 A

The first discovery of a planet in Gl 777 A (=HD 190360) was reported by Naef et al. [52] from the Geneva group of observers. This extrasolar planetary system is a wide binary with a very large separation (3000 AU); for the dynamical investigations of motions close to one star there was no need to take into account the perturbations of the very far companion. The central star is of spectral type G6 IV with $0.9 M_{\odot}$ and has a planet of minimum mass 1.33 $M_{\rm Jup}$ with a semimajor-axis of 4.8 AU. Because of the large eccentricity (e=0.48) the possible region of motion for additional planets is confined to $a < 2.4$ AU (= periastron). One can see in Fig. 64 that for this system the mean motion resonances are not important for motions in the habitable zone, because they are all well outside. This is sometimes quite different, as one may see in [2].

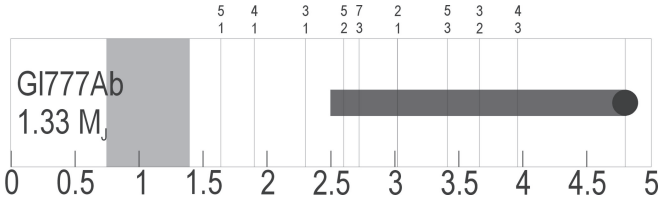

Fig. 64. Main characteristics of the extrasolar system Gl 777 A. The light grey region shows approximately the position of the HZ; the dark grey bar indicates how closely the planet approaches the central star in its orbit. The different numbers characterize the mean motion resonances

The interesting region of habitability (see Fig. 64), where planets could have temperature conditions to allow liquid water on the surface, corresponds roughly to $0.7 < a < 1.3$ AU, when we ignore the eccentricity of the terrestrial planet. The computations were started in a larger region ($0.5 < a < 1.3$) with a grid spacing of $\Delta a = 0.01$ AU and the eccentricity of the known planet was changed between $0.4 < e < 0.5$ with a gridsize of $\Delta e = 0.01$. The results (Fig. 65) of the direct analysis of the largest eccentricity (MEM- Maximum Eccentricity Method) achieved during the orbital evolution (1 Million years) show two main features: 1) strong vertical lines due to high order resonances and 2) more and more unstable orbits with larger semimajor axes of the terrestrial planet (red or yellow colors). We observe the same structure in the

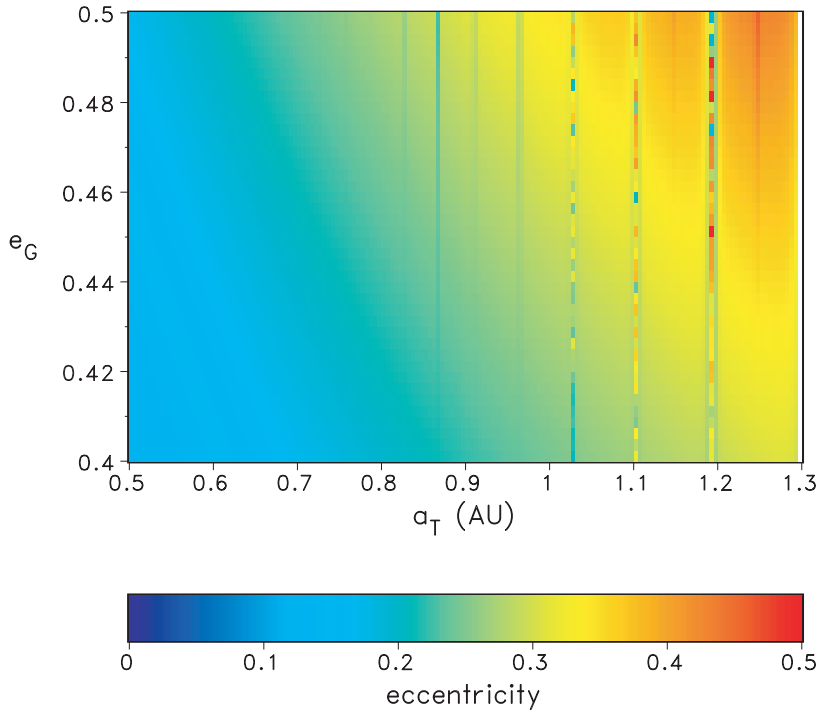

Fig. 65. Initial condition diagram of semi-major axis of the fictitious terrestrial planet versus the eccentricity of the GG of the extrasolar system Gl 777 A. The colors indicate the maximum eccentricity achieved during the integration of 1 million years

main belt of asteroids due to the perturbations of Jupiter. The eccentricity tells us directly the variable distance to the central star and consequently it is a direct measure of the differential energy flux (insolation) on the planet. We can therefore determine where the variation of this distance does not exceed 50 percent which corresponds roughly to an eccentricity of $e = 0.2$.

We also used a second method, namely the Rényi entropy[66] to distinguish between regular stable orbits and chaotic orbits (Fig. 66). One can see that this is a very sensitive instrument giving us a measure of the degree of chaos. High order resonances are also visible and one recognizes them also in regions where the MEM method does not show any features.

As a result for habitability of a terrestrial planet inside the orbit of the Jovian planet, we find that for the system Gl 777 A there is quite a good chance that planets will last long enough in the habitable zone to acquire the necessary conditions for life in the region with $a < 1$ AU.

[66] A detailed decription of the Rényi entropie which is the estimate of one of the dynamical invariants of Reccurence Plots can be found in [2], [68] and [69].

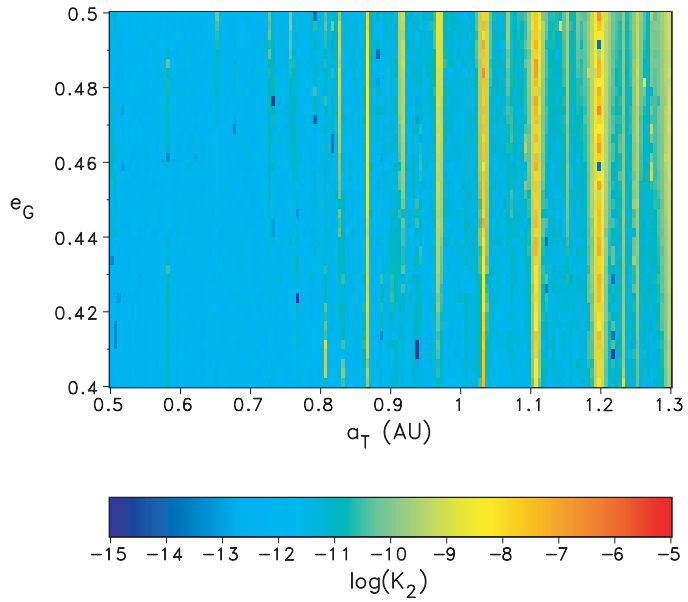

Fig. 66. Initial condition diagram of semi-major axis of the fictitious terrestrial planet versus the eccentricity of the GG of the extrasolar system Gl 777 A: The value of the entropy is marked by different colors; small values of the entropy stand for regular orbits

10.3 HD23079

As another example of possible stable orbits of terrestrial planets we discuss a Trojan planet (C4) in a specific ESP. HD 23079 has a central star of 1.1 $M_{\odot}$ and a GG with 2.61 M_{Jup} that moves on an orbit with a semimajor axis $a = 1.65$ AU and an eccentricity of $e = 0.1$ close to the habitable region. The initial conditions around the stable Lagrangian point L_4 were taken for a fine grid in the semimajor axis 1.56AU $< a <$ 1.76 AU with $\delta a = 0.001$ AU and for the synodic longitude $30^o < \lambda < 120^o$ with $\delta\lambda = 1^o$. For the eccentricity of the GG three different values were taken as initial conditions in order to see how the structure of the stable region around the equilibrium points diminishes (see Fig. 67). For the lowest value of $e_{\rm GG} = 0.05$ the eccentricity of the fictitious Trojan planet stays always smaller than $e < 0.05$; for the actual measured value of $e_{\rm GG} = 0.1$ the eccentricity of the fictitious Trojan planet still fulfills the requirement mentioned above of always being below $e < 0.2$. For a larger value $e_{\rm GG} = 0.15$ the respective large eccentricities of $e_{\rm max} > 0.2$ would – according to our hypothesis for the maximum variation of the distance to the central star – not allow conditions for habitability. It also confirms the ring of less stable motion around the Lagrangian point, which is

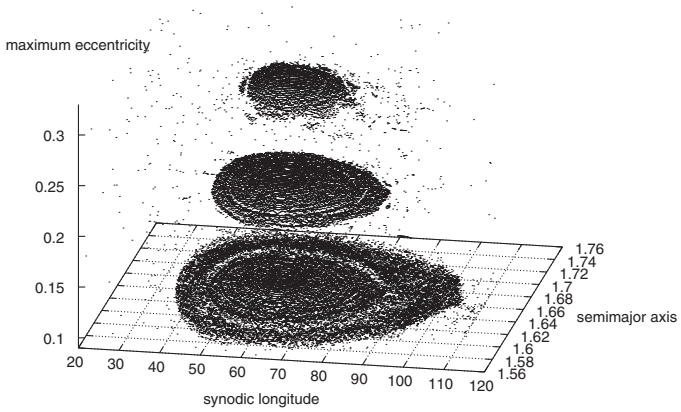

Fig. 67. Stability region around the Lagrangian point L_4 of the system HD 23079 for three different values of the eccentricity of the observed gas giant *(for details see in the text)*

well visible for $e = 0.05$. For the actual measured $e = 0.1$ one can recognize a bar like structure which is not visible with the MEM method.

This system was also investigated by another method, namely the RLIs (Relative Lyapunov Indicators, compare [20]) for the actually determined eccentricity of the primary's orbit $e_{\mathrm{GG}} = 0.1$. A comparison of the respective results show the good agreement of both results: it confirms the size of the stable region and also ring of less stable orbits around the Lagrangian point.

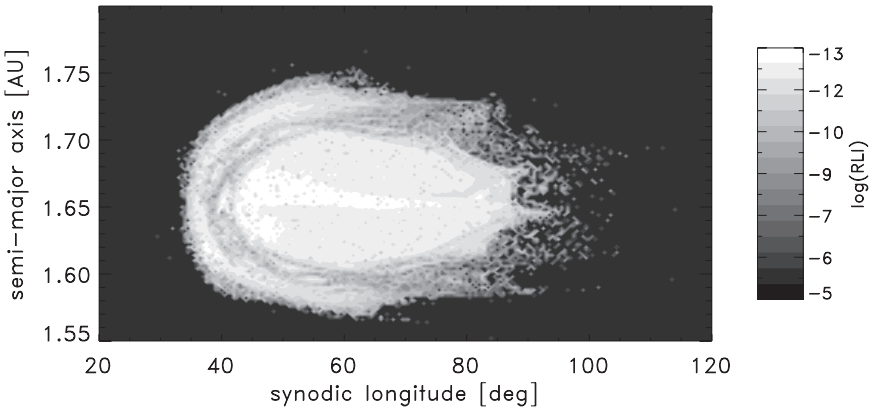

Fig. 68. RLIs around the Lagrangian point L_4 of HD23079: the degree of chaoticity is marked by the different grey-scales: white is regular and black is chaotic

11 Epilog: Back to the Golden Mean

11.1 Continuous Fractions and Golden Numbers

We started this course with the introduction of a very special number, the *golden mean*

$$\frac{1+\sqrt{5}}{2} = 1.618034\ldots \tag{431}$$

which plays an important role in number theory, because it is *the most irrational number*. This can be easily seen by representing the number of the golden mean as a *continued fraction*:

Definition 7. *An irrational number a can be written as a continued fraction of the form*

$$a = [a_1, a_2, a_3, ...] = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \ldots}}}. \tag{432}$$

where the a_is are positive integers.

Remark 3. The better a number can be approximated by a continued fraction, the larger the a_i will be (because if the approximation is good, the remaining part is small and thus the denominators have to be large).

Definition 8. *When $a_i = 1$ for all i above a certain value N, the number a is called a noble numbers of order N.*

For example [2,1,1,...] is of order one, [2,3,2,1,...] is of order three. In this sense the golden mean [1,1,1,1,1...] is of order zero and thus, keeping in mind remark 3, it is the most irrational of all numbers. There are some properties of the noble numbers which we recall briefly:

- For noble numbers of order 1 the following relations hold

$$[1, 1, ...] > [2, 1, ...] > [3, 1, 1, ...] > ... > [\infty, 1, ...] = 0 \tag{433}$$

- Noble numbers of order 2 are in between the first order noble numbers like

$$[1, 1, ...] > \frac{1}{2} = [2, \infty, 1, ...] > ... > [2, 3, 1, ...] > [2, 2, 1, ...] > [2, 1, 1, ...] \tag{434}$$

  which means that the noble numbers of order 2 with a number 2 in the first digit are larger than the first order noble number [2,1,...].
- The noble numbers of order 2 with the number 1 as first digit are larger than the golden number:

$$1 = [1, \infty, 1, ...] > ... > [1, 3, 1, ...] > [1, 2, 1, ...] > [1, 1, 1, ...] \tag{435}$$

- The noble numbers of order 3 are between the noble number of order 2:

$$[1,1,1,...] > [1,1,2,1,...] > [1,1,3,1,...] > ... > [1,1,\infty,1,...] = \frac{1}{2} \quad (436)$$

- An increase of a_2 of a noble number produces a larger number; on the contrary the increase of a_3 produces a smaller number.

This construction via continuous fraction is also valid for the computation of any rational number when we limit the number N of coefficients $a_i, i = 1, ..., N$ to a certain integer e.g.

$$\begin{aligned} \frac{1}{2} &= [1,1,1] \\ \frac{4}{9} &= [2,4,1] \\ \frac{5}{11} &= [2,4,1,1] \end{aligned} \quad (437)$$

There exists an interesting method of construction of rational numbers and the continued approximation of irrationals via the *Farey tree*, where we can arrange the different "generations" in the following way:

$$G_1 = \frac{0}{1}, \frac{1}{1} \quad (438)$$

$$G_2 = \frac{0}{1}, \frac{1}{2}, \frac{1}{1} \quad (439)$$

$$G_3 = \frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1} \quad (440)$$

$$G_4 = \frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1} \quad (441)$$

$$G_5 = \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{3}{8}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{4}{7}, \frac{3}{5}, \frac{5}{8}, \frac{2}{3}, \frac{5}{7}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1} \quad (442)$$

Two properties are visible, from which one easily can derive the whole tree like structure from generation to generation:

1. Each rational of a generation has as numerator and denominator the sum of the numerator and denominator of the two rationals of the previous generation (e.g. for G_5 the new rational in between $\frac{1}{4}$ and $\frac{1}{3}$ of the G_4 generation is $\frac{2}{7}$).
2. The numerator of a rational multiplied by the denominator of the following rational in a sequence minus the denominator of the first rational multiplied with the nominator of the following rational is always ± 1 (e.g. in G_5 we take the two neighboring rationals $\frac{2}{5}$ and $\frac{3}{7}$.

11.2 The KAM-Tori and Noble Numbers

In Sect. 2.4 where we investigated the Standard Mapping we have seen that an orbit is either *periodic* (same points on the plane), *quasiperiodic* (where the sequence of points fills densely the so called invariant curve or – in this case a one dimensional KAM Torus S^1) or chaotic (where no regularity can be seen in the sequence of points). On this KAM-Torus two vectors from the center (a periodic orbit) to two consecutive points have a certain rotational angle which is characteristic for every KAM-Torus (in fact the mean value of a great number of rotational angles). These points fill densely the curve and as a consequence this angle cannot be a rational number, because then the orbit would be a periodic one. Because it is an irrational number it can be approximated by a continuous fraction. In Fig. 69 we see a large part of the phase space where the initial conditions lead to invariant curves and will never leave the area, surrounded by a cloud of chaotic orbits. But on the edge of the "island of regular motion" we can see a quite dark region, where orbits are trapped for a long time, and then escape into the chaotic see. These "sticky" orbits are confined by Cantori, which are KAM-Tori which are destroyed with increasing strength of the non-linearity parameter (K in the case of the Standard Mapping) and have holes, where an orbit may escape through.

The connection to the noble numbers is the following: The KAM-Tori with the "most irrational" rotational numbers are destroyed and – as we have seen before – they are the ones having noble numbers as rotational numbers.

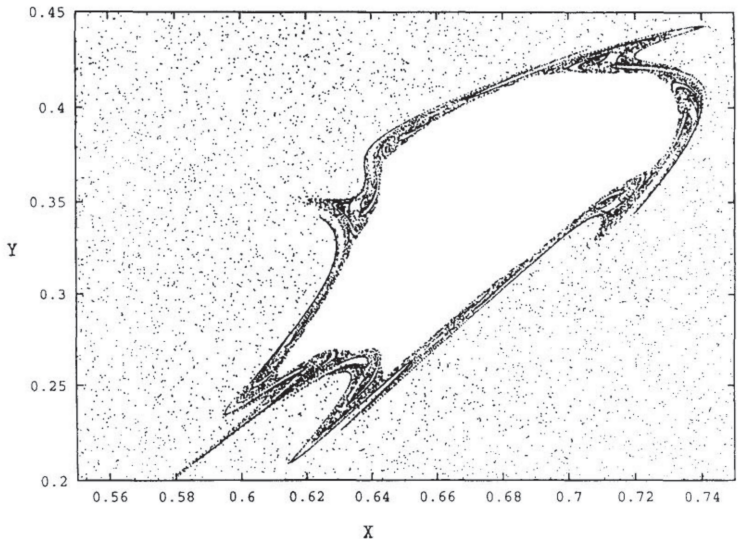

Fig. 69. Sticky region in the Standard Mapping

Because any noble number can be approximated by a continuous fraction with higher and higher orders, every KAM-torus with a special noble number as rotational number is surrounded from the outside and from the inside by periodic orbits characterized by rational numbers. This approximation is better and better with increasing generation G_n (compare [15]).

11.3 KAM-Theory and the Stability of the Solar System

What we have discussed before is not only true for a two dimensional mapping but also for n-dimensional dynamical systems. Also for our planetary system, when the motions would be quasiperiodic, the orbits that are confined to a high dimensional torus and thus stable for all times. Evidently the motion is not quasiperiodic as we have seen from the plots of the elements and from the computations by Laskar – as a consequence our planetary system is chaotic. But in the sense of the sticky orbits in the Standard Mapping we may conclude that the orbits of the planets are also sticky and that they are also confined by high dimensional Cantori, which hinder our planetary system to become unstable within the next Gigayears.

A Lie Integration Method

A.1 General Properties of the Lie-series

Gröbner [25] defined the Lie-operator D as follows:

$$D = \theta_1(z)\frac{\partial}{\partial z_1} + \theta_2(z)\frac{\partial}{\partial z_2} + \ldots + \theta_n(z)\frac{\partial}{\partial z_n} \tag{443}$$

D is a linear differential operator; the point $z = (z_1, z_2, \ldots, z_n)$ lies in the n-dimensional z-space; the functions $\theta_i(z)$ are holomorphic within a certain domain G, e.g. they can be expanded in converging power series. Let the function $f(z)$ be holomorphic in the same region as $\theta_i(z)$. Then D can be applied to $f(z)$:

$$Df = \theta_1(z)\frac{\partial f}{\partial z_1} + \theta_2(z)\frac{\partial f}{\partial z_2} + \ldots + \theta_n(z)\frac{\partial f}{\partial z_n} \tag{444}$$

If we proceed applying D to f we get

$$\begin{aligned} D^2 f &= D(Df) \\ &\vdots \\ D^n f &= D(D^{n-1} f) \end{aligned}$$

The **Lie-series** will be defined in the following way;

$$L(z,t) = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} D^\nu f(z) = f(z) + tDf(z) + \frac{t^2}{2!} D^2 f(z) + \ldots$$

Because we can write the Taylor-expansion of the exponential function

$$e^{tD} f = (1 + tD^1 + \frac{t^2}{2!} D^2 + \frac{t^3}{3!} D^3 + \ldots) f \tag{445}$$

$L(z,t)$ can be written in the symbolic form

$$L(z,t) = e^{tD} f(z) \tag{446}$$

The convergence proof of $L(z,t)$ is given in detail in Gröbner [25]

One of the most useful properties of Lie-series is the Vertauschungssatz:

Theorem 4. *Let $F(z)$ be a holomorphic function in the neighborhood of $(z_1, z_2, \ldots, z_n)$ where the corresponding power series expansion converges at the point $(Z_1, Z_2, \ldots, Z_n)$; then we have:*

$$F(Z) = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} D^\nu F(Z) \tag{447}$$

or

$$F(e^{tD}) z = e^{tD} F(z) \tag{448}$$

Making use of it we can demonstrate how Lie-series solve differential equations. Let us give the system of differential equations:

$$\frac{dz_i}{dt} = \theta_i(z) \tag{449}$$

with $(z_1, z_2, \ldots, z_n)$. The solution of (449) can be written as

$$z_i = e^{tD} \xi_i \tag{450}$$

where ξ_i are the initial conditions $z_i(t = 0)$ and D is the Lie-operator as defined in (443). In order to prove (450) we differentiate it with respect to time t:

$$\frac{dz_i}{dt} = D e^{tD} \xi_i = e^{tD} D \xi_i \tag{451}$$

Because of

$$D\xi_i = \theta_i(\xi_i) \tag{452}$$

we obtain the following result which turns out to be the original differential equation (449):

$$\frac{dz_i}{dt} = e^{tD} \theta_i(\xi_i) = \theta_i(e^{tD} \xi_i) = \theta_i(z_i) \tag{453}$$

A.2 A Simple Example

To demonstrate the principle of the Lie-integration, we will show, how one has to proceed in the simple case of the harmonic oscillator, a 2^{nd} order differential equation:

$$\frac{d^2x}{dt^2} + \alpha^2 x = 0 \tag{454}$$

The first step consists in separating (454) in two 1^{st} order differential equations such that

$$\frac{dx}{dt} = y = \theta_1(x, y)$$

$$\frac{dy}{dt} = -\alpha^2 x = \theta_2(x, y)$$

with the initial conditions $x(t = 0) = \xi$ and $y(t = 0) = \eta$. With this notation we find the Lie-operator of the form

$$D = \theta_1 \frac{\partial}{\partial \xi} + \theta_2 \frac{\partial}{\partial \eta} = \eta \frac{\partial}{\partial \xi} - \alpha^2 \xi \frac{\partial}{\partial \eta} \tag{455}$$

The solution can now be written as Lie-series

$$x = e^{\tau D}\xi \qquad \text{and} \qquad y = e^{\tau D}\eta \tag{456}$$

where $t - t_0 = \tau$ and the initial conditions can be obtained for $\tau = 0$. Being aware of the symbolic development of $e^{\tau D}$ we can compute the first terms:

$$D^1\xi = \eta = \theta_1$$

$$D^2\xi = D\eta = -\alpha^2\xi = \theta_2$$

$$D^3\xi = -\alpha^2 D\xi = -\alpha^2\eta$$

$$D^4\xi = -\alpha^2 D\eta = \alpha^4\xi$$

$$D^5\xi = \alpha^4 D\xi = \alpha^4\eta$$

$$D^6\xi = \alpha^4 D\eta = -\alpha^6\xi$$

$$\vdots$$

For the Lie-terms even (respectively odd) in the order consequently one can find

$$D^{2n}\xi = (-1)^n \alpha^{2n}\xi$$

$$D^{2n+1}\xi = (-1)^n \alpha^{2n}\eta$$

which leads to the solution for z:

$$z = \xi + \tau\eta - \frac{\tau^2}{2!}\alpha^2\xi - \frac{\tau^3}{3!}\alpha^2\eta + \frac{\tau^4}{4!}\alpha^4\xi \ldots$$

Finally we get after the factorization of ξ and of η

$$z = \xi \left(1 - \frac{\tau^2}{2!}\alpha^2 + \frac{\tau^4}{4!}\alpha^4 - \frac{\tau^6}{6!}\alpha^6 + \dots\right) + \frac{\eta}{\alpha}\left(\tau\alpha - \frac{\tau^3}{3!}\alpha^3 + \frac{\tau^5}{5!}\alpha^5 - \frac{\tau^7}{7!}\alpha^7 + \dots\right)$$

which is exactly the known solution of the harmonic oscillator:

$$z(t) = \xi \cos\alpha\tau + \frac{\eta}{\alpha}\sin\alpha\tau$$

A.3 Hénon Heiles System

The equations of motion for the coordinates and momenta of the Hénon-Heiles system are given by:

$$\begin{aligned} \dot{x} &= p_x \\ \dot{y} &= p_y \\ \dot{p_x} &= -x - 2xy \\ \dot{p_y} &= -y - x^2 + y^2 \end{aligned}$$

The initial conditions for a time τ are denoted as follows:

$$x(\tau) = \xi \qquad p_x(\tau) = p_\xi \tag{457}$$

$$y(\tau) = \eta \qquad p_y(\tau) = p_\eta \tag{458}$$

The system of equations needed to derive the Lie operator is now:

$$\dot{\xi} = p_\xi = \theta_1 \tag{459}$$

$$\dot{\eta} = p_\eta = \theta_2 \tag{460}$$

$$\dot{p_\xi} = -\xi - 2\xi\eta = \theta_3 \tag{461}$$

$$\dot{p_\eta} = -\eta - \xi^2 + \eta^2 = \theta_4 \tag{462}$$

and the Lie operator is given by:

$$\mathsf{D} = \theta_1\frac{\partial}{\partial\xi} + \theta_2\frac{\partial}{\partial\eta} + \theta_3\frac{\partial}{\partial p_\xi} + \theta_4\frac{\partial}{\partial p_\eta} \tag{463}$$

or

$$\mathsf{D} = p_\xi\frac{\partial}{\partial\xi} + p_\eta\frac{\partial}{\partial\eta} + (-\xi - 2\xi\eta)\frac{\partial}{\partial p_\xi} + \left(-\eta - \xi^2 + \eta^2\right)\frac{\partial}{\partial p_\eta} \tag{464}$$

Now the solution for a time $(\tau + \Delta\tau)$ can be derived by

$$x(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}\xi \tag{465}$$

$$y(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}\eta \tag{466}$$

$$p_x(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}p_\xi \tag{467}$$

$$p_y(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}p_\eta \tag{468}$$

Symbolically, these equations can be written in the form

$$x(\tau+\Delta\tau) = e^{\Delta\tau \mathsf{D}}\xi = \xi+\Delta\tau \mathsf{D}^1(\xi)+\frac{\Delta\tau^2}{2!}\mathsf{D}^2(\xi)+\frac{\Delta\tau^3}{3!}\mathsf{D}^3(\xi)+\frac{\Delta\tau^4}{4!}\mathsf{D}^4(\xi)+\mathsf{O}(5) \tag{469}$$

$$y(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}\eta = \eta + \Delta\tau \mathsf{D}^1(\eta) + \frac{\Delta\tau^2}{2!}\mathsf{D}^2(\eta) + \frac{\Delta\tau^3}{3!}\mathsf{D}^3(\eta) + + \frac{\Delta\tau^4}{4!}\mathsf{D}^4(\eta) + \mathsf{O}(5) \tag{470}$$

$$p_x(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}p_\xi = p_\xi + \Delta\tau \mathsf{D}^1(p_\xi) + \frac{\Delta\tau^2}{2!}\mathsf{D}^2(p_\xi) + \frac{\Delta\tau^3}{3!}\mathsf{D}^3(p_\xi) + \frac{\Delta\tau^4}{4!}\mathsf{D}^4(p_\xi) + \mathsf{O}(5) \tag{471}$$

$$p_x(\tau + \Delta\tau) = e^{\Delta\tau \mathsf{D}}p_\eta = p_\eta + \Delta\tau \mathsf{D}^1(p_\eta) + \frac{\Delta\tau^2}{2!}\mathsf{D}^2(p_\eta) + \frac{\Delta\tau^3}{3!}\mathsf{D}^3(p_\eta) + + \frac{\Delta\tau^4}{4!}\mathsf{D}^4(p_\eta) + \mathsf{O}(5) \tag{472}$$

Now the expressions for D^i have to be derived:

$$\mathsf{D}^1(\xi) = p_\xi \tag{473}$$

$$\mathsf{D}^2(\xi) = \mathsf{D}^1(p_\xi) = -\xi - 2\xi\eta \tag{474}$$

$$\mathsf{D}^3(\xi) = -\mathsf{D}^1(\xi) - \mathsf{D}^1(2\xi\eta) = -\mathsf{D}^1(\xi) - 2\left(\left(\mathsf{D}^1(\xi)\right)\eta + \left(\mathsf{D}^1(\eta)\right)\xi\right) \tag{475}$$

$$\mathsf{D}^4(\xi) = -\mathsf{D}^2(\xi) - 2\left(\left(\mathsf{D}^2(\xi)\right)\eta + 2\left(\mathsf{D}^1(\xi)\right)\left(\mathsf{D}^1(\eta)\right) + \left(\mathsf{D}^2(\eta)\right)\xi\right) \tag{476}$$

$$\mathsf{D}^1(\eta) = p_\eta \tag{477}$$

$$\mathsf{D}^2(\eta) = \mathsf{D}^1(p_\eta) = -\eta - \xi^2 + \eta^2 \tag{478}$$

$$\mathsf{D}^3(\eta) = -\mathsf{D}^1(\eta) - 2\left((\mathsf{D}^1(\xi))\xi - (\mathsf{D}^1(\eta))\eta\right) \tag{479}$$

$$\mathsf{D}^4(\eta) = -\mathsf{D}^2(\eta) - 2\left(\left(\mathsf{D}^2(\xi)\right)\xi + \left(\mathsf{D}^1(\xi)\right)^2 - \left(\mathsf{D}^2(\eta)\right)\eta - \left(\mathsf{D}^1(\eta)\right)^2\right) \tag{480}$$

Because of $\mathsf{D}^2(\eta) = \mathsf{D}^1(p_\eta)$ and $\mathsf{D}^2(\xi) = \mathsf{D}^1(p_\xi)$ the expressions for $\mathsf{D}^i(p_\eta)$ and $\mathsf{D}^i(p_x)$ can be obtained easily:

$$\mathsf{D}^1(p_\xi) = \mathsf{D}^2(\xi) \qquad \mathsf{D}^1(p_\eta) = \mathsf{D}^2(\eta) \tag{481}$$

$$\mathsf{D}^2(p_\xi) = \mathsf{D}^3(\xi) \qquad \mathsf{D}^2(p_\eta) = \mathsf{D}^3(\eta) \tag{482}$$

$$\mathsf{D}^3(p_\xi) = \mathsf{D}^4(\xi) \qquad \mathsf{D}^3(p_\eta) = \mathsf{D}^4(\eta) \tag{483}$$

$$\mathsf{D}^4(p_\xi) = \mathsf{D}^5(\xi) = \tag{484}$$

$$= -\mathsf{D}^3(\xi) - 2\left(\left(\mathsf{D}^3(\xi)\right)\eta + 3\left(\mathsf{D}^2(\xi)\right)\left(\mathsf{D}^1(\eta)\right) + 3\left(\mathsf{D}^2(\eta)\right)\left(\mathsf{D}^1(\xi)\right) + \left(\mathsf{D}^3(\eta)\right)\xi\right) \tag{485}$$

$$\mathsf{D}^4(p_\eta) = \mathsf{D}^5(\eta) = \tag{486}$$

$$= -\mathsf{D}^3(\eta) - 2\left(\left(\mathsf{D}^3(\xi)\right)\xi + 3\left(\mathsf{D}^2(\xi)\right)\left(\mathsf{D}^1(\xi)\right) - 3\left(\mathsf{D}^2(\eta)\right)\left(\mathsf{D}^1(\eta)\right) - \left(\mathsf{D}^3(\eta)\right)\eta\right) \tag{487}$$

Now, the equations of motion of the Hénon-Heiles system can be integrated numerically with the Lie integration method (up to order 4).

It appears the question why should we use another numerical integrator of differential equations when many of well tested methods are free on the "market"? Let us state five major points why we propose to make use of this method:

1. Because of the structure of the solution as power series expansions in the independent variable (we used t or τ in the examples), which is also the time step for the integration, we can choose it differently for every step.[67]
2. The desired precision can be fixed and controlled by two different parameters, namely the largeness of the "time step" and the number of Lie-terms we use for the Lie-series. A small step size needs less Lie-terms and vice-versa when we choose a large step-size (attention has to be drawn to the convergence radius of the series!) we need more Lie-terms to achieve the same precision.
3. The goal is to find out a recurrence form for the Lie-terms making use of the properties of the Lie-operator to be linear. If one can achieve it the integrator is very fast! Quite successful was the construction of an integrator of the elliptic restricted three body problem and also of the n-body problem (e.g. [28],[43], [7]).
4. Because of the convergence of the powerseries the Lie-integrator can be regarded as a kind of a symplectic integrator which conserves quite well the integrals of motion.
5. The construction of a proper tool to integrate a DE is very educative for students and avoids to take a black box as solver of the problem in question (it does not mean that the own Lie-integrator is always faster).

[67] It needs not to be the time step but in dynamical systems we are quite used to it. Precisely it is the step forward in the independent variable for which we want to get the solution of the DE in question.

The only disadvantage is that the derivation of reccurence in the Lie-terms may nevertheless lead to very lengthy expressions when the right hand sides are complicated and lengthy themselves. This is the only task where we propose to use a standard method. Most of the problems have simpler right hand sides of the equations; therefore – please try it out!

B Lie Perturbation Method

B.1 Lie Transformation

Lie Transformations are a very powerful technique in perturbation theory.[68] In general, a Lie Transformation simplifies a given Hamiltonian via a canonical transformation. Let us consider a Hamiltonian of the following type:

$$H = H\left(q_i, p_i, \epsilon\right) = H_0\left(q_i, p_i\right) + \epsilon H_1\left(q_i, p_i\right) + \epsilon^2 H_2\left(q_i, p_i\right) + O\left(\epsilon^3\right) \quad (488)$$

The equations of motion are then given in canonical form:

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \quad (489)$$

The coordinates p_i and q_i are now transformed to new coordinates P_i and Q_i via the functions F_{ji} and G_{ji}, which will be determined so that the transformation is a canonical one:

$$\begin{aligned} q_i &= Q_i + \epsilon F_{1i}\left(Q_i, P_i\right) + \epsilon^2 F_{2i}\left(Q_i, P_i\right) + \ldots \\ p_i &= P_i + \epsilon G_{1i}\left(Q_i, P_i\right) + \epsilon^2 G_{2i}\left(Q_i, P_i\right) + \ldots \end{aligned} \quad (490)$$

This gives a new Hamiltonian K (a *Kamiltonian*)[69] when we make use of (490):

$$\begin{aligned} K &= K\left(Q_i, P_i, \epsilon\right) \\ &= H\left(q_i, p_i, \epsilon\right) \\ &= H\left(Q_i + \epsilon F_{1i} + \epsilon^2 F_{2i} + \ldots, P_i + \epsilon G_{1i} + \epsilon^2 G_{2i} + \ldots\right) \end{aligned} \quad (491)$$

Also the Kamiltonian can be developed into a power-series with respect to the perturbing parameter ϵ:

$$K = K\left(Q_i, P_i, \epsilon\right) = K_0\left(Q_i, P_i\right) + \epsilon K_1\left(Q_i, P_i\right) + \epsilon^2 K_2\left(Q_i, P_i\right) + O\left(\epsilon^3\right) \quad (492)$$

[68] We follow the treatise on this problem by R. Rand [59].

[69] The expression Kamiltonian K is used by Goldstein in his textbook *Classical Mechanics* [24]. The argument for using another name and another notation was that it is shorter to use than transformed Hamiltonian $\tilde{H}$.

We emphasize that for $\epsilon = 0$ the Hamiltonian $H_0\,(q_i, p_i)$ and the Kamiltonian $K_0\,(Q_i, P_i)$ are the same and therefore $Q_i = q_i$ and $P_i = p_i$. The functions of the coordinate transformation have to be chosen in such a way, that the new Kamiltonian has the simplest possible form (i.e. by eliminating Q_i which is then a so-called ignorable coordinate), the resulting equations of motion can be solved more easily than in the original case.

First Order

The canonical transformation (490) yields equations of the Hamiltonian form for q_i and p_i depending on ϵ:

$$\begin{aligned} \frac{dq_i}{d\epsilon} &= \frac{\partial W}{\partial p_i} \\ \frac{dp_i}{d\epsilon} &= -\frac{\partial W}{\partial q_i} \end{aligned} \tag{493}$$

where $W = W\,(q_i, p_i, \epsilon)$ is the so-called *Lie-generating function*. Making use of the coordinate transformation (490) the function W can also be expanded in a power series in ϵ:

$$\begin{aligned} W &= W\left(Q_i + \epsilon F_{1i} + \epsilon^2 F_{2i} + \ldots, P_i + \epsilon G_{1i} + \epsilon^2 G_{2i} + \ldots\right) \\ &= W_1\,(Q_i, P_i) + \epsilon W_2\,(Q_i, P_i) + \epsilon^2 W_3\,(Q_i, P_i) + O(\epsilon^3) \end{aligned} \tag{494}$$

Additionally to the original equations of the time evolution of the coordinates for fixed ϵ – equation (488) – the evolution of the canonical transformation itself can be regarded as a Hamiltonian process as ϵ varies. Therefore we can make a Taylor expansion for the "old" coordinates q_i and also for p_i:

$$q_i(\epsilon) = q_i(0) + \epsilon \left.\frac{dq_i}{d\epsilon}\right|_{\epsilon=0} + \frac{\epsilon^2}{2} \left.\frac{d^2 q_i}{d\epsilon^2}\right|_{\epsilon=0} + \ldots \tag{495}$$

and find the following expression for the 1^{st} term using (493) and (495)

$$\left.\frac{dq_i}{d\epsilon}\right|_{\epsilon=0} = \left.\frac{\partial W}{\partial p_i}\right|_{\epsilon=0} = \left.\frac{\partial (W_1 + \epsilon W_2 + ...)}{\partial p_i}\right|_{\epsilon=0} = \frac{\partial W_1}{\partial P_i} \tag{496}$$

The same procedure can be accomplished for p_i, which leads to the following transformations from the old to the new coordinates:

$$\begin{aligned} q_i &= Q_i + \epsilon \frac{\partial W_1}{\partial P_i} + O\left(\epsilon^2\right) \\ p_i &= P_i - \epsilon \frac{\partial W_1}{\partial Q_i} + O\left(\epsilon^2\right) \end{aligned}$$

Given an arbitrary function $f(q_i, p_i)$, we can apply the following canonical transformation defined by the generating function W:

$$f\left(q_i, p_i, \epsilon\right) = f\left(Q_i + \epsilon\frac{\partial W_1}{\partial P_i} + O\left(\epsilon^2\right), P_i - \epsilon\frac{\partial W_1}{\partial Q_i} + O\left(\epsilon^2\right)\right)$$
$$= f(Q_i, P_i) + \epsilon\left.\frac{df}{d\epsilon}\right|_{\epsilon=0} + \frac{1}{2}\epsilon^2\left.\frac{d^2 f}{d\epsilon^2}\right|_{\epsilon=0} + O(\epsilon^3) \tag{497}$$

We now can calculate the first derivative with respect to ϵ in the following form when we use (495)

$$\left.\frac{df}{d\epsilon}\right|_{\epsilon=0} = \sum_i \left(\frac{\partial f}{\partial q_i}\frac{\partial q_i}{\partial \epsilon} + \frac{\partial f}{\partial p_i}\frac{\partial p_i}{\partial \epsilon}\right)$$
$$= \sum_i \left(\frac{\partial f}{\partial Q_i}\frac{\partial W_1}{\partial P_i} - \frac{\partial f}{\partial P_i}\frac{\partial W_1}{\partial Q_i}\right)$$
$$= [f, W_1] \tag{498}$$

Introducing furthermore the *Poisson Bracket* $[f, g]$:

$$[f, g] = \sum\left(\frac{\partial f}{\partial Q_i}\frac{\partial g}{\partial P_i} - \frac{\partial f}{\partial P_i}\frac{\partial g}{\partial Q_i}\right) \tag{499}$$

we can write for (497) up to the first order in ϵ

$$f(q_i, p_i) = f(Q_i, P_i) + \epsilon\,[f, W_1] + O(\epsilon^2) \tag{500}$$

We now replace f with H from (488) and obtain

$$H\left(q_i, p_i, \epsilon\right) = H_0\left(Q_i, P_i\right) + \epsilon\,[H_0, W_1] + \epsilon H_1\left(Q_i, P_i\right) + O\left(\epsilon^2\right) \tag{501}$$

By comparing equations (501) and (488) one can identify

$$K_0\left(Q_i, P_i\right) = H_0\left(Q_i, P_i\right)$$
$$K_1\left(Q_i, P_i\right) = H_1\left(Q_i, P_i\right) + [H_0, W_1] \tag{502}$$

and obtain the new Kamiltonian

$$K\left(Q_i, P_i\right) = H_0 + \epsilon\left(H_1 + [H_0, W_1]\right) + O(\epsilon^2) \tag{503}$$

We now can choose W in such a way, that K becomes considerably simpler than the original Hamiltonian H.

Higher Orders

We can use the development (497) up to the 2nd order and find for the 2nd derivative with respect to ϵ

$$\frac{d^2 f}{d\epsilon^2} = \frac{d}{d\epsilon}[f, W] = \left[\frac{df}{d\epsilon}, W\right] + \left[f, \frac{\partial W}{\partial \epsilon}\right] = [[f, W], W] + \left[f, \frac{\partial W}{\partial \epsilon}\right] \tag{504}$$

Differentiatingg W gives

$$\frac{\partial W}{\partial \epsilon} = W_2 \left(Q_i, P_i\right) + 2\epsilon W_3 \left(Q_i, P_i\right) + \ldots \tag{505}$$

and thus

$$\left.\frac{df}{d\epsilon}\right|_{\epsilon=0} = \left[f, W_1\right] \tag{506}$$

$$\left.\frac{d^2 f}{d\epsilon^2}\right|_{\epsilon=0} = \left[\left[f, W_1\right], W_1\right] + \left[f, W_2\right] \tag{507}$$

resulting in the transformation of f:

$$f\left(q_i, p_i\right) = f\left(Q_i, P_i\right) + \epsilon\left[f, W_1\right] + \frac{1}{2}\epsilon^2\left(\left[\left[f, W_1\right], W_1\right] + \left[f, W_2\right]\right) + O\left(\epsilon^3\right) \tag{508}$$

As a consequence the complete Kamiltonian up to the second order reads

$$K\left(Q_i, P_i\right) = H_0 + \epsilon\left(H_1 + \left[H_0, W_1\right]\right) + \frac{\epsilon^2}{2}\left(H_2 + \left[\left[H_0, W_1\right], W_1\right] + \left[H_0, W_2\right]\right) \tag{509}$$

Making the developments to even higher orders we obtain for the contribution of the third order Hamiltonian the following expression:

$$K_0 = H_0 \tag{510}$$

$$K_1 = H_1 + \left[H_0, W_1\right] \tag{511}$$

$$K_2 = H_2 + \left[H_1, W_1\right] + \frac{1}{2}\left[\left[H_0, W_1\right], W_1\right] + \frac{1}{2}\left[H_0, W_2\right] \tag{512}$$

$$\begin{aligned} K_3 = {} & H_3 + \left[H_2, W_1\right] + \frac{1}{2}\left[\left[H_1, W_1\right], W_1\right] + \\ & + \frac{1}{2}\left[H_1, W_2\right] + \frac{1}{6}\left[\left[\left[H_0, W_1\right], W_1\right], W_1\right] \\ & + \frac{1}{3}\left[H_0, W_3\right] + \frac{1}{3}\left[\left[H_0, W_2\right], W_1\right] + \frac{1}{6}\left[\left[H_0, W_1\right], W_2\right] \end{aligned} \tag{513}$$

From this formalism it is evident that the procedure of applying the Lie-transform as perturbation theory can be automized with an appropriate computer algebra program like *Maple* or *Mathematica*.

B.2 A two Degrees of Freedom System: the Hénon-Heiles System

We demonstrate the technique of Lie Transformation up to the 2^{nd} order for a two degrees of freedom system, the Hénon-Heiles system, where two linear oscillators are nonlinearly coupled. In the unperturbed (noncoupled case) both oscillators have the same frequency and this will – as we can see later – lead to terms which cannot be removed via a canonical transformation. Nevertheless the phase space structure in its regular parts can be represented

for small coupling in a satisfactory way. The Hamiltonian of the system is given by

$$H = H_0 + \epsilon H_1 = \frac{1}{2}p_x^2 + \frac{1}{2}p_y^2 + \frac{1}{2}x^2 + \frac{1}{2}y^2 + \epsilon\left(x^2 y - \frac{1}{3}y^3\right) \tag{514}$$

It is quite usual to transform to action-angle variables via

$$\begin{aligned} x &= \sqrt{2p_1}\sin q_1 \\ y &= \sqrt{2p_2}\sin q_2 \\ p_x &= \sqrt{2p_1}\cos q_1 \\ p_y &= \sqrt{2p_2}\cos q_2 \end{aligned} \tag{515}$$

which leads to the Hamiltonian

$$H\left(q_1, q_2, p_1, p_2\right) = p_1 + p_2 + 2\sqrt{2}\epsilon\left(p_1 p_2^{\frac{1}{2}}\sin^2 q_1 \sin q_2 - \frac{1}{3}p_2^{\frac{3}{2}}\sin^3 q_2\right) \tag{516}$$

This form is not practical for removing the periodic terms and as next step we have to replace the terms $\sin^2 q_1$ and $\sin^3 q_2$ by a well known tranformations which gives the possibility to remove the terms by an adequate choice of the Lie generating function W. After this transformation the Hamiltionian reads

$$\begin{aligned} H = p_1 + p_2 + 2\sqrt{2}\epsilon\Bigg(&p_1 p_2^{\frac{1}{2}}\left(\frac{1}{2}\sin q_2 - \frac{1}{4}\sin\left(q_2 + 2q_1\right) - \frac{1}{4}\sin\left(q_2 - 2q_1\right)\right) \\ &- \frac{1}{3}p_2^{\frac{3}{2}}\left(\frac{3}{4}\sin q_2 - \frac{1}{4}\sin 3q_2\right)\Bigg) \end{aligned} \tag{517}$$

With equations (497) and the canonical tranformation of the coordinates we get the Hamiltonian (respectively the Kamiltonian), where $K_0 = H_0 = P_1 + P_2$ and

$$\begin{aligned} K_1 &= H_1(Q_i, P_i) + [H_0, W_1] \\ &= 2\sqrt{2}\Bigg(P_1 P_2^{\frac{1}{2}}\left(\frac{1}{2}\sin Q_2 - \frac{1}{4}\sin(Q_2 + 2Q_1) - \frac{1}{4}\sin(Q_2 - 2Q_1)\right) \\ &\quad - \frac{1}{3}P_2^{\frac{3}{2}}\left(\frac{3}{4}\sin Q_2 - \frac{1}{4}\sin 3Q_2\right)\Bigg) - \frac{\partial W_1}{\partial Q_1} - \frac{\partial W_1}{\partial Q_2} \end{aligned} \tag{518}$$

We can now choose a Lie-generating function such, that K_1 becomes as simple as possible; we suggest the following ansatz for W_1

$$W_1 = A\cos Q_2 + B\cos\left(2Q_1 + Q_2\right) + C\cos\left(2Q_1 - Q_2\right) + D\cos 3Q_2 \tag{519}$$

We differentiate W_1 with respect to the coordinates Q_1 and Q_2 and choose the coefficients A, B and C to eliminate the terms in K_1

$$A = \frac{\sqrt{2}}{2}\left(P_2^{\frac{3}{2}} - 2P_1P_2^{\frac{1}{2}}\right)$$
$$B = \frac{\sqrt{2}}{6}\left(P_1P_2^{\frac{1}{2}}\right)$$
$$C = -\frac{\sqrt{2}}{2}\left(P_1P_2^{\frac{1}{2}}\right)$$
$$D = -\frac{\sqrt{2}}{18}\left(P_2^{\frac{3}{2}}\right). \tag{520}$$

With this choice *all* terms in K_1 cancel and consequently $K_1 = 0$! To proceed to the 2nd order theory we, first of all, make use of

$$K_1 = H_1 + [H_0, W_1] \tag{521}$$

and substitute in the respective equation for K_2 (512) $[H_0, W_1]$ with $K_1 - H_1$ and obtain $[[H_0, W_1], W_1] = [K_1, W_1] - [H_1, W_1]$. The 2nd order Kamiltonian K_2 therefore reads

$$K_2 = H_2 + \frac{1}{2}\left([K_1, W_1] + [H_1, W_1] + [H_0, W_2]\right) \tag{522}$$

Because $K_1 = 0$ and also $H_2 = 0$ the 2nd order Kamiltonian reduces to

$$K_2 = \frac{1}{2}[H_1, W_1] - \frac{\partial W_2}{\partial Q_1} - \frac{\partial W_2}{\partial Q_2} \tag{523}$$

We now have to choose W_2 appropriately to simplify the Hamiltonian; this time we cannot eliminate all terms in second order because through the Poisson bracket $[H_1, W_1]$ new resonances and also a constant term appear in K_1:

$$\begin{aligned}[H_1, W_1] = {} & \frac{1}{3}P_1P_2\cos(2Q_1 + 2Q_2) - \frac{7}{3}P_1P_2\cos(2Q_2 - Q_1) \\ & + \frac{1}{6}(P_2^2\cos 4Q_2 + P_1^2\cos 4Q_1 + \frac{2}{3}(P_1 + P_2)(P_2\cos 2Q_2 + P_1\cos 2Q_1) \\ & - \frac{5}{6}(P_1^2 + P_2^2) + \frac{2}{3}P_1P_2 \end{aligned} \tag{524}$$

The remaining terms lead to a Kamiltonian K_2

$$K_2 = \frac{1}{2}\left(-\frac{7}{3}P_1P_2\cos(2Q_2 - 2Q_1) + \frac{2}{3}P_1P_2 - \frac{5}{6}(P_1^2 + P_2^2)\right) \tag{525}$$

Nevertheless a further discussion leads to an interesting reduction of the Kamiltonian, because the two coordinates Q_1 and Q_2 appear only in the combination $Q_2 - Q_1$. Now we can pose a simple canonical transformation via

$$\begin{aligned} X_1 &= Q_2 - Q_1 \\ X_2 &= Q_2 \\ Y_1 &= P_1 \\ Y_2 &= P_1 + P_2 \end{aligned} \tag{526}$$

where the Kamiltonian is now given as

$$K = Y_2 + \frac{\epsilon^2}{2}\left(-\frac{7}{3}Y_1\left(Y_2 - Y_1\right)\cos 2X_1 - \frac{5}{6}Y_1^2 - \frac{2}{3}Y_1\left(Y_2 - Y_1\right) - \frac{5}{6}(Y_2 - Y_1)^2\right) + O(\epsilon^3) \tag{527}$$

and X_2 does not appear explicitly in (527). Thus, because of the Hamiltonian character of the equations, it follows that Y_2 is constant. As a consequence also the difference $K - Y_2 = G$ is a constant of motion and K_2 is – up to order 3 – a first integral of the system under consideration. We now transform this expression for G back to the original coordinates x, p_x, y and p_y, which gives us the following form for this second integral of motion

$$G = 5p_y^4 + (10y^2 + 10p_x^2 - 18x^2)p_y^2 + 56xyp_xp_y + 5y^4 + (10x^2 - 18p_x^2)y^2 + 5p_x^2 + 10x^2p_x^2 + 5x^4 \tag{528}$$

With the Hamiltonian as one integral and G as another integral of motion we can eliminate with the aid of H e.g. the coordinate p_x and replace it in the integral G. Still this would be a three dimensional subspace of the phase space, which is difficult to represent. We therefore choose a surface of section[70] and set x=0, which will then define curves in the $y - p_y$ plane depending on the value of the energy h (the value of the Hamiltonian which can be computed from the initial conditions). We first express p_x with the aid of the energy h and set $x = 0$

$$p_x^2 = 2h - p_y^2 - y^2 + \frac{2}{3}\epsilon y^3 \tag{529}$$

We also set $x = 0$ in the integral G and replace the value of p_x, which now leads to the following curves

$$\begin{aligned} &5(p_y^2 + y^2)^2 + (10p_y^2 - 18y^2)(2h - p_y^2 - y^2 + \frac{2}{3}\epsilon y^3) + \\ &+5(2h - p_y^2 - y^2 + \frac{2}{3}\epsilon y^3)^2 = const \end{aligned} \tag{530}$$

In Figs. 70 and 71 we show respective curves for $\epsilon = 0.1$ and $h = 0.15$ for the results obtained by the perturbation theory and numerical integrations.

[70] for details on how to construct a surface of section see Sect. 2.1 in the chapter on "Regular and Chaotic Motion in Hamiltonian Systems".

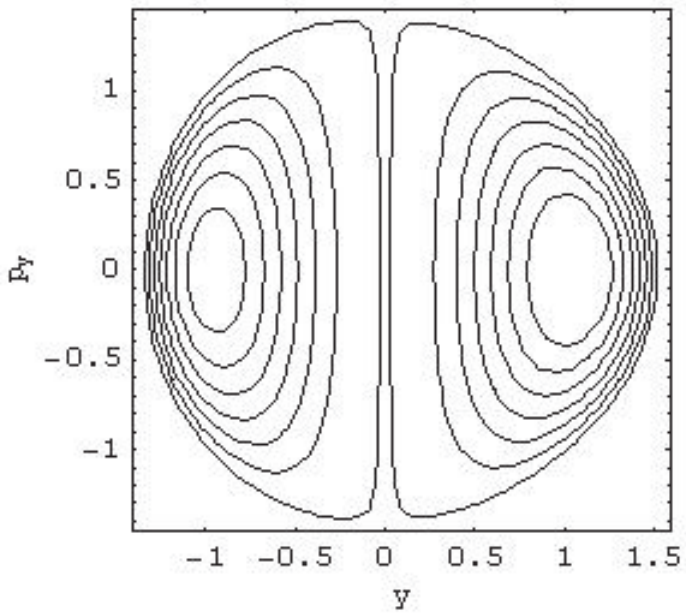

Fig. 70. Surface of Section for the Hénon Heiles system obtained from Lie Perturbation Theory *(for* $\epsilon = 0.1$ *and* $h = 0.15$*)*

C Laplace Coefficients

If we write the direct part of the perturbation function we obtain (see equation (330))

$$\frac{1}{\rho} = \frac{1}{r_2}\left[1 - 2\alpha\cos\phi + \alpha^2\right]^{\frac{1}{2}} \tag{531}$$

Expanding this expression in Fourier series gives

$$\frac{1}{r_2}\left[1 - 2\alpha\cos\phi + \alpha^2\right]^{\frac{1}{2}} = \frac{1}{2}b_{s/2}^{(0)} + \sum_{j=1}^{\infty} b_{s/2}^{(j)}\cos j\phi \tag{532}$$

where s is an positive odd integer. The coefficients $b_{s/2}^{(j)}$ where first studied by Laplace and therefore are called Laplace coefficients. They were and are studied intensively (see [66] for a recent work).

Table 4 shows some numerical values of the Laplace coefficients (also for the ratio of the semimajor axes of Jupiter and Saturn). They can be calculated by the following formulas (for details on the derivation see [64]):

$$\frac{1}{2}\left(b_{s+1}^{(j)} + b_{s+1}^{(j+1)}\right) = \frac{(j+s)\,b_s^{(j)} - (j-s+1)\,b_s^{(j+1)}}{2s\,(1-\alpha)^2} \tag{533}$$

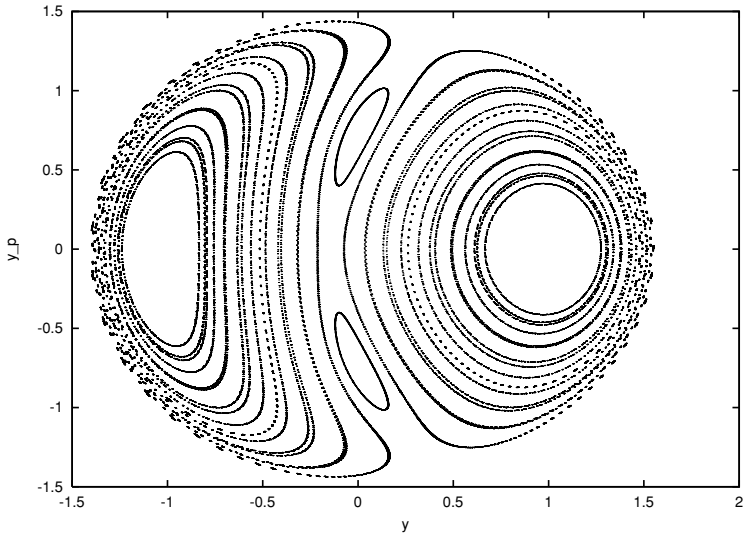

Fig. 71. Surface of Section for the Hénon Heiles system obtained from numerical integrations *(for* $\epsilon = 0.1$ *and* $h = 0.15$*)*

Table 4. Laplace coefficients

α	$\frac{1}{2}b_0^{(1/2)}$	$b_1^{(1/2)}$	$\frac{1}{2}b_0^{(3/2)}$	$b_1^{(3/2)}$	$b_2^{(3/2)}$
0.1	1.0025	0.1004	1.0228	0.3057	0.0381
0.2	1.0102	0.2031	1.0954	0.6519	0.1611
0.3	1.0237	0.3107	1.2349	1.0745	0.3983
0.4	1.0440	0.4267	1.4745	1.6660	0.8154
0.5	1.0732	0.5559	1.8908	2.5805	1.5580
0.54543	1.0902	0.6208	2.1800	3.1873	2.0837
0.6	1.1146	0.7060	2.6666	4.1867	2.9799
0.7	1.1750	0.8923	4.3323	7.5430	6.1179
0.8	1.2702	1.1443	9.0110	16.885	15.161

$$\frac{1}{2}\left(b_{s+1}^{(j)} - b_{s+1}^{(j+1)}\right) = \frac{(j+s)\, b_s^{(j)} + (j-s+1)\, b_s^{(j+1)}}{2s\,(1+\alpha)^2} \tag{534}$$

D Legendre Polynomials

The Legendre polynomials (or Legendre functions of first kind) are solutions of the Legendre differential equations and can also be used to develop the direct part of the perturbation function:

$$\left(1-x^2\right)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + l\,(l+1)\,y = 0 \tag{535}$$

These solutions are polynomials, if l is an integer and can be defined by

$$\wp_n(x) = \frac{1}{2\pi} \oint \left(1 - 2tx + t^2\right)^{\frac{1}{2}} t^{-n-1} dt \tag{536}$$

The Legendre polynomials up to order 7 are

$$\wp_0(x) = 1 \tag{537}$$
$$\wp_1(x) = x \tag{538}$$
$$\wp_2(x) = \frac{1}{2}\left(3x^2 - 1\right) \tag{539}$$
$$\wp_3(x) = \frac{1}{2}\left(5x^3 - 3x\right) \tag{540}$$
$$\wp_4(x) = \frac{1}{8}\left(35x^4 + 30x^2 + 3\right) \tag{541}$$
$$\wp_5(x) = \frac{1}{8}\left(65x^5 - 70x^3 + 15x\right) \tag{542}$$
$$\wp_6(x) = \frac{1}{16}\left(231x^6 - 315x^4 + 105x^2 - 5\right) \tag{543}$$
$$\wp_7(x) = \frac{1}{16}\left(429x^7 - 693x^5 + 315x^3 - 35x\right) \tag{544}$$

Acknowledgments

We want to thank the Helmholtz Institute for Supercomputational Physics in Potsdam and its directors, Profs. Kurths and Rüdiger, for inviting us to give lectures at the Summer School. We also acknowledge the support of Carmen Romano and Marco Thiel who excellently organized the Summer School. We want to thank Richard Neuteufel for providing us with Figs. 33-35 and the members of the ADG (AstroDynamicsGroup) in Vienna for their help in preparing this manuscript.

References

1. J. Alfaro and C. Chiralt: Invariant Rotational Curves in Sitnikov's Problem. CMDA **55**, 351 (1993) 79
2. N. Asghari, C. Broeg, L. Carone, R. Casas-Miranda, J.C. Castro Palacio, I. Csillik, R. Dvorak, F. Freistetter, G. Hadjivantsides, H. Hussmann, A. Khramova, M. Khristoforova, I. Khromova, I. Kitiashivilli, S. Kozlowski, T. Laakso, T. Laczkowski, D. Lytvinenko, O. Miloni, R. Morishima, A. Moro-Martin, V. Paksyutov, A. Pal, V. Patidar, B. Pečnik, O. Peles, J. Pyo, T. Quinn, A. Rodriguez, C. Romano, E. Saikia, J. Stadel, M. Thiel, N. Todorovic, D. Veras, E. Vieira Neto, J. Vilagi, W. von Bloh, R. Zechner, E. Zhuchkova: Stability of terrestrial planets in the habitable zone of Gl 777 A, HD 72659, Gl 614, 47 Uma and HD 4208. A&A **426**, 353 (2004) 68, 69, 113, 114, 115

3. D. Brouwer and G.M. Clemence: *Methods of Celestial Mechanics*, (Academic Press, New York and London 1961) 84
4. A. Celletti and A. Giorgilli: On the stability of the Lagrangian points in the spatial restricted problem of three bodies. CMDA **50**, 31 (1991) 113
5. G. Contopoulos : *Order and Chaos in Dynamical Astronomy*, Astronomy and Astrophysics Library, (Springer 2002) 34, 35
6. G. Contopoulos, R. Dvorak, M.Harsoula and F. Freistetter: Chaos and Birfucations, submitted (2004) 19
7. M. Delva: A Lie integrator program and test for the elliptic restricted three body problem. A&A Supp **60**, 277 (1985) 126
8. R. Dvorak, G. Contopoulos, C. Efthymiopoulos and N. Voglis: "Stickiness" in mappings and dynamical systems. Planetary and Space Science **46**, 1567 (1997) 19, 21, 22, 23
9. R. Dvorak, F. Vrabec and K. Wodnar: The Sitnikov Problem: A Short Review. In: *Sistema solare e sistemi stellari, Teorie perturbative, Dinamica del volo spaziale* ed by A. Celletti and E. Perozzi, 16 (1998) 80
10. R. Dvorak, F. Freistetter. B. Funk and G. Contopoulos: New insights on the chaotic behaviour of the standardmap. In: *Proceedings of the 3rd Austrian-Hungarian Workshop on Trojans and related Topic* ed by F. Freistetter, R. Dvorak and B.Érdi (Eötvös University Press 2003), 185 19, 113
11. R. Dvorak, E. Pilat-Lohinger, B. Funk and F. Freistetter: Planets in habitable zones: A study of the binary Gamma Cephei. A&A **398**, L1 (2003) 113
12. R. Dvorak, E. Pilat-Lohinger, B. Funk and F. Freistetter: A study of the stable regions in the planetary system HD 74156 - Can it host earthlike planets in habitable zones? A&A **410**, L13 (2003) 113
13. R. Dvorak: Chaos in Solar System Dynamics. In: *Galaxies and Chaos* ed by G. Contopoulos and N. Voglis, Lect. Notes in Phys. **626**, 395 (2003) 105
14. R. Dvorak, E. Pilat-Lohinger, R. Schwarz and F. Freistetter: Extrasolar Trojan planets close to habitable zones A&A **426**, L37 (2004) 66
15. Efthymiopoulos, C. Contopoulos, G. and N. Voglis: Cantori, Islands and Asymptotic Curves in the Stickiness Region, CMDA **73**, 221 (1999) 121
16. Ch. Efthymiopoulos: CMDA, (in press) (2005) 70, 77
17. B. Érdi: Long periodic perturbations of Trojan asteroids. CMDA **43**, 303 (1988) 67
18. B. Érdi: The Trojan problem. CMDA **65**, 149 (1997) 67
19. B. Érdi and A. Pál: Dynamics of resonant exoplanetary systems. In: *Proceedings of the 3rd Austrian-Hungarian Workshop on Trojans and related Topic* ed by F. Freistetter, R. Dvorak and B.Érdi (Eötvös University Press 2003), 3 113
20. B. Érdi, R. Dvorak, Zs. Sándor and E. Pilat-Lohinger: The dynamical structure of the habitable zone in the HD 38529, HD 168443 and HD 169830 systems. MNRAS **351**, 1943 (2004) 113, 117
21. E. Fermi: Phys. Rev. **75**, 1169 (1949) 14
22. T. Fliessbach: *Theoretische Mechanik* (Spektrum Akademischer Verlag 1996) 23
23. C. Froeschlé, E. Lega and R. Gonczi: Fast Lyapunov Indicators. Application to Asteroidal Motion. CMDA **67**, 41 (1997) 114
24. H. Goldstein, C. Poole and J. Safko: *Classical Mechanics*, (Pearson, Addison and Wesley 2002) 127
25. W. Gröbner: *Die Lie-Reihen und ihre Anwendungen*, (VEB, Deutscher Verlag der Wissenschaften, Berlin 1967) 121, 122

26. J. Hagel: A new analytic approach to the Sitnikov problem. CMDA **53**, 267 (1992) 78
27. J. Hagel and Th. Trenkler: A Computer-Aided Analysis of the Sitnikov Problem CMDA **56**, 81 (1993) 79
28. A. Hanslmeier and R. Dvorak: Numerical Integration with Lie Series. A&A **132**, 204 (1984) 105, 126
29. M. Hénon and C. Heiles: The applicability of the third integral of motion: some numerical experiments. AJ **69**, 73 (1964) 33
30. D.F. Hevia and A.F. Raada: Chaos in the three-body problem: the Sitnikov case. European Journal of Physics **17** 5, 295 (1995) 70
31. M.J. Holman and P.A. Wiegert: Long-Term Stability of Planets in Binary Systems. AJ **117**, 621 (1999) 113
32. T. Ito and K. Tanikawa: Long-term integrations and stability of planetary orbits in our Solar system. MNRAS **336**, 483 (2002) 105
33. J.F. Kasting, D.P. Whitmire and R.T. Reynolds: Habitable Zones around Main Sequence Stars. Icarus, **101**, 108 (1993) 112
34. H. Lammer, R. Dvorak, E. Pilat-Lohinger, B. Funk, F. Freistetter, I. Ribas, F. Selsis, E.F. Guinan, W.W. Weiss and S.J. Bauer: Atmosphere and orbital stability of exosolar planets orbiting gamma Cephei. In: *Proceedings EGS-AGU-EUG Joint Assembly, Nice* (2003) 112
35. J. Laskar: The chaotic motion of the solar system - A numerical estimate of the size of the chaotic zones. Icarus, **88**, 266 (1990) 108, 109
36. J. Laskar, T. Quinn and S. Tremaine: Confirmation of resonant structure in the solar system, Confirmation of resonant structure in the solar system. Icarus **95**, 148 (1992) 110
37. J. Laskar: Large-scale chaos in the solar system. A&A **287**, L9 (1994) 108
38. J. Laskar: Large Scale Chaos and Marginal Stability in the Solar System. CMDA**64**, 115 (1996) 108, 110
39. G. Laughlin and J.E. Chambers: Extrasolar Trojans: The Viability and Detectability of Planets in the 1:1 Resonance. AJ **124**, 592 (2002) 66, 113
40. M. Lecar, F.A. Franklin, M.J. Holman and N.J. Murray: Chaos in the Solar System. Ann. Rev. of Astronomy and Astrophysics **39**, 581 (2001) 105
41. Ch. Lhotka: High oder perturbation theory for the Sitnikov problem. MA Thesis, University of Vienna, (2004) 76, 79
42. A.J. Lichtenberg and M.A. Lieberman: *Regular and Stochastic Motion*, (Springer 1983) 14, 33, 60
43. H. Lichtenegger: The dynamics of bodies with variable masses. CMDA **34**, 357 (1984) 126
44. M.A. Liebermann and A.J. Lichtenberg: Phys. Rev **A5**, 1852 (1972) 15
45. W.D. MacMillan: An integrable case in the restricted problem of three bodies. AJ **27**, 11 (1911) 71
46. C. Marchal: *The Three Body Problem*, (Elseveer 1991) 113
47. K. Menou and S. Tabachnik: Dynamical Habitability of Known Extrasolar Planetary Systems. AJ **583**, 473 (2003) 113
48. A. Milani: The Trojan asteroid belt: Proper elements, stability, chaos and families. CMDA **57**, 59 (1993) 68
49. A. Milani, A. Nobili and Z. Knežević: Stable Chaos in the Asteroid Belt. Icarus **125**, 13 (1997) 68

50. J. Moser: Stable and Random Motions in Dynamical Systems with Special Emphasis on Celestial Mechanics. In: *Annals of Mathematical Studies Number 77*, (Princeton University Press, Princeton, New Jersey 1971) 70
51. C.D. Murray and S.F. Dermott: *Solar System Dynamics* (Cambridge University Press 1999) 84, 95, 102, 103
52. D. Naef, M. Mayor, S.G. Korzennik, D. Queloz, S. Udry, P. Nisenson, R.W. Noyes, T.M. Brown, J.L. Beuzit, C. Perrier and J.P. Sivan: The ELODIE survey for northern extra-solar planets. II. A Jovian planet on a long-period orbit around GJ 777 A. A&A **410**, 1051 (2003) 114
53. G. Pavanini: Annali di Mathematica, Serie III, Tomo XIII (1907) 71
54. E. Pilat-Lohinger, R. Dvorak and Ch. Burger: Trojans in Stable Chaotic Motion. CMDA **73**, 117 (1999) 68
55. E. Pilat-Lohinger and R. Dvorak: Stability of planetary orbits in double stars. CMDA **82**, 143 (2002) 113
56. W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery: *Numerical Recipes*, (Cambridge University Press 1992) 113
57. E. Rabe: Third-order stability of the long-period Trojan librations. AJ **72**, 10 (1967) 113
58. G. Rabl and R. Dvorak: Satellite-type planetary orbits in double stars - A numerical approach. A&A **191**, 385 (1988) 113
59. R.H. Rand: *Topics in Nonlinear Dynamics with Computer Algebra*, (Gordon and Breach Science Publishers 1994) 127
60. A.E. Roy: *Orbital Motion*, (Adam Hilger, Bristol and Philadelphia 1988) 84
61. Z. Sándor, B. Érdi, A. Széll and B. Funk: The Relative Lyapunov Indicator: An Efficient Method of Chaos Detection. CMDA. **90**, 127 (2004) 114
62. Ch. Skokos and A. Dokoumetzidis: Effective stability of the Trojan asteroids. A&A **367**, 729 (2001) 70
63. K. Stumpff: *Himmelsmechanik I, Das Zweikörperproblem und die Methoden der Bahnbestimmung der Planeten und Kometen*, (VEB, Deutscher Verlag der Wissenschaften, Berlin 1959) 38
64. K. Stumpff: *Himmelsmechanik II, Das Dreikörperproblem*, (VEB, Deutscher Verlag der Wissenschaften, Berlin 1965) 51, 72, 134
65. K. Stumpff and J. Meffroy: *Himmelsmechanik III, Allgemeine Störungen*, (VEB, Deutscher Verlag der Wissenschaften, Berlin 1974) 65, 84
66. A. Süli, B. Èrdi and Andras Pál: A New Method to Determine the Derivatives of the Laplace Coefficients. CMDA **88**, 259 (2004) 134
67. V. Szebehely: *Theory of Orbits, The Restricted problem of Three Bodies*, (Academic Press, 1967) 51
68. M. Thiel, M.C. Romano and J. Kurths: Applied Nonlinear Dynamics 11, N3 115
69. M. Thiel, M.C. Romano, J. Kurths and P. Read: Chaos (in press) 115
70. K. Tsiganis, R. Dvorak and E. Pilat-Lohinger: Thersites: a 'jumping' Trojan? A&A **354**, 1091 (2000) 68
71. K. Tsiganis, H. Varvoglis and R. Dvorak: Chaotic Diffusion and Effective Stability of Jupiter Trojans, CMDA, in press (2005) 68
72. S. Ulam: In: *Proceedings of the 4th Berkely Symp. on Math. Stat. and Probability* (University of California Press 1961), p. 315 15
73. J.P. van der Weele, H.W. Capel, T.P. Valkering and T. Post: The squeeze effect in non-integrable Hamiltonian systems, Physica A, **147**, 499 (1987) 77

74. H. Varvoglis, K. Tsiganis and J.D. Hadjidemetriou: The "Third" Integral in the Restricted Three-Body Problem Revisited. In: *Galaxies and Chaos* ed by G. Contopoulos and N. Voglis, Lect. Notes Phys. **626**, 433 (2003) 71
75. H. Varvoglis, Ch. Vozikis and K. Wodnar: The Two Fixed Centers: An Exceptional Integrable System. CMDA **89**, 343 (2004) 71
76. K. Wodnar: New formulations of the Sitnikov Problem, In: *Predictibility, Stability and Chaos in N-Body Dynamical Systems* ed by A. Roy, (Plenum Press, New York 1992), 457 78

Regular and Chaotic Motion in Hamiltonian Systems

Harry Varvoglis[1]

Section of Astrophysics Astronomy & Mechanics, Department of Physics, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece
(varvogli@physics.auth.gr)

1 Introduction

All laws that describe the time evolution of a *continuous* system are given in the form of differential equations, ordinary (if the law involves one independent variable) or partial (if the law involves two or more independent variables). Historically the first law of this type was Newton's second law of motion. Since then Dynamics, as it is customary to name the branch of Mechanics that studies the motion of a body as the result of a force acting on it, has become the "typical" case that comes into one's mind when a system of ordinary differential equation is given, although this system might as well describe any other system, e.g. physical, chemical, biological, financial etc. In particular the study of "conservative" dynamical systems, i.e. systems of ordinary differential equations that originate from a time-independent Hamiltonian function, has become a thoroughly developed area, because of the fact that mechanical energy is very often conserved, although many other physical phenomena, beyond motion, can be described by Hamiltonian systems as well. In what follows we will restrict ourselves exactly to the study of Hamiltonian systems, as typical dynamical systems that find applications in many scientific disciplines.

The method used traditionally in the study of a conservative dynamical system, whose Hamiltonian leads to a complex set of differential equations, is to follow a sequence of approximations. In the beginning, we try to solve exactly the equations derived from a simplified form of the Hamiltonian, the so-called zero-order approximation, omitting complex or non-linear terms. Then we revert to the full Hamiltonian and consider the omitted terms as a "perturbation" of the zero order approximation. We then modify the initial solution by a sequence of "corrections", which are expected to describe better and better the real system.

This method is based on the implicit assumption that the sequence of successive approximations converges to the "real" solution. Since the method is based on the fact that the solution of the zero order dynamical system is known exactly, it is closely connected to the theory of *integrable*, or else *regular* dynamical systems. By this term, we denote the Hamiltonian dynamical systems whose differential equations of motion can be solved at least by quadratures, i.e. they can be expressed as integrals of functions of one

H. Varvoglis: *Regular and Chaotic Motion in Hamiltonian Systems*, Lect. Notes Phys. **683**, 141–184 (2005)
www.springerlink.com

variable. Under some rather general conditions and when the phase space is compact, i.e. neither of the generalized co-ordinates or momenta do not become infinite, the phase space is isomorphic, as we show in Sect. 3, to a n-torus, when n is the number of degrees of freedom of the system. In that case the "natural" selection for the generalized co-ordinates is a set of n angles and the system can be solved exactly.

Since the nineteenth century it was understood, however, that the method of beginning by an integrable Hamiltonian and then add perturbation terms is in some cases ill-behaved. The reason is that, if the series representing the solution of the "real" system is to converge to the "real" solution, the successive terms generated by the successive approximations have to get smaller and smaller. But it was found out that these terms have, in their denominators, linear combinations of the n angles which can become, for appropriate sets of coefficients, arbitrarily small. Therefore the terms of the series do not get ever smaller, and its convergence is not guaranteed. This is the famous problem of the "small divisors", which appeared first when people working on Celestial Mechanics tried to find the general solutions of the three-body problem.

In the beginning the problem of the small divisors was considered to be a "technicality", which could be overcome through an "appropriate" mathematical handling. It so happened because the partial sums of the first few terms of the series solutions were giving very good approximations for the actual motion of the planets and their satellites. At the end of the nineteenth century, however, Poincaré proved that the problem of small divisors is inherent to the dynamical system of the three-body problem, so that there is no way to avoid it. Moreover he showed that this problem entails a very complex evolution of trajectories that approach an unstable periodic orbit, a phenomenon that today we call *chaos*. However Poincaré's result was not particularly appreciated at his time, since the analytical solutions, in formal power series, were working in Celestial Mechanics as expected.

Things started to change after the appearance of the electronic computer, with the aid of which it became possible to attempt the numerical solution of non-linear equations of motion. Hénon and Heiles [17], in their pioneering work, showed that chaos may dominate the phase space of very simple dynamical systems. In the following years it became evident that non-integrable dynamical systems, in which appears the phenomenon discovered by Poincaré, are the rule rather than the exception in nature. It is evident that the approximation of a chaotic dynamical system by an integrable one is not useful. Indeed, as we will show in Sects. 2.2 and 4.1, a solution of a chaotic dynamical system deviates exponentially not only from the solution of any integrable approximation, but from any other nearby solution of the full system as well. Therefore methods devised for the study of regular systems cannot be applied to chaotic ones, at least not without a justification on a case by case basis.

Extensive analytical and numerical work since the paper by Hénon and Heiles has shown that almost all real dynamical systems are neither purely ordered nor purely chaotic, but their phase space consists of closely intermingled ordered and chaotic regions. The systems that have this property are said to posses a "divided" phase space. These systems can be considered as ordered or chaotic, depending on the measure of the chaotic and ordered regions of phase space.

This article is divided into four sections:

1. Elementary tools for the study of dynamical systems,
2. Integrable Hamiltonian systems,
3. Non-integrable (or else chaotic) Hamiltonian systems and
4. Transition from order to chaos and the Chirikov-Contopoulos stochasticity criterion.

In the first section we present the two most basic tools that are used in the numerical study of Hamiltonian systems, and in particular for the determination of the ordered or chaotic nature of trajectories. These are the *surface of section technique* and the *Lyapunov Characteristic Numbers*.

In the second section we introduce the notion of an integrable Hamiltonian system, through the existence of integrals of motion and canonical action-angle variables. We find these variables in two simple Hamiltonian systems, which are very often used as simple models of more complicated systems, the harmonic oscillator and the pendulum, and in three real Hamiltonian systems, the *two fixed centers problem*, the *Toda lattice* and the *Kepler's problem*. We introduce also the concept of *resonance*.

In the third section we present the mathematical definitions and properties of five classes of dynamical systems that are used as models of "real" chaotic Hamiltonian systems with divided phase space: *ergodic*, *mixing*, *K*- and *C-systems* and *Bernoulli shifts*.

In the fourth and last section we present the Contopoulos-Chirikov *stochasticity criterion* and we give a complete working example on how this criterion is implemented in a specific dynamical system.

There exist excellent textbooks in most of the topics covered in the present mini-course, such as those by Arnold [3], Contopoulos [10], Gutzwiller [16], Lichtenberg and Lieberman [21] or Ott [23], to name a few in alphabetical order. As the careful reader may notice, our presentation is, in some points, influenced by their style. There exist, also, many fine review articles, such as, in alphabetical order as well, those by Berry [6], Chirikov [8] and Ford [12] [13].

2 Elementary Tools

2.1 Surface of Section

The evolution of a dynamical Hamiltonian system can be visualized by drawing its trajectories in phase space. Unfortunately this can be done only for one-degree-of-freedom (1-D) systems, whose phase space is two dimensional and can, therefore, be plotted on a surface, using the conjugate variables (co-ordinate, q, vs. momentum, p) as cartesian co-ordinates.[1] Poincaré proposed a method that enables us to plot the evolution of a 2-D Hamiltonian system as well. This is the famous *surface of section technique*.

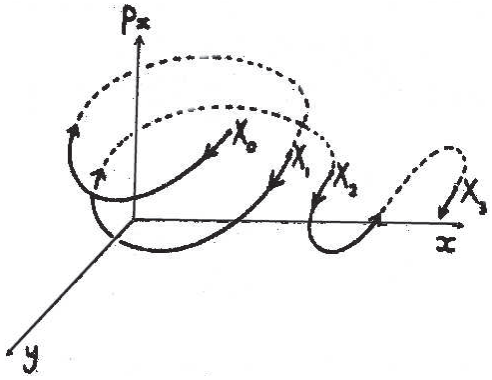

Fig. 1. The consecutive intersection of the trajectory with the $y = 0$ plane lie on a surface of section *(from [6])*

We select one pair of conjugate variables, say (q_1, p_1), and we plot on the corresponding surface q_1p_1 (as in the 1-D case) a point whenever the value of the other co-ordinate, q_2, takes a preselected value, usually $q_2 = 0$. In this way we represent the trajectory on the *Poincaré's surface of section, PSS* q_1p_1 by a *sequence* of points, which are called *consequents* (Fig. 1). In other words we have replaced the evolution of a continuous dynamical system, i.e. a *flow*, with the evolution of a discrete dynamical system, i.e. a *mapping*, with two degrees of freedom less (Fig. 2). In order to make the representation unique and avoid ambiguities, we plot only the consequents corresponding to either positive or negative values of the momentum i.e. either

[1] A standard result of Hamiltonian Mechanics is that the *exterior product* of the unit vectors along the conjugate axes $\hat{p}_i \wedge \hat{q}_j$ is given by the delta function of Kronecker, $\delta_{i,j}$, while $\hat{q}_i \wedge \hat{q}_j = \hat{p}_i \wedge \hat{p}_j = 0$. In this way, following the usual Euclidean metric, the conjugate variables q_ip_i may be considered as orthogonal to each other.

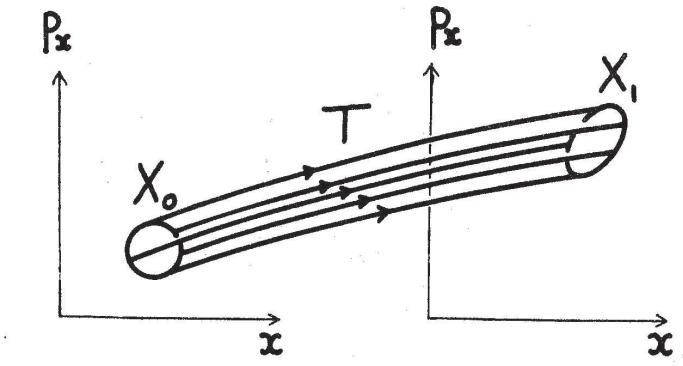

Fig. 2. The surface of section technique maps a region of phase space, X_0, to another region, X_1 *(from [6])*

$$p_2 > 0 \quad \text{or} \quad p_2 < 0. \tag{1}$$

In this way we define a *mapping*

$$(q_i, p_i) \quad \rightarrow \quad (q_{i+1}, p_{i+1}) \tag{2}$$

which is a single-valued representation of the trajectory. In the general case of an arbitrary Hamiltonian system, it is not always easy to select a surface of section, since there is the possibility that some trajectories do not intersect this surface very often, or even at all. If, however, we are dealing with an *integrable* dynamical system with a *compact* phase space, then the selection is rather easy. As we will see in the next section, in this case the variables of choice are *actions* and *angles*, J_1, θ_1 and J_2, θ_2, and we know that the trajectories are helices, winding on nested tori. The "natural" selection for a PSS is then $\theta_2 =$ const. and the consequents lie on smooth curves, named invariant curves (e.g. the lines $J_1 - \theta_1$ in Fig. 3). As we will see in Sect. 4, non-integrable dynamical systems possess trajectories whose intersections with at least a region of the surface of section look scattered (see Sect. 4). It should be noted that a PSS can be defined for dynamical systems with more than two degrees-of-freedom. In these cases, however, it is not easy to get an easy visualization of the corresponding mapping, since it has more than two dimensions.

2.2 Lyapunov Characteristic Numbers

The surface of section technique, used by Hénon and Heiles for the differentiation between ordered and chaotic trajectories, cannot be applied (at least in a straightforward way) to dynamical systems with more than two degrees of freedom. Therefore soon appeared the need for a tool that could be used for

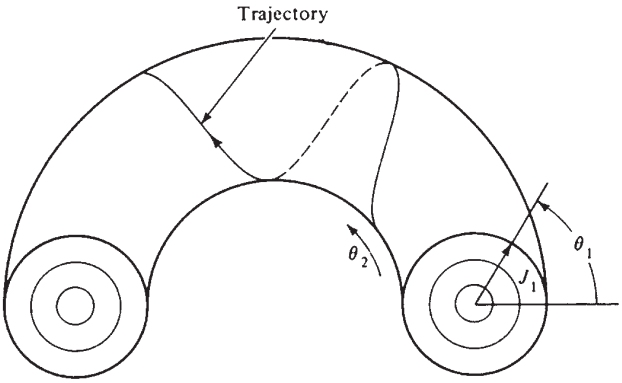

Fig. 3. The two actions of a 2-degrees of freedom dynamical system correspond to the two radii of a torus, $J_{1,2} = (2I_{1,2})^{1/2}$. The two angles correspond to the two angle coordinates on the toroidal surface

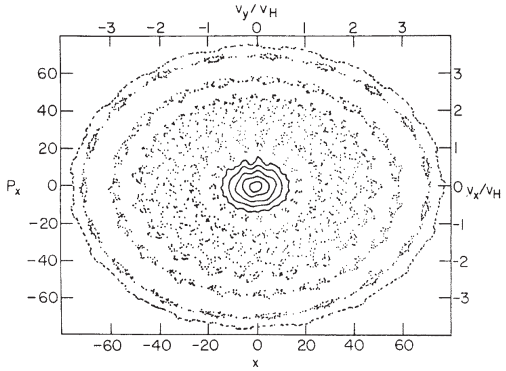

Fig. 4. Typical surface of section of a 1 1/2 D dynamical system (1-D plus time), describing the motion of a charged particle in a magnetic field and an electrostatic wave propagating perpendicular to the field. Note that the consequents near the origin lie on smooth curves, while far from it look scattered, an indication of chaos

the study of dynamical systems with any number of degrees of freedom. The first such tool, proposed already in 1968 by Oseledec [22], is the calculation of the Lyapunov Characteristic Exponents (LCE) or Lyapunov Characteristic Numbers (LCN), which not only enables us to discriminate between ordered and chaotic trajectories, but can as well "measure" the *degree of stochasticity* of a chaotic trajectory.

Hénon and Heiles, in their pioneering work on the numerical study of the stochastic region of a conservative dynamical system, made the remark that "Experience has shown previously that ... in a region occupied by (smooth

invariant) curves the distance (between orbits) increases only slowly, about linearly, but ... in the ergodic region the distance increases rapidly, roughly exponentially." Roughly speaking, the LCNs of a trajectory characterize the mean exponential rate of divergence of other trajectories surrounding it and, therefore, it is the mathematical aspect of the experimental observation by Hénon and Heiles.

In the years that followed, the concept of LCNs was founded on solid mathematical ground by many authors so that the LCNs have become the "industry standard" in assessing the *degree of stochasticity* of a dynamical system. It should be mentioned that the calculation of LCNs has some weak points, related to the fact that they are defined, as we will see, in the limit $t \to \infty$. Therefore finite time numerical calculations can be considered only as "useful approximations" to the *unknown* "real" LCNs of a trajectory. This problem becomes more pronounced in the cases where the values of the LCNs are extremely small. Several authors have proposed other tools, through which one can asses the chaotic or ordered nature of a trajectory in a faster or more efficient way. Examples are the stretching numbers, the fast Lyapunov indicators, the frequency analysis, the methods PSOD, MEGNO etc. For a concise review the reader is referred to . However the ultimate criterion to evaluate the efficiency of such a method is the confirmation of its results by the calculation of the LCNs, so that the understanding of this method is absolutely necessary in the study of dynamical systems. Before proceeding further we should point out that LCNs may be defined for conservative as well as dissipative systems; in the present text, however, we will restrict ourselves to conservative systems.

The procedure for computing the LCNs can be found in any standard textbook. There are basically two methods for their calculation. The first, the "classical" one, is based on the simultaneous numerical integration of a "main" trajectory, described by $\mathbf{x}$, and a "nearby" trajectory, described by $\mathbf{x} + \boldsymbol{\Delta}\mathbf{x}$. Then we calculate the limit

$$\sigma(\mathbf{x}_0, \Delta\mathbf{x}) = \lim_{t\to\infty, \|\Delta x(0)\|\to 0} \left[\frac{1}{t} \ln \frac{\|\Delta\mathbf{x}(t)\|}{\|\Delta\mathbf{x}(0)\|}\right] \tag{3}$$

Equation (3) is a vector equation. It can be shown that the limits of its components exist and are finite. If we select a co-ordinate system, we can write these limits as numbers, which are as many as the number of first order differential equations describing the evolution of the dynamical system. These numbers are the LCNs characterizing the "main" trajectory. In practice we calculate only the largest of them, called *maximal characteristic Lyapunov number*, since it is the one that dominates the evolution of the dynamical system. It should be pointed out that the values of the LCNs are independent of the choice for the metric of phase space used in (3). This is really important, since the phase space does not posses a "natural metric", which therefore would be "preferred" in the calculation of the norms in (3), as the Euclidian metric is "preferred" in configuration space.

The second method is based on the numerical solution of the variational equations of the system, i.e the linearized equations of motion. It has the advantage that we do not have to take the limit $\Delta\mathbf{x}(0) \to 0$ in (3). It also does not have the problem of re-normalization (i.e. the re-definition of $\mathbf{x}_0$, when the separation between the two trajectories reaches the physical dimensions of the dynamical system) neither the problem of migration of each trajectory to a region of phase space with different properties. On the other hand, in many real applications it is difficult to write down the variational form of complex equations of motion. A note of caution should be included here. In systems with $n = 2$ degrees of freedom a positive Maximal Lyapunov Number implies complete chaos . However in systems with $n > 2$ this is not so, as there is always the possibility that one or more degrees of freedom correspond either to Lyapunov numbers equal to zero or very close to it (i.e they are *mildly* chaotic), as it is the case with stable chaos in the Solar system (e.g. see [26]). Since numerically these two cases cannot be differentiated efficiently, one should not consider a dynamical system with $n > 2$ as *completely chaotic* solely by the fact that it has a positive Maximal Lyapunov Number, unless he has computed *all* its Lyapunov exponents.

The algebraic sum of all Lyapunov exponents is related to the evolution of volume in phase space. In a system with dissipation this sum is negative, while in a conservative dynamical system it is exactly equal to zero, due to the conservation of phase space volume. Moreover in any conservative dynamical system at least one of the LCNs is zero because, in the direction along the trajectory, $\Delta\mathbf{x}$ grows only linearly with time. If, in addition, the dynamical system is Hamiltonian (i.e. it is written in canonical co-ordinates, which define a *special* co-ordinate system), then the Lyapunov exponents have a particular symmetry

$$\sigma_i = -\sigma_{2N-i+1} \tag{4}$$

where N is the number of degrees of freedom. It follows that there are at least two Lyapunov exponents equal to zero in the case of a conservative Hamiltonian dynamical system.

LCNs in Maps andi Flows

Besides dynamical systems defined by differential equations, i.e. flows, Lyapunov Characteristic Numbers may be defined for maps as well. For example, several authors calculate LCNs not for the flow itself but for the evolution of consequents on a PSS. Sometimes there is a confusion on the dimensions and the units, in which LCNs are measured in these two cases. Here we show the relationship between them. For the M-dimensional map

$$\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n) \tag{5}$$

we can define σ^{map} through (3) with t replaced by n, which leads to the M Lyapunov exponents ${\sigma_i}^{\text{map}}$. If the Poincaré's map of dimension M is

generated by an autonomous Hamiltonian flow of dimension $M+2$, then the LCNs of the map are proportional to the LCNs of the flow

$$\sigma_i{}^{\text{map}} = \tau\sigma_i, \quad i = 1, ..., M \tag{6}$$

The constant of proportionality, τ, is the mean time between successive intersections of the trajectory with the surface of section. Note that, since two of the LCNs of an autonomous Hamiltonian flow are known to be equal to zero, the number of non-trivial LCNs for a map equals that of the flow, from which the map was derived.

3 Integrable Systems

Since integrable systems are described in detail in all advanced textbooks on Classical Mechanics, this topic is highly developed. We will try, therefore, to explain the salient features without proofs or elaborate calculations, focusing our presentation on *conservative* dynamical systems.

3.1 Constants of Motion

The basic property of integrable *conservative* dynamical systems is that they posses *constants of motion*, in addition to the "energy" (which is by definition constant in a conservative system). Sometimes it is difficult to find explicit expressions of these constants as functions of coordinates and momenta, but their existence can be recognized by a special structure they impose on phase space, which is described as a *foliation in invariant tori*. The definition of a constant of motion, of an invariant torus and of the foliation in invariant tori will be among the topics of the present and the next section.

Conservative integrable systems and their properties are usually discussed in the framework of *Hamiltonian mechanics* and *canonical variables*. We will assume, therefore, that we consider an autonomous dynamical system (i.e. $\partial H/\partial t = 0$) and that we have written its Hamiltonian function, $H(p_i, q_i)$, as well as the equations of motion, in the canonical variables (p_i, q_i). Suppose then that we have discovered a (single valued) function, $F(p_i, q_i)$, with the following property: The value of F does not change with time, t, when its arguments, p_i and q_i, are replaced by any solution of the equations of motion

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{7}$$

Such a function is called a *constant of motion* of the dynamical system.

The constant value of such a function, along a trajectory, imposes the condition

$$\frac{dF(p_i,q_i)}{dt} = 0 = \sum_i \left(\frac{\partial F}{\partial p_i}\frac{dp_i}{dt} + \frac{\partial F}{\partial q_i}\frac{dq_i}{dt} \right) =$$
$$\sum_i \left(\frac{\partial H}{\partial p_i}\frac{\partial F}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial F}{\partial p_i} \right) = [F,H] \tag{8}$$

It should be emphasized that the *Poisson bracket*, $[F,H]$, can be computed for *any* two functions of the canonical variables, not only the Hamiltonian and a constant of motion. However, the necessary and sufficient condition for a function F to be a constant of motion is that the Poisson bracket between the Hamiltonian, H, and F is zero $\forall q_i,\ p_i$.

Now suppose that we have found a set of *independent* constants of motion for the dynamical system in study and let us call them F_1, F_2, etc. According to what has been said, they have to be single-valued functions of the canonical coordinates and momenta and each one should have a vanishing Poisson bracket with the Hamiltonian, in accordance with (8). Moreover being independent implies that we cannot express, say, F_3 as a function of F_1 and F_2. A specific trajectory then lies on the intersection of all the hypersurfaces $F_j(p_i,q_i) = \text{const.}$, when these constants take the values corresponding to the initial conditions of the trajectory.

However, the above three properties (e.g. single-valued, constant and independent) are not sufficient to guarantee the solution of the equations of motion of the corresponding dynamical system *in a straightforward way*. But if we could impose some additional conditions on the constants of motion, we would be able to construct a *special* system of canonical coordinates, the so-called *action-angle variables*, which portray in a very simple and elegant way the special properties of integrable systems.

3.2 Action-Angle Variables

A Hamiltonian dynamical system with n degrees of freedom has a phase space of $2n$ dimensions. Let's assume that we have already found k constants of motion with the above properties, including the energy $H(p_i,q_i) = E$. Any trajectory of this dynamical system is then restricted to a $(2n-k)$- dimensional subspace of the whole phase space. In order to construct a canonical coordinate system using the above constants as level surfaces, these constants should be "in involution", which means that the Poisson bracket of any two functions should vanish, i.e. $[F_i,F_j] = 0\ ,\forall i \neq j$. This imposes some restrictions on the topology of the subspace, on which the motion takes place, and ensures that the constants F_i can, in principle, be used as a set of momenta in some canonical set of variables.

The most favorable case, for the explicit solution of the equations of motion of a conservative dynamical system, is when there exist at least n constants of motion (i.e. as many as there are degrees of freedom) with all the above mentioned four properties. If, moreover, all trajectories lie on

a *connected* subset of phase space which is *compact* (i.e. neither the positions nor the momenta take infinite values), the trajectory is confined to a n-dimensional manifold. Then, by a well-known theorem due to Liouville (e.g. see [3], beginning of Chapter 10), this manifold is isomorphic to an n-dimensional torus, i.e. a "hypersurface" on which the position may be defined by a set of n angles. Each trajectory of the dynamical system lies on such a torus, which is therefore called *invariant*. Since two different trajectories do not intersect in phase space, so do any two invariant tori. Therefore the invariant tori are *nested*, one inside the other, and the phase space is said to be *foliated*. Such a dynamical system is called *integrable*. A few remarks on the notion of integrability are necessary at this point.

- It is generally accepted that a dynamical system is integrable because it has symmetries, which are responsible for the existence of the constants (or *integrals*) of motion (the famous Noether's theorem, e.g. see Arnold [3], p. 88). But in most integrable dynamical systems these symmetries, and the corresponding constants, are not obvious at first sight. This is exactly the case with the Toda lattice (Sect. 3.4), where the symmetry corresponding to the additional integral is not trivial at all.
- Integrability, in the Liouville sense, is sufficient for the solution of the equations of motion by *quadratures*, i.e. as integrals of elementary functions. However it is not a necessary condition. A trivial dynamical system which is not integrable, in the strict Liouville sense, but whose equations of motion are solvable in closed form, is Kepler's problem for non-negative energy values and/or zero angular momentum. The reason is that the available region of phase space is not compact, since in the first case the position and in the second the momentum go to infinity.
- There exist Hamiltonian dynamical systems for which we know as many constants of the motion as the number of degrees of freedom, but we cannot find the action-angle variables in explicit form, because the corresponding integrals cannot be computed in terms of elementary functions (e.g. the non-linear pendulum, Sect. 3.4 or the problem of the two fixed centers, Sect. 3.4). A dynamical system, for which we can find a co-ordinate system in which the equations of motion for each one degree of freedom do not depend on the other degrees, is called *separable* (e.q. see Sect. 3.4).

It is intuitively helpful to realize why, in a system with two degrees of freedom and two constants of motion in involution, the trajectories cannot lie on a 2-sphere (or a surface with the same topology): it is impossible to define on a sphere a coordinate system that does not have, at least, one singular point somewhere. On the surface of a 2-torus, however, there is an obvious system of coordinates without any singularity. It consists of two sets of circles, one winding around the symmetry axis that passes through the "hole" of the torus (i.e. parallel to the "equator" of the torus) and the other one winding parallel to this axis (i.e. the "meridians" of the torus) (see Fig. 5).

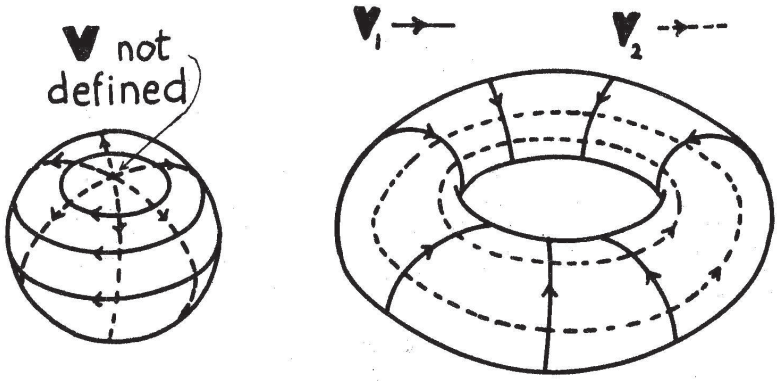

Fig. 5. A co-ordinate system on a sphere has at least one singular point, where one co-ordinate is not defined. This is not so on a torus *(from [6])*

On a n-torus the position is defined by a set of n co-ordinates, $(\theta_1, \theta_2, ..., \theta_n)$ which are named *angle variables*, since each one has a 2π-periodicity. If we select the angle variables as generalized co-ordinates of a new system of canonical co-ordinates of an integrable dynamical system, then by the same above theorem we know that their time derivatives $(\omega_1, \omega_2, ..., \omega_n)$, called *frequencies* of the dynamical system, are constant on each torus. The frequencies ω_i depend on the values of the constants of motion characterizing the particular invariant torus. The conjugate momenta of the angles are called *actions* and are denoted by the symbols $(I_1, I_2, ..., I_n)$. From what was said in the previous paragraph, it becomes evident that the angle variables are linear functions of the time, since their derivatives with respect to time, the frequencies ω_i, are constant on each invariant torus. From the second part of (7) then it follows that the Hamiltonian cannot depend on the angles but only on the actions. Therefore the Hamiltonian in the new coordinates becomes $H^*(I_1, I_2, ..., I_n)$ (note that we have put an asterisk in the new Hamiltonian, because it is a function different from H). This means that, in action-angle variables, Hamilton's equations of motion given by (7) take the trivial form

$$\frac{dI_i}{dt} = -\frac{\partial H^*}{\partial \theta_i} = 0, \quad \frac{d\theta_i}{dt} = \frac{\partial H^*}{\partial I_i} = \omega_i \tag{9}$$

We see that the frequencies, ω_i, are simply the derivatives of the Hamiltonian with respect to the actions, I_i, and that the actions are constants of motion, something that should be expected, since they are functions of the original constants of motion, F_i. Therefore the trajectories become straight lines on each (I_i, θ_i) plane. We recall that the principal property of a Hamiltonian system is that the *action integral* of $\mathbf{p}\ d\mathbf{q}$ is invariant under a canonical transformation. This means that $\sum p_i dq_i = \sum I_i d\theta_i$. From the latter condition we can explicitly calculate the I_is, in the following way. We assume that

all other angles are kept constant except for θ_i, which varies from 0 to 2π. In this way the trajectory is a closed loop, C_i, which can be expressed either in the original coordinates, (p_i, q_i), or in the action-angle variables, (I_i, θ_i). Since the action integrals in the two coordinate systems have to be equal (as stated above), we find immediately the formula for the calculation of the actions

$$\int_{C_i} I_i d\theta_i = \int_{C_i} p_i dq_i \;\; \Rightarrow \;\; I_i = \frac{1}{2\pi} \int_{C_i} p_i dq_i \tag{10}$$

We emphasize that, by definition, *action-angle variables* can be constructed only for integrable Hamiltonian systems.

In order to get a geometrical picture of the action variables, we start with the one-degree-of-freedom (1-D) case. We recall that, since the action integral is invariant under a canonical transformation, so is its differential, $dpdq$, which expresses the elementary area on the surface (pq). Therefore the elementary area in action-angle variables is written as $dE = dId\theta$. On the other hand it is known that conjugate canonical co-ordinates may be considered as being orthogonal (see footnote 1). Then the elementary area on the surface $(I\theta)$, in polar coordinates, is expressed as $dE = rdrd\theta = d(\frac{r^2}{2})d\theta$. By comparing the last two relations we find that $r = \sqrt{2I}$. In a similar way, in a two-degrees-of-freedom (2-D) integrable Hamiltonian system, the two action variables are, up to a factor 1/2, the squares of the two (constant) radii of a torus: the first is the radius of a "meridian" (a section parallel to the axis of symmetry) and the second is the radius of the "equator", which is orthogonal to the meridian (Fig. 3). Action-angle variables are of particular interest in Hamiltonian Mechanics, not only because they are the variables of choice for integrable systems, but because they are used, as well, in the study of chaotic Hamiltonian systems and the calculation of the, so-called, resonance overlap criterion, which gives an estimate on when a perturbed integrable Hamiltonian system shows prominent chaotic behavior (see Sect. 5).

3.3 Periodic, Quasi-Periodic and Ergodic Motion

If we are able to find the Hamiltonian H^*, then we have arrived, as well, at the complete solution of the equations of motion. This is so because the general solution of (9) is given by $\theta_i = \omega_i t + \phi_i$. In this expression the frequencies, ω_i, are functions of the actions I_i only and, together with the phase angles, ϕ_i, are the integration constants (they correspond to the initial conditions!). There are many possible sets of angle-type variables in any special problem (e.g. the true or the eccentric anomaly in Kepler's problem). They are, however, not linear functions of time in general, so that they are not *the* angle variables of an *action-angle* variables set. In contrast, the mean anomaly in Kepler's problem is a linear function of time and, therefore, it can be considered as the angle of a corresponding action-angle pair. The difficulty in integrating the

equations of motion consists exactly in finding the angle-type variables whose time-dependence is linear. But if we manage to do so, then the geometrical form of the trajectories is remarkably simple: they are helices winding on the surface of a torus. In particular, if all n frequencies are *independent in the rationals*, i.e. if the *resonance relation*

$$m_1\omega_1 + m_2\omega_2 + ... + m_n\omega_n = 0$$

is not satisfied for *any* set of the integers $m_1, m_2, ..., m_n$, then the helix is open and covers densely (and, hence, ergodically, see Sect. 4.1) the surface of the (n-dimensional) torus. If only one frequency is independent (i.e. if $n-1$ resonance relations are satisfied), the trajectory is closed and, thus, periodic (Fig. 6). In between lie all other cases, in which the number of independent frequencies is between 1 and n and the trajectory covers densely a submanifold of the n-torus. The motion in these intermediate cases is said to be *conditionally periodic* or *quasi-periodic*.

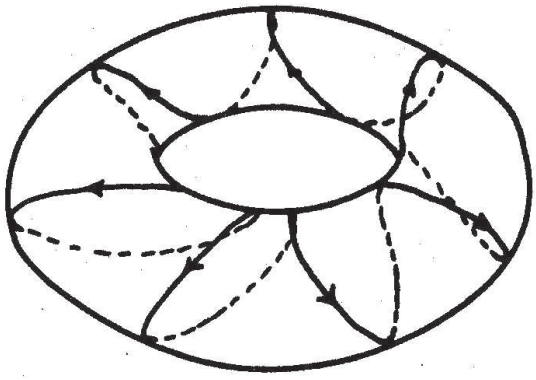

Fig. 6. In a 2-D system with only one independent frequency, the corresponding trajectories are periodic *(from [6])*

3.4 Examples of Integrable Systems

The Harmonic Oscillator (or Linear Pendulum)

The simplest one-dimensional (and thus integrable by quadratures) system is the harmonic oscillator, described by the Hamiltonian

$$H = \frac{1}{2}p^2 + \frac{1}{2}\omega^2 q^2 = h \tag{11}$$

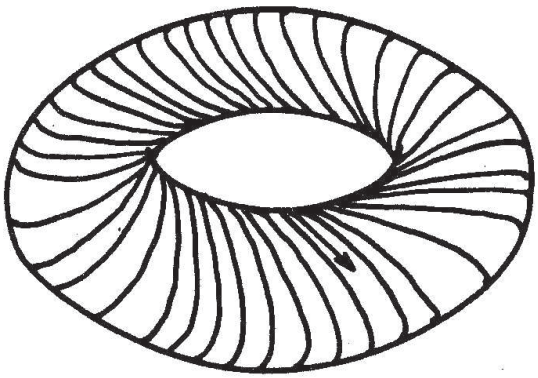

Fig. 7. In a 2-D system with two independent frequencies, the corresponding trajectories are conditionally periodic *(from [6])*

We will try to write this Hamiltonian in action-angle variables, using the rule given by (10). We solve the Hamiltonian for p,

$$p = p(q, h) = \sqrt{2h - \omega^2 q^2} \tag{12}$$

substitute it in (10) and calculate the integral. The result is

$$I = \frac{h}{\omega} \tag{13}$$

which, of course, could be found directly by substituting the well know solutions of the equation of motion

$$x = \cos(\omega t), \quad p = \omega \sin(\omega t) \tag{14}$$

in (11). The angle θ can be found by simple inspection: Since the argument in the trigonometric functions of the solution is linear in time, it is exactly the angle sought, i.e. $\theta = \omega t$. The action, I, as a function of the value of the Hamiltonian, h, can be found by squaring the solutions, $p(t)$ and $q(t)$ and adding them. The result is

$$H = \omega I \tag{15}$$

from which the action variable is given immediately as

$$I = H/\omega \tag{16}$$

The Non-Linear Pendulum

The harmonic oscillator is a trivial application of the method to calculate action-angle variables for a one-dimensional dynamical system. A more interesting case is the (non-linear) pendulum, described by the equation of

motion:

$$\ddot{q} = -\sin(q) \tag{17}$$

which is derived from the Hamiltonian

$$H = \frac{1}{2}p^2 - \cos(q) \tag{18}$$

where for convenience and without loss of generality we have set ω, i.e. the frequency of low-amplitude oscillations, equal to 1 (in dimensionless units).

Equation (17) has two families of solutions, one for $h > 1$ and the other for $-1 < h < 1$. Since the trajectories of the family $h > 1$ are unbounded, a case of limited interest in Physics and Astronomy, we focus on the case $-1 < h < 1$. We note, however, that the energy $h_{\rm crit} = 1$ corresponds to a particular set of trajectories of this system, observed in practically all Hamiltonian dynamical systems, whose union is named "separatrix". A separatrix of an integrable dynamical system is a self-intersecting curve in phase space, composed of three trajectories. One trajectory consists of only the point of intersection and it is essentially a point of equilibrium. The two other trajectories start in the neighborhood of this point and return to it after an infinite time interval. Seperatrices play an important role in the appearance of chaos in non-integrable dynamical systems, which are perturbation of integrable ones, since the first signs of stochastic behavior appear in their neighborhood as the perturbation increases.

By expressing p as a function of h through (18) and then by substituting this function to (10), we arrive at

$$I = \frac{1}{2\pi}\oint (H + 2\cos(q))^{1/2} dq \tag{19}$$

The integral in (19) is non-trivial, but it can be found in tables of integrals. The solution is well known, and we give it without derivation

$$I = I(h) = \frac{8}{\pi}\left[E(\pi/2;\kappa) - (1-\kappa^2)F(\pi/2;\kappa)\right] \tag{20}$$

where F and E are elliptic integrals of the first and second kind, respectively, and the parameter κ is related to the energy, h, through the relation $\kappa^2 = \frac{1+h}{2}$. In order to write the Hamiltonian, (18), as a function of the new action, we have to invert (20). However this is highly non-trivial and the result is expressed in terms of elliptic integrals as well, something that does not allow us to obtain an easy "insight" in the properties of the solutions. We will see in Sect. 3.4 that this problem appears in another integrable dynamical Hamiltonian system, the problem of the two attracting fixed centers, and this is the reason why some unexpected properties of this latter system, which was solved a long time ago, were not understood until recently.

For reasons of completeness we note here that, although the complete solution of the problem, which includes the calculation of the Hamiltonian

$H(I)$ as well as the expression of the "old" variable as a function of time, $q(t)$, is not very illuminating, the derivatives of these functions are relatively simple:

$$\omega(I) = \frac{dH}{dI} = \frac{\pi}{2F(\pi/2;\kappa)} \quad \dot{q} = \frac{dq}{dt} = 2\kappa \mathrm{cn}(t;\kappa) \tag{21}$$

where cn is the elliptic cosine and $\kappa = \sqrt{\frac{1+h}{2}}$.

The Toda Lattice

The above theory of integrable systems will now be illustrated with one example of a 2-D dynamical system of a form slightly more complicated kind than the ones ordinarily found in textbooks, the Toda lattice. The "complication" arises from the fact that the frequencies ω_i, in this example, depend on the actions I_i in a non-trivial manner. This means that the matrix of derivatives of ω_i with respect to I_i (i.e. the *Hessian* of the Hamiltonian) is not singular (in other words it is different from zero; the corresponding dynamical system is said to be *non-degenerate*). This condition is not satisfied in the usual textbook examples, such as the Kepler's problem or the Hénon-Heiles system (or any other system of two harmonic oscillators coupled through a perturbing term), which have a Hessian equal to zero and are, therefore, *degenerate*. For example the Hénon-Heiles Hamiltonian can be written in the form

$$H = \frac{1}{2}(p_x^2+p_y^2)+\frac{1}{2}(x^2+y^2)+\varepsilon\left(x^2y-\frac{1}{3}y^3\right) = \omega_1 I_1+\omega_2 I_2+O(\varepsilon) = H_0+O(\varepsilon) \tag{22}$$

so that H_0 is integrable and its second derivatives, with respect to the actions, are zero. It should be pointed out that the original KAM (Kolmogorov-Arnold-Moser) theorem, which guarantees the existence of ordered trajectories of positive measure in any perturbed integrable Hamiltonian system, does not apply to degenerate systems [3]. Therefore, Kepler's problem, as well as the zero-th order term of the Hénon-Heiles system, although integrable, present special difficulties when perturbations are applied.

The Toda lattice (or chain) consists of n particles, which move along a line and are coupled by non-linear springs. If by p_i and q_i we denote, respectively, the momentum and the position of the i-th particle, then the Hamiltonian describing the Toda lattice is written as

$$H = \frac{1}{2m}\sum_1^n p_i^2 + V_0 \sum_1^{[n]} \exp\left(\frac{q_i - q_{i+1}}{a}\right) \tag{23}$$

The symbol $[n]$ in the second summation may take two values: for an open lattice it is $[n] = n-1$, while for a periodic (or circular) lattice it becomes $[n] = n$, with the identification $q_{n+1} = q_1$.

We observe that there appear three physical parameters in the above Hamiltonian, the *mass*, m, of each particle, the *strength* of the spring, denoted by the potential V_0, and the *scale*, a, of the (short range) force of the springs. By taking these parameters as new units (i.e. by setting $m = V_0 = a = 1$), the Hamiltonian is written in dimensionless units as a function of generalized co-ordinates and momenta only.

The Hamiltonian in (23) does not present any explicit symmetry, so it seems reasonable to believe that it does not possess any further constants of motion, besides the energy and the momentum. Therefore it should be non-integrable. The numerical calculation of surfaces of section for the periodic lattice with $n = 3$, however, showed exactly the opposite. This lattice is a three degrees of freedom Hamiltonian system that has two integrals of motion: the total momentum and the total energy. Using the momentum integral and by a suitable change of variables, one can write down the problem as a two degrees of freedom dynamical system, described by the Hamiltonian

$$H = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{24}\left[e^{2y+2\sqrt{3}x} + e^{2y-2\sqrt{3}x} + e^{-4y}\right] - \frac{1}{8} \tag{24}$$

This dynamical system has only one obvious integral of motion, the Hamiltonian itself. Moreover, if we expand the exponentials to third order in the co-ordinates, we get the famous Hénon-Heiles Hamiltonian, which is well known to be non-integrable. However the numerical results indicated that the Hamiltonian in (24) is integrable! The "second" integral was constructed by Hénon and, almost simultaneously, Flashka discovered that the Toda lattice is integrable for any value of n and that the integrals of motion, in the original variables, are polynomials of order $1, 2, ..., n$ in the momenta. This is very interesting, since in the previously known integrable systems the additional constants of motion were at most quadratic in the momenta. It should be noted, however, that the action-angle variables of a Toda lattice cannot be written in the form of elementary functions of the the energy and the other constants of motion.

The Kepler's Problem

It would be interesting, also, to see how we can introduce action-angle variables to a well known integrable dynamical system, namely the motion of a *test particle* in the gravitational field of a *fixed* spherically symmetric body. Hamilton's function is written, in polar co-ordinates, as

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2}\right) - \frac{GMm}{r} \tag{25}$$

where by m we denote the particle's mass, by M the central body's mass and G is the gravitational constant. The corresponding dynamical system is 2-D and has two independent integrals of motion, namely the energy, h,

and the angular momentum, p_θ. The compactness of the available phase space region is guaranteed by assuming that the value of the Hamiltonian, h, is negative, (in order to have bounded motion in configuration space), and that the angular momentum is different from zero (in order to have bounded motion in momentum space, since otherwise the particle is on a collision course with the central body!). Therefore this dynamical system is integrable in the Liouville sense and can be written in action-angle variables.

The angular momentum, p_θ, can be taken as the "new" action, I_θ, since θ is an ignorable co-ordinate, as

$$\frac{\partial H}{\partial \theta} = 0 \tag{26}$$

and, therefore, it can be considered to be a (trivially) linear function of time. The question is how one can find the second action. The answer is: follow the rule given by (10):

$$2\pi I_r = \int_C p_r dr \tag{27}$$

From (25) we express p_r as a function of r and we finally have

$$I_r = \frac{1}{2\pi}\int_C \left[2m\left(h + \frac{GMm}{r}\right) - \frac{I_\theta^2}{r^2}\right]^{1/2} dr \tag{28}$$

It is not a straightforward problem to calculate this integral, but we can find its solution in tables of integrals (or in Goldstein!). It is

$$I_r = -I_\theta + \frac{GMm}{2}\left(\frac{2m}{-h}\right)^{1/2} \tag{29}$$

By solving this relation for h we get the "new" Hamiltonian function, H^*, (expressed now in the actions I_r and I_θ)

$$H^* = -\frac{(GMm)^2 m}{2(I_r + I_\theta)^2} = h \tag{30}$$

From the last relation it is obvious that the two frequencies of this problem, ω_r and ω_θ, are not independent, since

$$\omega_r = \frac{\partial H}{\partial I_r} = \frac{(GMm)^2 m}{(I_r + I_\theta)^3} = \omega_\theta = \frac{\partial H}{\partial I_\theta} \tag{31}$$

According to what we have said in the previous section, Sect. 3.4, we understand that the problem of Kepler is a degenerate dynamical system. As a consequence of the equality of the two frequencies, all trajectories are winding periodically on the torus defined by the two actions and are, therefore, periodic in time. This is actually the well known result of Newton, that all

bounded trajectories in Kepler's problem are ellipses. For reasons that will become clear in the next section, people working on Celestial Mechanics prefer to use another set of actions in Kepler's problem, the so-called *Delaunay actions* L and G, which are related to I_r and I_θ through the relations $L = I_r + I_\theta$ and $G = I_\theta$. In these variables the Hamiltonian becomes simply

$$H^* = -\frac{(GMm)^2 m}{2L^2} = h \tag{32}$$

The Two Fixed-Centers Problem (2FCP)

The motion of a point mass, moving in the gravitational field of two fixed attracting centers, is a problem first posed by Euler in the 18th century, as an intermediate step towards the solution of the famous three-body problem. Euler himself was able to integrate the equations of motion for the two-dimensional (2-D) case, i.e. the case where the point mass moves on a plane containing the two attracting centers. Almost a century later Jacobi showed that the corresponding potential of the full 3-D case is separable in prolate spheroidal coordinates. Another century later Erikson and Hill found, in explicit form, the third integral of motion of the full three-dimensional (3-D) case (besides the other two "classical" ones, i.e. the total energy and the angular momentum about the axis passing through the two centers). Since then, the problem has been considered as a non-exciting example of a separable potential and it is included, as such, in many textbooks of Theoretical Mechanics.

Although the 2FCP is a separable dynamical system, the qualitative behavior of its solutions was, up to now, not very well understood, probably due to the fact that, as we have mentioned already in Sect. 3.4, the solutions are expressed in the form of elliptic functions.[2] Here we focus our interest on the (simpler to study) 2-D case of the problem, where the trajectory of the third body lies on a plane and we assume, without loss of generality, that the third body has unit mass. Then the Hamiltonian of the dynamical system is written, in cartesian co-ordinates, x-y, as

$$H(x, y, p_x, p_y) = \frac{p_x^2}{2} + \frac{p_y^2}{2} - \frac{\alpha_1}{r_1} - \frac{\alpha_2}{r_2} \equiv E \tag{33}$$

where $\alpha_1 = 2\mu$, $\alpha_2 = 2(1-\mu)$ ($\mu \in [0, 1]$ is the mass parameter), $r_1 = \sqrt{(x+1)^2 + y^2}$ and $r_2 = \sqrt{(x-1)^2 + y^2}$. We note that, in the variables used in (33), the distance between the two centers is equal to $\alpha_1 + \alpha_2 = 2 = -\alpha$. For later use we define also the *asymmetry mass parameter*, β, which is equal to $\beta = \alpha_1 - \alpha_2 = 4\mu - 2$. We note however that in the numerical examples

[2] For an extensive presentation of the history of the problem and its properties, see [27] and references therein.

we restrict our attention to the case of equal masses, so that $\mu = 0.5$ and, therefore, $\beta = 0$.

Following Euler and Jacobi, we write the above Hamiltonian in elliptic-hyperbolic coordinates, through the canonical (point-)transformation:

$$\xi = \frac{r_1 + r_2}{2}, \quad \eta = \frac{r_1 - r_2}{2} \tag{34}$$

Then the "old" variables, as functions of the "new", are given by

$$x = \xi\eta \tag{35}$$

$$y = (\operatorname{sign} y)\sqrt{(\xi^2 - 1)(1 - \eta^2)} \tag{36}$$

$$p_x = \frac{\eta(\xi^2 - 1)p_\xi + \xi(1 - \eta^2)p_\eta}{\xi^2 - \eta^2} \tag{37}$$

$$p_y = (\operatorname{sign} y)\frac{\sqrt{(\xi^2 - 1)(1 - \eta^2)}(\xi p_\xi - \eta p_\eta)}{\xi^2 - \eta^2} = \\ = \frac{y}{\xi^2 - \eta^2}(\xi p_\xi - \eta p_\eta) \tag{38}$$

The Hamiltonian H, in the new variables $(\xi, \eta, p_\xi, p_\eta)$, becomes

$$H = \frac{1}{\xi^2 - \eta^2}\left[\frac{1}{2}(\xi^2 - 1)^2 p_\xi^2 + \alpha\xi + \frac{1}{2}(1 - \eta^2)^2 p_\eta^2 + \beta\eta\right] \tag{39}$$

Finally we change the time scale, by multiplying the Hamiltonian function by $\xi^2 - \eta^2$, noting that this quantity is positive everywhere except when the moving body collides with one of the two centers, in which case it is equal to zero. At the same time we switch to the extended phase space, where the additional co-ordinate is time, t, and the corresponding momentum, p_t, is equal to $-E$, where by E we denote the numerical value of the Hamiltonian function, (39). The new Hamiltonian, $K(\xi, \eta, \tau, p_\xi, p_\eta, p_\tau)$,

$$K = \frac{1}{2}(\xi^2 - 1)^2 p_\xi^2 + \alpha\xi + p_t\xi^2 + \frac{1}{2}(1 - \eta^2)^2 p_\eta^2 + \beta\eta - p_t\eta^2 \tag{40}$$

has in the extended phase space a numerical value equal to zero.

We observe that the Hamiltonian is the sum of two parts, K_ξ and K_η, the first depending only on ξ and p_ξ and the second only on η and p_η. Since the value of the Hamiltonian is by definition zero, the two parts should have opposite values, and satisfy the relation

$$K_\xi \equiv -K_\eta \equiv \gamma \tag{41}$$

In this way we have separated the variables and, at the same time, we have found the second integral of motion

$$G = -\frac{1}{\xi^2 - \eta^2}(\eta^2 K_\xi + \xi^2 K_\eta) \tag{42}$$

However, although we have separated the variables in the Hamiltonian (40), we have not written it in action-angle variables. The solutions can be found by quadratures, but they are not computable in elementary functions. Therefore some uncommon properties of the solutions were not known until the work of Varvoglis et al. [27]. One such property is that the quasi-periodic orbits of this dynamical system approach arbitrarily close the two attracting centers and, if the centers are bodies with physical dimensions, the orbits collide with one of them in a finite time.

4 Chaotic Systems

Since, according to the reasoning in Sect. 3.2, the presence of symmetries implies integrability and, hence, regular behavior, it is obvious that their absence is equivalent to non-integrability. In this sense most dynamical systems are non-integrable and, therefore, chaotic. However this does not mean that we cannot get useful information from the study of a chaotic system. The trick is to change our strategy, looking for a *statistical description* of a chaotic dynamical system, instead of a complete solution of the motion.

A statistical description of a dynamical system usually implies that we try to follow the time evolution of certain *average* quantities, instead of following specific trajectories corresponding to given initial conditions. This procedure is not new to physicists, since it is the essence of Statistical Mechanics. The methodology followed in seeking a statistical description of a dynamical system, which is discussed in the present section, depends on a number of rather involved mathematical concepts, such as ergodicity, mixing and K-entropy. We will try to discuss these concepts, as well as their role in such a statistical description, in a rather heuristic, non-rigorous way.

4.1 Definitions and Basic Concepts

Ergodicity

The concept of ergodicity was introduced in Physics by Ludwig Boltzmann, in his attempt to prove the second law of Thermodynamics, i.e that the entropy of a closed system is a non-decreasing function. According to Boltzmann, an ergodic system is the mathematical model of a gas: it has the property that any initial fluctuations of the density or the temperature are quickly "wiped out" and the system, after a while, becomes "homogeneous". According to Boltzmann, a dynamical system is ergodic if the time average of any smooth (i.e. infinitely differentiable) function of the canonical coordinates is equal to the space average. In order to understand this definition, we should first give exact definitions for the notions of the time and space average of a function in the phase space of a N-degrees of freedom Hamiltonian system. For any smooth function $f(p_i, q_i)$ of the generalized coordinates, $q_i = q_1, ..., q_n$, and

momenta $p_i = p_1, ..., p_n$, we can define the *time average* $\langle f(p,q)\rangle_t$ through the relation

$$\langle f(p,q)\rangle_t \;=\; \lim_{T\to\infty}\frac{1}{T}\int_t^{t+T} f(p(t'),q(t'))dt' \tag{43}$$

For the same function we can define also the *space average*, $\langle f(p,q)\rangle_\Gamma$, through the relation

$$\langle f(p,q)\rangle_\Gamma \;=\; \int_\Gamma f(p,q)d\Gamma \tag{44}$$

In (44) the domain of integration, Γ, is the phase space and the differential, dV, is given by $d\Gamma = dp_1dp_2...dp_ndq_1dq_2...dq_n$.

If the two mean values (temporal and spatial) are equal and the equality does not depend on the choice of the time t, we say that the motion is ergodic. As we will se, this definition is not directly applicable to Hamiltonian systems. We will, therefore, introduce a more mathematically oriented definition.
Let $\mathbf{x}$ be a vector in phase space. Then it can be shown that for almost all $\mathbf{x}$:

- $\langle f(\mathbf{x})\rangle_t$ exists;
- $\langle f(\mathbf{x})\rangle_t$ is an invariant function, i.e. it is independent of the initial conditions on a given orbit,

$$\langle f(T^n\mathbf{x})\rangle_t \;=\; \langle f(\mathbf{x})\rangle_t\,,\ \forall n; \tag{45}$$

- the space-mean value of $\langle f(\mathbf{x})\rangle_t$ is equal to the space-mean value of $f(\mathbf{x})$. An equivalent to (44) way of defining the space average of $f(\mathbf{x})$ is

$$\langle f(\mathbf{x})\rangle_{\mathbf{x}} \;=\; \int_M f(\mathbf{x})d\mu \tag{46}$$

In (46) the manifold M, having dimension M, is the phase space of the system and μ is an invariant measure, given by $d\mu = P(\mathbf{x})d^M\mathbf{x}$, where $P(\mathbf{x})$ is the probability density function. Thus $d\mu$ gives the probability that the value of $\mathbf{x}$ is in the "interval" between $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$. In the special case of a Hamiltonian system, where $\mathbf{x}$ denotes the set of canonical variables, $P(\mathbf{x}) = 1$, due to Liouville's theorem on the conservation of phase space volume along a trajectory.

A dynamical system is defined, according to the ideas of Boltzmann, as being ergodic if

$$\langle f(\mathbf{x})\rangle_{\mathbf{x}} \;=\; \langle f(\mathbf{x})\rangle_t \tag{47}$$

for *almost all* $\mathbf{x}$ (this means that $\mathbf{x}$ belongs to the manifold M minus any subset A of zero measure). From this definition it is clear that the time mean of an ergodic system does not depend on $\mathbf{x}$. Since the observable function,

f, can be chosen arbitrarily,[3] an ergodic system has the property that every trajectory passes arbitrarily close to any point in phase space, infinitely many times. It is then evident that a system possessing invariant tori, in addition to chaotic regions, is not ergodic; we say that it is *decomposable*. For example, any *non-degenerate* perturbed integrable system, to which the famous KAM theorem may be applied, is decomposable, since this theorem guarantees that, for *some* range of the perturbation parameter, invariant tori do exist. Note that the class of decomposable systems is more general that the one discussed above, since this class contains perturbed *degenerate* dynamical systems as well (where the KAM theorem cannot be directly applied), such as the Hénon-Heiles system. A delicate point in deciding on the ergodicity or not of a system is the (sub)space where the motion is considered. An autonomous Hamiltonian system does not posses an ergodic set of trajectories on the *entire* phase space, because the energy is an integral of motion and, therefore, a trajectory cannot visit *all* phase space. However a set of trajectories may be ergodic on a constant energy surface, i.e. a submanifold of M. In particular, an integrable system with n-degrees of freedom is ergodic on any non-resonant $n-$torus (on which, remember, all n frequencies are independent). Then the time spent by a trajectory in a subset of the surface of the torus is proportional to the measure of this subset. If other integrals, besides the energy, do exist, then the system can be ergodic only on a subspace of the phase space, which consists of the intersection of the level surfaces of all these integrals.

In order to clarify the above definitions, we use the example of translations on a torus, given by the mapping

$$T(\theta_1, \theta_2) = (\theta_1 + \omega_1, \theta_2 + \omega_2) \ (\text{mod } 2\pi), \tag{48}$$

If the frequency ratio $\alpha = \frac{\omega_1}{\omega_2}$ is irrational, the orbit is everywhere dense on the torus and, therefore, its density can be given by an invariant function $\langle f(\mathbf{x})\rangle_{\mathbf{x}}$. It can be proven that $\langle f(\mathbf{x})\rangle_{\mathbf{x}}$ is a constant independent of the initial conditions, a fact that implies ergodicity. Physically, (48) corresponds to the twist mapping on the surface of section $\theta_1 - I_1$ of a two-degrees of freedom (2-D) Hamiltonian system. Omitting the subscript 1, this mapping is:

$$\begin{aligned} J_{n+1} &= J_n \\ \theta_{n+1} &= \theta_n + 2\pi\alpha \end{aligned} \tag{49}$$

If α is rational, all trajectories of this mapping are periodic. Since a periodic trajectory consists of a finite number of points on a circle and, therefore, it is not everywhere dense in the available region of phase space, this system cannot be ergodic. On the other hand for α irrational the entire circle is covered by the orbit, and $\langle f(\mathbf{x})\rangle_{\mathbf{x}} \rightarrow \langle f(\mathbf{x})\rangle_t$, which is a property implying ergodicity.

[3] E.g. it might be one of the generalized co-ordinates or momenta.

Now, the mapping in (49) corresponds to the motion described by an integrable 2-D Hamiltonian system on a given KAM torus, where θ_1 and θ_2 are the angle variables. Thus, in a restricted sense, we can say that the motion is ergodic on any irrational torus, while non-ergodic within the entire phase space. From the above example of regular quasi-periodic motion, it is clear that an ergodic dynamical system is not necessarily chaotic. On the other hand the "careful" definition of the subspace where the motion is considered lets us hope that a chaotic region, such as a separatrix layer (defined by the invariant curves that bound it) is also ergodic. However, this may be of somewhat limited usefulness, as a separatrix layer is not an open set, since it contains (according to the KAM theorem) surviving invariant tori of finite measure.

Mixing

The concept of mixing is very simple. Following Arnold, if we want to prepare a cocktail named *rum-cola*, we put in a shaker one part of rum and four parts of cola. This setting represents the initial distribution of the "incompressible fluid" in phase space. If we then stir or shake the liquid repeatedly, we expect that, after the fluid has been stirred sufficiently often (n stirrings with $n \to \infty$), every part of the shaker, however small, will contain "approximately" 20% rum and 80% cola. In a mathematically rigorous way we say that a dynamical system is mixing if

$$\lim_{n\to\infty} \mu\left[\phi_n\left(A \bigcap B\right)\right] = \mu(A)\mu(B) \tag{50}$$

A mixing system approaches an equilibrium; $\langle f(\mathbf{x})\rangle_{\mathbf{x}} \to \langle f(\mathbf{x})\rangle_t$, as $t \to \infty$.

Mixing is a notion more subtle than ergodicity. It was introduced into statistical physics already by Gibbs, but it found its way into non-linear dynamics only relatively recently. The definition given above can be understood in more physical terms if we introduce the notion of the autocorrelation function. Since there are more than one nomenclature conventions in the corresponding literature, we will try to give detailed definitions on the concepts introduced here. Let X be any co-ordinate or momentum (or a function of them) of a dynamical system. Then we can define a function of X, which is called *covariance* function of X, and is represented, here, by the symbol R. The covariance function can be defined for flows as well as for mappings. The definition for flows is, as a rule, used in theoretical calculations, while the definition for mappings is used in numerical exploration of dynamical systems, in which the value of any variable is recorded at regular intervals of time. If the dynamical system is ergodic, then, according to (43) and (44) the temporal mean value of X, $\mu = \langle X\rangle_t$, exists and is independent of the selection of t_0. Then the covariance function of X in the discrete case, $R(n)$, is defined through the relation

$$R(n) = \langle [(X_j - \mu)(X_{j+n} - \mu)] \rangle \tag{51}$$

and is independent of j. Note that $R(0) = Var(X) = \sigma_X^2$, where by $Var(X)$ we denote the variance of X and by σ_X its standard deviation. Note also that the function $R(n)$ is even, i.e.

$$R(-n) = \langle [(X_j - \mu)(X_{j-n} - \mu)] \rangle = \langle [(X_{j-n} - \mu)(X_j - \mu)] \rangle = R(n) \tag{52}$$

so that $R(n) = R(-n) = R(|n|)$. Then the function $C(n)$, defined through the relation

$$C(n) = \frac{R(n)}{{\sigma_X}^2} \tag{53}$$

is called the *autocorrelation function* of X. Note that $C(0) = 1$ and that $-1 \leq C(n) \leq 1$. In the continuous case we are led to exactly analogous relations, only that the discrete time n is replaced by the continuous time, t. It should be noted that the above definitions can be generalized by introducing in (51) and (52) *two* functions, X and Y, instead of one, in which case the corresponding function, analogous to (53), is named *cross-correlation function*. Also the definition can be generalized for vector functions as well. However these generalizations are outside the scope of the present article.

The autocorrelation function is closely related to the concept of *mixing*. We say that an ergodic dynamical system is mixing if the autocorrelation function of its co-ordinates and momenta satisfies the relation

$$\lim_{\tau \to \infty} C(\tau) = 0 \tag{54}$$

which is equivalent to (50) and is well known in physics as the condition of decay of *time correlations* (see e.g. [28]). The process of mixing may be visualized by considering the time evolution of a droplet of "phase-space liquid". In the case of a mixing motion its evolution in phase space is extremely complicated (Fig. 8). The boundary of the droplet quickly assumes an irregular shape with long protruding "filaments" and deep "fjords". As time goes on, the shape of the boundary gets more and more complicated, and the filaments extend to ever new regions of phase space. By virtue of Liouville's theorem, the volume of the droplet is conserved, and the phase space is gradually being filled by the liquid, owing to the stretching and thinning of the filaments of the droplet. It is no surprise that the ergodicity conditions (43–45) automatically follow from the condition of mixing, (50), since the equivalent definition of mixing, (54), was derived under exactly this assumption.

The difference of a "simply" ergodic motion from a "mixing" ergodic motion is illustrated in Fig. (9). In the case of "simple" ergodicity (i.e. non-mixing), a trajectory fills the phase space consecutively in a regular manner, e.g. like a two-dimensional Lissajous curve on the $x - y$ plane when the ratio of the $x-$ and $y-$ frequencies is irrational. Mixing results in a quite different scheme of filling the phase space. First the system covers the whole phase

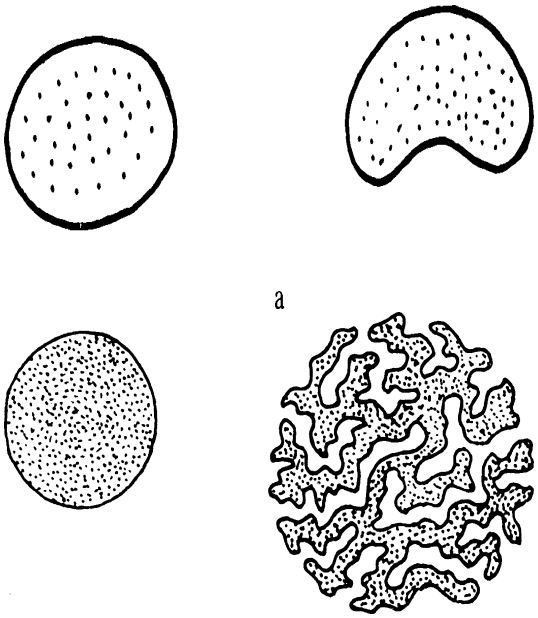

Fig. 8. As time goes by, a droplet of initial conditions of a mixing dynamical system spreads in phase space in a way that depends on whether the corresponding dynamical system is **(a)** non-mixing and **(b)** mixing *(from [28])*

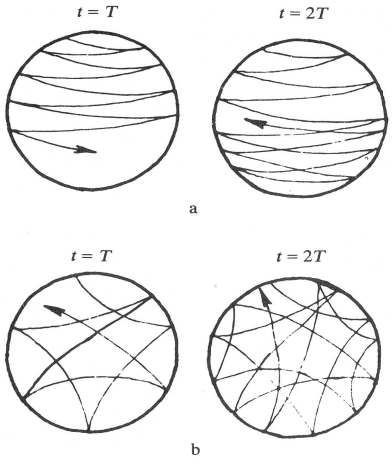

Fig. 9. The difference between **(a)** purely ergodic motion and **(b)** mixing motion *(from [28])*

space nearly uniformly over a certain time interval T. In a time $2T$ this process is approximately repeated in such a way, so that the size of the cells of this network becomes approximately twice as small, and so on.

The differences between mixing and non-mixing ergodic systems may be understood in even another way, which sometimes is more intuitive to a physicist. Instead of studying the evolution of the system in the *time domain*, one might switch to the *frequency domain*. The *spectral density function*, or simply *spectrum*, $f(\nu)$, of any dynamical variable of a system is the counterpart (in the frequency domain) of the *covariance function*, $R(n)$. In other words, $f(\nu)$ is the Fourier transform of $R(n)$ and vice versa. In the case of practical interest to dynamical systems, the trajectory is computed by the numerical integration of the equations of motion and, thus, it is known as a time series. If we assume that the time series consists of N points at intervals $\Delta t = 1$, the function $R(n)$ consists of N points as well. The largest ν will be $\frac{1}{2\Delta t} = \frac{1}{2}$ (remember that $\Delta t = 1$), since we do not resolve oscillations with periods smaller than $2\Delta t$, and the smallest ν will be $\frac{1}{N}$, since we have not enough information for oscillations with period longer than $N\Delta t$. The νs then lie in the interval $-N/2 < \nu < N/2$ and are equally spaced, differing by $\frac{1}{N}$. As as a result then, $f(\nu)$ and $R(n)$ are related through the equations:[4]

$$f(\nu_j) = \sum_{k=0}^{N-1} R(\tau_k) e^{(-2\pi i \nu_j \tau_k)} \tag{55}$$

and

$$R(\tau_k) = \frac{1}{N} \sum_{j=-\frac{N}{2}}^{\frac{N}{2}} f(\nu) e^{(2\pi i \nu_j \tau_k)} \tag{56}$$

Since from (52) we know that $R(k) = R(-k)$, it turns out that $f(\nu)$ is a real and even function. Therefore it is sufficient to consider the spectrum only for positive values of ν. It should be emphasized that conventions differ in the literature. Some authors use different normalization factors in (55–56). Others use the correlation function, $C(n)$, instead of the covariance function, $R(n)$, in (55)[5] In our nomenclature, the last case should be named *normalized spectrum*. If now we use the spectrum as a tool, the difference between mixing and non-mixing dynamical systems becomes clear. In the case of non-mixing motion, the spectrum is discrete, that is

[4] For reasons of completeness, we should note that the pair of conjugate functions $f(\nu)$ and is defined in a different way if the "signal" is a continuous function of time defined in a finite interval and yet in another way if it is a continuous function defined over the real numbers. These cases, however, are outside the scope of the present chapter.

[5] Note that in (55) the lower limit of summation is 0, since negative ks have no new information.

$$f(\nu) = \sum_k f_k \delta(\nu - \nu_k) \tag{57}$$

while in the case of mixing motion the spectrum is continuous. Therefore *in principle* one could differentiate between a mixing and a non-mixing dynamical system by just calculating the spectrum of a dynamical variable. In reality things are not so simple because, due to the phenomenon known as *leakage* in the calculation of a Fast Fourier Transform (FFT), closely spaced spectral lines give the impression of a continuum spectrum. Another delicate point in the calculation of the spectrum is the phenomenon called *aliasing*. If we under-sample the signal, i.e. the signal contains still information at frequencies above our highest one (the so-called *Nyquist frequency*), then the power corresponding to these modes will wrongly appear at lower frequencies, and therefore will contaminate the power spectrum that we get. This problem is more dangerous in dynamical systems that have characteristic frequencies that differ by orders of magnitude. A typical example is the motion of planets and asteroids in our solar system. Due to the degeneracy of the Hamiltonian, the motion has a set of characteristic frequencies related to the orbital motion, which are of the order of $1 - 10^2$ yrs, and another one related to the secular change of the orbits due to the mutual perturbations between the planets, which is of the order of $10^3 - 10^5$ yrs. As a result we face the following dilemma: either we record the output of the numerical integration at time intervals pertaining to the orbital motion, in which case we are flooded with an enormous volume of data points, or we record the output at time intervals pertaining to the secular motion, in which case the higher orbital frequencies contaminate the spectrum through aliasing. The solution is to "smooth" the high frequencies before applying the FFT, e.g. through the use of a running mean filter. It should be noted, however, that in this case the high frequency information is lost.

Mixing implies a coarse-graining of the phase space. With this we mean the following. If we want to model time-irreversible processes, as it is assumed for the evolution of a mixing dynamical system, we should assume that we can know the final position of the moving particle(s) only within the uncertainty of a small but finite box in phase space. Otherwise any trajectory traced backwards in time would return exactly to its initial conditions and the system would be exactly reversible. The assumption that phase space is coarse-grained guarantees that information is lost somewhere in our dynamical system, as required by the existence of a positive Kolmogorov-Sinai entropy (see Sect. 4.1).

It is easy to see that mixing implies ergodicity. However, the converse is not true: ergodicity does not imply mixing. This can be understood at once from a counterexample concerning the translations on the torus, described by (48). If we paint one bundle of trajectories black and a neighboring bundle grey, clearly they will maintain their relationship to one another as they ergodically cover the torus, so that they do not mix to an equilibrium. This

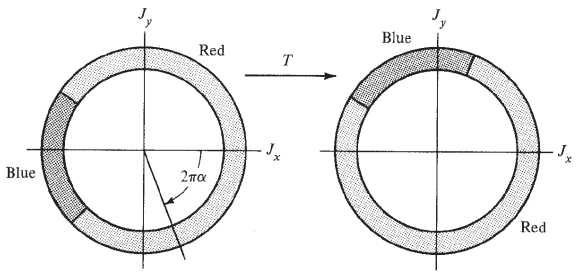

Fig. 10. The rotation by an irrational fraction, a, of 2π is an ergodic dynamical system *(from [21])*

property is illustrated in Fig. 10 for the twist mapping, where the "black" and "grey" portions of the circle do not mix, even when α is irrational.

We intuitively would expect that the chaotic trajectories of non-integrable Hamiltonian systems are mixing within the phase space region they explore. However it is hard to prove the mixing property of dynamical systems rigorously. Up to now there are only a few cases where this has been achieved. One such case is the "coarse-grained" observable function of the mapping known as *baker's transformation* (see Sect. 3.4), which approaches an equilibrium state that is uniform, independent of the initial conditions, a fact implying that this mapping is mixing. Another system proven to be mixing is the completely ergodic system of the hard sphere gas. This system is equivalent to the motion of a point mass on the surface of a 2-torus having a circular hole, with the "rule" that the angle of incidence on the circumference of the hole equals that of reflection (Fig. 11). The proof depends on the continual defocusing of a bundle of trajectories, as they are reflected on the boundary of the hole, because the boundary has negative curvature (Fig. 12). These results justify, in a way, the assumptions made when attempting the statistical description of non-integrable dynamical systems. Note, however, that both refer to dynamical systems in which the "force" is non-differentiable, and not to analytic, infinitely differentiable flows, which are of main interest in physical applications.

K-Systems

K-systems, so called after Kolmogorov who was the first to study them, are systems that have positive *metric* or *Kolmogorov-Sinai* (KS) entropy, h_{KS}. The KS entropy can be thought of as a number measuring the *rate of flow of information*, as a chaotic orbit evolves in time.[6] Therefore the *KS entropy*

[6] Note that whether this flow is considered as loss or gain depends on the point of view we observe this process.

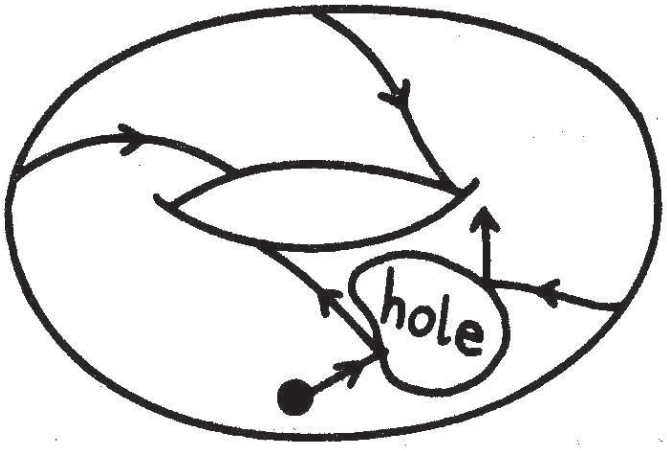

Fig. 11. The motion on the surface of a 2-torus with a hole is mixing *(from [6])*

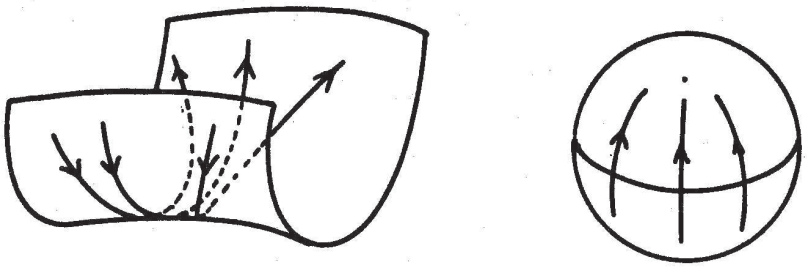

Fig. 12. Trajectories on a surface with negative curvature have positive Lyapunov numbers. On such a surface two trajectories, that are initially at some point close and parallel, separate exponentially both in the past and in the future. This is not so for trajectories on a spherical surface *(from [6])*

should be related, in some sense, to the average rate of exponential divergence of nearby trajectories, i.e. to the *Lyapunov exponents*. In Hamiltonian systems the LCNs are related to the KS entropy through the relation

$$h_{\mathrm{KS}} = \int_M \left[\sum_{\sigma_i(x)>0} \sigma_i(x) \right] d\mu \tag{58}$$

where the sum is over all positive Lyapunov exponents and the integral is over a specified region of phase space, M. If we restrict our study to a single connected chaotic region of phase space (excluding regular regions with KAM tori), the σs are independent of x and the integral over M is unity,

$$h_{\mathrm{KS}} = \Sigma_{\sigma_i} \sigma_i \tag{59}$$

For a two-degrees of freedom autonomous Hamiltonian, only σ_1 is greater than zero and

$$h_{\mathrm{KS}} = \sigma_1 \tag{60}$$

It should be emphasized that the KS-entropy, h_{KS}, differs from the usual thermodynamic quantity (as appearing in Boltzmann's theory), S, in that it gives the *rate* of flow (gain or loss) of information, while the usual entropy describes the *total* information of the system.

Properties of K-Systems

The only systems that can be described in a rigorous mathematical way by the tools of Statistical Physics are those belonging to the class of *Markovian processes*. A discrete dynamical system is said to be a Markovian process if the value of any variable at a time t_{i+1} depends only on the value at t_i and not on $t < t_i$, i.e. not on the "history" of the trajectory up to this time. A similar definition applies to flows. Therefore, a Markovian process has no memory. However this property is in direct contradiction with the fact that "real" dynamical systems are described by differential (or difference) equations, so that the initial conditions of a trajectory define uniquely its future evolution. The next best choice for the mathematical study of a "real" chaotic dynamical system is to model it as a C-system (see next paragraph). But this class of dynamical systems is very restrictive as well, since it assumes that the system has a dense set of periodic orbits and that all of them are unstable. It is therefore evident that the next best approximation of a chaotic dynamical system should be by modeling it as a system with positive K-entropy. In what follows we will prove, in a heuristic way, the principal property of a K-system, which is a positive K-entropy, and find its relations with the mixing property, the Lyapunov Characteristic Exponent and the coarse- graining of the phase space.

As mentioned already, the Lyapunov characteristic exponent, λ, measures the exponential divergence of nearby trajectories in a mixing dynamical system; its inverse, $T_L = 1/\lambda$, gives the time interval over which a trajectory of the dynamical system "forgets" its initial conditions. A large Lyapunov time implies a weekly chaotic motion, while a short Lyapunov time implies a strong one. Of course whether a time interval can be considered either as "long" or "short" depends on the typical time scale of a particular dynamical system. A Lyapunov time of 100 years corresponds to a very chaotic trajectory of a body in our Solar System but to a very stable trajectory of a particle in an accelerator. A Lyapunov time of 10^8 years corresponds to a very chaotic trajectory of a star in our Galaxy but to a weekly chaotic trajectory of an asteroid in our Solar System.

The autocorrelation time, t_C, gives the time interval over which a trajectory keeps memory of its past and it is estimated by calculating the autocorrelation function, (53). The "physical" meaning of the autocorrelation

function becomes clear if we note that the autocorrelation function, up to a normalization constant, is a function of the time-delay, τ, through the integral

$$C(\tau) = \int_0^\infty f(t) f(t-\tau) \tag{61}$$

If the trajectory is periodic or quasi-periodic, that is regular, the autocorrelation function oscillates about zero and, in particular, between the values -1 and $+1$ for a proper normalization of the integral. If, on the other hand, the trajectory is chaotic, the autocorrelation function has a damped wavy form, whose envelope is a decreasing function of time. The value of the autocorrelation time, t_C, is then defined as the first time at which the value of the autocorrelation function drops below $1/e$ and *never rises above this value at later times*. As discussed in [28], in a mixing dynamical system the Lyapunov time and the autocorrelation time are of the same order of magnitude, $T_L \approx t_C$.

Now, as discussed already, a K-system is mixing as well. Therefore there is a relation between the main property of a K-system, on one hand, which is $h > 0$, and T_L and t_C on the other. This relation was introduced in [24], but it was based on the work of Kolmogorov, who was able to relate the mixing properties of a chaotic system with the notion of entropy, introduced by Clausius and Boltzmann in thermodynamics. As it is well known, we can consider the evolution of a system of n particles as the motion of a point in a $6n$-dimensional phase space, Γ. If we divide Γ in a number of cells, then the entropy of the system of the particles, S_i, is given by the relation (the Boltzmann equation,

$$S_i = k \ln W_i \tag{62}$$

where W_i is the probability that the point representing the system is in the i-th cell and k is Boltzmann's constant.

Now, as argued by [28], the states of the system are distributed uniformly within a region of phase space $\Delta\Gamma$. Hence

$$W \approx \frac{\Delta\Gamma}{\Gamma_0} \tag{63}$$

so that Boltzmann's equation becomes

$$S = k \ln \Delta\Gamma - k \ln \Gamma_0 \tag{64}$$

At this point it will become clear how the introduction of the coarse-graining of the phase space, in combination with the above formula, leads naturally to the notion of the K-entropy. As shown by Liouville, a set of trajectories of complete measure in phase space evolves as an incompressible fluid, i.e. the volume of an element of phase space is conserved. If we consider the evolution of a volume element $\Delta\Gamma(0)$ in phase space, we see that, by virtue of Liouville's theorem,

$$\Delta\Gamma(t) = \Delta\Gamma(0) \tag{65}$$

The shape, however, of this volume element changes with time. If the system is non-mixing, the shape is distorted but nothing spectacular happens. In contrast, if the system is mixing, the volume element is elongated in certain directions and contracted in some other, so that the volume is filamented (see Fig. 9). Since the phase space is finite, the filaments after a while "turn back" and intermingle more and more, becoming at the same time finer and finer, while their surface area becomes larger and larger.

Lets now introduce a quantity ε, with the dimensionality $6n$ of the phase space, and extend the boundary of the filaments by ε. In this way the volume of phase space, occupied by the volume element, increases and the increase is proportional to the total area of the surface. In the case of a non-mixing system this increase is unimportant, while in the case of a mixing system not only the increase is important, but it is a function of time as well. Since the volume of a filament is equal to the minimum "width" of the filaments, ε, times their "length", we see that

$$\Delta\Gamma(t) = \Delta\Gamma_0 e^{ht} \tag{66}$$

where h is an appropriate average of the maximal Lyapunov exponent, characterizing the mixing system, over the phase space volume. If we now substitute (66) into (64), we find

$$S = k\ln(\Delta\Gamma_0 e^{ht}) - k\ln\Gamma_0 = ht + k\ln(\Delta\Gamma_0) - k\ln\Gamma_0 \tag{67}$$

Now without loss of generality we can take $\Gamma_0 = 1$, select units so that $k = 1$ and take $\Delta\Gamma_0 = \varepsilon$, since through the coarse graining of phase space we cannot "measure" volumes with a value less than ε. Then the *average rate of change of entropy*, S, is given by taking the limit $\varepsilon \to \infty$

$$\left\langle \frac{dS}{dt} \right\rangle = \lim_{\varepsilon\to 0}\lim_{t\to\infty}\frac{1}{t}\ln(\Delta\Gamma(t)) = \lim_{\varepsilon\to 0}\lim_{t\to\infty}\frac{1}{t}(ht + \ln\varepsilon) = h \tag{68}$$

From its definition it is obvious that the K-entropy, h, gives the average rate of change of Boltzmann's entropy, S, due to the mixing of trajectories in phase space. It can be proven that h is a metric invariant of a dynamical system, i.e. its value does not depend neither on the way the phase space is divided into cells nor on the way the phase space is coarsened. Moreover from (66) it is evident that $h \sim \lambda$, so that finally we have that $h \sim \lambda \sim 1/t_C$.

Before we move to the next topic, it is interesting to briefly discuss a apparent disagreement between the definitions of Boltzmann's and Kolmogorov's entropy: If the rate of change of S is a constant, h, how is it possible for S to reach a maximum, as postulated by Clausius for a closed system? The answer lies in the fact that we cannot measure phase space volume with an infinite accuracy! Then if we fix a smallest coarsening of

the phase space, ε_0, we can find from (66) a time t_0, after which the region $\Delta\Gamma_0 = \varepsilon_0$ expands to the value $\delta\Gamma = 1$:

$$t_0 = \frac{1}{h} \ln \frac{1}{\varepsilon_0} \tag{69}$$

The value of S corresponding to the time t_0 is the maximum value of the entropy of this dynamical system.

Anosov-Systems

Anosov-systems are *abstract* dynamical systems, which are assumed to be as chaotic as a Hamiltonian dynamical system can be. Loosely speaking we can say that an Anosov system has a phase space characterized by two directions: one where the trajectories diverge exponentially and one where the trajectories converge exponentially. Thus the region around *any* point of phase space has a topological structure similar to that around a hyperbolic fixed point, where chaotic motion develops in Hamiltonian systems. In other words, in a C-system all periodic trajectories are unstable. This is a very strong condition, which "ordinary" Hamiltonian systems never satisfy, since in all real systems studied up to now arbitrarily close to any hyperbolic point there are always elliptic (stable) fixed points. It has been shown that Anosov-systems are an open set in the superset of all dynamical systems and, therefore, they are *structurally stable*, i.e. a perturbed Anosov-system remains an Anosov-system. An example of a simple dynamical system belonging to the class of Anosov-systems is the following area preserving mapping, known as *Arnold's cat map* (from the figure Arnold used to depict its action, see Fig. 13):

$$(x_{n+1}, y_{n+1}) = ((x_n + y_n), (x_n + 2y_n)) \pmod 1 \tag{70}$$

From Fig. 13 it is clear that the procedure of mapping back onto the unit square continually filaments the cat (the phase space) and, furthermore, mixes the filaments. In the case of this map it is possible to calculate analytically the Lyapunov numbers. We first find the eigenvalues of the mapping

$$\lambda_{1,2} = \frac{3 \pm \sqrt{5}}{2} \tag{71}$$

by solving the characteristic equation

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0 \tag{72}$$

We can see that, as it should be for an area preserving mapping, $\lambda_1 = \frac{1}{\lambda_2}$ and that

$$\lambda_{1,2} = e^{\pm\sigma} \tag{73}$$

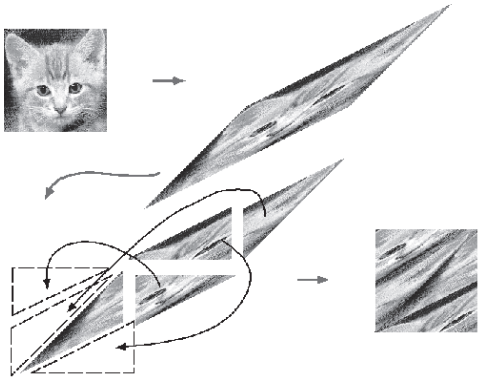

Fig. 13. Graphical representation of the Arnold's cat map

where $\sigma = \ln \frac{(3+\sqrt{5})}{2}$. It is easy to see that this mapping has one direction which is always expanding and another one which is contracting. Moreover $\sigma_1 > 0 \ , \forall (x_n, y_n)$, so that the cat mapping is a K-system. It has been shown [4] that C-systems are a subset of K-systems. For the cat mapping of (70) we have the particularly simple result

$$h_{\mathrm{KS}} = \sigma_1 = \ln \left[\frac{1}{2} (3 + \sqrt{5}) \right] \tag{74}$$

Bernoulli Shifts

There are systems even "more" chaotic than C-systems, known as *Bernoulli shifts*. These systems are equivalent to *Markov processes* (see Sect. 4.1) in the sense that, if we are presented with two sequences of 0s and 1s, one sequence which is a "trajectory" of a Bernoulli shift and another sequence, which is the result of a coin toss, we cannot tell the difference between the two. Or, in other words, in an infinite sequence of 0s and 1s, produced by a coin toss, we can find always a sub-sequence, which is identical to a "trajectory" of a Bernoulli shift.

If, in the toss of a coin, we denote "head" by "1" and "tail" by "0", then the "motion" is described by a random binary sequence, such as

0111011

It can be proven that the KS entropy of such a sequence, consisting of the permutations of n symbols from a set of M different symbols ($n > M$), is given by

$$h_{\mathrm{KS}} = \ln M \tag{75}$$

As an example, the "trajectory" given above has a KS entropy equal to $\ln 2$. On the contrary the "usual" (i.e. thermodynamic) entropy is given, in general,

by the relation

$$S = nh_{\mathrm{KS}} \tag{76}$$

which, for the "trajectory" given above, is equal to $7 \ln 2$ (since the sequence 0111011 consists of 2 symbols and 7 digits). Therefore the KS entropy represents the rate of growth of S with the "time" n. As a result the thermodynamic entropy, S, increases without bound as $n \to \infty$.

A typical example of a Bernoulli shift is the well known *baker's transformation* (Fig. 14)

$$(x_{n+1}, y_{n+1}) = \begin{cases} (2x_n, \frac{1}{2}y_n), & 0 < x_n < \frac{1}{2} \\ (2x_n - 1, \frac{1}{2}(y_n + 1)), & \frac{1}{2} < x_n < 1 \end{cases} \tag{77}$$

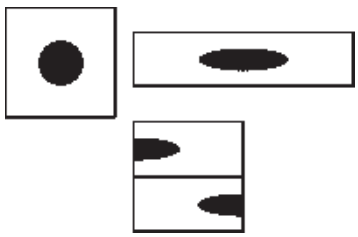

Fig. 14. Graphical representation of the baker's transformation

It can be proven that the baker's transformation is isomorphic to a Bernoulli shift with equal probabilities for "head" and "tail", $p = q = 1/2$.

Hierarchy of Chaos - Real World

To wrap up things, from the five classes of chaotic behavior four belong to an hierarchy of ideal dynamical systems, in which each successive class is a subset of the previous

$$\text{ergodic systems} \supset \text{mixing systems} \supset \mathrm{K-systems} \supset \mathrm{C-systems}. \tag{78}$$

The weakest chaotic condition is *ergodicity*, by which we mean that any single orbit visits all accessible phase space. This condition is met even by the ordered trajectories of integrable systems and, therefore, is not related to the phenomenon of chaos in non-integrable systems. The next more chaotic class consists of *mixing* systems, which approach equilibrium in the sense that Boltzmann had in mind when he defined ergodicity. An even more chaotic dynamical system is a *K-system*, which has a positive Kolmogorov-Sinai entropy, i.e. a positive average divergence rate of trajectories. This means that

a trajectory selected at random has at least one positive Lyapunov exponent, while there may be also a set of periodic orbits of zero measure with all Lyapunov numbers equal to zero. Finally in a *C-system* all trajectories diverge exponentially in every part of phase space (and not just those of complete measure).

All the above classes represent, in a sense, ideal dynamical systems which are not really stochastic, in the Markovian sense. The fifth class of chaotic dynamical systems however, the Bernoulli shifts, are as chaotic as the toss of a coin and, therefore, can be described as Markov chains. We could say that Bernoulli shifts are "more chaotic" than C-systems but it is not yet proven whether they are a subset of them or not. From the above it becomes clear that, strictly speaking, only Bernoulli shifts can be described by the tools of Statistical Physics, e.g. as random walks. However, if we consider the "position" of a trajectory at consecutive "steps" of finite length, then under some not very restrictive conditions we may *assume* that a chaotic dynamical system, such as a K-system or a C-system, is described by a Markov chain, although this is not strictly true.

Here we give some examples, in order to differentiate among the various classes of chaotic behavior. A typical example of an ergodic dynamical system, which is not mixing, is the motion on a torus of an integrable system, characterized by an irrational winding number. Presently a "physical" example of a mixing system, that is not a K-system as well, is not available. Typical examples of K-systems, which are neither C-systems nor Bernoulli shifts, are the hard sphere gas of Sinai and the stadium billiard, in which a ball moves in a horizontal billiard made from two semi-circles connected with straight lines. A typical example of a C-system is Arnold's cat. Finally a typical example of a Bernoulli shift is the baker's transformation.

Perturbed integrable systems such as the Hénon-Heiles one, in which regular and chaotic regions are dense in phase space, do not fall into any of the above categories. That is why we prefer to describe them as *chaotic*, a word that has not any well defined meaning in Mathematics. However, the motion near a homoclinic point in a chaotic region is locally equivalent to the baker's transformation and, therefore, shows the random behavior of a Bernoulli shift. Furthermore, there is strong numerical evidence that each connected stochastic region, such as a separatrix layer, has positive KS entropy, and thus it has the properties of a K-system.

5 Resonance Overlap Criterion

If one perturbs an integrable dynamical system more and more, then more and more ordered trajectories become chaotic. However one of the conclusions of Hénon and Heiles was that, the measure of phase space regions dominated by chaotic trajectories is not a "smooth" function of the strength of the perturbation. Instead the, initially, very slow increase, during which the chaotic

regions cannot be detected easily, is followed by an explosive one, after which the chaotic regions dominate more or less the available phase space. Therefore macroscopically it appears that a *threshold* value of the perturbation exists, below which the measure of chaotic regions in phase space is vanishingly small and above which becomes of order unity. This value is called *stochasticity threshold* and is of importance if we are attempting a statistical description of a "physical" dynamical system, since this description usually holds only for values of the perturbation considerably larger than the stochasticity threshold. The calculation of this threshold is performed by the so-called *resonance overlap criterion*. This criterion is based on the assumption that chaos appears as a result of the "interaction of resonances". In the simplest case of a 2-D conservative dynamical system, the stochasticity threshold is computed by equating the distance between the centers of two neighboring islands on a surface of section with the sum of the "radii" of these islands. We show here in detail how this calculation is done, following the "standard" procedure analyzed in detail in [21] and [13].

The simplest class of dynamical systems, to which we can apply the resonance overlap criterion in order to estimate the stochasticity threshold, is the class consisting of 1-D time-dependent Hamiltonians. Sometimes these dynamical systems are referred to as $1\frac{1}{2}$*-D systems* since, as we will see at the end of this section, by a trivial canonical transformation they can be written as 2-D systems, whose Hamiltonian function is identically equal to zero. From the class of all time-dependent 1-D Hamiltonian systems we select to study those who can be considered as 1-D autonomous systems (which, by definition, are integrable), perturbed by a time-dependent function. Let's assume that we have written the Hamiltonian of the autonomous system, i.e. the unperturbed part, in action-angle variables, $I-\theta$. Then the most general form of the full, perturbed Hamiltonian is

$$H = H_0(I) + \varepsilon \sum_{m,n} V_{m,n} \cos(m\theta - n\Omega t) \tag{79}$$

As we have already discussed, for $\varepsilon = 0$ the phase space trajectories of this system lie on nested toroidal surfaces. If the integrable system is non-degenerate, the unperturbed frequency

$$\omega(I) = \frac{dH_0}{dI} \tag{80}$$

has a non-zero second derivative (see Sect. 3.4). When the rotation number, i.e. the ratio $\frac{\omega}{\Omega}$, is rational, i.e. when

$$\frac{\omega(I)}{\Omega} = \frac{P}{Q} \tag{81}$$

where P and Q are integers, the trajectories of the unperturbed system are periodic and close after making P revolutions in the toroidal direction, which

we take to represent the time variable, and Q revolutions in the poloidal direction of the torus, which we take to represent the angle variable. When the rotation number is irrational, the trajectories cover ergodically the nested toroidal surfaces.

Now let's study the effect of the perturbation term to the integrable part of the Hamiltonian, through the surface of section of the nested toroidal surfaces, say at the value of the toroidal angle $\Omega t = 2k\pi, k = 0, 1, 2, \ldots, n$. For $\varepsilon = 0$ the surface of section is covered by topological circles; those corresponding to an irrational rotation number consist of the consequents of only one trajectory, while those corresponding to a rational rotation number consist of an infinite number of periodic orbits. The introduction of the perturbation term has an important qualitative change on the appearance of the surface of section. As the integrable systems are of zero measure in the set of all dynamical systems, the perturbed system, as a rule, is not any more integrable. The result is that the invariant curves corresponding to rational rotation numbers cease to exist. In general they are replaced by a finite number of periodic orbits, half of them stable and half of them unstable, which appear on the surface of section as *fixed points*.[7] Around each one of the stable periodic orbits "resonant" invariant curves appear, while, if the system would be integrable, the unstable invariant curves emanating from the unstable invariant point (see [21]) would join smoothly. However, since the perturbed system is non-integrable, the consequents around the unstable periodic orbits do not lie on smooth curves and chaos appears in this region. We loosely name the set of the resonant islands around the stable periodic orbits, together with the chaotic strip around the unstable periodic orbits, a *resonant structure*. Resonant structures appear at all regions where the rotation number is rational, but they are separated by invariant curves, as guaranteed by the KAM theorem. If, however, the strength of the perturbation (i.e. ε in (79)) is increased, the width of the resonant structure becomes wider, as it was found by Hénon and Heiles, and so does the chaotic strip around the unstable periodic orbits. At some point the invariant curves, separating two neighboring chaotic strips, cease to exist and the two chaotic strips join together. One way to look at this process is to postulate that the last invariant curve separating two consecutive resonance structures is "squeezed out" when the sum of the widths of the two resonance structures is equal to the distance (in i-space) between them. This is the resonance overlap criterion. In order to apply the resonance overlap criterion we should be able to estimate (i) the distance between two resonant structures and (ii) the width of each resonant structure, as functions of the strength of the perturbation, ε. First we start by selecting a rational value of the rotation number, say P/Q. From (81) we see that the resonance structure is close to the torus $I = I_0$, where $\omega(I_0)/\Omega = P/Q$. The dominant terms in the perturbation part of the Hamiltonian are those with

[7] The terminology comes from the fact that, on the surface of section, these points are mapped, through the equations of motion, on themselves.

$n/m = P/Q$ since, if we construct the solution of the equations of motion in a formal series form, they will correspond to the terms with the small divisor $Qn - Pm$. Assuming that the coefficients of the higher harmonics fall off rapidly with the order of the harmonics, we can approximate the motion near the resonant structure by the *resonant* Hamiltonian

$$H_R = H_0(I) + \varepsilon V_{Q,P}(I)\cos(Q\theta - P\Omega t) \tag{82}$$

The resonant Hamiltonian contains only one trigonometric term and it is integrable, since by a suitable canonical transformation we can "remove" the time. Indeed, through the generating function of the third kind

$$F_3 = -(I - I_0)\frac{\psi + P\Omega t}{Q} \tag{83}$$

we arrive at the canonical transformation from (I, θ) to (p, ψ)

$$\theta = -\frac{\partial F_3}{\partial I} = \frac{\psi + P\Omega t}{Q} \quad p = -\frac{\partial F_3}{\partial \psi} = \frac{(I - I_0)}{Q} \tag{84}$$

or, in other words, the argument of the cosine function is the new angle, while the new action is the difference of the old action from the value corresponding to the unperturbed torus. The Hamiltonian is then written in the new variables

$$\begin{aligned} H_R &= H(I(p), \theta(\psi)) + \frac{\partial F_3}{\partial t} \\ &= H_0(I(p)) + \varepsilon V_{Q,P}(I(p))\cos(\psi) - \frac{(I - I_0)P\Omega}{Q} \\ &= H_0(I(p)) + \varepsilon V_{Q,P}(I(p))\cos(\psi) - pP\Omega \end{aligned} \tag{85}$$

This Hamiltonian is 1-D and autonomous (i.e. time-independent) and, therefore, integrable. This of course does not mean that by a "suitable" canonical transformation we have changed a chaotic dynamical to a regular one. We simply approximate the phase portrait of the initial Hamiltonian close to the center of an island, where the chaotic nature of the initial system is not apparent. But now we can estimate analytically the width of the island as a function of the perturbation's strength, i.e. ε. To do so we expand H_R around $p = 0$ in a Taylor series

$$H_R = H_R(p = 0) + \frac{dH_R}{dI}\frac{dI}{dp}p + \frac{d^2H_R}{dI^2}\left(\frac{dI}{dp}\right)^2 p^2 + \varepsilon V_{Q,P}(I(p))\cos(\psi) - P\Omega p \tag{86}$$

But according to (80)

$$\left(\frac{dH}{dI}\right)\left(\frac{dI}{dp}\right)p = P\Omega p \tag{87}$$

so that, keeping terms up to second order, H_R becomes

$$H_R = H_0(I_0) + \frac{1}{2}q^2 H''(I_0)p^2 + \varepsilon V_{Q,P}(I_0)\cos(\psi) = h_R \tag{88}$$

which is identical to the Hamiltonian of the pendulum discussed in Sect. 3.4, if we move the constant term $H_0(I_0)$ to the right hand side. We know that the separatrix corresponds to the value of ε for which

$$h_R - H_0(I_0) = \varepsilon V_{P,Q}(I_0) \tag{89}$$

and its width is

$$\Delta p = 4\sqrt{\frac{\varepsilon V_{P,Q}(I_0)}{Q^2 H_0''(I_0)}} \tag{90}$$

or, going back to the old variables, we find that the width of the island, ΔI is given by the relation

$$\Delta I = 4\sqrt{\frac{\varepsilon V_{P,Q}(I_0)}{H_0''(I_0)}} \tag{91}$$

On the other hand the overlap criterion depends on the distance between islands too. Here enters the non-degeneracy condition we have assumed in the beginning. By differentiating (80) we find that the distance between two tori characterized by a difference $\Delta\omega$ in the rotation number is given by the relation

$$(\delta I)_\omega = \frac{\Omega \Delta R}{H_0''} \tag{92}$$

which cannot be defined if the second derivative of the zero-th order Hamiltonian is zero. Now we have to decide which value of $\Delta\omega$ we will use in (92). This can be done if we express the difference in rotation numbers through the theory of the continued fractions. The expansion of the rotation number ω as a continued fraction is

$$\frac{\omega}{\Omega} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \ldots}} = [a_0, a_1, a_2, \ldots] \tag{93}$$

It is well known that the successive truncations of the above continued fraction give, at every order, the best rational approximation to ω. In this way we can write

$$\frac{\omega}{\Omega} = R = \frac{P}{Q} = [a_1, a_2, a_3, \ldots, a_n] \tag{94}$$

and

$$\frac{\omega + \Delta\omega}{\Omega} = R + \Delta R = \frac{P_+}{Q_+} = [a_1, a_2, a_3, \ldots, a_n + 1] \tag{95}$$

From the above relations we find

$$\frac{\Delta\omega}{\Omega} = \Delta R = \frac{1}{Q(Q+Q_-)} \tag{96}$$

where

$$\frac{P_-}{Q_-} = [a_1, a_2, a_3, \ldots, a_{n-1}] \tag{97}$$

By combining (90-97) we find that, according to the resonance overlap criterion, chaotic motion sets on in the neighborhood of the torus with rotation number equal to P/Q when

$$\beta = \frac{(\Delta I)_R}{(\Delta I)_\omega} > \frac{4Q}{\Omega}(Q+Q_-)\sqrt{\varepsilon V_{P,Q} H_0''(I_0)} \approx \frac{8Q^2}{\Omega}\sqrt{\varepsilon V_{P,Q} H_0''(I_0)} \tag{98}$$

The extension of this method to the case of a "pure" 2-D Hamiltonian system is straightforward, assuming that the Hamiltonian of the integrable, non-perturbed, part has been written in action-angle variables. Simply the degeneracy condition is slightly more complicated and the role of the second angle is played by the time in the above example. The role of the second action is played by the numerical value of the Hamiltonian, H_{Res}, which in the above example enters in the "new" Hamiltonian through the term $\frac{\partial F}{\partial t}$ of the transformation, while in the 2-D case enters directly from the canonical transformation.

The situation is considerably more complicated if the degeneracy condition does not hold. In that case the two neighboring resonances, the interaction of which creates the chaotic behavior, are not both appearing explicitly in the Fourier expansion of the original Hamiltonian. At least one of them (and some times both) is a "secondary" resonance, one which appears at a higher order (with respect to the "strength of the perturbation", ε), when one tries to find a perturbative solution of the problem, using series expansions of the variables. A good discussion on this subject appears in [21].

References

1. D.V. Anosov: Sov. Math. Dokl. **3**, 1068 (1962)
2. D.V. Anosov: Sov. Math. Dokl. **4**, 1153 (1963)
3. V.I. Arnold: *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1989) 143, 151, 157
4. V.I. Arnold and A. Avez: *Ergodic problems of Classical Mechanics* (Benjamin, New York, 1968) 176
5. G. Benettin, L. Galgani, A. Giorgilli and J.M. Strelcyn: *Meccanica*, March, 21 (1980)
6. M.V. Berry: Regular and irregular motion. In: *Topics in nonlinear dynamics*, vol 46, ed by S. Jorna (AIP, New York, 1978) pp. 16-120 143, 144, 145, 152, 154, 155, 171
7. B.V. Chirikov: Sov. Phys. Dokl. **4**, 390 (1959)

8. B.V. Chirikov: Phys. Rep. **52**, 265, (1979) 143
9. G. Contopoulos: Bull. Astron. 3e ser. 2, Fasc. 1, 223 (1967)
10. G. Contopoulos: *Order and Chaos in Dynamical Astronomy* (Springer, Berlin, 2002) 143
11. J. Ford, S.D. Stoddard and J.S. Turner: Progr. Theoret. Phys. **50**, 1547 (1973)
12. J. Ford: The Statistical Mechanics of Classical Analytical Dynamics. In: *Fundamental Problems in Statistical Mechanics*, vol 3, ed by E.D.G. Cohen (North Holland, Netherlands, 1975) pp. 215-255 143
13. J. Ford: A Picture Book in Stochasticity. In: *Topics in nonlinear dynamics*, vol 46, ed by S. Jorna (AIP, New York, 1978), pp. 121-146 143, 179
14. C. Froeschlé and J.P Scheidecker: Astrophys. Space Sci. **25**, 373, 1973
15. J. Greene: Ann. N.Y. Acad. Sci. **357**, 80 (1980)
16. Gutzwiller, M.C., *Chaos in Classical and Quantum Mechanics* (Springer, Berlin, 1990) 143
17. M. Hénon and C. Heiles: AJ **69**, 73 (1964) 142
18. J. Klafter, M.F. Shlesinger and G. Zumofen: Physics Today **49**, 33 (1996)
19. J. Klafter, M.F. Shlesinger and G. Zumofen: Physics Today **55**, 48 (2001)
20. A.N. Kolmogorov: Dokl. Akad. Nauk SSSR **124**, 754 (1959)
21. A.J. Lichtenberg M.A. and Lieberman: *Regular and Chaotic Dynamics* (Springer, Berlin, 1992) 143, 170, 179, 180, 183
22. V.I. Oseledec: Trans. Moscow Math. Soc. **19**, 197 (1968) 146
23. E. Ott: *Chaos in Dynamical Systems* (Cambridge University Press) 1993 143
24. Ya.B. Pesin: Russ. Math. Surveys **32**, 55 (1977) 173
25. Ya.G. Sinai: Sov. Math. Dokl. **4**, 1818, (1963)
26. K. Tsiganis, H. Varvoglis and J. Hadjidemetriou: Icarus **146**, 240 (2000) 148
27. H. Varvoglis, Ch. Vozikis and K. Wodnar: CMDA **89**, 343, 2004 160, 162
28. G.M. Zaslavsky: *Chaos in Dynamic Systems*, 2nd edn (Harwood, London, 1987) pp 6–45 166, 167, 173

Part II

Extrasolar Planetary Systems

Planet Formation

Thomas Quinn

Astronomy Department, University of Washington, Box 351580, Seattle WA 98195-1580, USA
(trq@astro.washington.edu)

1 Introduction

Motivating the study of planet formation is not difficult for any curious audience. One of the fundamental human questions is that of origins: "where did I come from?". Breaking this down into constituents produces a series of questions. How did the Universe begin? How did stars form? *How did planets form?* How did life begin? How did intelligent life develop? It is therefore obvious how the origins of planetary systems is a central issue to origins in general.

Theories for the formation of the Solar System have been around since Laplace and Kant, but in the past 10 years, theoretical research has become more focused for several reasons. These include:

1. Observations of planetary systems other than our own. The discovery of a Jupiter size planet in a close orbit around 51 Peg [29] gave us the first clue about the variety of planetary systems. Since then over 100 extra-solar planets have been discovered, and we can now talk statistically about the properties of these systems.
2. Surveys of planet forming regions. Infra-red and millimeter observations of the interstellar medium have allowed us to peer into the dusty star forming regions and get constraints on the conditions there. High resolution observations have even allowed us to estimate protoplanetary disk masses and temperatures.
3. Predictive theories of planet formation pushed forward by computational advances. The physics of planet formation is complex and non-linear. Numerical simulation has been necessary to fully model it. The possibility of doing computationally intensive calculations has had particular impact on the long term dynamics of planetary systems both for small bodies and planets.

1.1 Major Issues for Solar System Formation

Theories of planet formation need to come to grips with a number of issues which are summarized here.

T. Quinn: *Planet Formation*, Lect. Notes Phys. **683**, 187–217 (2005)
www.springerlink.com

- What caused it to form?
  We certainly need to understand this if we are to know anything about planet formation in other parts of the galaxy or in other parts of the Universe. Since this is intimately connected with star formation, this question not only has implications for planet formation, but also for the entire history of the visible Universe.
- What was the nebula really like in terms of pressure, temperature, and the timescales for processes?
  For planet and star formation theories, this is the question of initial conditions. What are the properties of the ISM in which planets form?
- What are the processes involved in planet formation?
  We are all familiar with the Kant nebular hypothesis: a disk of gas fragments into disk of planets, but what are the actual physical processes? Clearly gravity as the dominant astrophysical force is significant; however, a variety of other forces play rôles as well. These range from magnetic fields and hydrodynamic shocks on large scales down to the material properties of dust particles.
- What is the rôle of condensation?
  The planets (particularly the terrestrial planets) are not of cosmological abundance. How did they get enhanced? This is the primary evidence that non-gravitational forces played a significant part in planet formation.
- How Universal are the processes?
  Here we introduce issues that relate to the uniqueness of our own solar system and the ubiquity of planets around other stars. The possible questions are numerous: Are the terrestrial planets always near the Star? What is the rôle of the Giant Planets? How do the small bodies, Comets and Asteroids, fit in? What determines the planets' spacing? What determines their orbits? Are all planetary systems similar? How are compositions and atmospheres determined?

2 Planet Formation: the Standard Model

Following is an overview of the Standard, or planetesimal, model broken down into stages. We will then describe each stage in more detail. For a more complete review see [23].

1. Initial stage:
   This is the connection to star formation: one assumes a circumstellar disk with solar abundance around the Sun. It gives a natural explanation for the coplanar motions of the planets. The total mass of the nebula should also give a rough estimate of the mass of the planets. As the formation of this disk has more to do with star formation and the collapse of the protostar within the molecular cloud, it will not be covered further here.

2. Early Stage:
   As the disk cools (because the infall slows down) materials can condense. One expects (for solar composition) silicates and iron compounds to condense first. In the outer regions (asteroid belt and beyond) ices can form. This is a natural explanation for the compositions of the planets. Rocky bodies like the terrestrial planets will form close to the parent star where only silicates and iron can condense, while the gas giants and ice giants have a much larger fraction of light gases.
3. Middle Stage:
   At this stage the condensed dust collects in km-sized objects via a gravitational instability (see below). We will call these bodies *planetesimals*. They will interact with each other gravitationally, and occasionally collide, building up into larger bodies. This provides a natural connection between the planets and the small bodies. Planetesimals that do not accrete onto larger bodies remain and provide the origin of comets and asteroids.
4. Late Stage:
   The middle stage ends with dozens of lunar-size bodies, which we refer to as protoplanets, in the terrestrial region. These interact over a long time scale, occasionally colliding to form large bodies, and clearing the region of the small bodies. At the end of this stage, we have a few bodies whose masses and orbital separations match those of the terrestrial planets. The rotation of these planets can also be explained via the angular momentum imparted during the last major impact.
   In the outer regions, the end-state of this process produces a core of tens of Earth masses onto which gas and cometary material accretes. Hence the gas giants have a core-envelope structure, and the ice giants are presumably cores that did not form in time to accrete any nebular gas.

2.1 Initial Stage: the Solar Nebula

Protoplanetary Disk Structure

Although protoplanetary disks tend to be modeled as smooth and axisymmetric they are probably not uniform, having both radial (i.e. gaps) and and vertical structure. They are heated on the surface by stellar radiation, and internally by accretion energy. Observations of disk disk spectral energy distributions (SEDs) indicate that they are flared so that a significant portion can be heated by the star. Observations also constrain the size to be a few to 100 AU, and the mass of the disk to be 0.001 to 0.1 $M_\odot$. However, the observations tend to be of the dust emission, and one has to assume a dust to gas ratio in order to estimate the total masses of these disks. Nevertheless, these mass estimates nicely match the "minimum mass solar nebula" that can be gotten from our own Solar System in the following manner. Enhance

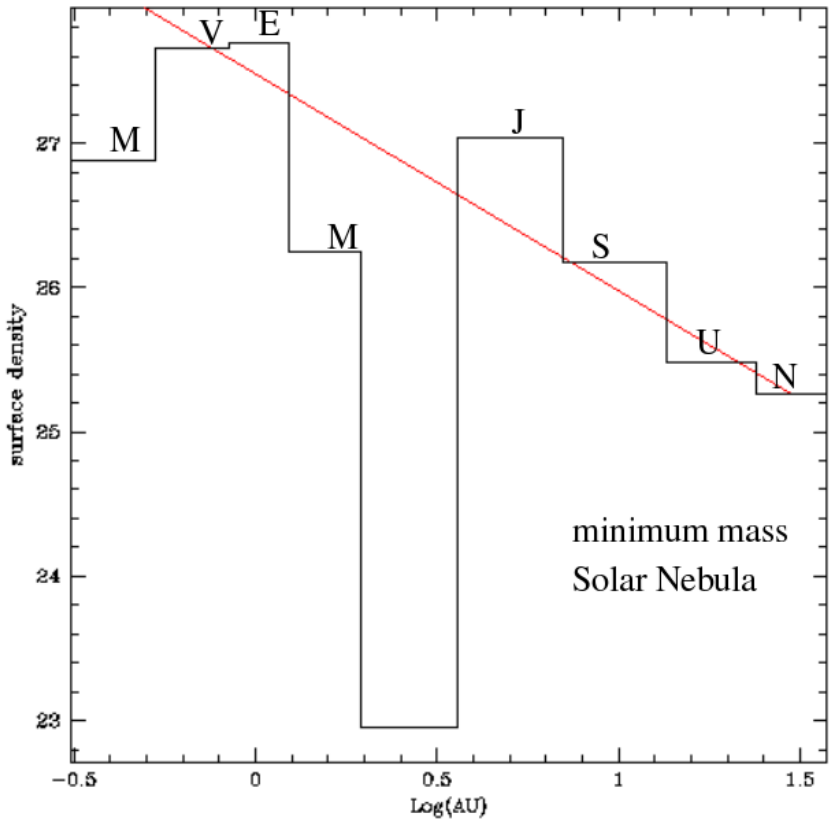

Fig. 1. Mass surface density of the planets after being enhanced to solar abundance *(the line is an $r^{-3/2}$ power law)*

all the planets' masses so that their composition is solar, then spread that mass in annuli with widths given by the separations in semi-major axis. The result is shown in Fig. 1 and the total mass obtained is about 0.01 solar masses. The disks are optically thick, and the opacity is determined mainly by dust. Because the dust properties are a strong function of composition, and the dust composition is determined by what can condense, the opacity is a strong function of temperature. [31] Hence radiative transport is important for determining disk temperatures. Of particular note is that the possible strong drops in temperature as one moves outward in the disk due to opacity changes could lead to convective instabilities. For example there is a large drop in opacity at 500 K due to the evaporation of "refractory organics".

Temperatures

As well as the instabilities mentioned above, MHD instabilities, in particular, the magneto-rotational instability can develop if the gas becomes ionized. Therefore it is important to understand the temperature structure of the protoplanetary disks. For a fuller review of constraints on disk temperature profiles see [7]. The observed SEDs give us a constraint on the surface temperatures of the disk. This is done by fitting power law disk models, and it is typically found that $T_s \propto r^{-1/2}$, with $T_{1\mathrm{AU}} \sim 150\mathrm{K}$. The model temperatures (at least at $r < 1\mathrm{AU}$) seem to be constant over $\sim 10^6$ years, that is, over most of the time that planets take to form. However, it should be noted that the midplane temperatures are higher.

Temperatures in the midplane can be understood theoretically from viscous accretion models. In these models, the gravitational potential energy released as gas slowly flows inward due to viscous forces is turned into thermal energy. Such models predict a steeper temperature profile: $T \propto r^{-3/4}$.

Disk Chemistry

Other constraints come from the composition of planets and small bodies, and therefore the disk chemistry needs to be understood. Condensation arguments imply that the $T \sim r^{-1.1}$. Since this more closely matches the midplane temperature profile, it implies that the grains that ultimately end up in the planetesimals formed in an optically thick region.

The most reliable determination of temperature is the low abundances of volatile elements (such as potassium and sulfur) compared to more refractory elements. This implies temperatures in the range of 1200-1400K in the terrestrial region. In contrast, the ice-rock cores of the giant planets require temperatures less than 160 K at 5 AU. Comets contain HS, CO, methanol, N2 and H2O. This suggests that their formation temperatures are even colder, $\sim 25 - 50K$.

Calcium Aluminum Inclusions (CAIs) in meteorites (in particular the primitive Allende meteorite) are dated to be the oldest objects in the Solar System. They are where the evidence for the short lived isotope, ^{26}Al, is found giving further evidence for being formed very early. These meteorites also indicate condensation temperatures of 1400K from chemical and textural properties. In particular, some CAIs have rounded shapes indicative of partial melting. This requires heating to at least 1700K for a short period of time. Crystal patterns indicate that they cooled quickly at the rate of 2-50 K/hour. Some have multiple rims that indicate they experienced more than one flash heating event.

Chondrules are millimeter sized igneous spherules in primitive meteorites. They contain more volatiles and require formation at temperatures lower than 650K. But they also appear to be flash heated (they are igneous after all),

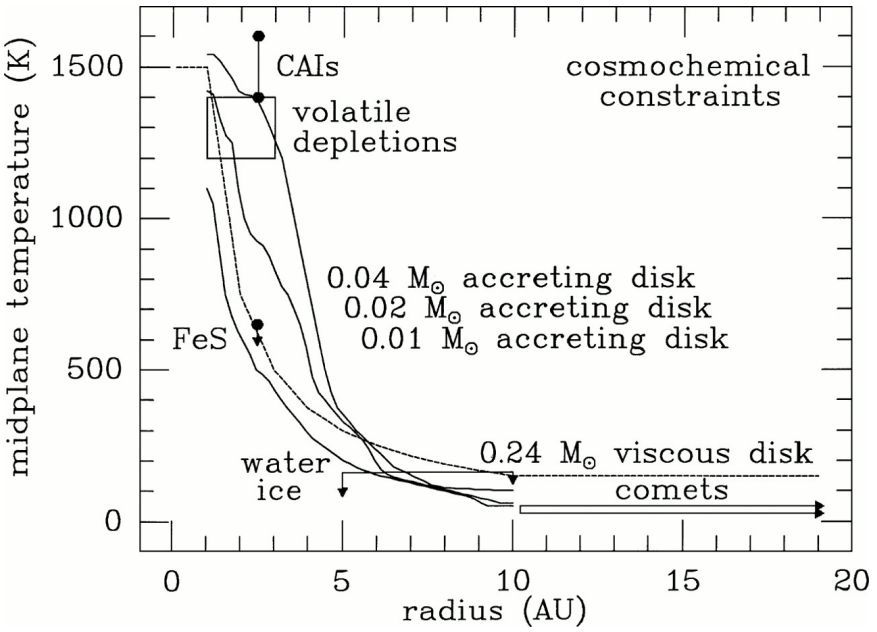

Fig. 2. Temperature vs radius is plotted for several models and compared with cosmochemical constraints *(taken from [7])*

but cooled at 100-1000K/hour. It therefore appears that CAIs formed first in a hot nebula, and then chondrules in a warm nebula. How they end up together in the same meteorite is a problem, perhaps indicating some degree of radial mixing.

These temperature constraints are compared with disk models in Fig. 2 taken from [7]. The accretion models are Boss's models with different mass accretion rates. These models are axisymmetric collapse calculations. A radiative hydro code was used to determine the thermal structure, but the radial density profile was constrained to be $\Sigma \propto r^{-1/2}$. The viscous accretion disk is a calculation of the style of Lynden-Bell and Pringle [25]. The jumps are associated with sublimation and condensation of grains at the different temperatures. Note again, that some such jumps would be unstable to convection. However, convective turbulence seems to stop accretion. Also note that there may be a thermostat effect going on: if the gas is cool enough that grains condense, then it gets optically thick, and the temperature goes back up. Hence the disk could be kept right at the grain condensation temperature.

On the other hand, the surface temperatures are slightly less sensitive to disk mass. This is because the disk is optically thick. This also means that the midplane disk temperature is decoupled from the stellar radiation.

In summary, the models predict a relatively uniform midplane temperature which is matched by volatile depletions. In particular there are no

systematic variations with r within the terrestrial region. Also, the gradual cooling of the disk by becoming optically thinner matches volatile depletion in chondritic meteorites. This gives the obvious conclusion: we expect silicate and iron rich planets to form inside 4 AU, and ice rich planets outside 5 AU. Furthermore, there could have been a cold trap at 5 AU, if water vapor could circulate radially [38]. As water vapor circulates inside 5 AU, it freezes and is trapped right at the ice line. The resulting enhancement of icy solids there would make Jupiter the dominant planet.

One caveat to keep in mind when considering the disk chemistry is that the densities are very low. Therefore, it is only at temperatures above 2000 K that the chemistry will be in equilibrium. In that case, the reaction rates are fast compared to the cooling rate of the disk. Otherwise, the reactions "freeze-out" and the equilibrium abundances are never reached. For example, we see N_2 and CO ices on Triton and Pluto, whereas all the N and C should be in the form of CH_4 and NH_3 at low temperatures. Therefore, the outer parts of the nebula did not reach chemical equilibrium.

So the chemical evolution is quite complex. As well as the long reaction times one must keep in mind the fact that there were primordial interstellar grains, evaporation of grains by shocks and temperature discontinuities, and ionization of gas by cosmic rays and UV radiation that in turn affects the molecular chemistry. Nevertheless, the basic picture seems to hold up as evidenced by our ability to explain the basic compositional characteristics of the planets.

Disk Clearing

At some point the gaseous disk gets cleared away; however, the process by which this happens is not clear. One possible explanation is that the accreting sun had a luminosity 20-30 times larger than it currently has. This could drive a solar wind which could clear out the disk, or if the star were significantly hotter, UV radiation could photoevaporate the faces of the disk.

Recent observations have shed light on the fact that stars almost always form in groups. Hence UV photons from nearby massive stars are a viable candidate for clearing the disk. The stunning HST observations of the Eagle nebula seem to show this process in action. Such a process gives a nice explanation for Uranus and Neptune. All giant planets have the same amount of rock and ice forming elements, but Uranus and Neptune have a factor of 100 less H and He. If the UV came from the outside, as in this scenario, Uranus and Neptune's accretion could have been interrupted before that of Saturn or Jupiter.

2.2 Early Stage: Growth of Grains into Planetesimals

Dust growth

Data from soot and numerical simulations [35] indicate that dust can grow into loosely packed fractal structures held together via van der Waals forces. The existence of many chondrules (i.e. small igneous inclusions 1mm in size) in meteorites means that these fluffy aggregates had to be rapidly heated and cooled before they were incorporated into larger bodies.

Dust Dynamics

To get an understanding of dust dynamics in the protoplanetary nebula, let us review the relevant forces that apply to dust grains.

Radiation Force

For micron sized particles, the radiation pressure force can be significant. It is given by:

$$\boldsymbol{F}_{rad} = \frac{L_\odot A}{4\pi c r^2} Q_{pr} \hat{r}$$

where A is cross section of the particle, r is the distance to the Sun, and Q is a radiation pressure coefficient that accounts for scattering and absorption. If we define the parameter β as the ratio of the radiation force to the gravitational force then

$$\beta \equiv \frac{F_{rad}}{F_g} = 5.7 \times 10^{-5} \frac{Q}{\rho R}$$

where ρ is the density of the grain material in gm/cc and R is the grain radius in *cm*. Given the typical grain densities, this instantly shows that F_{rad} can only be important for sub micron size objects. Furthermore, if the grains get much smaller than that, they become comparable to or smaller than the wavelength of the light and Q starts to become small.

Note that the ratio of forces, β, is independent of the distance from the Sun because both gravity and light drop off as $1/r^2$. So if $\beta > 1$, then the net force will always be outward and the dust will obviously be ejected. If $\beta > 0.5$ then a dust grain will be on a hyperbolic orbit if it is released on what would be a circular orbit in the absence of radiation pressure and therefore also be ejected.

Poynting–Robertson Drag

As grains become of order a centimeter in size, direct radiation pressure becomes less important, but a secular effect occurs because as radiation is absorbed and re-radiated by the particle, the re-radiated radiation is anisotropic

in the rest frame of the sun. This results in a drag force opposite the direction of motion of the particle referred to as Poynting–Robertson Drag.

The total radiation force including this effect to first order in v/c is:

$$\boldsymbol{F}_{\mathrm{rad}} = \frac{L_{\odot} Q_{\mathrm{pr}} A}{4\pi c r^2} \left[\left(1 - \frac{2v_r}{c} \right) \hat{r} - \frac{v_\theta}{c} \hat{\theta} \right]$$

where $L_{\odot}$ is the solar luminosity, r is the radius of the particle, and v_r and v_θ are the radial and tangential component respectively of the particle velocity in the frame of the Sun. The first term in this expression is the radiation pressure, and the second and third terms are the Poynting–Robertson drag. There is a factor of 2 in the second radial term because of the combination of Doppler shift in absorbing radiation and Doppler shift in emission.

Plugging in the current solar luminosity, one can obtain the decay time for a particle on a circular orbit:

$$\tau_{\mathrm{pr}} \sim 400 \frac{r^2 (AU)}{\beta} \mathrm{years}.$$

An obvious application of this formula is source of the zodiacal light. It is emitted from 20 to 200 μm particles, and by the above formula their lifetimes are 10^5 years. Therefore there needs to be a source to replenish these particles. Possibilities include dust from comets and collisions in the asteroid belt.

The Yarkovski Effect

Radiation based drag forces can also be relevant for meter to kilometer sized bodies. For a rotating body, the evening surface temperature is warmer than the morning surface. The resulting anisotropy in the thermal radiation from the body results in a net force called the Yarkovski effect. Since the force per unit area from thermal emission is:

$$dF = \frac{2\sigma T^4 dA}{3c}$$

where T is the temperature and σ is the Stefan-Boltzmann constant, we have

$$F_Y = \frac{8}{3} \pi R^2 \frac{\sigma T^4}{c} \frac{\Delta T}{T} \cos \psi$$

where ψ is the obliquity of the body, and ΔT is the evening-morning temperature difference. The force is therefore in the direction of motion for direct rotators and a drag force for retrograde rotators.

The exact magnitude of this effect is hard to calculate. It depends on detailed surface properties of the asteroids which are not well known. Nevertheless it is postulated as a mechanism for moving asteroids through the belt over long timescales.

Gas Drag

Unlike gas drag on small particles in the lab, where the small particles eventually reach the speed of the gas, gas drag on dust in a Keplerian disk usually results in long term evolution. This is because the drag decreases the energy of the particle, which will ultimately increase the speed of the particle in a Keplerian potential. There are several limiting regimes to the drag force. Stokes drag is given by

$$F_D = -\frac{C_D A \rho v^2}{2}$$

where C_D is a dimensionless coefficient, A is the cross-sectional area of the particle, and ρ is the density of the gas, and is appropriate for high Reynolds number flows. Epstein drag is appropriate when the mean free path of the gas particles is greater than the physical size of the dust particle, and is given by

$$F_D = -A \rho v c_s,$$

where c_s is the sound speed of the gas. In the protoplanetary disk, the drag can have long term effects on intermediate size particles since the gas is not rotating at Keplerian speed. This is because the force that the gas feels is

$$g_{\rm eff} = -\frac{GM_\odot}{r^2} - \frac{1}{\rho}\frac{dP}{dr},$$

where P is the gas pressure. Since the pressure is rising inward, the pressure gradient opposes gravity, and the gas moves slower than Keplerian. Therefore any particle that is only loosely coupled to the gas will experience a headwind, and spiral inward due to the drag forces given above.

Application to the Planetesimal Disk

The typical timescales for inward motion are slow for small bodies because they are tightly coupled to the gas. It is also slow for large bodies, because their momentum is relatively large compared to the gas drag force. For solar nebula conditions, meter size bodies have the shortest decay times of ~ 100 years. Bodies smaller or larger than this will have longer decay times.

For dust grains, settling in the vertical direction is important because it may lead to a gravitational instability (the Goldreich-Ward instability, see below) in the dust. The equation of motion in the vertical direction for the dust with Epstein drag is

$$\frac{dv_z}{dt} = -\frac{\rho_g c_s}{R\rho} v_z - n^2 z$$

where ρ_g is the density of the gas, ρ is the density of the dust, R is the dust radius and $n = \sqrt{\frac{GM}{r^3}}$ is the mean motion. If the drag is large then there is a terminal velocity of

$$v_z = \frac{n^2 z \rho R}{\rho_g c_s}$$

For conditions in the terrestrial portion of the protoplanetary nebula ($T = 500-800K$, $c_s = 2.5km/s$) this velocity gives a settling timescale, z/v_z, of 10^6 years. Furthermore, we need many settling times to concentrate the dust into larger bodies. This settling rate gets shorter as the grains get larger, although if they grow into fluffy aggregates their cross-section to volume ratio would stay large. Nevertheless, grain growth may shorten lifetimes to as little as 10^4 years.

Goldreich-Ward Instability

There are a couple prevalent theories for getting to larger (100 m to 1 km) size bodies. One is that aggregation of dust can occur within turbulent eddies [39]. The more popular theory is that if the dust settles into a cold disk in the plane, it would be susceptible to gravitational instabilities, known as the Goldreich-Ward instability [14].

This instability is essentially the Toomre instability for rotating disks. For the Toomre instability, there is a critical wavelength

$$\lambda = \frac{4\pi^2 G \Sigma}{\kappa^2},$$

where Σ is the surface density and κ is the epicyclic frequency. This wavelength is set by the scale at which the rotational shear is able to overcome gravity independent of the velocity dispersion of the dust. Therefore, if the dust is dynamically cold enough, it will fragment into clumps of about this size. We therefore expect a critical mass of

$$M_p \sim \Sigma \lambda^2 \sim \frac{16\pi^4 G^2 \Sigma^3}{\kappa^4}$$

If the density of the forming planetesimal is about 1 gm/cc then this critical mass would produce a body of roughly 1 km in size in the inner regions of the solar nebula.

This is a rather elegant result in that the final size of the objects formed is relatively independent of the conditions in the solar nebula other than the dust surface density, and presumably this can be directly constrained by SEDs of protoplanetary disks. Secondly, it gives relatively simple initial conditions for the next stage of planet formation: a cold disk of mono-disperse bodies. However, getting the dust to this thin layer has a lot of unknowns. What with turbulence, solid body collisions, and large radial motions, we don't understand this era of planet formation very well.

2.3 Middle Stage: Planetesimal Accretion

Now we come to the crux of the "planetesimal hypothesis": the building of large bodies out of the 1 km size bodies produced at the end of the early stage. As well as providing a mechanism for building the planets, it also makes the connection between those planets and the small bodies.

The simplest accretion picture was first quantified by Safranov [37]. Consider a sphere of radius s moving with velocity v through a uniform medium of density ρ. Then the accretion rate is

$$\frac{dm}{dt} = \pi s^2 \rho v.$$

However, the gravity of the body plays a rôle by gravitationally enhancing the cross-section. This gravitational cross-section can be calculated as follows. Consider the particle that just grazes the surface of the planetesimal as it goes by. From angular momentum conservation we have: $v_i s = v_0 d$, where v_i is the velocity at impact, v_0 is the encounter velocity and d is the impact parameter. From energy conservation we have $-v_{\rm esc}^2 + v_i^2 = v_0^2$ where $v_{\rm esc}$ is the escape velocity from the surface of the planetesimal. Solving for d^2/s^2, we have

$$\frac{d^2}{s^2} = 1 + \left(\frac{v_{\rm esc}}{v_0}\right)^2.$$

Since all particles with an impact parameter less than d will strike the body, πd^2 is an effective cross-section, and the above accretion rate should be multiplied by this factor to account for the gravitational focusing.

Clearly the growth rate will be a strong function of the ratio of the encounter velocity to the escape velocity from the surface of the planetesimals. Therefore, the velocity distribution of planetesimals is a key ingredient to the outcome of this stage. Work in this field has progressed using the kinetic theory of gases, and for that we turn to the Boltzmann equation.

Boltzmann Equation

The evolution of the distribution function $f(\underline{\mathrm{r}}, \underline{\mathrm{v}})$ can be described by the collisional Boltzmann equation,

$$\frac{\partial f}{\partial t} + \underline{\mathrm{v}} \cdot \nabla_r f - \left[\frac{GM_\odot \underline{\mathrm{r}}}{r^3} + (\underline{\mathrm{v}} + \underline{\mathrm{v}}_K) \cdot \nabla \underline{\mathrm{v}}_K\right] \cdot \nabla_v f = \left.\frac{df}{dt}\right|_{\rm coll},$$

where we have split the velocity into a Keplerian, $\underline{\mathrm{v}}_K$, and random component. One typically solves this equation by assuming a form for f, and working out the evolution rate. For the planetesimal case at a particular radius, one usually assumes that the planetesimals start out with a Rayleigh distribution. That is, we assume that planetesimals are distributed according to

$$f(e,i) = 4\frac{\Sigma}{m}\frac{ei}{\langle e^2\rangle\langle i^2\rangle}\exp\left[-\frac{e^2}{\langle e^2\rangle} - \frac{i^2}{\langle i^2\rangle}\right].$$

where Σ is the surface mass density, m is the mass of the planetesimals, e is the orbital eccentricity and i is the inclination. In the *Epicycle approximation*, that is, small eccentricities and inclinations, this is equivalent to

$$f(z,\underline{v}) = \frac{\Omega\Sigma}{2\pi^2 c_r^2 c_z^2 m}\exp\left[-\frac{v_r^2 + 4v_\theta^2}{2c_r^2} - \frac{v_z^2 + \Omega^2 z^2}{2c_z^2}\right]$$

where $2c_r = \langle e^2\rangle v_K^2$, $2c_z^2 = \langle i^2\rangle v_K^2$, $v_K = \sqrt{GM_\odot/r}$, and $\Omega = v_K/r$. In the absence of collisions c_z and c_r would be completely independent, but numerical simulations of the accretion process show that $c_z/c_r = 0.5$.

There are a number of possible contributions to the collision term, but the dominant one that we will consider is stirring from gravitational scattering. The natural scale for the gravitational scattering is set by the escape velocity of the particles, so we define

$$\left(\frac{v_{esc}}{v_0}\right)^2 \equiv 2\theta$$

where θ is the *Safronov number*.

Now the scale height of the disk is $h = v/\Omega$ because the vertical motion is dominated by the gravity of the Sun. Therefore, the density can be gotten from $\Sigma = \rho h$, and we can derive the accretion rate in terms of the Safronov number as

$$\frac{dm}{dt} = \pi s^2 \Sigma\Omega(1 + 2\theta)$$

Note that the flux is now no longer explicitly a function of v, but is constant at a given radius.

If we now assume (as Safronov [37] did), that θ is a constant, we can substitute for the mass of the planetesimal $m = 4/3\pi R_{pl}^3\rho_{pl}$ and obtain the growth rate of the radius as

$$\frac{dR_{\rm pl}}{dt} = \frac{\Sigma\Omega}{4\rho_{pl}}(1 + 2\theta).$$

Turning this into a growth rate gives us

$$t_{\rm grow} = \frac{R_{pl}4\rho_{\rm pl}}{\Sigma\Omega(1 + 2\theta)}.$$

Now let's apply this formula to the terrestrial planet region. Assuming the radius of the Earth $R_\oplus$, the density of the Earth, $\rho_\oplus = 6$ gm/cc, a density of $\Sigma = 10$ gm/cm^2, and $\theta = 3$, one gets a growth timescale of $t_{\rm grow} = 3 \times 10^7$ years. This is a bit long. Furthermore, for constant θ, $\frac{dR}{dt} \propto r^{-3}$, where r is the distance from the Sun. Hence at Saturn's radius of 10 AU, we have

$\tau \sim 3 \times 10^{10}$ years. We could never grow the outer planets or their cores this way.

There are other collisional terms that can contribute to the Boltzmann equation. These include stirring from collisional scattering, damping from inelastic collisions, damping from gas drag and dynamical friction. Of these terms, dynamical friction may play the most significant rôle. One way to think of dynamical friction is that a large body will create a wake of smaller bodies behind it as it moves through them. The gravitational attraction between the large body and this wake will act as a drag to slow the large body down. Another way to think of dynamical friction is as a mechanism to establish equipartition: the large bodies will increase the velocity dispersion of the small bodies. Thought of in this way, one gets that the velocity dispersion should scale as $v^2 \sim 1/m$.

Runaway Growth

We expect a larger planetesimal to have a larger growth rate because: 1) gravitational focusing is greater and 2) dynamical friction on smaller planetesimals decreases their velocities. Hence we will modify the growth timescale given above for the largest planetesimals. This will lead to a phenomenon referred to as runaway growth.

A simple way to look at this is to note that if the escape velocity $v_{\rm esc}^2 \sim G\rho R^2$ grows large, while the typical encounter speed with the small bodies, v_0, remains constant then the growth rate will scale as R^2, rather than R as derived above. This larger growth rate will only be valid for the largest bodies in the region, and so we have the result that the "rich get richer", and large bodies will break away from the general mass distribution.

This breakaway has been seen in numerical simulations such as those in [19] and [36]. The results from [36] are shown in Fig. 3. Note how at the last timestep, the most massive particles have significantly smaller eccentricities than the general population, presumably because they have been damped by dynamical friction. Also note how quickly the large masses have grown. With runaway accretion, the timescales for growth reduce to $10^5 yrs$.

Oligarchic Growth

However, the simulations indicate that there are limits to runaway growth. They finish with a bi-modal distribution of bodies with the larger bodies being separated by about 4 Hill radii. This is referred to as Oligarchic growth, and the final mass of the large bodies is determined by the mass available within the feeding zone.

$$M_{iso} \sim 2\pi r(8R_H)\Sigma,$$

where R_H is the Hill radius (see below) of the body. Once the body has eaten up this mass, it has depleted the feeding zone and becomes isolated. The size

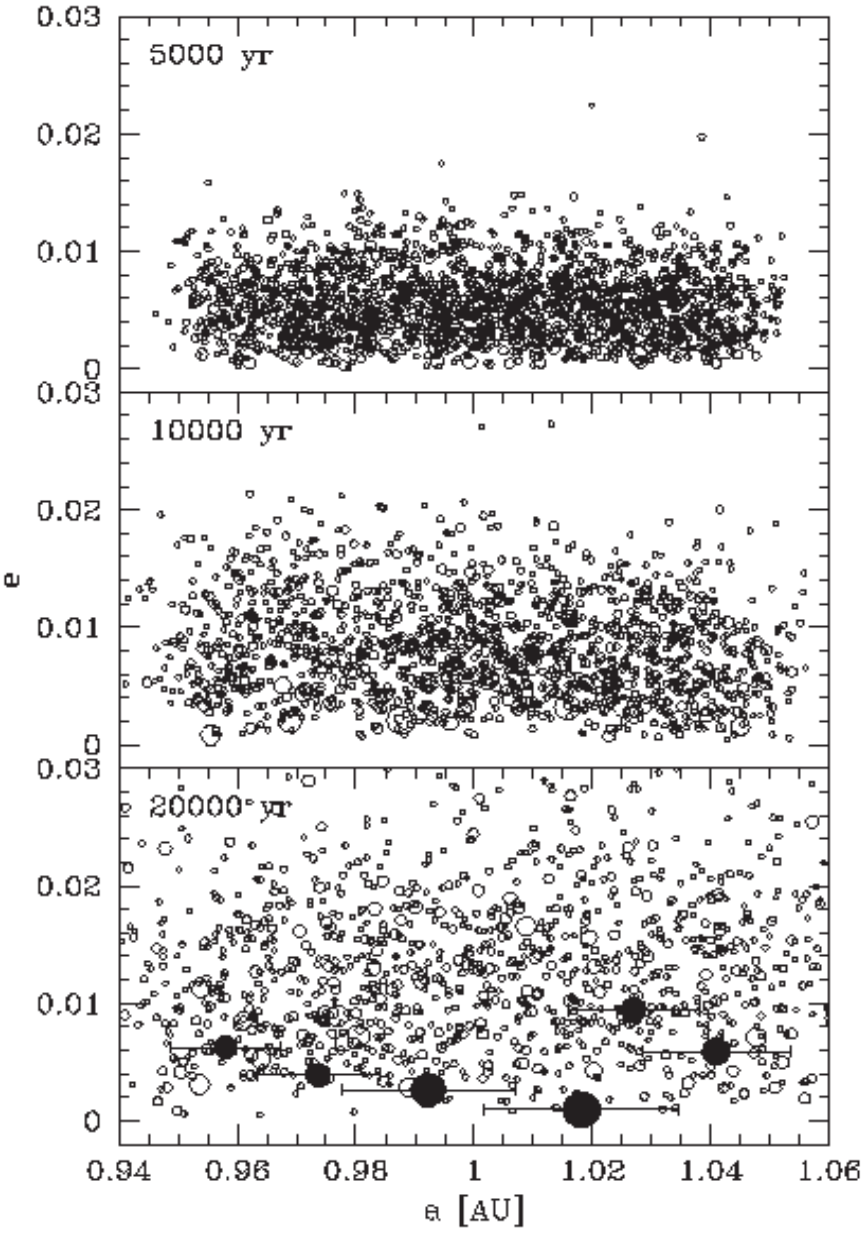

Fig. 3. Evolution of the eccentricities and masses of planetesimals in a simulation of the middle stage from [36]. Eccentricity is plotted against semimajor axis for 3 different times in an N-body simulation of planetesimal accretion. The size of the points is proportional the mass of the particles. For the snapshot at 20000 years, 5 Hill radii are plotted for the most massive bodies

of this feeding zone can be determined by considering the 3-body problem for nearly circular orbits.

The 3-Body Problem and the Hill Sphere

If we scale the distances in the 3-body problem by the *Hill radius*

$$R_H = \left(\frac{m_p}{3M_\odot}\right)^{1/3} a,$$

then the 3-body problem is independent of $m_p/M_\odot$ (if it is small). We can define

$$e_H \equiv \frac{ea}{R_H}, \quad i_H \equiv \frac{ia}{R_H}, \quad b_H \equiv \frac{(a_2 - a_1)}{R_H}.$$

where a_1 and a_2 refer to the semimajor axes of nearby bodies.

In terms of these quantities the 3-body integral of motion is

$$\Gamma \equiv \frac{3}{4} b_H^2 - e_H^2 - i_H^2.$$

A particle can't enter a protoplanet's Hill sphere if $\Gamma > 9$. Therefore, if a planetesimal starts out with a circular orbit, it can not be accreted onto a neighboring protoplanet unless its semimajor axis is such that $b_H < 4$. Hence we derive the size of the feeding zone given above.

Substituting in the value for the Hill radius in the isolation mass, we have:

$$M \sim 16\pi r^2 \left(\frac{M}{3\,\mathrm{M}_\odot}\right)^{1/3} \Sigma$$

or

$$M_{\rm iso} \sim \frac{(16\pi r^2 \Sigma)^{3/2}}{(3\,\mathrm{M}_\odot)^{1/3}}$$

or if we consider solar values

$$M_{\rm iso} \sim 1.6 \times 10^{25} gm (r_{AU}^2 \Sigma_{\rm gm/cc})^{3/2}$$

which gives a few lunar masses for the minimum mass solar nebula at the Earth's radius, and roughly an Earth mass at Jupiter.

Therefore at the end of runaway growth we get proto-planets, not planets. Can we fix this up so that we obtain larger masses? One possibility to get larger masses is to start with more than the minimum solar nebula of material; however, the excess solids will need to be cleared out. This can easily happen in the neighborhood of Jupiter as that planet is large enough to eject the bodies, but in the terrestrial region it is not clear how the extra material would be lost.

This picture also gives an explanation for the existence of the asteroid belt: if the velocity dispersion is increased by an early Jupiter, then growth

rates will remain small. Of course this means having a significant mass in Jupiter before runaway can get started in the asteroid belt. At first this would seem implausible because the growth timescales tend to increase as one moves outward in the solar nebula; however, the existence of an ice line at 5 AU could make Jupiter grow quickly because of the resulting enhanced surface density of solids.

Is there a way to continue growth once the isolation mass has been reached? Two possibilities have been brought up.

- The diffusion of more material into the feeding zone from neighboring regions. This could be caused by the perturbations of neighboring protoplanets.
- The migration of the protoplanet into a new feeding zone via either gravitational torques or gas drag. We will discuss planet migration in general below.

2.4 Late Stage

At the end of the middle stage we have a regular distribution in a of protoplanets with roughly constant mass. There is also a swarm of smaller planetesimals, which may or may not be relevant. Since most of the mass is in the large bodies, dynamical friction is no longer playing a significant rôle in their orbital evolution. These are the initial conditions for the late stage.

This is the few body problem. It isn't analytically tractable, and there are too few bodies to use statistical methods as could be done in the middle stage. There isn't much else to do but to numerically follow the orbits of these bodies for 10^8 years. In such a calculation, there are several things that need to be considered. First, the full N^2 forces among the bodies need to be calculated. This problem is similar to the planetary stability problem, so all the techniques described elsewhere in this book are applicable here. Using these types of codes for the planetesimal case was pioneered by [11]. Secondly, every so often collisions occur. Almost all simulations performed to date assume that the collisions are inelastic, (a reasonable assumption) and all the material in the colliding bodies ends up in the collision remnant. This last assumption is rather wishful, given that detailed SPH simulations of colliding protoplanets show significant ejecta. [10, 3]

Typical results from the simulations are shown in Fig. 4 from [34]. As can be seen from the figure, a few planets of roughly terrestrial-planet mass are indeed formed. However, the general trend is that eccentricities and inclinations are gradually increased by the mutual gravitational interactions. Therefore the eccentricities coming out at the end of the simulations tend to be larger than observed in the Solar System.

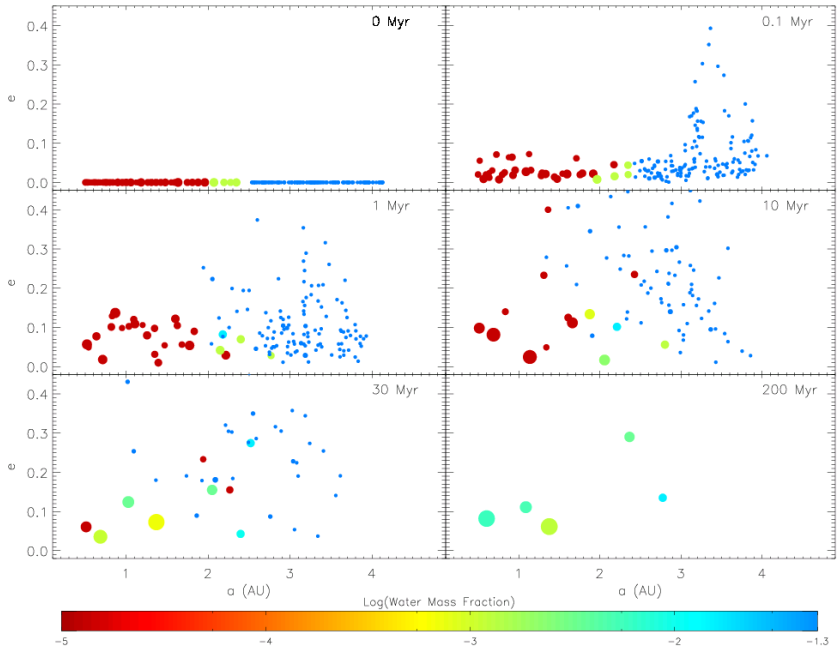

Fig. 4. Eccentricity is plotted against semimajor axes for a simulation of gravitationally interacting and colliding protoplanets during the late stage as they evolve through time. The size of the points is proportional to the cube root of the masses of the protoplanets. The color scale indicates the mass fraction of water *(taken from [34])*

2.5 Outer Planets: Gas Envelope Accretion

The giant planets are enhanced in the heavy elements as well. Even with the uncertainty in the core mass, the mass of elements in the envelope leads to an enhancement in heavy elements over solar abundances of at least 5 for Jupiter. This suggests that the planetesimal accretion model also plays a part for the outer planets, but augmented with a later gas accretion stage.

Therefore we expect the following sequence for the stages of giant planet formation:

- Runaway growth of a solid core which clears out a few R_H.
- A gas envelope starts to accrete.
- Planetesimals interact with gas envelope, depositing energy and metals.

These processes have been modeled in detail by several groups [42, 32]. The way these models work is very similar to stellar interior modeling. One self-consistently solves for density and temperature as a function of radius, assuming that pressure gradients are balanced by gravity. As the radiation

can diffuse out, the temperature structure changes, and one iterates to a new solution. The difference from standard stellar modeling is that the energy comes from gravitational energy deposited by planetesimals, not nuclear burning.

The models are then updated to a later time based on the relevant information. First, mass and energy are injected by accreting planetesimals (estimated from the feeding zone). Orbits of the planetesimals with the the envelope including gas drag, and vaporization rate are calculated to determine where in the envelope the mass and energy are deposited. Second, the envelope contractions due to energy lost due to radiative transfer. As the planet contracts, nebular gas is added to the outer part. The outer radius is the minimum of the Hill radius, and an accretion radius

$$R_a = \frac{GM_p}{c_s^2},$$

where c_s is the sound speed of the gas in the nebula.

An issue that has already been mentioned concerns Uranus and Neptune. What happened to their envelopes? Several solutions have been offered. The gaseous nebula could have been dissipated before the exponential phase of envelope accretion. This brings up a fine tuning problem: why did the timescales match so that Jupiter and Saturn got envelopes, but Uranus and Neptune did not? Wuchterl [42] proposes another solution. A dynamical instability similar to the "κ" mechanism in Cepheids would drive an "overstability" that could remove the outer envelope.

A key result from these accretion calculations is that the growth of the envelope requires millions to tens of millions of years. This is long compared to observational constraints on lifetimes of protoplanetary nebula. Furthermore, an initial nebula with 3 to 4 times the mass of the minimum mass solar nebula is needed to get even these timescales. However, there are obvious improvements to the calculations that may modify these results. These include better opacities, a better treatment of convection, and perhaps a small amount of migration of the planet as it grows.

3 Small Bodies and Planet Formation

Within the planetesimal hypothesis, the small bodies we observe today play a very important rôle. They are essentially the "fossils" of the planet formation process, and as such provide valuable clues to the history of the Solar System. Particularly the comets and the small bodies in the outer Solar System are helpful since they have been relatively unprocessed since their formation.

3.1 Theoretical Considerations

A key theoretical consideration is taken from the stability of the orbits of small bodies. Long term orbit integrations been done for small bodies starting

with circular orbits at a variety of semimajor axes [17]. These have shown that only bodies that start their orbits with semimajor axes between 2 and 3.5 AU or beyond 35 AU are able to remain on stable orbits for the age of the Solar System. Hence it is only Main Belt Asteroids and Kuiper belt objects that are on stable orbits and can be in their original location. All other small bodies must be injected into their current orbits from some longer-lasting reservoir. Furthermore the delivery mechanism from that reservoir must operate on a timescale of billions of years. If it were any shorter, the reservoir would be depleted.

3.2 Near Earth Asteroids

Objects with perihelia less than 1.3 AU and aphelia greater than 0.983 AU are classified as Near Earth Asteroids (NEAs). The orbital lifetimes of these objects are only tens of millions of years so they must originate from another source. Possibilities include the Main Belt and extinct comets. Some of the bodies may also be ejecta from the Moon or Mars when they are impacted by other asteroids. Of course these bodies also impact the Earth and so are very relevant to Earth's geological and biological history. Since they are also the source of meteorites, these bodies are invaluable for determining the composition and early thermal history of the Solar System.

Determining their complete orbital distribution is very model dependent because the observations are so incomplete. Bottke [9] has developed a sophisticated procedure for doing this. First orbits of small bodies are integrated starting from a number of proposed source regions. These are different regions of the main belt and the short period comets. A best fit model is then created by varying the relative contributions of these sources in order to match the distribution of the known orbits. They find that 61% of NEAs come from the inner Asteroid belt, 24% from the middle, and 8% from the outer belt. Extinct comets provide 6% of the NEAs.

3.3 Main Belt Asteroids

Most of the known asteroids are in the Main Belt with semi-major axes between 2.1 AU and 3.3 AU. Their distribution in semi-major axis is remarkable because of a number of gaps referred to as Kirkwood gaps. These coincide with mean motion resonances with Jupiter indicating that Jupiter has played a significant rôle in the dynamical evolution of the belt.

A detailed study of the orbital elements show that there are a large number of families that are concentrated around particular orbital elements. Furthermore [18] has shown that these families have similar colors indicating that their composition is identical. This leads to the hypothesis that these families are created by collisions or break-ups of larger bodies, and generally indicates that there is a significant amount of collisional evolution in the Asteroid Belt.

3.4 Centaurs and Trans-Neptunian Objects

Between the orbits of Jupiter and Neptune are the Centaur objects. There are only a handful of these known. They typically have orbits that cross the semi-major axis of one or more planets, therefore they have a very short dynamical lifetime. Orbital integrations indicate that they are in transit between the Kuiper belt and short-period comets [21]. The Kuiper belt refers to objects in low inclination orbits with semi-major axes beyond Neptune. Their dynamical lifetimes are very long so they are presumably remnants from the formation of the Solar System. That is, they are the remaining "planetesimals" discussed above. They are also postulated to be the source of the short period comets [33].

One interesting question concerning the Kuiper Belt objects is the status of Pluto. Pluto has an orbit similar to a number of the Kuiper Belt objects leading to the speculation that it is simply the largest body of this population. Although there is a considerable gap between Pluto, Chiron and the largest Kuiper belt object, models of the mass distribution predict roughly one Pluto size body.

3.5 Comets

Comets are divided into a number of classes based on their orbital elements. There are several ways to make the division into classes; one way is as follows. Jupiter family comets have periods less than 20 years; Halley family comets have periods between 20 and 200 years; long period comets have periods greater than 200 years but semi-major axes less than 10,000 AU, and Oort cloud comets have semi-major axes greater than 10,000 AU. There may be an inner Oort cloud with smaller semi-major axes. Because comets have spent most of their lifetime far from the Sun, they are relatively pristine objects with connections to Solar System formation.

Short Period Comets

Short period comets have inclinations such that they are concentrated in the ecliptic plane. As shown in Fig. 5, there are several peaks in the semi-major axis distribution with the largest at 3 AU. The peaks are more pronounced in the aphelion distribution where they correspond to the semi-major axes of the outer planets. The argument of perihelion distribution peaks at 0 and 180 degrees, meaning that as the comets go through aphelion they are in the plane of the planets. Because this orbital configuration favors close approaches to Jupiter and Saturn, the short period comets are dynamically controlled by the giant planets

This strong interaction with the outer planets means that their dynamical lifetime is very short, of order 1 million years. They are not in a stable distribution, and must be injected into their current orbits from some larger reservoir

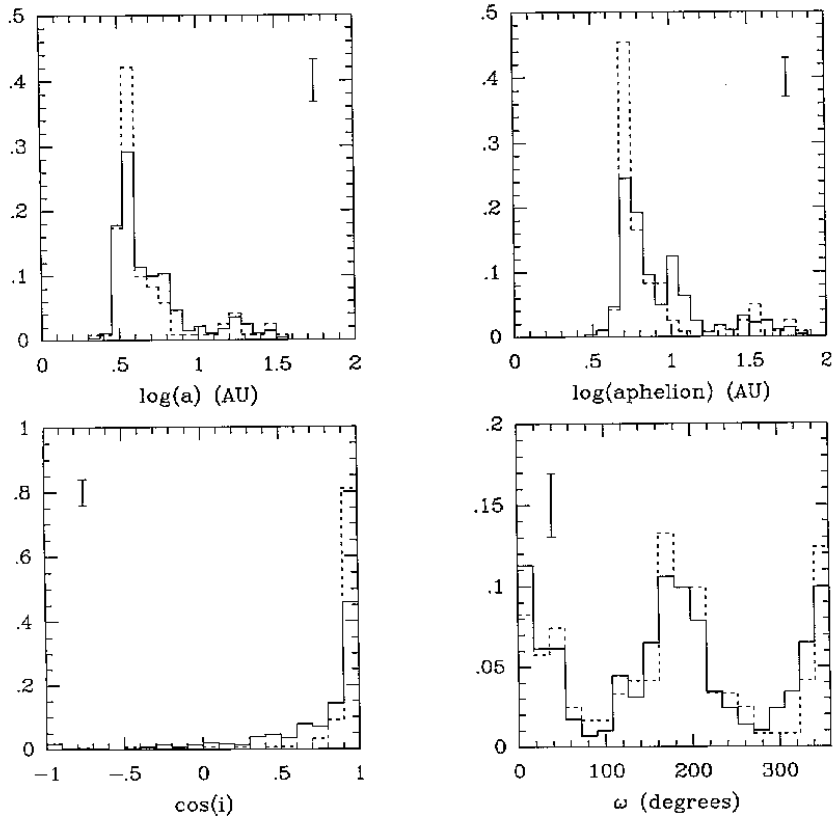

Fig. 5. The distribution of the short period comets in semi-major axis, aphelion, inclination and argument of perihelion is plotted *(taken from [33])*

that can survive the age of the Solar System. It was originally thought that the long period comets could be this source. However, simulations [12] show that such a scenario is incompatible with their current orbital distribution. First, such a source would produce a much broader inclination distribution, and second the argument of perihelion would peak at 90 and 270 degrees due to a secular evolution of the comet orbits. The conclusion is that a low inclination belt of bodies beyond Neptune could produce short period comets with a much better match to the observed orbital distribution. The first of these Kuiper belt objects was discovered several years after these simulations were published.

Long Period Comets and Oort Cloud

The distribution of longer period comets is remarkable because of a sharp peak at a semi-major axis of about 20,000 AU. The inclinations are isotropic

and the perihelion distribution seems uniform. The energy distribution is consistent with all comets being marginally bound within the observational errors. The implication is that the observed comets are the very eccentric subsample of comets in a large reservoir at $a = 10^4$ AU: the Oort cloud.

The likely origin of the Oort cloud is also a disk in the outer Solar System [12]. The scenario is as follows, and a typical trajectory is shown in Fig. 6. Planetesimals between the planets in the outer Solar System are initially perturbed by interactions with those planets. If the planetesimals are scattered outward, the planetary perturbations tend to cause random walks in a at fixed perihelion and inclination. They therefore diffuse outward in semi-major axis until the orbits are large enough that Galactic tides and passing stars start to influence their orbits. The tides and passing stars tend to change the angular momentum without changing energy. The change in angular momentum will lift the perihelion out of the planetary region, and the comets will start to evolve at constant semi-major axis, and the orbits will become isotropic. Of course, those same tides will bring comets back into high eccentricity orbits which gives us the comets we see today.

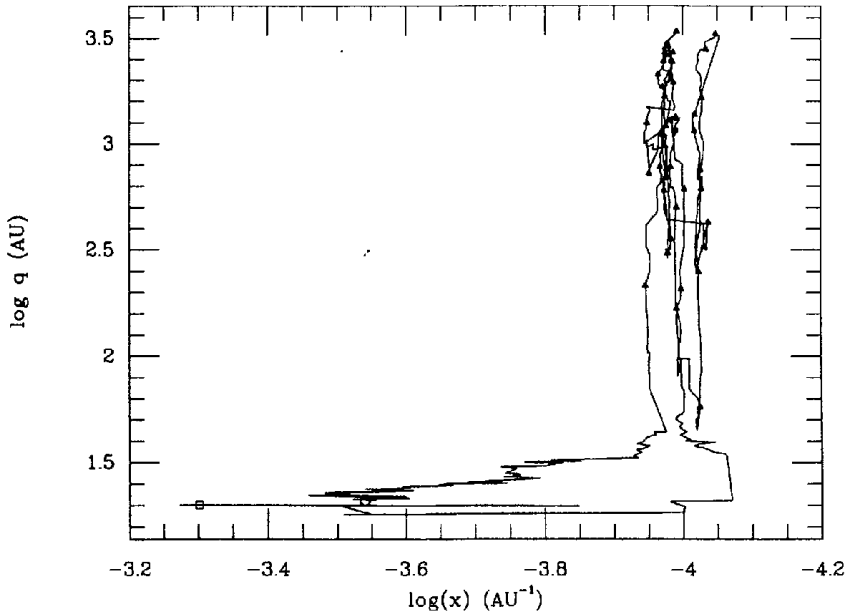

Fig. 6. Evolution in $x \equiv 1/a$ and q of a typical Oort cloud comet from [12]. The comet first diffuses in a because of perturbations by the planets. When a becomes large, q gets increased because of galactic tides and stellar encounters

3.6 Unifying Small Body and Planet Origins

In summary the planetesimal hypothesis gives us a mechanism to link together all the small body populations in the context of planet formation. During the middle stage we have many icy and rocky bodies. As the outer planets grow via accretion, they scatter most of the original planetesimals out of the region. Planetesimals scattered to higher semi-major axis slowly diffuse outward until tides lift them out of the planetary region and they become Oort cloud comets. Planetesimals outside Neptune do not feel strong enough planetary perturbations to diffuse rapidly and remain as the Kuiper belt objects. They subsequently slowly supply the short period comets. The only other planetesimals that remain are the Main Belt asteroids, which become a source for the Near Earth Asteroids.

4 Planet Formation: the Fragmentation Model

The model outlined in the previous two sections has had great success in explaining the characteristics of our own Solar System; however, observations of planets around other stars, and observations of protoplanetary disks have raised some challenges. The first thing that these observations indicate is that planet formation (or at least, formation of giant planets) is fairly ubiquitous. Roughly 10% of the surveyed stars have had planets discovered around them [26], and when considering selection effects due to inclinations and masses, the real fraction must be higher than this. This indicates that whatever the formation mechanism, it must be efficient.

The second observational evidence impacting our theories of planet formation are the studies of star forming regions. On the one hand, large scale millimeter surveys indicate that most stars form in OB associations [2]. However, studies of the circumstellar disks in these regions indicate that the lifetimes of those disk are only .1 million years. Presumably the strong UV flux from the nearby bright stars quickly ablate the disks.

Since the formation of the gas giants must happen before the disk is dissipated, and the core-accretion model described above generally happens on a longer timescale than typical disk lifetimes, we must conclude that either planet formation is rare, or there is another quicker mechanism for it to happen. Combining this argument with the fact that giant planets are relatively common pushes us to consider other theories. There are a couple of other issues with the core-accretion model, such as gap formation (not included in the above calculations) prematurely terminating accretion, and ice giants have a hard time forming in the outer solar system because the escape velocity from the forming cores becomes comparable to the orbital velocity so that most material gets scattered out instead of accreting. This could be solved by forming Uranus in the Jupiter-Saturn region [40].

Gravitationally driven fragmentation of a gas disk is an idea that goes back to Kuiper [20], but has recently been revived [5, 6, 8]. The idea is that if the disk is massive enough, the disk can fragment gravitationally in a short time. The condition for local instability to gravitational collapse is the Toomre criterion:

$$Q = \frac{c_s \kappa}{\pi G \Sigma} < 1,$$

where κ is the epicyclic frequency and Σ is the surface density. Furthermore, numerical simulations show that $1 < Q < 2$ allows global instabilities, and these may fragment. If we take the temperature constraints above and use densities corresponding to the "minimum solar nebula", we obtain a protosolar disk with a mass of $.01\,\mathrm{M}_\odot$. This would imply $Q > 2$ everywhere, so we would need more mass than this to drive such an instability. The mass of disks around other stars can be obtained from radiative transfer modeling, however the gas/dust ratio must be assumed. If we assume an interstellar gas/dust ratio of ~ 100 then the implied disk mass is $M \sim 0.05$ to $0.1\,\mathrm{M}_\odot$. With typical temperatures this will give a $Q \sim 1.5$ for $R > 5$ AU. This is not that much more massive than the masses needed for envelope accretion models.

Whether the spiral arms that develop within the unstable disks actually collapse into protoplanets is an outstanding issue, the answer to which is limited by the quality of simulations that can be performed. Since the modeling has to follow scales of many AU down to the Hill sphere of the planet, the dynamic range is quite challenging, and most simulations have been limited by their spatial resolution.

Recently simulations with Smooth Particle Hydrodynamics (SPH) have been attempted [27, 28]. These simulations fixed the temperature profile to that obtained from static radiative transfer simulations [4, 6]. If the disk was massive enough such that $Q \sim 1.4$, then the spiral arms did indeed fragment into protoplanets. Furthermore the masses and eccentricities of the formed planets are similar to the masses of observed extra solar planets.

The fixed temperature profile is a significant limitation to these simulations. Clearly as the gas gets denser, radiative diffusion out of the clumps will become more difficult. Other simulations with constant cooling time show significant clump formation only when the cooling time is of order the orbital time [13]. Modifying the equation of state so that it becomes adiabatic after the clumps become sufficiently dense still results in protoplanet formation, but the issue will not be fully resolved until there are high resolution simulations that include radiative transfer.

4.1 Long Term Evolution

The long term evolution of planets formed by disk fragmentation is significantly different than the standard model. This is because in the fragmentation model a handful of planets will form at the same radius from a single spiral

arm, whereas in the core-accretion model, it is assumed that the planets are well separated when they form. The result is that there are many mergers and scatterings during the evolution. This can be exacerbated via "seeded" planet formation [1]. Once a single giant planet forms, it excites strong spiral waves at its Linblad resonances which in turn fragment into more giant planets.

This seeding may also have implications for terrestrial planet formation. It could be that the perturbations from the gas giants are needed to get the terrestrial protoplanets on crossing orbits so that they collide to build larger objects. Furthermore if we are to explain the asteroid belt by planet formation being suppressed by the presence of Jupiter, the gas giant needs to have formed quickly before runaway growth in the asteroid belt can get going. This is natural in the fragmentation model where the giant planets form quickly.

After hundreds of years of evolution the distribution of planets has the following characteristics.

1. The eccentricity distribution matches the "non-hot Jupiter" (that is $a >$ .25 AU) distribution observed in the extrasolar planets. Eccentricities can be as high as 0.7.
2. The entire mass range from "super-Jupiters" down to Saturn mass objects are generated, depending on the initial mass of the nebula.
3. Because of the mergers, we expect the obliquity distribution to have some outliers with very large obliquities as that of Uranus.
4. There are high accretion rates onto the star. The strong interactions among the planets and between the planets and the gas drives a significant amount of mass inwards. This a mechanism to clear out the disk quickly. Episodic flows onto the star could explain the FU Orionis outburst phenomenon.

4.2 Arguments against Fragmentation

The fragmentation model is attractive because it neatly solves the timescale problem. However, there are a number of problems which make the core-accretion model more favorable. First there is the evidence that the outer planets contain cores. Observations of Jupiter do not rule out a zero mass core mostly due to uncertainties in the equation of state of hydrogen at very high densities and temperatures. Nevertheless, a zero mass core is not the favored fit. In the fragmentation model one would not expect to see an enhanced core since high Z material added after the collapse should be thoroughly mixed in the envelope. Settling could happen if the planet was non-turbulent (i.e. non-convective); there are disagreements as to whether this is the case. However, Haghighipour and Boss [16] show that small solids of size centimeters to meters can migrate quickly (1000 years) to the pressure maximum and thereby enhance the abundances of the giant planets in the

fragmentation model. Note that even the envelopes of gas giants are enriched, implying that we need the solid condensation picture anyway.

The ice giants, Neptune and Uranus are problems in both formation models; they don't naturally occur in the fragmentation model. However, they could be formed via fragmentation as Jupiter and Saturn are, and then their envelopes get ablated away via EUV radiation from nearby massive stars. Jupiter and Saturn keep their envelopes because they are more deeply embedded in the protoplanetary nebula.

Whether the fragmentation model works depends significantly on the details of the gas dynamics and radiative transfer. This is difficult to calculate, and a lot of detailed numerical work will be needed to confirm or rule out this model.

5 Planet Migration

The "hot Jupiters" that comprise about one quarter of the extrasolar planets observed pose a problem in either of the planet formation scenarios discussed above. They can not be formed in the core accretion model because the ices (and indeed many rocky compounds) don't condense at the temperatures seen so close to the star. In the fragmentation model Q tends to be high near the star both because the temperature, and hence c_s is high, and the shear, i.e. κ, is large. The obvious explanation is that planets form further out where the theories predict they do, but then migrate significantly due to their interaction with a disk.

The interaction between the planet and the disk can be calculated using spiral density wave theory [15]. In this theory, the coupling to the disk is expected to be strongest at resonances. Of particular importance are the *Linblad resonances:*

$$m(\Omega - \Omega_{ps}) = \pm\kappa$$

where plus refers to the inner and minus refers to the outer Linblad resonance and m is the azimuthal wave number ($m = 1$ for a one arm spiral). Consider a frame rotating with the Keplerian velocity of the outer Linblad resonance (OLR). If the gas moves in an elliptical orbit, it is equivalent to an epicycle in this frame. Once per epicycle, the pattern comes by, so it is resonantly perturbed by always seeing the perturbation at the same point in its epicycle and a secular transfer of energy and angular momentum can happen.

5.1 Type I Migration

The inner Linblad resonance (ILR) increases the angular momentum of the planet, while the OLR decreases the angular momentum. However, these two effects don't quite cancel out, and there is a net loss of angular momentum to the disk by the planet [41]. This is referred to as Type I migration.

The migration timescale can be calculated from the linear theory and is:

$$\tau = \frac{1}{C_a \Omega} \left(\frac{\mathrm{M}_\odot}{M_p} \right) \left(\frac{\mathrm{M}_\odot}{\Sigma r^2} \right) \left(\frac{c_s}{r\Omega} \right)^2$$

where C_a is of order unity. Note that the migration rate scales linearly with the planet mass, M_p, and the disk mass (proportional to Σ), and inversely with the temperature of the disk. The scaling with planet mass assumes that the strength of the perturbation in the disk is proportional to the mass of the planet. However, numerical simulations [30] indicate that this perturbation may saturate, and then the migration rate becomes independent of planet mass.

This phenomenon is fundamentally equivalent to dynamical friction: the gravitational interaction of the planet produces a wake, and the wake gravitationally back-reacts on the planet.

5.2 Type II Migration

If the planet is massive enough then a gap forms in the disk. This occurs if the torquing of the disk by the planet exceeds the "viscous torquing" of the disk by itself. That is the planet pushes material away from it faster than disk viscous forces can bring material in.

In contrast to Type I, the planet is now fixed in the disk, and evolves along with it. Therefore the migration rate is purely determined by disk properties:

$$\dot{r} = c_{\mathrm{II}} \alpha \left(\frac{c_s}{r\Omega} \right)^2 r\Omega$$

where the kinematic viscosity of the disk is parameterized by $\nu = \alpha c_s H$ where H is the mixing length, and c_{II} is a parameter of order unity. These timescales tend to be much longer than the timescales associated with type I migration.

5.3 Stopping Migration

If we are to end up with hot Jupiters, we need to have migration terminate before the planet falls into star. In type I, the ILR gets lost because there is no more disk in which it can be raised. Since the ILR contributes an outward torque, this means the migration should speed up onto the star. If the OLR can be lost, then migration will halt. Perhaps the X wind region around the star could clear out a large enough region around the star for this to happen.

Tidal interaction with the star can also stop migration [22], but the star must in general be spinning faster than the orbital period. Eventually the star should slow down, and then the migration would continue.

On the other hand migration may only stop when the disk disappears. Planets continue to form and migrate, and the ones we observe are the ones

that happen to be left when the gas disk disappears. Such a migration probably did not happen in our own solar system for two reasons. First the asteroid belt seems to be chemically stratified in the radial direction. One would expect radial migration of Jupiters to stir it up and thoroughly mix the asteroid types. Secondly, it would be difficult to form the terrestrial planets if the planetesimals were scattered away by a migrating Jupiter.

5.4 Migration and Disk Fragmentation

As pointed out above, the migration rate depends on the disk mass. Therefore, the migration rate indirectly depends on Q the stability criterion. Simulations show [24] that for typical mass planets (of order a Jupiter mass) disks with lower Q, (less than 7) will have gaps, but higher Q disks will experience type I migration. Therefore, in disks unstable to fragmentation, migration would naively be very slow. However, as discussed above, multiple planets can form and significant "migration" can occur due to the interaction between these planets.

References

1. P.J. Armitage and B.M.S. Hansen: Early planet formation as a trigger for further planet formation. Nature **402**, 663 (1999) 212
2. J. Bally, L. Testi, A. Sargent and J. Carlstrom: Disk Mass Limits and Lifetimes of Externally Irradiated Young Stellar Objects Embedded in the Orion Nebula. AJ **116**, 854 (1998) 210
3. W. Benz, W.L. Slattery and A.G.W. Cameron: The origin of the moon and the single-impact hypothesis I. Icarus **66**, 515 (1986) 203
4. A.P. Boss: Evolution of the Solar Nebula. III. Protoplanetary Disks Undergoing Mass Accretion. ApJ **469**, 906 (1996) 211
5. A.P. Boss: Giant planet formation by gravitational instability. Science **276**, 1836 (1997) 211
6. A.P. Boss: Evolution of the Solar Nebula. IV. Giant Gaseous Protoplanet Forma tion. ApJ **503**, 923 (1998) 211
7. A.P. Boss: Temperatures in Protoplanetary Disks. Annual Review of Earth and Planetary Sciences **26**, 53 (1998) 191, 192
8. A.P. Boss: Gas Giant Protoplanet Formation: Disk Instability Models with Thermodynamics and Radiative Transfer. ApJ **563**, 367 (2001) 211
9. W.F. Bottke, A. Morbidelli, R. Jedicke, J. Petit, H.F. Levison, P. Michel and T.S. Metcalfe: Debiased Orbital and Absolute Magnitude Distribution of the Near-Earth Objects. Icarus **156**, 399 (2002) 206
10. R.M. Canup and L.W. Esposito: Accretion of the Moon from an Impact-Generated Disk. Icarus **119**, 427 (1996) 203
11. J.E. Chambers and G.W. Wetherill: Making the Terrestrial Planets: N-Body Integrations of Planetary Embryos in Three Dimensions. Icarus **136**, 304 (1998) 203

12. M. Duncan, T. Quinn and S. Tremaine: The formation and extent of the solar system comet cloud. AJ **94**, 1330 (1987) 208, 209
13. C.F. Gammie: Nonlinear Outcome of Gravitational Instability in Cooling, Gaseous Disks. ApJ **553**, 174 (2001) 211
14. P. Goldreich and W.R. Ward: The Formation of Planetesimals. ApJ **183**, 1051 (1973) 197
15. P. Goldreich and S. Tremaine: Disk-satellite interactions. ApJ **241**, 425 (1980) 213
16. N. Haghighipour and A.P. Boss: On Pressure Gradients and Rapid Migration of Solids in a Nonuniform Solar Nebula. ApJ **583**, 996 (2003) 212
17. M.J. Holman and J. Wisdom: Dynamical stability in the outer solar system and the delivery of short period comets. AJ **105**, 1987 (1993) 206
18. Ž. Ivezić, R.H. Lupton, M. Jurić, S. Tabachnik, T. Quinn, J.E. Gunn, G.R. Knapp, C.M. Rockosi and J. Brinkmann: Color Confirmation of Asteroid Families. AJ **124**, 2943 (2002) 206
19. E. Kokubo and S. Ida: Formation of Protoplanets from Planetesimals in the Solar Nebula. Icarus **143**, 15 (2000) 200
20. G.P. Kuiper: On the Origin of the Solar System. In: *Proceedings of a topical symposium, commemorating the 50th anniversary of the Yerkes Observatory and half a century of progress in astrophysics* ed by J.A. Hynek (New York: McGraw-Hill 1951), p. 357 211
21. H.F. Levison and M.J. Duncan: From the Kuiper Belt to Jupiter-Family Comets: The Spatial Distribution of Ecliptic Comets. Icarus **127**, 13 (1997) 207
22. D.N.C Lin, P. Bodenheimer and D.C. Richardson: Orbital migration of the planetary companion of 51 Pegasi to its present location. Nature **380**, 606 (1996) 214
23. J.J. Lissauer: Planet formation. Annual Review of Astronomy and Astrophysics **31**, 31 (1993) 188
24. G. Lufkin, T. Quinn, J. Wadsley, J. Stadel and F. Governato: Simulations of gaseous disc-embedded planet interaction. MNRAS **347**, 421 (2004) 215
25. D. Lynden-Bell and J.E. Pringle: The evolution of viscous discs and the origin of the nebular variables. MNRAS **168**, 603 (1974) 192
26. G.W. Marcy and R.P. Butler: Planets Orbiting Other Suns. PASP **112**, 137 (2000) 210
27. L. Mayer, T. Quinn, J. Wadsley and J. Stadel: Formation of Giant Planets by Fragmentation of Protoplanetary Disks. Science **298**, 1756 (2002) 211
28. L. Mayer, T. Quinn, J. Wadsley and J. Stadel: The Evolution of Gravitationally Unstable Protoplanetary Disks: Fragmentation and Possible Giant Planet Formation. ApJ **609**, 1045 (2004) 211
29. M. Mayor, D. Queloz, G. Marcy, P. Butler, R. Noyes, S. Korzennik, M. Krockenberger, P. Nisenson, T. Brown, T. Kennelly, C. Rowland, S. Horner, G. Burki, M. Burnet and M. Kunzli: 51 Pegasi. IAU Circ. **6251**, 1 (1995) 187
30. R.P. Nelson, J.C.B. Papaloizou, F. Masset and W. Kley: The migration and growth of protoplanets in protostellar discs. MNRAS **318**, 18 (2000) 214
31. J.B. Pollack, D. Hollenbach, S. Beckwith, D.P. Simonelli, T. Roush and W. Fong: Composition and radiative properties of grains in molecular clouds and accretion disks. ApJ **421**, 615 (1994) 190
32. J.B. Pollack, O. Hubickyj, P. Bodenheimer, J.J. Lissauer, M. Podolak and Y. Greenzweig: Formation of the Giant Planets by Concurrent Accretion of Solids and Gas. Icarus **124**, 62 (1996) 204

33. T. Quinn, S. Tremaine and M. Duncan: Planetary perturbations and the origins of short-period comets. ApJ **355**, 667 (1990) 207, 208
34. S.N. Raymond, T. Quinn and J.I. Lunine: Making other earths: dynamical simulations of terrestrial planet formation and water delivery. Icarus **168**, 1 (2004) 203, 204
35. D.C. Richardson: A Self-Consistent Numerical Treatment of Fractal Aggregate Dynamics. Icarus *115*, 320 (1995) 194
36. D.C. Richardson, T. Quinn, J. Stadel and G. Lake: Large-Scale N-body Simulations of Planetesimal Dynamics. Icarus **143**, 45 (2000) 200, 201
37. V.S. Safronov: *Evoliutsiaa doplanetnogo oblaka*, (1969) 198, 199
38. D.J. Stevenson and J.I. Lunine: Rapid formation of Jupiter by diffuse redistribution of water vapor in the solar nebula. Icarus **75**, 146 (1988) 193
39. P. Tanga, P. Michel and D.C. Richardson: Planetesimal clusters in a Keplerian disk. I. gravitational evolution. A&A **395**, 613 (2002) 197
40. E.W. Thommes, M.J. Duncan and H.F. Levison: The Formation of Uranus and Neptune among Jupiter and Saturn. AJ **123**, 2862 (2002) 210
41. W.R. Ward: Density waves in the solar nebula - Differential Lindblad torque. Icarus **67**, 164 (1986) 213
42. G. Wuchterl, T. Guillot and J.J. Lissauer: Giant Planet Formation. Protostars and Planets **IV**, 1081 (2000) 204, 205

Part II

Extrasolar Planetary Systems

Planet Formation

Thomas Quinn

Astronomy Department, University of Washington, Box 351580, Seattle WA 98195-1580, USA
(trq@astro.washington.edu)

1 Introduction

Motivating the study of planet formation is not difficult for any curious audience. One of the fundamental human questions is that of origins: "where did I come from?". Breaking this down into constituents produces a series of questions. How did the Universe begin? How did stars form? *How did planets form?* How did life begin? How did intelligent life develop? It is therefore obvious how the origins of planetary systems is a central issue to origins in general.

Theories for the formation of the Solar System have been around since Laplace and Kant, but in the past 10 years, theoretical research has become more focused for several reasons. These include:

1. Observations of planetary systems other than our own. The discovery of a Jupiter size planet in a close orbit around 51 Peg [29] gave us the first clue about the variety of planetary systems. Since then over 100 extra-solar planets have been discovered, and we can now talk statistically about the properties of these systems.
2. Surveys of planet forming regions. Infra-red and millimeter observations of the interstellar medium have allowed us to peer into the dusty star forming regions and get constraints on the conditions there. High resolution observations have even allowed us to estimate protoplanetary disk masses and temperatures.
3. Predictive theories of planet formation pushed forward by computational advances. The physics of planet formation is complex and non-linear. Numerical simulation has been necessary to fully model it. The possibility of doing computationally intensive calculations has had particular impact on the long term dynamics of planetary systems both for small bodies and planets.

1.1 Major Issues for Solar System Formation

Theories of planet formation need to come to grips with a number of issues which are summarized here.

T. Quinn: *Planet Formation*, Lect. Notes Phys. **683**, 187–217 (2005)
www.springerlink.com

- What caused it to form?
  We certainly need to understand this if we are to know anything about planet formation in other parts of the galaxy or in other parts of the Universe. Since this is intimately connected with star formation, this question not only has implications for planet formation, but also for the entire history of the visible Universe.
- What was the nebula really like in terms of pressure, temperature, and the timescales for processes?
  For planet and star formation theories, this is the question of initial conditions. What are the properties of the ISM in which planets form?
- What are the processes involved in planet formation?
  We are all familiar with the Kant nebular hypothesis: a disk of gas fragments into disk of planets, but what are the actual physical processes? Clearly gravity as the dominant astrophysical force is significant; however, a variety of other forces play rôles as well. These range from magnetic fields and hydrodynamic shocks on large scales down to the material properties of dust particles.
- What is the rôle of condensation?
  The planets (particularly the terrestrial planets) are not of cosmological abundance. How did they get enhanced? This is the primary evidence that non-gravitational forces played a significant part in planet formation.
- How Universal are the processes?
  Here we introduce issues that relate to the uniqueness of our own solar system and the ubiquity of planets around other stars. The possible questions are numerous: Are the terrestrial planets always near the Star? What is the rôle of the Giant Planets? How do the small bodies, Comets and Asteroids, fit in? What determines the planets' spacing? What determines their orbits? Are all planetary systems similar? How are compositions and atmospheres determined?

2 Planet Formation: the Standard Model

Following is an overview of the Standard, or planetesimal, model broken down into stages. We will then describe each stage in more detail. For a more complete review see [23].

1. Initial stage:
   This is the connection to star formation: one assumes a circumstellar disk with solar abundance around the Sun. It gives a natural explanation for the coplanar motions of the planets. The total mass of the nebula should also give a rough estimate of the mass of the planets. As the formation of this disk has more to do with star formation and the collapse of the protostar within the molecular cloud, it will not be covered further here.

2. Early Stage:
   As the disk cools (because the infall slows down) materials can condense. One expects (for solar composition) silicates and iron compounds to condense first. In the outer regions (asteroid belt and beyond) ices can form. This is a natural explanation for the compositions of the planets. Rocky bodies like the terrestrial planets will form close to the parent star where only silicates and iron can condense, while the gas giants and ice giants have a much larger fraction of light gases.
3. Middle Stage:
   At this stage the condensed dust collects in km-sized objects via a gravitational instability (see below). We will call these bodies *planetesimals*. They will interact with each other gravitationally, and occasionally collide, building up into larger bodies. This provides a natural connection between the planets and the small bodies. Planetesimals that do not accrete onto larger bodies remain and provide the origin of comets and asteroids.
4. Late Stage:
   The middle stage ends with dozens of lunar-size bodies, which we refer to as protoplanets, in the terrestrial region. These interact over a long time scale, occasionally colliding to form large bodies, and clearing the region of the small bodies. At the end of this stage, we have a few bodies whose masses and orbital separations match those of the terrestrial planets. The rotation of these planets can also be explained via the angular momentum imparted during the last major impact.
   In the outer regions, the end-state of this process produces a core of tens of Earth masses onto which gas and cometary material accretes. Hence the gas giants have a core-envelope structure, and the ice giants are presumably cores that did not form in time to accrete any nebular gas.

2.1 Initial Stage: the Solar Nebula

Protoplanetary Disk Structure

Although protoplanetary disks tend to be modeled as smooth and axisymmetric they are probably not uniform, having both radial (i.e. gaps) and and vertical structure. They are heated on the surface by stellar radiation, and internally by accretion energy. Observations of disk disk spectral energy distributions (SEDs) indicate that they are flared so that a significant portion can be heated by the star. Observations also constrain the size to be a few to 100 AU, and the mass of the disk to be 0.001 to 0.1 $M_\odot$. However, the observations tend to be of the dust emission, and one has to assume a dust to gas ratio in order to estimate the total masses of these disks. Nevertheless, these mass estimates nicely match the "minimum mass solar nebula" that can be gotten from our own Solar System in the following manner. Enhance

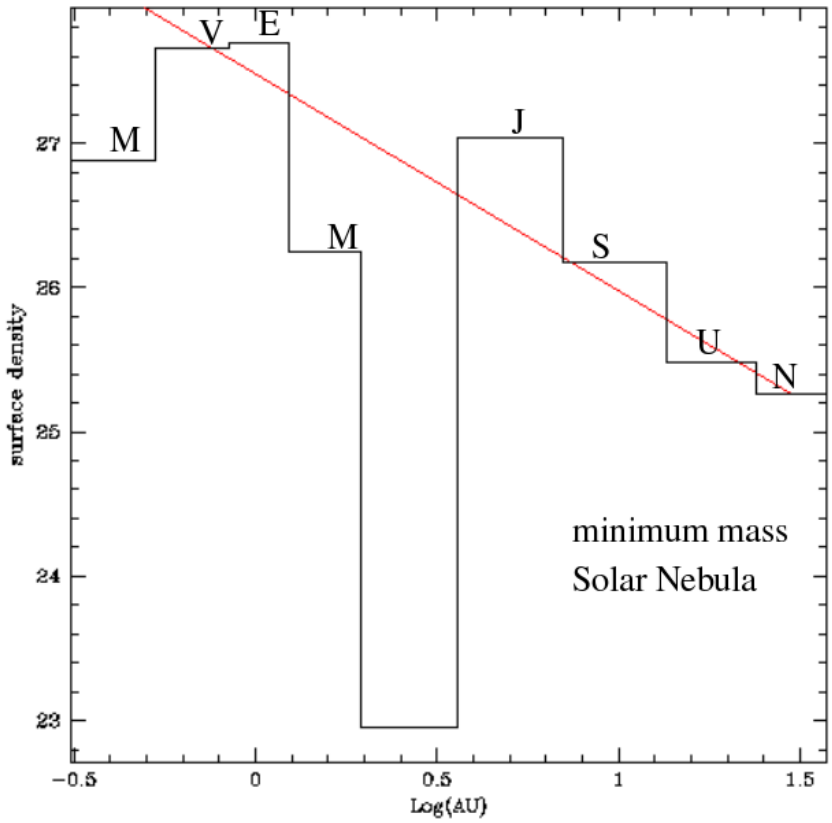

Fig. 1. Mass surface density of the planets after being enhanced to solar abundance *(the line is an $r^{-3/2}$ power law)*

all the planets' masses so that their composition is solar, then spread that mass in annuli with widths given by the separations in semi-major axis. The result is shown in Fig. 1 and the total mass obtained is about 0.01 solar masses. The disks are optically thick, and the opacity is determined mainly by dust. Because the dust properties are a strong function of composition, and the dust composition is determined by what can condense, the opacity is a strong function of temperature. [31] Hence radiative transport is important for determining disk temperatures. Of particular note is that the possible strong drops in temperature as one moves outward in the disk due to opacity changes could lead to convective instabilities. For example there is a large drop in opacity at 500 K due to the evaporation of "refractory organics".

Temperatures

As well as the instabilities mentioned above, MHD instabilities, in particular, the magneto-rotational instability can develop if the gas becomes ionized. Therefore it is important to understand the temperature structure of the protoplanetary disks. For a fuller review of constraints on disk temperature profiles see [7]. The observed SEDs give us a constraint on the surface temperatures of the disk. This is done by fitting power law disk models, and it is typically found that $T_s \propto r^{-1/2}$, with $T_{1\mathrm{AU}} \sim 150\mathrm{K}$. The model temperatures (at least at $r < 1\mathrm{AU}$) seem to be constant over $\sim 10^6$ years, that is, over most of the time that planets take to form. However, it should be noted that the midplane temperatures are higher.

Temperatures in the midplane can be understood theoretically from viscous accretion models. In these models, the gravitational potential energy released as gas slowly flows inward due to viscous forces is turned into thermal energy. Such models predict a steeper temperature profile: $T \propto r^{-3/4}$.

Disk Chemistry

Other constraints come from the composition of planets and small bodies, and therefore the disk chemistry needs to be understood. Condensation arguments imply that the $T \sim r^{-1.1}$. Since this more closely matches the midplane temperature profile, it implies that the grains that ultimately end up in the planetesimals formed in an optically thick region.

The most reliable determination of temperature is the low abundances of volatile elements (such as potassium and sulfur) compared to more refractory elements. This implies temperatures in the range of 1200-1400K in the terrestrial region. In contrast, the ice-rock cores of the giant planets require temperatures less than 160 K at 5 AU. Comets contain HS, CO, methanol, N2 and H2O. This suggests that their formation temperatures are even colder, $\sim 25 - 50K$.

Calcium Aluminum Inclusions (CAIs) in meteorites (in particular the primitive Allende meteorite) are dated to be the oldest objects in the Solar System. They are where the evidence for the short lived isotope, ^{26}Al, is found giving further evidence for being formed very early. These meteorites also indicate condensation temperatures of 1400K from chemical and textural properties. In particular, some CAIs have rounded shapes indicative of partial melting. This requires heating to at least 1700K for a short period of time. Crystal patterns indicate that they cooled quickly at the rate of 2-50 K/hour. Some have multiple rims that indicate they experienced more than one flash heating event.

Chondrules are millimeter sized igneous spherules in primitive meteorites. They contain more volatiles and require formation at temperatures lower than 650K. But they also appear to be flash heated (they are igneous after all),

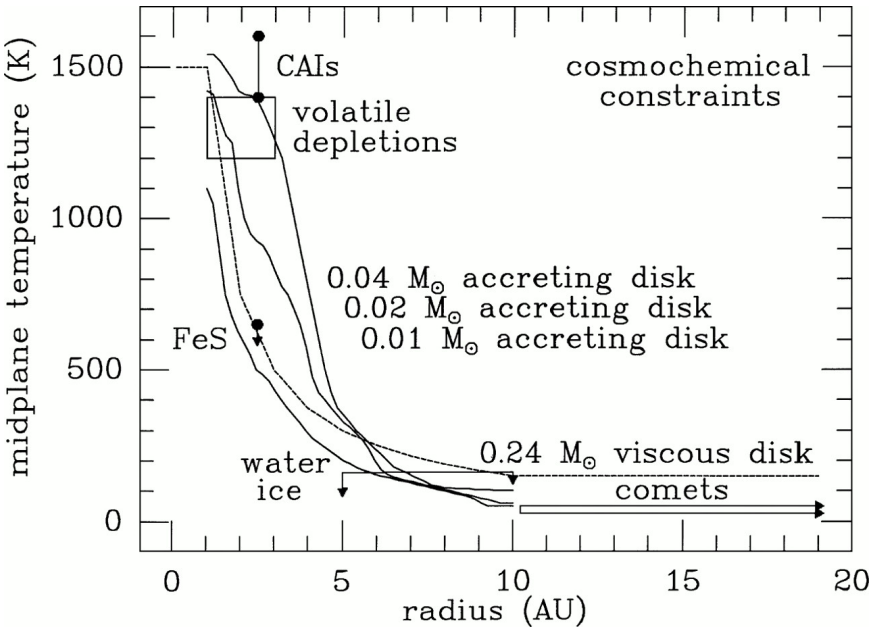

Fig. 2. Temperature vs radius is plotted for several models and compared with cosmochemical constraints *(taken from [7])*

but cooled at 100-1000K/hour. It therefore appears that CAIs formed first in a hot nebula, and then chondrules in a warm nebula. How they end up together in the same meteorite is a problem, perhaps indicating some degree of radial mixing.

These temperature constraints are compared with disk models in Fig. 2 taken from [7]. The accretion models are Boss's models with different mass accretion rates. These models are axisymmetric collapse calculations. A radiative hydro code was used to determine the thermal structure, but the radial density profile was constrained to be $\Sigma \propto r^{-1/2}$. The viscous accretion disk is a calculation of the style of Lynden-Bell and Pringle [25]. The jumps are associated with sublimation and condensation of grains at the different temperatures. Note again, that some such jumps would be unstable to convection. However, convective turbulence seems to stop accretion. Also note that there may be a thermostat effect going on: if the gas is cool enough that grains condense, then it gets optically thick, and the temperature goes back up. Hence the disk could be kept right at the grain condensation temperature.

On the other hand, the surface temperatures are slightly less sensitive to disk mass. This is because the disk is optically thick. This also means that the midplane disk temperature is decoupled from the stellar radiation.

In summary, the models predict a relatively uniform midplane temperature which is matched by volatile depletions. In particular there are no

systematic variations with r within the terrestrial region. Also, the gradual cooling of the disk by becoming optically thinner matches volatile depletion in chondritic meteorites. This gives the obvious conclusion: we expect silicate and iron rich planets to form inside 4 AU, and ice rich planets outside 5 AU. Furthermore, there could have been a cold trap at 5 AU, if water vapor could circulate radially [38]. As water vapor circulates inside 5 AU, it freezes and is trapped right at the ice line. The resulting enhancement of icy solids there would make Jupiter the dominant planet.

One caveat to keep in mind when considering the disk chemistry is that the densities are very low. Therefore, it is only at temperatures above 2000 K that the chemistry will be in equilibrium. In that case, the reaction rates are fast compared to the cooling rate of the disk. Otherwise, the reactions "freeze-out" and the equilibrium abundances are never reached. For example, we see N_2 and CO ices on Triton and Pluto, whereas all the N and C should be in the form of CH_4 and NH_3 at low temperatures. Therefore, the outer parts of the nebula did not reach chemical equilibrium.

So the chemical evolution is quite complex. As well as the long reaction times one must keep in mind the fact that there were primordial interstellar grains, evaporation of grains by shocks and temperature discontinuities, and ionization of gas by cosmic rays and UV radiation that in turn affects the molecular chemistry. Nevertheless, the basic picture seems to hold up as evidenced by our ability to explain the basic compositional characteristics of the planets.

Disk Clearing

At some point the gaseous disk gets cleared away; however, the process by which this happens is not clear. One possible explanation is that the accreting sun had a luminosity 20-30 times larger than it currently has. This could drive a solar wind which could clear out the disk, or if the star were significantly hotter, UV radiation could photoevaporate the faces of the disk.

Recent observations have shed light on the fact that stars almost always form in groups. Hence UV photons from nearby massive stars are a viable candidate for clearing the disk. The stunning HST observations of the Eagle nebula seem to show this process in action. Such a process gives a nice explanation for Uranus and Neptune. All giant planets have the same amount of rock and ice forming elements, but Uranus and Neptune have a factor of 100 less H and He. If the UV came from the outside, as in this scenario, Uranus and Neptune's accretion could have been interrupted before that of Saturn or Jupiter.

2.2 Early Stage: Growth of Grains into Planetesimals

Dust growth

Data from soot and numerical simulations [35] indicate that dust can grow into loosely packed fractal structures held together via van der Waals forces. The existence of many chondrules (i.e. small igneous inclusions 1mm in size) in meteorites means that these fluffy aggregates had to be rapidly heated and cooled before they were incorporated into larger bodies.

Dust Dynamics

To get an understanding of dust dynamics in the protoplanetary nebula, let us review the relevant forces that apply to dust grains.

Radiation Force

For micron sized particles, the radiation pressure force can be significant. It is given by:

$$\boldsymbol{F}_{rad} = \frac{L_\odot A}{4\pi c r^2} Q_{pr} \hat{r}$$

where A is cross section of the particle, r is the distance to the Sun, and Q is a radiation pressure coefficient that accounts for scattering and absorption. If we define the parameter β as the ratio of the radiation force to the gravitational force then

$$\beta \equiv \frac{F_{rad}}{F_g} = 5.7 \times 10^{-5} \frac{Q}{\rho R}$$

where ρ is the density of the grain material in gm/cc and R is the grain radius in *cm*. Given the typical grain densities, this instantly shows that F_{rad} can only be important for sub micron size objects. Furthermore, if the grains get much smaller than that, they become comparable to or smaller than the wavelength of the light and Q starts to become small.

Note that the ratio of forces, β, is independent of the distance from the Sun because both gravity and light drop off as $1/r^2$. So if $\beta > 1$, then the net force will always be outward and the dust will obviously be ejected. If $\beta > 0.5$ then a dust grain will be on a hyperbolic orbit if it is released on what would be a circular orbit in the absence of radiation pressure and therefore also be ejected.

Poynting–Robertson Drag

As grains become of order a centimeter in size, direct radiation pressure becomes less important, but a secular effect occurs because as radiation is absorbed and re-radiated by the particle, the re-radiated radiation is anisotropic

in the rest frame of the sun. This results in a drag force opposite the direction of motion of the particle referred to as Poynting–Robertson Drag.

The total radiation force including this effect to first order in v/c is:

$$\boldsymbol{F}_{\mathrm{rad}} = \frac{L_{\odot} Q_{\mathrm{pr}} A}{4\pi c r^2} \left[\left(1 - \frac{2v_r}{c} \right) \hat{r} - \frac{v_\theta}{c} \hat{\theta} \right]$$

where $L_{\odot}$ is the solar luminosity, r is the radius of the particle, and v_r and v_θ are the radial and tangential component respectively of the particle velocity in the frame of the Sun. The first term in this expression is the radiation pressure, and the second and third terms are the Poynting–Robertson drag. There is a factor of 2 in the second radial term because of the combination of Doppler shift in absorbing radiation and Doppler shift in emission.

Plugging in the current solar luminosity, one can obtain the decay time for a particle on a circular orbit:

$$\tau_{\mathrm{pr}} \sim 400 \frac{r^2 (AU)}{\beta} \mathrm{years}.$$

An obvious application of this formula is source of the zodiacal light. It is emitted from 20 to 200 μm particles, and by the above formula their lifetimes are 10^5 years. Therefore there needs to be a source to replenish these particles. Possibilities include dust from comets and collisions in the asteroid belt.

The Yarkovski Effect

Radiation based drag forces can also be relevant for meter to kilometer sized bodies. For a rotating body, the evening surface temperature is warmer than the morning surface. The resulting anisotropy in the thermal radiation from the body results in a net force called the Yarkovski effect. Since the force per unit area from thermal emission is:

$$dF = \frac{2\sigma T^4 dA}{3c}$$

where T is the temperature and σ is the Stefan-Boltzmann constant, we have

$$F_Y = \frac{8}{3} \pi R^2 \frac{\sigma T^4}{c} \frac{\Delta T}{T} \cos \psi$$

where ψ is the obliquity of the body, and ΔT is the evening-morning temperature difference. The force is therefore in the direction of motion for direct rotators and a drag force for retrograde rotators.

The exact magnitude of this effect is hard to calculate. It depends on detailed surface properties of the asteroids which are not well known. Nevertheless it is postulated as a mechanism for moving asteroids through the belt over long timescales.

Gas Drag

Unlike gas drag on small particles in the lab, where the small particles eventually reach the speed of the gas, gas drag on dust in a Keplerian disk usually results in long term evolution. This is because the drag decreases the energy of the particle, which will ultimately increase the speed of the particle in a Keplerian potential. There are several limiting regimes to the drag force. Stokes drag is given by

$$F_D = -\frac{C_D A \rho v^2}{2}$$

where C_D is a dimensionless coefficient, A is the cross-sectional area of the particle, and ρ is the density of the gas, and is appropriate for high Reynolds number flows. Epstein drag is appropriate when the mean free path of the gas particles is greater than the physical size of the dust particle, and is given by

$$F_D = -A \rho v c_s,$$

where c_s is the sound speed of the gas. In the protoplanetary disk, the drag can have long term effects on intermediate size particles since the gas is not rotating at Keplerian speed. This is because the force that the gas feels is

$$g_{\rm eff} = -\frac{GM_\odot}{r^2} - \frac{1}{\rho}\frac{dP}{dr},$$

where P is the gas pressure. Since the pressure is rising inward, the pressure gradient opposes gravity, and the gas moves slower than Keplerian. Therefore any particle that is only loosely coupled to the gas will experience a headwind, and spiral inward due to the drag forces given above.

Application to the Planetesimal Disk

The typical timescales for inward motion are slow for small bodies because they are tightly coupled to the gas. It is also slow for large bodies, because their momentum is relatively large compared to the gas drag force. For solar nebula conditions, meter size bodies have the shortest decay times of ~ 100 years. Bodies smaller or larger than this will have longer decay times.

For dust grains, settling in the vertical direction is important because it may lead to a gravitational instability (the Goldreich-Ward instability, see below) in the dust. The equation of motion in the vertical direction for the dust with Epstein drag is

$$\frac{dv_z}{dt} = -\frac{\rho_g c_s}{R\rho} v_z - n^2 z$$

where ρ_g is the density of the gas, ρ is the density of the dust, R is the dust radius and $n = \sqrt{\frac{GM}{r^3}}$ is the mean motion. If the drag is large then there is a terminal velocity of

$$v_z = \frac{n^2 z \rho R}{\rho_g c_s}$$

For conditions in the terrestrial portion of the protoplanetary nebula ($T = 500-800K$, $c_s = 2.5km/s$) this velocity gives a settling timescale, z/v_z, of 10^6 years. Furthermore, we need many settling times to concentrate the dust into larger bodies. This settling rate gets shorter as the grains get larger, although if they grow into fluffy aggregates their cross-section to volume ratio would stay large. Nevertheless, grain growth may shorten lifetimes to as little as 10^4 years.

Goldreich-Ward Instability

There are a couple prevalent theories for getting to larger (100 m to 1 km) size bodies. One is that aggregation of dust can occur within turbulent eddies [39]. The more popular theory is that if the dust settles into a cold disk in the plane, it would be susceptible to gravitational instabilities, known as the Goldreich-Ward instability [14].

This instability is essentially the Toomre instability for rotating disks. For the Toomre instability, there is a critical wavelength

$$\lambda = \frac{4\pi^2 G \Sigma}{\kappa^2},$$

where Σ is the surface density and κ is the epicyclic frequency. This wavelength is set by the scale at which the rotational shear is able to overcome gravity independent of the velocity dispersion of the dust. Therefore, if the dust is dynamically cold enough, it will fragment into clumps of about this size. We therefore expect a critical mass of

$$M_p \sim \Sigma \lambda^2 \sim \frac{16\pi^4 G^2 \Sigma^3}{\kappa^4}$$

If the density of the forming planetesimal is about 1 gm/cc then this critical mass would produce a body of roughly 1 km in size in the inner regions of the solar nebula.

This is a rather elegant result in that the final size of the objects formed is relatively independent of the conditions in the solar nebula other than the dust surface density, and presumably this can be directly constrained by SEDs of protoplanetary disks. Secondly, it gives relatively simple initial conditions for the next stage of planet formation: a cold disk of mono-disperse bodies. However, getting the dust to this thin layer has a lot of unknowns. What with turbulence, solid body collisions, and large radial motions, we don't understand this era of planet formation very well.

2.3 Middle Stage: Planetesimal Accretion

Now we come to the crux of the "planetesimal hypothesis": the building of large bodies out of the 1 km size bodies produced at the end of the early stage. As well as providing a mechanism for building the planets, it also makes the connection between those planets and the small bodies.

The simplest accretion picture was first quantified by Safranov [37]. Consider a sphere of radius s moving with velocity v through a uniform medium of density ρ. Then the accretion rate is

$$\frac{dm}{dt} = \pi s^2 \rho v.$$

However, the gravity of the body plays a rôle by gravitationally enhancing the cross-section. This gravitational cross-section can be calculated as follows. Consider the particle that just grazes the surface of the planetesimal as it goes by. From angular momentum conservation we have: $v_i s = v_0 d$, where v_i is the velocity at impact, v_0 is the encounter velocity and d is the impact parameter. From energy conservation we have $-v_{\rm esc}^2 + v_i^2 = v_0^2$ where $v_{\rm esc}$ is the escape velocity from the surface of the planetesimal. Solving for d^2/s^2, we have

$$\frac{d^2}{s^2} = 1 + \left(\frac{v_{\rm esc}}{v_0}\right)^2.$$

Since all particles with an impact parameter less than d will strike the body, πd^2 is an effective cross-section, and the above accretion rate should be multiplied by this factor to account for the gravitational focusing.

Clearly the growth rate will be a strong function of the ratio of the encounter velocity to the escape velocity from the surface of the planetesimals. Therefore, the velocity distribution of planetesimals is a key ingredient to the outcome of this stage. Work in this field has progressed using the kinetic theory of gases, and for that we turn to the Boltzmann equation.

Boltzmann Equation

The evolution of the distribution function $f(\underline{\mathrm{r}}, \underline{\mathrm{v}})$ can be described by the collisional Boltzmann equation,

$$\frac{\partial f}{\partial t} + \underline{\mathrm{v}} \cdot \nabla_r f - \left[\frac{GM_\odot \underline{\mathrm{r}}}{r^3} + (\underline{\mathrm{v}} + \underline{\mathrm{v}}_K) \cdot \nabla \underline{\mathrm{v}}_K\right] \cdot \nabla_v f = \left.\frac{df}{dt}\right|_{\rm coll},$$

where we have split the velocity into a Keplerian, $\underline{\mathrm{v}}_K$, and random component. One typically solves this equation by assuming a form for f, and working out the evolution rate. For the planetesimal case at a particular radius, one usually assumes that the planetesimals start out with a Rayleigh distribution. That is, we assume that planetesimals are distributed according to

$$f(e,i) = 4\frac{\Sigma}{m}\frac{ei}{\langle e^2\rangle\langle i^2\rangle}\exp\left[-\frac{e^2}{\langle e^2\rangle} - \frac{i^2}{\langle i^2\rangle}\right].$$

where Σ is the surface mass density, m is the mass of the planetesimals, e is the orbital eccentricity and i is the inclination. In the *Epicycle approximation*, that is, small eccentricities and inclinations, this is equivalent to

$$f(z,\underline{v}) = \frac{\Omega\Sigma}{2\pi^2 c_r^2 c_z^2 m}\exp\left[-\frac{v_r^2 + 4v_\theta^2}{2c_r^2} - \frac{v_z^2 + \Omega^2 z^2}{2c_z^2}\right]$$

where $2c_r = \langle e^2\rangle v_K^2$, $2c_z^2 = \langle i^2\rangle v_K^2$, $v_K = \sqrt{GM_\odot/r}$, and $\Omega = v_K/r$. In the absence of collisions c_z and c_r would be completely independent, but numerical simulations of the accretion process show that $c_z/c_r = 0.5$.

There are a number of possible contributions to the collision term, but the dominant one that we will consider is stirring from gravitational scattering. The natural scale for the gravitational scattering is set by the escape velocity of the particles, so we define

$$\left(\frac{v_{esc}}{v_0}\right)^2 \equiv 2\theta$$

where θ is the *Safronov number*.

Now the scale height of the disk is $h = v/\Omega$ because the vertical motion is dominated by the gravity of the Sun. Therefore, the density can be gotten from $\Sigma = \rho h$, and we can derive the accretion rate in terms of the Safronov number as

$$\frac{dm}{dt} = \pi s^2 \Sigma\Omega(1 + 2\theta)$$

Note that the flux is now no longer explicitly a function of v, but is constant at a given radius.

If we now assume (as Safronov [37] did), that θ is a constant, we can substitute for the mass of the planetesimal $m = 4/3\pi R_{pl}^3\rho_{pl}$ and obtain the growth rate of the radius as

$$\frac{dR_{\rm pl}}{dt} = \frac{\Sigma\Omega}{4\rho_{pl}}(1 + 2\theta).$$

Turning this into a growth rate gives us

$$t_{\rm grow} = \frac{R_{pl}4\rho_{\rm pl}}{\Sigma\Omega(1 + 2\theta)}.$$

Now let's apply this formula to the terrestrial planet region. Assuming the radius of the Earth $R_\oplus$, the density of the Earth, $\rho_\oplus = 6$ gm/cc, a density of $\Sigma = 10$ gm/cm^2, and $\theta = 3$, one gets a growth timescale of $t_{\rm grow} = 3 \times 10^7$ years. This is a bit long. Furthermore, for constant θ, $\frac{dR}{dt} \propto r^{-3}$, where r is the distance from the Sun. Hence at Saturn's radius of 10 AU, we have

$\tau \sim 3 \times 10^{10}$ years. We could never grow the outer planets or their cores this way.

There are other collisional terms that can contribute to the Boltzmann equation. These include stirring from collisional scattering, damping from inelastic collisions, damping from gas drag and dynamical friction. Of these terms, dynamical friction may play the most significant rôle. One way to think of dynamical friction is that a large body will create a wake of smaller bodies behind it as it moves through them. The gravitational attraction between the large body and this wake will act as a drag to slow the large body down. Another way to think of dynamical friction is as a mechanism to establish equipartition: the large bodies will increase the velocity dispersion of the small bodies. Thought of in this way, one gets that the velocity dispersion should scale as $v^2 \sim 1/m$.

Runaway Growth

We expect a larger planetesimal to have a larger growth rate because: 1) gravitational focusing is greater and 2) dynamical friction on smaller planetesimals decreases their velocities. Hence we will modify the growth timescale given above for the largest planetesimals. This will lead to a phenomenon referred to as runaway growth.

A simple way to look at this is to note that if the escape velocity $v_{\rm esc}^2 \sim G\rho R^2$ grows large, while the typical encounter speed with the small bodies, v_0, remains constant then the growth rate will scale as R^2, rather than R as derived above. This larger growth rate will only be valid for the largest bodies in the region, and so we have the result that the "rich get richer", and large bodies will break away from the general mass distribution.

This breakaway has been seen in numerical simulations such as those in [19] and [36]. The results from [36] are shown in Fig. 3. Note how at the last timestep, the most massive particles have significantly smaller eccentricities than the general population, presumably because they have been damped by dynamical friction. Also note how quickly the large masses have grown. With runaway accretion, the timescales for growth reduce to $10^5 yrs$.

Oligarchic Growth

However, the simulations indicate that there are limits to runaway growth. They finish with a bi-modal distribution of bodies with the larger bodies being separated by about 4 Hill radii. This is referred to as Oligarchic growth, and the final mass of the large bodies is determined by the mass available within the feeding zone.

$$M_{iso} \sim 2\pi r(8R_H)\Sigma,$$

where R_H is the Hill radius (see below) of the body. Once the body has eaten up this mass, it has depleted the feeding zone and becomes isolated. The size

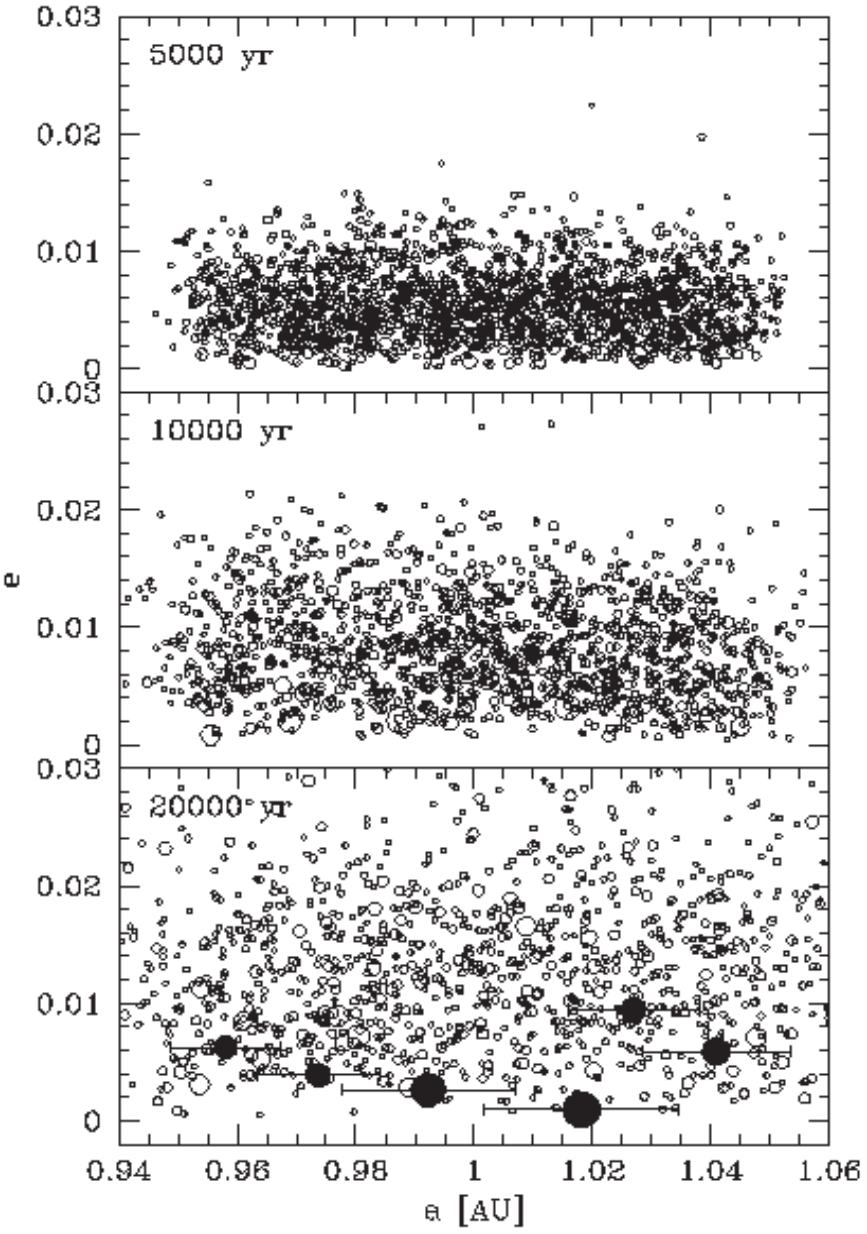

Fig. 3. Evolution of the eccentricities and masses of planetesimals in a simulation of the middle stage from [36]. Eccentricity is plotted against semimajor axis for 3 different times in an N-body simulation of planetesimal accretion. The size of the points is proportional the mass of the particles. For the snapshot at 20000 years, 5 Hill radii are plotted for the most massive bodies

of this feeding zone can be determined by considering the 3-body problem for nearly circular orbits.

The 3-Body Problem and the Hill Sphere

If we scale the distances in the 3-body problem by the *Hill radius*

$$R_H = \left(\frac{m_p}{3M_\odot}\right)^{1/3} a,$$

then the 3-body problem is independent of $m_p/M_\odot$ (if it is small). We can define

$$e_H \equiv \frac{ea}{R_H}, \quad i_H \equiv \frac{ia}{R_H}, \quad b_H \equiv \frac{(a_2 - a_1)}{R_H}.$$

where a_1 and a_2 refer to the semimajor axes of nearby bodies.

In terms of these quantities the 3-body integral of motion is

$$\Gamma \equiv \frac{3}{4} b_H^2 - e_H^2 - i_H^2.$$

A particle can't enter a protoplanet's Hill sphere if $\Gamma > 9$. Therefore, if a planetesimal starts out with a circular orbit, it can not be accreted onto a neighboring protoplanet unless its semimajor axis is such that $b_H < 4$. Hence we derive the size of the feeding zone given above.

Substituting in the value for the Hill radius in the isolation mass, we have:

$$M \sim 16\pi r^2 \left(\frac{M}{3\,\mathrm{M}_\odot}\right)^{1/3} \Sigma$$

or

$$M_{\rm iso} \sim \frac{(16\pi r^2 \Sigma)^{3/2}}{(3\,\mathrm{M}_\odot)^{1/3}}$$

or if we consider solar values

$$M_{\rm iso} \sim 1.6 \times 10^{25} gm (r_{AU}^2 \Sigma_{\rm gm/cc})^{3/2}$$

which gives a few lunar masses for the minimum mass solar nebula at the Earth's radius, and roughly an Earth mass at Jupiter.

Therefore at the end of runaway growth we get proto-planets, not planets. Can we fix this up so that we obtain larger masses? One possibility to get larger masses is to start with more than the minimum solar nebula of material; however, the excess solids will need to be cleared out. This can easily happen in the neighborhood of Jupiter as that planet is large enough to eject the bodies, but in the terrestrial region it is not clear how the extra material would be lost.

This picture also gives an explanation for the existence of the asteroid belt: if the velocity dispersion is increased by an early Jupiter, then growth

rates will remain small. Of course this means having a significant mass in Jupiter before runaway can get started in the asteroid belt. At first this would seem implausible because the growth timescales tend to increase as one moves outward in the solar nebula; however, the existence of an ice line at 5 AU could make Jupiter grow quickly because of the resulting enhanced surface density of solids.

Is there a way to continue growth once the isolation mass has been reached? Two possibilities have been brought up.

- The diffusion of more material into the feeding zone from neighboring regions. This could be caused by the perturbations of neighboring protoplanets.
- The migration of the protoplanet into a new feeding zone via either gravitational torques or gas drag. We will discuss planet migration in general below.

2.4 Late Stage

At the end of the middle stage we have a regular distribution in a of protoplanets with roughly constant mass. There is also a swarm of smaller planetesimals, which may or may not be relevant. Since most of the mass is in the large bodies, dynamical friction is no longer playing a significant rôle in their orbital evolution. These are the initial conditions for the late stage.

This is the few body problem. It isn't analytically tractable, and there are too few bodies to use statistical methods as could be done in the middle stage. There isn't much else to do but to numerically follow the orbits of these bodies for 10^8 years. In such a calculation, there are several things that need to be considered. First, the full N^2 forces among the bodies need to be calculated. This problem is similar to the planetary stability problem, so all the techniques described elsewhere in this book are applicable here. Using these types of codes for the planetesimal case was pioneered by [11]. Secondly, every so often collisions occur. Almost all simulations performed to date assume that the collisions are inelastic, (a reasonable assumption) and all the material in the colliding bodies ends up in the collision remnant. This last assumption is rather wishful, given that detailed SPH simulations of colliding protoplanets show significant ejecta. [10, 3]

Typical results from the simulations are shown in Fig. 4 from [34]. As can be seen from the figure, a few planets of roughly terrestrial-planet mass are indeed formed. However, the general trend is that eccentricities and inclinations are gradually increased by the mutual gravitational interactions. Therefore the eccentricities coming out at the end of the simulations tend to be larger than observed in the Solar System.

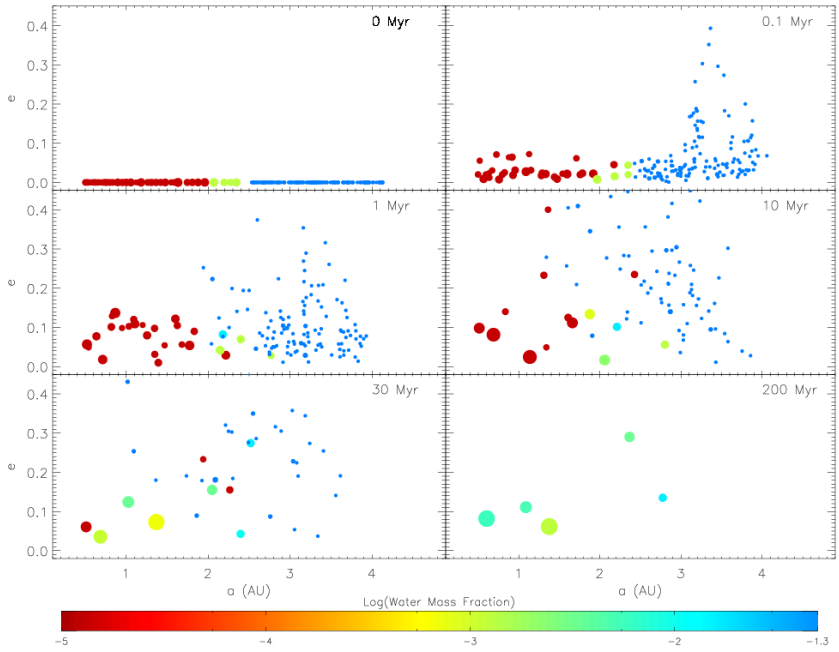

Fig. 4. Eccentricity is plotted against semimajor axes for a simulation of gravitationally interacting and colliding protoplanets during the late stage as they evolve through time. The size of the points is proportional to the cube root of the masses of the protoplanets. The color scale indicates the mass fraction of water *(taken from [34])*

2.5 Outer Planets: Gas Envelope Accretion

The giant planets are enhanced in the heavy elements as well. Even with the uncertainty in the core mass, the mass of elements in the envelope leads to an enhancement in heavy elements over solar abundances of at least 5 for Jupiter. This suggests that the planetesimal accretion model also plays a part for the outer planets, but augmented with a later gas accretion stage.

Therefore we expect the following sequence for the stages of giant planet formation:

- Runaway growth of a solid core which clears out a few R_H.
- A gas envelope starts to accrete.
- Planetesimals interact with gas envelope, depositing energy and metals.

These processes have been modeled in detail by several groups [42, 32]. The way these models work is very similar to stellar interior modeling. One self-consistently solves for density and temperature as a function of radius, assuming that pressure gradients are balanced by gravity. As the radiation

can diffuse out, the temperature structure changes, and one iterates to a new solution. The difference from standard stellar modeling is that the energy comes from gravitational energy deposited by planetesimals, not nuclear burning.

The models are then updated to a later time based on the relevant information. First, mass and energy are injected by accreting planetesimals (estimated from the feeding zone). Orbits of the planetesimals with the the envelope including gas drag, and vaporization rate are calculated to determine where in the envelope the mass and energy are deposited. Second, the envelope contractions due to energy lost due to radiative transfer. As the planet contracts, nebular gas is added to the outer part. The outer radius is the minimum of the Hill radius, and an accretion radius

$$R_a = \frac{GM_p}{c_s^2},$$

where c_s is the sound speed of the gas in the nebula.

An issue that has already been mentioned concerns Uranus and Neptune. What happened to their envelopes? Several solutions have been offered. The gaseous nebula could have been dissipated before the exponential phase of envelope accretion. This brings up a fine tuning problem: why did the timescales match so that Jupiter and Saturn got envelopes, but Uranus and Neptune did not? Wuchterl [42] proposes another solution. A dynamical instability similar to the "κ" mechanism in Cepheids would drive an "overstability" that could remove the outer envelope.

A key result from these accretion calculations is that the growth of the envelope requires millions to tens of millions of years. This is long compared to observational constraints on lifetimes of protoplanetary nebula. Furthermore, an initial nebula with 3 to 4 times the mass of the minimum mass solar nebula is needed to get even these timescales. However, there are obvious improvements to the calculations that may modify these results. These include better opacities, a better treatment of convection, and perhaps a small amount of migration of the planet as it grows.

3 Small Bodies and Planet Formation

Within the planetesimal hypothesis, the small bodies we observe today play a very important rôle. They are essentially the "fossils" of the planet formation process, and as such provide valuable clues to the history of the Solar System. Particularly the comets and the small bodies in the outer Solar System are helpful since they have been relatively unprocessed since their formation.

3.1 Theoretical Considerations

A key theoretical consideration is taken from the stability of the orbits of small bodies. Long term orbit integrations been done for small bodies starting

with circular orbits at a variety of semimajor axes [17]. These have shown that only bodies that start their orbits with semimajor axes between 2 and 3.5 AU or beyond 35 AU are able to remain on stable orbits for the age of the Solar System. Hence it is only Main Belt Asteroids and Kuiper belt objects that are on stable orbits and can be in their original location. All other small bodies must be injected into their current orbits from some longer-lasting reservoir. Furthermore the delivery mechanism from that reservoir must operate on a timescale of billions of years. If it were any shorter, the reservoir would be depleted.

3.2 Near Earth Asteroids

Objects with perihelia less than 1.3 AU and aphelia greater than 0.983 AU are classified as Near Earth Asteroids (NEAs). The orbital lifetimes of these objects are only tens of millions of years so they must originate from another source. Possibilities include the Main Belt and extinct comets. Some of the bodies may also be ejecta from the Moon or Mars when they are impacted by other asteroids. Of course these bodies also impact the Earth and so are very relevant to Earth's geological and biological history. Since they are also the source of meteorites, these bodies are invaluable for determining the composition and early thermal history of the Solar System.

Determining their complete orbital distribution is very model dependent because the observations are so incomplete. Bottke [9] has developed a sophisticated procedure for doing this. First orbits of small bodies are integrated starting from a number of proposed source regions. These are different regions of the main belt and the short period comets. A best fit model is then created by varying the relative contributions of these sources in order to match the distribution of the known orbits. They find that 61% of NEAs come from the inner Asteroid belt, 24% from the middle, and 8% from the outer belt. Extinct comets provide 6% of the NEAs.

3.3 Main Belt Asteroids

Most of the known asteroids are in the Main Belt with semi-major axes between 2.1 AU and 3.3 AU. Their distribution in semi-major axis is remarkable because of a number of gaps referred to as Kirkwood gaps. These coincide with mean motion resonances with Jupiter indicating that Jupiter has played a significant rôle in the dynamical evolution of the belt.

A detailed study of the orbital elements show that there are a large number of families that are concentrated around particular orbital elements. Furthermore [18] has shown that these families have similar colors indicating that their composition is identical. This leads to the hypothesis that these families are created by collisions or break-ups of larger bodies, and generally indicates that there is a significant amount of collisional evolution in the Asteroid Belt.

3.4 Centaurs and Trans-Neptunian Objects

Between the orbits of Jupiter and Neptune are the Centaur objects. There are only a handful of these known. They typically have orbits that cross the semi-major axis of one or more planets, therefore they have a very short dynamical lifetime. Orbital integrations indicate that they are in transit between the Kuiper belt and short-period comets [21]. The Kuiper belt refers to objects in low inclination orbits with semi-major axes beyond Neptune. Their dynamical lifetimes are very long so they are presumably remnants from the formation of the Solar System. That is, they are the remaining "planetesimals" discussed above. They are also postulated to be the source of the short period comets [33].

One interesting question concerning the Kuiper Belt objects is the status of Pluto. Pluto has an orbit similar to a number of the Kuiper Belt objects leading to the speculation that it is simply the largest body of this population. Although there is a considerable gap between Pluto, Chiron and the largest Kuiper belt object, models of the mass distribution predict roughly one Pluto size body.

3.5 Comets

Comets are divided into a number of classes based on their orbital elements. There are several ways to make the division into classes; one way is as follows. Jupiter family comets have periods less than 20 years; Halley family comets have periods between 20 and 200 years; long period comets have periods greater than 200 years but semi-major axes less than 10,000 AU, and Oort cloud comets have semi-major axes greater than 10,000 AU. There may be an inner Oort cloud with smaller semi-major axes. Because comets have spent most of their lifetime far from the Sun, they are relatively pristine objects with connections to Solar System formation.

Short Period Comets

Short period comets have inclinations such that they are concentrated in the ecliptic plane. As shown in Fig. 5, there are several peaks in the semi-major axis distribution with the largest at 3 AU. The peaks are more pronounced in the aphelion distribution where they correspond to the semi-major axes of the outer planets. The argument of perihelion distribution peaks at 0 and 180 degrees, meaning that as the comets go through aphelion they are in the plane of the planets. Because this orbital configuration favors close approaches to Jupiter and Saturn, the short period comets are dynamically controlled by the giant planets

This strong interaction with the outer planets means that their dynamical lifetime is very short, of order 1 million years. They are not in a stable distribution, and must be injected into their current orbits from some larger reservoir

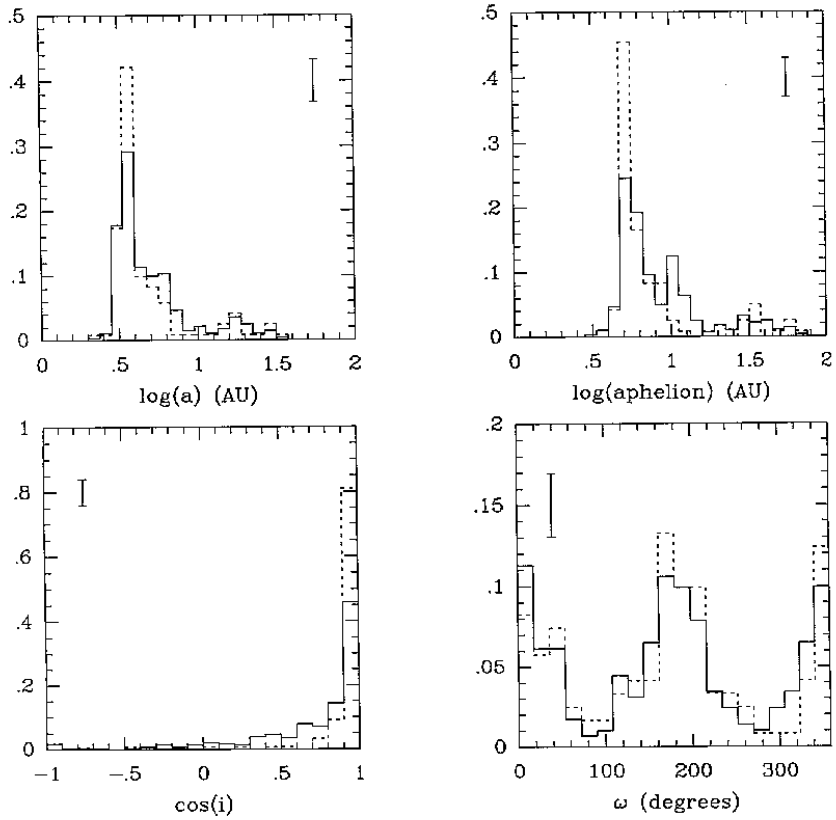

Fig. 5. The distribution of the short period comets in semi-major axis, aphelion, inclination and argument of perihelion is plotted *(taken from [33])*

that can survive the age of the Solar System. It was originally thought that the long period comets could be this source. However, simulations [12] show that such a scenario is incompatible with their current orbital distribution. First, such a source would produce a much broader inclination distribution, and second the argument of perihelion would peak at 90 and 270 degrees due to a secular evolution of the comet orbits. The conclusion is that a low inclination belt of bodies beyond Neptune could produce short period comets with a much better match to the observed orbital distribution. The first of these Kuiper belt objects was discovered several years after these simulations were published.

Long Period Comets and Oort Cloud

The distribution of longer period comets is remarkable because of a sharp peak at a semi-major axis of about 20,000 AU. The inclinations are isotropic

and the perihelion distribution seems uniform. The energy distribution is consistent with all comets being marginally bound within the observational errors. The implication is that the observed comets are the very eccentric subsample of comets in a large reservoir at $a = 10^4$ AU: the Oort cloud.

The likely origin of the Oort cloud is also a disk in the outer Solar System [12]. The scenario is as follows, and a typical trajectory is shown in Fig. 6. Planetesimals between the planets in the outer Solar System are initially perturbed by interactions with those planets. If the planetesimals are scattered outward, the planetary perturbations tend to cause random walks in a at fixed perihelion and inclination. They therefore diffuse outward in semi-major axis until the orbits are large enough that Galactic tides and passing stars start to influence their orbits. The tides and passing stars tend to change the angular momentum without changing energy. The change in angular momentum will lift the perihelion out of the planetary region, and the comets will start to evolve at constant semi-major axis, and the orbits will become isotropic. Of course, those same tides will bring comets back into high eccentricity orbits which gives us the comets we see today.

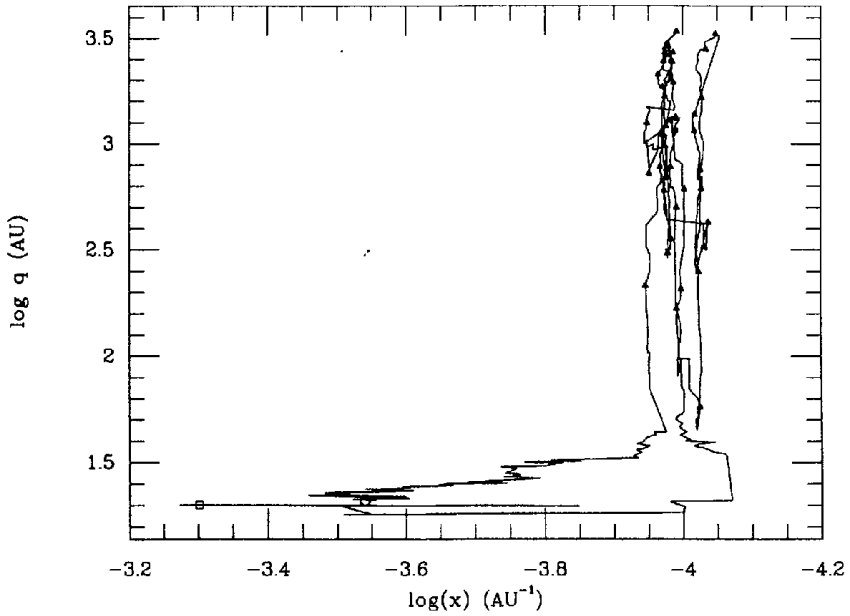

Fig. 6. Evolution in $x \equiv 1/a$ and q of a typical Oort cloud comet from [12]. The comet first diffuses in a because of perturbations by the planets. When a becomes large, q gets increased because of galactic tides and stellar encounters

3.6 Unifying Small Body and Planet Origins

In summary the planetesimal hypothesis gives us a mechanism to link together all the small body populations in the context of planet formation. During the middle stage we have many icy and rocky bodies. As the outer planets grow via accretion, they scatter most of the original planetesimals out of the region. Planetesimals scattered to higher semi-major axis slowly diffuse outward until tides lift them out of the planetary region and they become Oort cloud comets. Planetesimals outside Neptune do not feel strong enough planetary perturbations to diffuse rapidly and remain as the Kuiper belt objects. They subsequently slowly supply the short period comets. The only other planetesimals that remain are the Main Belt asteroids, which become a source for the Near Earth Asteroids.

4 Planet Formation: the Fragmentation Model

The model outlined in the previous two sections has had great success in explaining the characteristics of our own Solar System; however, observations of planets around other stars, and observations of protoplanetary disks have raised some challenges. The first thing that these observations indicate is that planet formation (or at least, formation of giant planets) is fairly ubiquitous. Roughly 10% of the surveyed stars have had planets discovered around them [26], and when considering selection effects due to inclinations and masses, the real fraction must be higher than this. This indicates that whatever the formation mechanism, it must be efficient.

The second observational evidence impacting our theories of planet formation are the studies of star forming regions. On the one hand, large scale millimeter surveys indicate that most stars form in OB associations [2]. However, studies of the circumstellar disks in these regions indicate that the lifetimes of those disk are only .1 million years. Presumably the strong UV flux from the nearby bright stars quickly ablate the disks.

Since the formation of the gas giants must happen before the disk is dissipated, and the core-accretion model described above generally happens on a longer timescale than typical disk lifetimes, we must conclude that either planet formation is rare, or there is another quicker mechanism for it to happen. Combining this argument with the fact that giant planets are relatively common pushes us to consider other theories. There are a couple of other issues with the core-accretion model, such as gap formation (not included in the above calculations) prematurely terminating accretion, and ice giants have a hard time forming in the outer solar system because the escape velocity from the forming cores becomes comparable to the orbital velocity so that most material gets scattered out instead of accreting. This could be solved by forming Uranus in the Jupiter-Saturn region [40].

Gravitationally driven fragmentation of a gas disk is an idea that goes back to Kuiper [20], but has recently been revived [5, 6, 8]. The idea is that if the disk is massive enough, the disk can fragment gravitationally in a short time. The condition for local instability to gravitational collapse is the Toomre criterion:

$$Q = \frac{c_s \kappa}{\pi G \Sigma} < 1,$$

where κ is the epicyclic frequency and Σ is the surface density. Furthermore, numerical simulations show that $1 < Q < 2$ allows global instabilities, and these may fragment. If we take the temperature constraints above and use densities corresponding to the "minimum solar nebula", we obtain a protosolar disk with a mass of $.01\,\mathrm{M}_\odot$. This would imply $Q > 2$ everywhere, so we would need more mass than this to drive such an instability. The mass of disks around other stars can be obtained from radiative transfer modeling, however the gas/dust ratio must be assumed. If we assume an interstellar gas/dust ratio of ~ 100 then the implied disk mass is $M \sim 0.05$ to $0.1\,\mathrm{M}_\odot$. With typical temperatures this will give a $Q \sim 1.5$ for $R > 5$ AU. This is not that much more massive than the masses needed for envelope accretion models.

Whether the spiral arms that develop within the unstable disks actually collapse into protoplanets is an outstanding issue, the answer to which is limited by the quality of simulations that can be performed. Since the modeling has to follow scales of many AU down to the Hill sphere of the planet, the dynamic range is quite challenging, and most simulations have been limited by their spatial resolution.

Recently simulations with Smooth Particle Hydrodynamics (SPH) have been attempted [27, 28]. These simulations fixed the temperature profile to that obtained from static radiative transfer simulations [4, 6]. If the disk was massive enough such that $Q \sim 1.4$, then the spiral arms did indeed fragment into protoplanets. Furthermore the masses and eccentricities of the formed planets are similar to the masses of observed extra solar planets.

The fixed temperature profile is a significant limitation to these simulations. Clearly as the gas gets denser, radiative diffusion out of the clumps will become more difficult. Other simulations with constant cooling time show significant clump formation only when the cooling time is of order the orbital time [13]. Modifying the equation of state so that it becomes adiabatic after the clumps become sufficiently dense still results in protoplanet formation, but the issue will not be fully resolved until there are high resolution simulations that include radiative transfer.

4.1 Long Term Evolution

The long term evolution of planets formed by disk fragmentation is significantly different than the standard model. This is because in the fragmentation model a handful of planets will form at the same radius from a single spiral

arm, whereas in the core-accretion model, it is assumed that the planets are well separated when they form. The result is that there are many mergers and scatterings during the evolution. This can be exacerbated via "seeded" planet formation [1]. Once a single giant planet forms, it excites strong spiral waves at its Linblad resonances which in turn fragment into more giant planets.

This seeding may also have implications for terrestrial planet formation. It could be that the perturbations from the gas giants are needed to get the terrestrial protoplanets on crossing orbits so that they collide to build larger objects. Furthermore if we are to explain the asteroid belt by planet formation being suppressed by the presence of Jupiter, the gas giant needs to have formed quickly before runaway growth in the asteroid belt can get going. This is natural in the fragmentation model where the giant planets form quickly.

After hundreds of years of evolution the distribution of planets has the following characteristics.

1. The eccentricity distribution matches the "non-hot Jupiter" (that is $a >$ .25 AU) distribution observed in the extrasolar planets. Eccentricities can be as high as 0.7.
2. The entire mass range from "super-Jupiters" down to Saturn mass objects are generated, depending on the initial mass of the nebula.
3. Because of the mergers, we expect the obliquity distribution to have some outliers with very large obliquities as that of Uranus.
4. There are high accretion rates onto the star. The strong interactions among the planets and between the planets and the gas drives a significant amount of mass inwards. This a mechanism to clear out the disk quickly. Episodic flows onto the star could explain the FU Orionis outburst phenomenon.

4.2 Arguments against Fragmentation

The fragmentation model is attractive because it neatly solves the timescale problem. However, there are a number of problems which make the core-accretion model more favorable. First there is the evidence that the outer planets contain cores. Observations of Jupiter do not rule out a zero mass core mostly due to uncertainties in the equation of state of hydrogen at very high densities and temperatures. Nevertheless, a zero mass core is not the favored fit. In the fragmentation model one would not expect to see an enhanced core since high Z material added after the collapse should be thoroughly mixed in the envelope. Settling could happen if the planet was non-turbulent (i.e. non-convective); there are disagreements as to whether this is the case. However, Haghighipour and Boss [16] show that small solids of size centimeters to meters can migrate quickly (1000 years) to the pressure maximum and thereby enhance the abundances of the giant planets in the

fragmentation model. Note that even the envelopes of gas giants are enriched, implying that we need the solid condensation picture anyway.

The ice giants, Neptune and Uranus are problems in both formation models; they don't naturally occur in the fragmentation model. However, they could be formed via fragmentation as Jupiter and Saturn are, and then their envelopes get ablated away via EUV radiation from nearby massive stars. Jupiter and Saturn keep their envelopes because they are more deeply embedded in the protoplanetary nebula.

Whether the fragmentation model works depends significantly on the details of the gas dynamics and radiative transfer. This is difficult to calculate, and a lot of detailed numerical work will be needed to confirm or rule out this model.

5 Planet Migration

The "hot Jupiters" that comprise about one quarter of the extrasolar planets observed pose a problem in either of the planet formation scenarios discussed above. They can not be formed in the core accretion model because the ices (and indeed many rocky compounds) don't condense at the temperatures seen so close to the star. In the fragmentation model Q tends to be high near the star both because the temperature, and hence c_s is high, and the shear, i.e. κ, is large. The obvious explanation is that planets form further out where the theories predict they do, but then migrate significantly due to their interaction with a disk.

The interaction between the planet and the disk can be calculated using spiral density wave theory [15]. In this theory, the coupling to the disk is expected to be strongest at resonances. Of particular importance are the *Linblad resonances:*

$$m(\Omega - \Omega_{ps}) = \pm\kappa$$

where plus refers to the inner and minus refers to the outer Linblad resonance and m is the azimuthal wave number ($m = 1$ for a one arm spiral). Consider a frame rotating with the Keplerian velocity of the outer Linblad resonance (OLR). If the gas moves in an elliptical orbit, it is equivalent to an epicycle in this frame. Once per epicycle, the pattern comes by, so it is resonantly perturbed by always seeing the perturbation at the same point in its epicycle and a secular transfer of energy and angular momentum can happen.

5.1 Type I Migration

The inner Linblad resonance (ILR) increases the angular momentum of the planet, while the OLR decreases the angular momentum. However, these two effects don't quite cancel out, and there is a net loss of angular momentum to the disk by the planet [41]. This is referred to as Type I migration.

The migration timescale can be calculated from the linear theory and is:

$$\tau = \frac{1}{C_a \Omega} \left(\frac{\mathrm{M}_\odot}{M_p} \right) \left(\frac{\mathrm{M}_\odot}{\Sigma r^2} \right) \left(\frac{c_s}{r\Omega} \right)^2$$

where C_a is of order unity. Note that the migration rate scales linearly with the planet mass, M_p, and the disk mass (proportional to Σ), and inversely with the temperature of the disk. The scaling with planet mass assumes that the strength of the perturbation in the disk is proportional to the mass of the planet. However, numerical simulations [30] indicate that this perturbation may saturate, and then the migration rate becomes independent of planet mass.

This phenomenon is fundamentally equivalent to dynamical friction: the gravitational interaction of the planet produces a wake, and the wake gravitationally back-reacts on the planet.

5.2 Type II Migration

If the planet is massive enough then a gap forms in the disk. This occurs if the torquing of the disk by the planet exceeds the "viscous torquing" of the disk by itself. That is the planet pushes material away from it faster than disk viscous forces can bring material in.

In contrast to Type I, the planet is now fixed in the disk, and evolves along with it. Therefore the migration rate is purely determined by disk properties:

$$\dot{r} = c_{\mathrm{II}} \alpha \left(\frac{c_s}{r\Omega} \right)^2 r\Omega$$

where the kinematic viscosity of the disk is parameterized by $\nu = \alpha c_s H$ where H is the mixing length, and c_{II} is a parameter of order unity. These timescales tend to be much longer than the timescales associated with type I migration.

5.3 Stopping Migration

If we are to end up with hot Jupiters, we need to have migration terminate before the planet falls into star. In type I, the ILR gets lost because there is no more disk in which it can be raised. Since the ILR contributes an outward torque, this means the migration should speed up onto the star. If the OLR can be lost, then migration will halt. Perhaps the X wind region around the star could clear out a large enough region around the star for this to happen.

Tidal interaction with the star can also stop migration [22], but the star must in general be spinning faster than the orbital period. Eventually the star should slow down, and then the migration would continue.

On the other hand migration may only stop when the disk disappears. Planets continue to form and migrate, and the ones we observe are the ones

that happen to be left when the gas disk disappears. Such a migration probably did not happen in our own solar system for two reasons. First the asteroid belt seems to be chemically stratified in the radial direction. One would expect radial migration of Jupiters to stir it up and thoroughly mix the asteroid types. Secondly, it would be difficult to form the terrestrial planets if the planetesimals were scattered away by a migrating Jupiter.

5.4 Migration and Disk Fragmentation

As pointed out above, the migration rate depends on the disk mass. Therefore, the migration rate indirectly depends on Q the stability criterion. Simulations show [24] that for typical mass planets (of order a Jupiter mass) disks with lower Q, (less than 7) will have gaps, but higher Q disks will experience type I migration. Therefore, in disks unstable to fragmentation, migration would naively be very slow. However, as discussed above, multiple planets can form and significant "migration" can occur due to the interaction between these planets.

References

1. P.J. Armitage and B.M.S. Hansen: Early planet formation as a trigger for further planet formation. Nature **402**, 663 (1999) 212
2. J. Bally, L. Testi, A. Sargent and J. Carlstrom: Disk Mass Limits and Lifetimes of Externally Irradiated Young Stellar Objects Embedded in the Orion Nebula. AJ **116**, 854 (1998) 210
3. W. Benz, W.L. Slattery and A.G.W. Cameron: The origin of the moon and the single-impact hypothesis I. Icarus **66**, 515 (1986) 203
4. A.P. Boss: Evolution of the Solar Nebula. III. Protoplanetary Disks Undergoing Mass Accretion. ApJ **469**, 906 (1996) 211
5. A.P. Boss: Giant planet formation by gravitational instability. Science **276**, 1836 (1997) 211
6. A.P. Boss: Evolution of the Solar Nebula. IV. Giant Gaseous Protoplanet Forma tion. ApJ **503**, 923 (1998) 211
7. A.P. Boss: Temperatures in Protoplanetary Disks. Annual Review of Earth and Planetary Sciences **26**, 53 (1998) 191, 192
8. A.P. Boss: Gas Giant Protoplanet Formation: Disk Instability Models with Thermodynamics and Radiative Transfer. ApJ **563**, 367 (2001) 211
9. W.F. Bottke, A. Morbidelli, R. Jedicke, J. Petit, H.F. Levison, P. Michel and T.S. Metcalfe: Debiased Orbital and Absolute Magnitude Distribution of the Near-Earth Objects. Icarus **156**, 399 (2002) 206
10. R.M. Canup and L.W. Esposito: Accretion of the Moon from an Impact-Generated Disk. Icarus **119**, 427 (1996) 203
11. J.E. Chambers and G.W. Wetherill: Making the Terrestrial Planets: N-Body Integrations of Planetary Embryos in Three Dimensions. Icarus **136**, 304 (1998) 203

12. M. Duncan, T. Quinn and S. Tremaine: The formation and extent of the solar system comet cloud. AJ **94**, 1330 (1987) 208, 209
13. C.F. Gammie: Nonlinear Outcome of Gravitational Instability in Cooling, Gaseous Disks. ApJ **553**, 174 (2001) 211
14. P. Goldreich and W.R. Ward: The Formation of Planetesimals. ApJ **183**, 1051 (1973) 197
15. P. Goldreich and S. Tremaine: Disk-satellite interactions. ApJ **241**, 425 (1980) 213
16. N. Haghighipour and A.P. Boss: On Pressure Gradients and Rapid Migration of Solids in a Nonuniform Solar Nebula. ApJ **583**, 996 (2003) 212
17. M.J. Holman and J. Wisdom: Dynamical stability in the outer solar system and the delivery of short period comets. AJ **105**, 1987 (1993) 206
18. Ž. Ivezić, R.H. Lupton, M. Jurić, S. Tabachnik, T. Quinn, J.E. Gunn, G.R. Knapp, C.M. Rockosi and J. Brinkmann: Color Confirmation of Asteroid Families. AJ **124**, 2943 (2002) 206
19. E. Kokubo and S. Ida: Formation of Protoplanets from Planetesimals in the Solar Nebula. Icarus **143**, 15 (2000) 200
20. G.P. Kuiper: On the Origin of the Solar System. In: *Proceedings of a topical symposium, commemorating the 50th anniversary of the Yerkes Observatory and half a century of progress in astrophysics* ed by J.A. Hynek (New York: McGraw-Hill 1951), p. 357 211
21. H.F. Levison and M.J. Duncan: From the Kuiper Belt to Jupiter-Family Comets: The Spatial Distribution of Ecliptic Comets. Icarus **127**, 13 (1997) 207
22. D.N.C Lin, P. Bodenheimer and D.C. Richardson: Orbital migration of the planetary companion of 51 Pegasi to its present location. Nature **380**, 606 (1996) 214
23. J.J. Lissauer: Planet formation. Annual Review of Astronomy and Astrophysics **31**, 31 (1993) 188
24. G. Lufkin, T. Quinn, J. Wadsley, J. Stadel and F. Governato: Simulations of gaseous disc-embedded planet interaction. MNRAS **347**, 421 (2004) 215
25. D. Lynden-Bell and J.E. Pringle: The evolution of viscous discs and the origin of the nebular variables. MNRAS **168**, 603 (1974) 192
26. G.W. Marcy and R.P. Butler: Planets Orbiting Other Suns. PASP **112**, 137 (2000) 210
27. L. Mayer, T. Quinn, J. Wadsley and J. Stadel: Formation of Giant Planets by Fragmentation of Protoplanetary Disks. Science **298**, 1756 (2002) 211
28. L. Mayer, T. Quinn, J. Wadsley and J. Stadel: The Evolution of Gravitationally Unstable Protoplanetary Disks: Fragmentation and Possible Giant Planet Formation. ApJ **609**, 1045 (2004) 211
29. M. Mayor, D. Queloz, G. Marcy, P. Butler, R. Noyes, S. Korzennik, M. Krockenberger, P. Nisenson, T. Brown, T. Kennelly, C. Rowland, S. Horner, G. Burki, M. Burnet and M. Kunzli: 51 Pegasi. IAU Circ. **6251**, 1 (1995) 187
30. R.P. Nelson, J.C.B. Papaloizou, F. Masset and W. Kley: The migration and growth of protoplanets in protostellar discs. MNRAS **318**, 18 (2000) 214
31. J.B. Pollack, D. Hollenbach, S. Beckwith, D.P. Simonelli, T. Roush and W. Fong: Composition and radiative properties of grains in molecular clouds and accretion disks. ApJ **421**, 615 (1994) 190
32. J.B. Pollack, O. Hubickyj, P. Bodenheimer, J.J. Lissauer, M. Podolak and Y. Greenzweig: Formation of the Giant Planets by Concurrent Accretion of Solids and Gas. Icarus **124**, 62 (1996) 204

33. T. Quinn, S. Tremaine and M. Duncan: Planetary perturbations and the origins of short-period comets. ApJ **355**, 667 (1990) 207, 208
34. S.N. Raymond, T. Quinn and J.I. Lunine: Making other earths: dynamical simulations of terrestrial planet formation and water delivery. Icarus **168**, 1 (2004) 203, 204
35. D.C. Richardson: A Self-Consistent Numerical Treatment of Fractal Aggregate Dynamics. Icarus *115*, 320 (1995) 194
36. D.C. Richardson, T. Quinn, J. Stadel and G. Lake: Large-Scale N-body Simulations of Planetesimal Dynamics. Icarus **143**, 45 (2000) 200, 201
37. V.S. Safronov: *Evoliutsiaa doplanetnogo oblaka*, (1969) 198, 199
38. D.J. Stevenson and J.I. Lunine: Rapid formation of Jupiter by diffuse redistribution of water vapor in the solar nebula. Icarus **75**, 146 (1988) 193
39. P. Tanga, P. Michel and D.C. Richardson: Planetesimal clusters in a Keplerian disk. I. gravitational evolution. A&A **395**, 613 (2002) 197
40. E.W. Thommes, M.J. Duncan and H.F. Levison: The Formation of Uranus and Neptune among Jupiter and Saturn. AJ **123**, 2862 (2002) 210
41. W.R. Ward: Density waves in the solar nebula - Differential Lindblad torque. Icarus **67**, 164 (1986) 213
42. G. Wuchterl, T. Guillot and J.J. Lissauer: Giant Planet Formation. Protostars and Planets **IV**, 1081 (2000) 204, 205

Extrasolar Planetary Systems

S. Ferraz-Mello[1], T.A. Michtchenko[1], C. Beaugé[2] and N. Callegari Jr.[3]

[1] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Brasil (sylvio@astro.iag.usp.br)
[2] Observatorio Astronómico, Universidad Nacional de Córdoba, Argentina
[3] Departamento de Matemática, Universidade Federal de São Carlos, Brasil.

Abstract. This paper is an updated version of lectures given at the Helmholtz Summer School on Extrasolar Planetary Systems (Potsdam, 2003). It includes five sections: Orbit determination, the known planetary systems, chaos, transition from secular to resonant dynamics and planetary migration. The first section is mainly devoted to problems arising in the orbit determination of systems with 2 or more planets from data on the radial velocity of the central star. In the second section, an attempt of classification is done in which the known planetary systems are grouped in 4 classes, one of which includes 47 UMa and our Solar System (orbits with low period ratio and small eccentricities). The phenomena related to mean-motion resonances (MMR) are studied with some detail in the remaining sections, since they may directly affect the stability of many systems. In the appendix, the FFT techniques used in the paper to assess the stability of a system are reviewed.

1 Introduction

This paper deals mainly with systems of planets. More than a dozen planetary systems with 2 to 4 planets are currently known and others are just waiting confirmation. Fig. 1 shows four examples of extra-solar planetary systems and compares the orbits of their planets to those of our own planetary system. The two shown in the top figures have planets in large orbits: υ Andromedae and 55 Cancri. υ And was the first system with several planets to be found around a main-sequence star [9]. The outer planet of 55 Cnc has the widest orbit among known planets [47]. 55 Cnc is an example of a hierarchical system in which some planets lie very close to the star ($a < 0.24$ AU) while the other is far away ($a \sim 6$ AU). Many of the discovered systems are hierarchical. The orbit of the innermost planet of 55 Cnc, at ~ 0.04 AU from the star, is not shown because in the scale of the figure it would appear just as a spot in the center. The two planetary systems shown in the bottom figures are much smaller. The orbits of the two planets of GJ 876 [46] are smaller than the orbit of Mercury. The peculiarity of this system is that the two planets move in orbits with commensurable periods. It is the paradigm of stable resonant planetary systems. In GJ 876, the two apsidal lines almost coincide and the two periastra are aligned. In addition, the common apsidal line rotates slowly in such a way that at each 61 days the two planets come close to a symmetric

S. Ferraz-Mello et al.: *Extrasolar Planetary Systems*, Lect. Notes Phys. **683**, 219–271 (2005)
www.springerlink.com

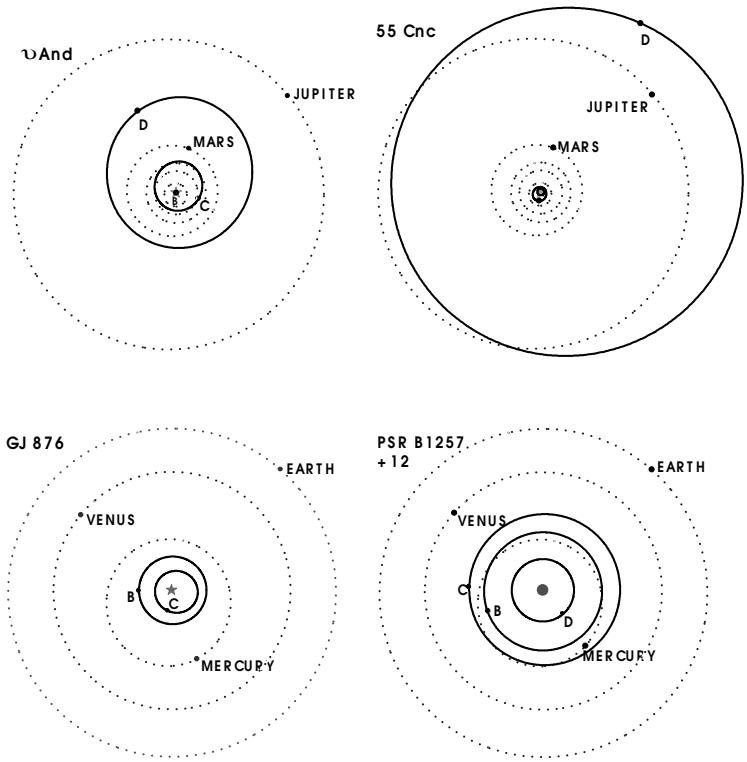

Fig. 1. Orbits in four extra-solar planetary systems compared to Solar System orbits

conjunction in which the two planets pass almost simultaneously through the aligned periastra.

The other planetary system in Fig. 1 is included to remind us that long before the first planet around a main-sequence star was discovered, a system of planets has been discovered around the pulsar PSR B1257+12 [70].

Even today, we do not know how a system of planets could have been formed around a pulsar. This is a very exceptional event and only one more pulsar planet is known in addition to the three orbiting PSR B1257+12. The planets of PSR B1257+12 are much smaller than the planets known around main-sequence stars. While those planets have masses generally between 0.1 and 10 Jupiter masses, the mass of the largest planet around PSR B1257+12 is only 4.3 Earth masses.

At variance with planets in hierarchical systems, the planets in GJ 876 and PSR B1257+12 (and the innermost planets of 55 Cnc) have comparable semi-major axes and may come close one to another, allowing important gravitational interactions between them.

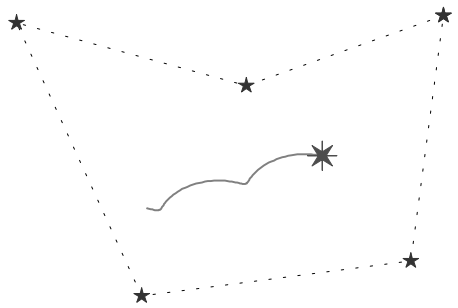

Fig. 2. The sinuous motion of one star due to the presence of a planetary companion. The faint field stars are used as a reference frame

The discovery of planets around main-sequence stars was one of the greatest achievements of Astronomy in the past ten years. Planets are billions of times less luminous than the stars around which they move. The techniques used to discover them consist in observing the star and detecting the effects of the presence of a planet. The most obvious way is the observation of the stellar motion around the common barycenter. We may either observe the displacement of the star or its changing velocity.

The astrometrical technique developed half-century ago to search planetary companions of stars is conceptually very simple [69]. Initially, a set of faint stars is chosen around the star whose displacement is to be studied. Faint stars are very distant objects and can be assumed as fixed during a long time. The field is imaged with the help of a long-focus astrometrical telescope (which allows a large scale on the focal plane) and the star under study is located with respect to the selected field stars (Fig. 2). Relative measurements show the displacement of the star due to its proper motion and, if it is the case, the existence of a planetary companion.[1]

This technique was able to detect the existence of several dark companions in orbit around some stars of the solar neighborhood. Some of these companions were dwarf stars, some were not confirmed by later observations and one of them, Lalande 21185, is still on the list of possible, but unconfirmed planets. Recently, the astrometrical technique was used to measure the displacement of GJ 876, due to its outermost planet, with the fine guidance sensor of the Hubble Space Telescope [5].

If the displacement of stars due to planets and planetary companions could only be observed in a very few cases, the use of new technology spectrographs allowing the radial velocity of the stars to be measured with precisions down to a few m/s revealed the presence of more than one hundred planets (Fig. 3).

The importance of the new spectrographs is clearly seen in Fig. 4, which shows the distribution of the discovered planets (updated December 2004) in

[1] The sinuous motions of Sirius and Procyon were observed by Bessel and Auwers, respectively, many decades before the discovery of their stellar companions.

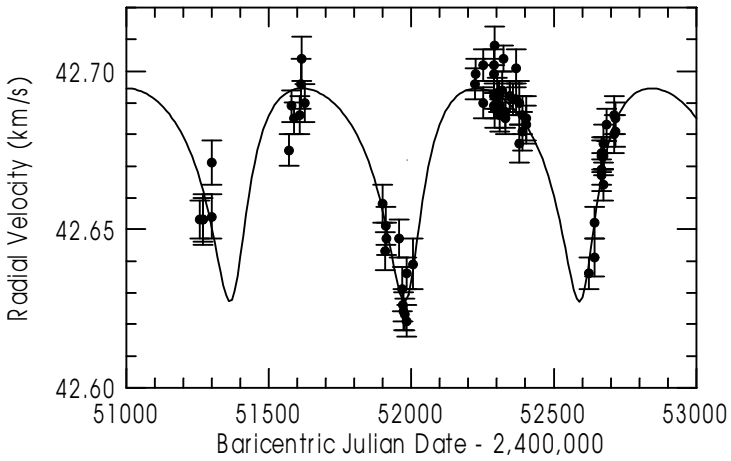

Fig. 3. Radial velocity measurements of the star HD 65216 showing variations due to a Jupiter-size planet *(taken from [48])*

a graph whose axes correspond to the mass of the planet and its distance to the central stars, weighted by the sine of the inclination (unknown in all but a very few cases) and the star mass, respectively.[2] Some amplitudes of variation of the radial velocity of the central stars are also shown. In Fig. 4, the planets discovered before June 1999 are shown by circles, planets discovered after May 2003 are shown by open squares and all others are shown by triangles. This differentiation serves to show how the techniques have progressed since the discovery of the first exoplanet around a main-sequence star in 1995. It is worth stressing that, at variance with many recent astronomical discoveries, the discovery of extra-solar planets was not a product of the use of large telescopes or space telescopes, but of the use of new technology spectrographs.

Figure 5 shows the distribution of the semi-major axes and eccentricities, of the known exoplanets. The histogram in the right panel shows a peak in the interval $0-0.1$ which is due to the fact that all exoplanets with $a < 0.06$ AU have eccentricities less than 0.1 and also by the fact that in ill determined cases, a small eccentricity and even $e = 0$ is assumed. In the interval 0.1-0.5 the eccentricities are more or less equally distributed with an average: 20 ± 5 per class. The decrease starts at $e = 0.4 - 0.5$ and becomes sharp at $e = 0.7$.

It is still too soon to get general taxonomic conclusions about the planet orbits and masses since the discovered sample is strongly biased by the main technique used, the measurement of star radial velocities, and its current capability: high planetary masses and orbital periods not larger than a few years (see Fig. 4).

[2] The coordinates in Fig. 4 are the parameters entering in the definition of the half-amplitude of variation of the radial velocities, $\mathbb{K}$, to be introduced in the next section.

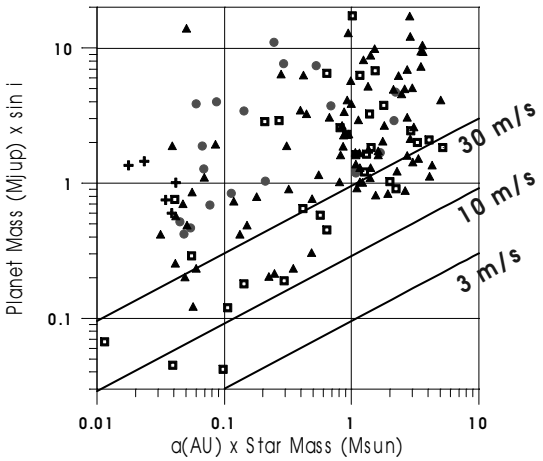

Fig. 4. Distribution of the discovered exoplanets. Crosses show those discovered by photometry (transits) and the other symbols those discovered by radial velocity measurements. Circles are discoveries done before June 1999, open squares are discoveries done after May 2003 and triangles discoveries done between 1999 and 2003. Solid lines: half-amplitudes of variation of the stellar radial velocity due to the exoplanets (updated December 2004)

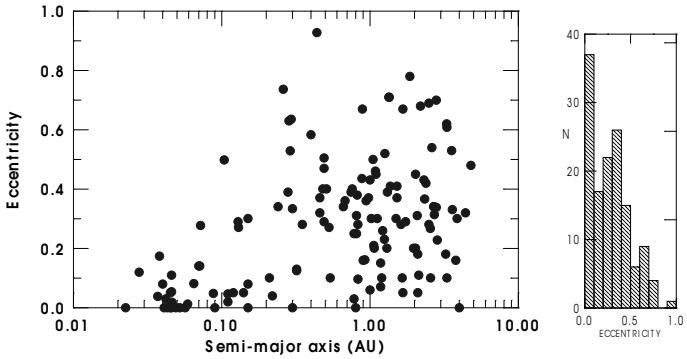

Fig. 5. Distribution of the orbital elements of the discovered exoplanets

The first planets inducing radial velocity variations below the limit of 10 m/s have just been discovered. In addition, at this very moment, discoveries of planets as big as Jupiter, very close to the central star (the so-called "very-hot Jupiters") are being announced. They were not discovered using radial velocity measurements as almost all others, but through the slight diminution of the observed star luminosity when the planet passes in front of it (see Fig. 6). The planets discovered in this way are shown by crosses in Fig. 4. The "transit" technique is only at its beginning, but when used from space telescopes such as CoRoT and Kepler, to be placed in orbit around the

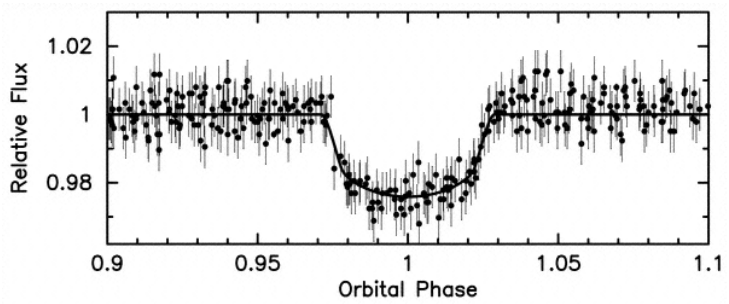

Fig. 6. Light-curve of the star OGLE-TR-113 showing the decrease of the measured light when the planet passes in front of the star *(taken from [37])*

Earth in the forthcoming years, it may completely change our knowledge of the extra-solar systems.

Another detection technique, the timing of radio-pulses, has allowed the first extra-solar planetary system ever known to be discovered around pulsar PSR B1257+12, but it is only restricted to this kind of objects. At last, for completeness, we mention the occasional microlensing by a star with a planet, whose observational signature allows the detection of the planet (see [59]). One planet was discovered in this way around the star OGLE-235, but is not plotted in Fig. 4 because the star type and mass are unknown.

2 Orbit Determination

The basic techniques used to derive orbital elements and masses of extra-solar planets from radial velocity measurements were established long ago. We recall that the first spectroscopic binary stars were discovered by Pickering (Harvard) and Vogel (Potsdam) in 1889-1890. The principles developed for the study of these stars may be found in many classical texts. See for instance [60].

Doppler measurements give the value of the projection on the line of sight of the velocity of the observed star with respect to the observer. Thus, we have to consider the motion in both ends of the light path.

2.1 The Motion of the Observer

To reduce the observations to an inertial frame, it is necessary to consider the motion of the observer with respect to that frame. Some steps are easily described. First, the observer is on the Earth and the Earth is rotating. The Earth rotation carries the observer with a velocity $\approx 460\,\mathrm{m/s}\,\cos\phi$ (ϕ is the latitude of the observing site) along the Earth's geographic parallel. Second, we have to consider that the Earth belongs to a binary system and is moving

with a strongly perturbed elliptic motion around the barycenter of the Earth-Moon system. Its velocity with respect to the system barycenter is $\approx 13\,\mathrm{m/s}$ with a $\approx$ 28-day period. Third is the elliptic motion of the Earth-Moon system around the Sun with an average velocity of $\approx 29.8\,\mathrm{km/s}$. However, because of the eccentricity of the Earth, this velocity is not constant and has a main oscillation of $\approx 0.5\,\mathrm{km/s}$. Next, we have to consider that the Sun is itself moving around the barycenter of the Solar System mainly due to the giant planets. Jupiter, for instance, causes a motion of the Sun around the barycenter of the Solar System of $\approx 13\,\mathrm{m/s}$.

The standard error of a radial velocity measurement, during most of the 20^{th} century, was not better than 1 km/s and it was enough to consider a very simplified model to correct the measurements from the motion of the observer with respect to the barycenter of the Solar System.

Nowadays, relative measurements are getting close to 1 m/s, and simplified models are no longer enough to guarantee a good result. Several public routines exist for that sake. The more precise routines are based on several versions of Bretagnon's VSOP planetary theory [7] and the lunar theory ELP2000/82 of Chapront-Touzé and Chapront [12]. They are the routine `VSOP87E`, by Bretagnon and Francou [8], downloadable from the Centre de Données Stellaires (ref: `VizieR VI/81`) and the expansion published by Ron and Vondrák [63]. Ron and Vondrák's expansion is founded on VSOP 82 and has the nominal precision of 10^{-7} AU/d (that is, $\sim$ 2cm/s). The routine `VSOP87E` is founded on VSOP 87 and its comparison to JPL ephemerides DE 403 confirms the authors estimate that the precision is around 0.5 cm/s (Francou, pers.comm.). Other available routines are `bcvcorr`, developed by the SAO Telescope Data Center, and `SLA_EVP`, found in the SLALIB Positional Astronomy Library (Starlink Project). These two routines use the planetary perturbation codes `BARVEL` written by Stumpff [66], whose comparison to the JPL ephemerides DE96 showed differences less than 42 cm/s in the velocity components. At last, it is worth mentioning the routine `rvcorrect` of the IRAF software package, which does not consider the planetary perturbations and whose nominal precision is only 5 m/s. An alternative to the use of the above-mentioned routines is the direct calculation of the barycentric velocity of the Earth from JPL ephemerides DE 405. In this case, the barycentric coordinates of the Earth must be interpolated from tables and used to obtain the velocity components. The precision in this case is better than a few mm/s (E. M. Standish, pers.comm.).

In the transition of the old $\mathcal{O}(\mathrm{km/s})$ measurements to the new $\mathcal{O}(\mathrm{m/s})$ measurements, we mention the discovery, around 1990, of one "planet" with a period $\approx$ 180 days, a period just half the period of the Earth around the Sun. It was certainly only the consequence of some inaccuracy in the adopted second harmonic of the Earth motion around the Sun. The current routines allow a precise reduction of radial velocity measurements to the barycenter of the Solar System.

A second correction to be introduced comes from the fact that the light takes about 1000 seconds ($\sim$ 0.01 d) to go from one point in the Earth's orbit to the point diametrically opposed to it. This means that the time of one phenomenon will not be the same if the Earth is in one or another point of its orbit. In order to circumvent this problem, it is important to reduce the date of a given observation to a virtual clock located in the Sun, or, better, in the barycenter of the Solar System.

The "heliocentric" and "barycentric" date of one phenomenon are the dates in which it would have been observed by a virtual observer placed in the Sun or in the barycenter of the Solar System, respectively. This correction is easily obtained: it is equal to the time spent by the light to have a displacement equal to the projection of the Earth's radius vector on the line of sight, that is,

$$\Delta t = \frac{1}{c} \begin{pmatrix} \cos\delta\cos\alpha \\ \cos\delta\sin\alpha \\ \sin\delta \end{pmatrix}^T \begin{pmatrix} x_E \\ y_E \\ z_E \end{pmatrix} \tag{1}$$

where α, δ are the right ascension and declination of the star and x_E, y_E, z_E the components of the Earth's position vector in the same equatorial system of reference (astrocentric or barycentric). Dates in published radial velocity measurements are given in either "barycentric" or "heliocentric" Julian date. For data on exoplanetary systems, both dates are acceptable since the consideration of the motion of the Sun around the barycenter of the Solar System has a maximum difference 10^{-4} d, that is, ten times smaller than the current time accuracy of the observations (10^{-3} d). It is worth stressing that the correction considered here just accounts for the differences due to the finite velocity of the light and is not relativistic.

Before closing this introduction, we should mention the more critical case of the reduction of the observed times of arrival of radio pulses coming from a pulsar. The measurement of the time elapsed between the arrival of two consecutive pulses is akin to a radial velocity measurement. However, the high accuracy of clocks makes these measurements equivalent to cm/s in radial velocity scale. In that case, we have to know the motion of the Earth with respect to the Solar System barycenter with very high accuracy. It is necessary to use routines founded on accurate barycentric ephemerides of the planets taking into account all planets. Besides, it is not enough to reduce the observed date to the system barycenter as in the case of Doppler measurements. The clock is also moving with 10^{-4} of light speed, and the relativistic difference between the measured time and the time given by an inertial clock can no longer be neglected. (For a complete list of the effects to be considered in the reduction, in this case, see [11]).

2.2 One-Planet Kinematics

Would the star be isolated, its radial velocity would be constant (at least for times short as compared to the galactic motions time scale). However, if it has one or more companions, no matter if stars or planets, the star will move around the common barycenter (B) of the system. The velocity of this motion (with respect to the observer) is the quantity that is measured from Earth. Let us use the kinematics of planetary motion to derive it.

In a reference frame centered in the star, the position and velocity of one planet are given by

$$\mathbf{r} = r\cos f\ \mathbf{i} + r\sin f\ \mathbf{j} \tag{2}$$

$$\mathbf{v} = -\frac{2\pi a}{T\sqrt{1-e^2}}[\sin f\ \mathbf{i} - (e+\cos f)\ \mathbf{j}] \tag{3}$$

where

$$r = \frac{a(1-e^2)}{1+e\cos f}, \qquad \frac{df}{dt} = \frac{2\pi a^2}{Tr^2}\sqrt{1-e^2}, \tag{4}$$

$r = |\mathbf{r}|$, f is the true anomaly, a is the astrocentric semi-major axis, T is the orbital period, e is the eccentricity, $\mathbf{i}$ and $\mathbf{j}$ are two unit vectors in the plane of the motion: $\mathbf{i}$ is in the direction of the periastron and $\mathbf{j}$ is orthogonal to it.

The astrocentric position vector $\mathbf{r}$ of the planet may be trivially converted into the barycentric position vector $\mathbf{R}$ of the star (see Fig. 7). The conversion of Eq. (3) to the velocity of the star with respect to the barycenter is done in the same way:

$$\mathbf{V} = \frac{m}{\mathbb{M}}\frac{2\pi a}{T\sqrt{1-e^2}}[\sin f\ \mathbf{i} - (e+\cos f)\ \mathbf{j}] \tag{5}$$

where m is the mass of the planet, M is the mass of the star and $\mathbb{M} = M+m$ is the total mass of the system.

The next step is to project this velocity on the line of sight. It is convenient, beforehand, to introduce a reference frame to be used in these studies. The first reference system used in the study of visual double stars was Herschel's astrocentric frame where the $x-$ and $y-$axes are tangent to the celestial sphere and directed toward North and East, respectively. The $z-$axis

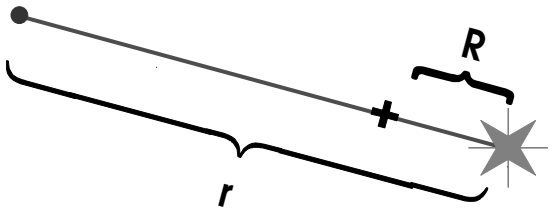

Fig. 7. Barycentric position vector of the star ($\mathbf{R}$) and astrocentric position vector of the planet ($\mathbf{r}$). The cross shows the barycenter of the system

is directed along the line of sight, away from the observer. For the sake of studying spectroscopic binaries, astronomers modified Herschel astrocentric frame rotating it around the $z-$axis and placing the $x-$axis along the nodal line defined by the intersection of the orbital plane with the sky tangent plane (see [32]). This change was introduced because radial velocity measurements do not allow the position of the intersection line to be known. This system is being used as reference for the orbit of the extra-solar planets. The $x-$axis is taken along the intersection line and directed towards the point crossed by the planet when moving towards the observer. The $y-$axis is defined so that the system is right-handed.

The adopted definition of the $x-$ axis corresponds to the practice of measuring the position angle of the periastron from the intersection (γ) where the planet is moving towards the observer (the star is receding from the observer and near the radial velocity maximum).[3] For instance, the planet of the star HD 65216, whose radial velocity curve is shown in Fig. 3, is at periastron on JD 2,452,601, a short time after the minimum of the light curve, and the position angle of the periastron is $\omega = 198°$.

In the adopted system, the vectors **i**, **j** and **k** are given by:

$$\mathbf{i} = \begin{pmatrix} \cos\omega \\ \sin\omega\ \cos i \\ -\sin\omega\ \sin i \end{pmatrix} \qquad \mathbf{j} = \begin{pmatrix} -\sin\omega \\ \cos\omega\ \cos i \\ -\cos\omega\ \sin i \end{pmatrix} \qquad \mathbf{k} = \begin{pmatrix} 0 \\ \sin i \\ \cos i \end{pmatrix} \tag{6}$$

These vectors are the same classically used to give the unit vectors associated with the orbital plane on an arbitrary reference frame, with $\Omega = 180°$, inclination i and argument of periastron $\omega + 180°$. The use of $\omega + 180°$ (instead of ω as in classical formulas) is due to the use of the point γ located in the opposite side of the ascending node **N** as origin.

In the dynamical study of systems with two or more non co-planar planets, we no longer use the plane of the sky as the fundamental reference frame. A new reference plane, closer to the orbital planes (the invariable plane, for instance) is a better choice in this case. New angles are then defined. It is important to recognize from the very beginning that the position angle ω is measured from the intersection with the sky plane and is akin to a longitude. Even if a new reference frame is used, the distance from the periastron to the sky plane should be ω as given in the orbit determination. If the longitude of a hypothetical node is added to ω, as sometimes done, the resulting orbit is no longer the orbit issued from the observations.

The measured radial velocity is the component of the vector **V** along the $z-$axis plus the radial velocity of the system barycenter, that is:

$$V_z = \mathbb{K}\,[\cos(f + \omega) + e\cos\omega] + V_r \tag{7}$$

where

[3] The radial velocity is considered positive if the star is receding from the observer and negative if it is approaching him.

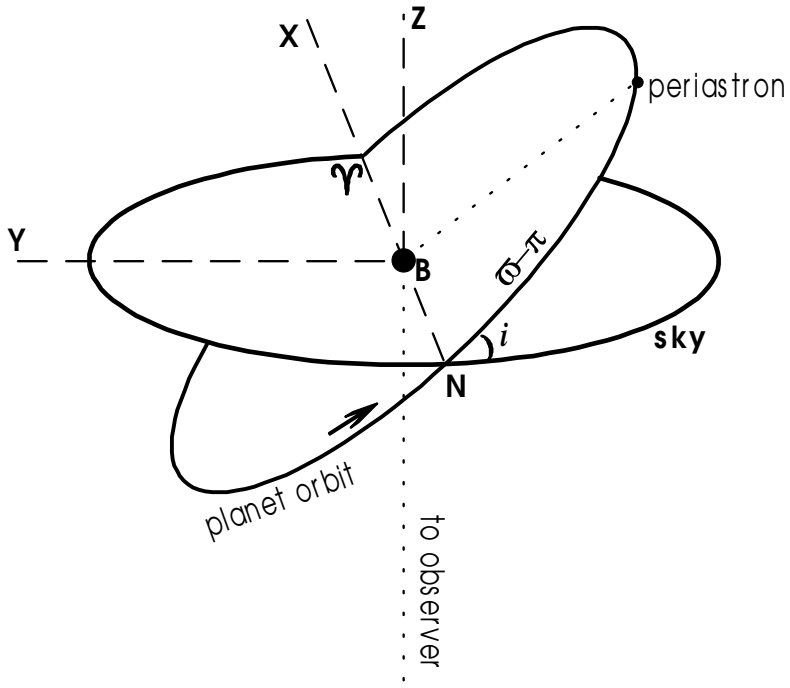

Fig. 8. The rotated astrocentric reference frame showing the planet orbit plane, the plane tangent to the celestial sphere (sky) and their intersection. The origin of reckoning of the angles is the intersection (γ) in which the planet is moving towards the observer

$$\mathbb{K} = \frac{m}{\mathbb{M}} \frac{2\pi a}{T} \frac{\sin i}{\sqrt{1-e^2}}. \tag{8}$$

$\mathbb{K}$ is the half-amplitude of variation of the radial velocity.

Two typical examples are obtained using data from our own planetary system: For a Jupiter-like planet in a Jupiter-like orbit around a Sun-like star (and $i \approx 90°$), $\mathbb{K} = 12.7$ m/s; in the same case, but for an Earth-like planet in an Earth-like orbit we have instead, $\mathbb{K} = 9.2$ cm/s. For a Jupiter-like planet in an Earth-like orbit, we have $\mathbb{K} = 29\,\mathrm{m/s}$. It is worth stressing that the variation of f is not uniform and, therefore, the curve $V_z(t)$ is not a sinusoid, except when $e = 0$. When $e \neq 0$ the curve presents a skewness that allows us to determine the eccentricity and the position angle of the periastron. Eq. (7) has 6 unknowns: V_r, $\mathbb{K}$, T, e, ω and the time of periastron (necessary to obtain the true anomaly f at the date of each observation). There are some classical formulations to determine these parameters from the curve of radial velocity (see [60]), however, it is generally easy to use a best-fit routine starting from simple guesses to determine them.

We may use the third Kepler law:

$$\frac{a^3}{T^2} = \frac{G(M+m)}{4\pi^2}$$

to relate a and T and separate, in Eq. (8), the known and unknown parts. We obtain,

$$\frac{m^3}{\mathbb{M}^2}\sin^3 i = \frac{\mathbb{K}^3 T}{2\pi G}(1-e^2)^{3/2}. \tag{9}$$

The left-hand side:

$$f(M,m,i) = \frac{m^3}{\mathbb{M}^2}\sin^3 i = \left(\frac{m}{\mathbb{M}}\sin i\right)^3 \mathbb{M} \tag{10}$$

is the so-called mass-function.

In the 1-planet problem, there is no way to separate the three quantities involved in the mass function. Usually, the mass of the star is taken from sophisticated models. However, one must keep in mind that even for Hipparcos stars having the best available spectroscopy and astrometry, the more accurate models do not allow to know the masses better than ≈ 8 percent [1]. This fact supersedes some discussions on the nature of the published planetary elements, if astrocentric or barycentric. The difference between coordinates in these systems is much smaller than the uncertainty in our knowledge of the star mass.

It is worth repeating that it is impossible to know the orientation of the intersection line on the sky tangent plane from spectroscopic (and photometric) observations only.

2.3 Primary and Derived Parameters. Planetary Aberration

We define as primary parameters those issued from the fit of the observations to a Keplerian model. Table 1 shows the parameters determined from the radial velocities of the star HD 65216 (the same star whose radial velocity curve is shown in Fig. 3).

The parameters shown in Table 2 are deduced from those given in Table 1. For their derivation, we have to adopt a mass for the star. In the values given in the first column (Case I), we adopted the same star mass given by Mayor et al. [48]. In the last column (Case II), we considered exactly 1 solar mass, a value about 10 percent larger than the given one. This is perhaps an exaggeration of the incertitude on the star mass, but it was chosen to allow the reader to see how the adopted mass affects the results. For the same reason, results are given with one digit beyond what would be reasonable. It is worth noting that some parameters show only a weak dependence on the chosen star mass, while others are affected in almost the same proportion.

We also have to consider that the star is moving and, because of that motion, the time scale of the motion measured with a clock moving with

Table 1. Parameters determined from the observed radial velocities of HD 65216 [48]

parameter	planet HD 65216 b
$\mathbb{K}$ (m/s)	33.7 ± 1.1
Apparent Period (days)	613.1 ± 11.4
Periastron time(†)	50,762 ± 25
Eccentricity	0.41 ± 0.06
ω (degrees)	198 ± 6
V_r (km/s)	42.674 ± 0.002

† barycentric Julian date) − 2,400,000.

Table 2. Parameters derived from those in Table 1. Epoch= BJD 2,452,000

parameter	case I†	case II‡
Mean Longitude at the epoch (degrees)	204.9	204.9
a (AU) (osculating)	1.374	1.413
$f(M, m, i)$ $(10^{-9} M_\odot)$	1.845	1.845
$m \sin i$ $(M_{\rm Jup})$	1.216	1.286
$m \sin i$ $(10^{-3} M_{\rm star})$	1.262	1.227

† assuming for the star mass $M = 0.92 M_\odot$.
‡ assuming for the star mass $M = 1.0 M_\odot$.

the star is not the same as that resulting from measurements made from Earth. This is the same effect known as planetary aberration, which affects the astrometrical observations of a solar-system body. It is easy to see that a phenomenon whose duration measured with a clock moving with the star is Δt_s, has a duration

$$\Delta t_\oplus = \Delta t_s \left(1 + \frac{V_r}{c}\right) \tag{11}$$

when measured by a clock on Earth. For instance, if the star is receding ($V_r > 0$), the period of the motion measured from Earth will be larger than the actual one. Typical radial velocities in nearby stars are of the order of some tens of km/s, thus affecting the fourth digit of the period. In some cases, they may reach 100 km/s and the relative correction to the periods is 3×10^{-4}. Many stars already have nominal periods determined with a relative precision better than 10^{-4} what means that the aberration of light has to be taken into account. In the case of the star HD 65216, the actual osculating period of the planet around the star is smaller than the apparent period by ~ 0.08 days. In this case, the difference is much smaller than the incertitude in the period determination. However, as the precision in the period determination increases with a longer observation time span, the time scale correction will become important even for stars with moderate radial velocities.

2.4 The Case of Two Planets

In an isolated system formed by one star with mass M and two planets with masses m_1, m_2, the barycentric velocities are such that $M\mathbf{V} + m_1\mathbf{V}_1 + m_2\mathbf{V}_2 = 0$. Substituting the barycentric velocities of the planets by their astrocentric velocities $\mathbf{v}_k = \mathbf{V}_k - \mathbf{V}$, there follows

$$\mathbf{V} = -\frac{m_1}{\mathbb{M}}\mathbf{v}_1 - \frac{m_2}{\mathbb{M}}\mathbf{v}_2 \tag{12}$$

$\mathbb{M}$ is the sum of all masses in the system. The measured radial velocity is the component of the vector $\mathbf{V}$ along the z-axis, plus the radial velocity of the system barycenter. If the mutual attraction of the planets is disregarded, we have

$$V_z = \sum_k \mathbb{K}_k[\cos(f_k + \omega_k) + e_k \cos\omega_k] + V_r \tag{13}$$

where $\mathbb{K}_k$ is, for each planet, the same parameter obtained in the 1-planet case (Eq. 8). The mass function is now

$$f(M, m_k, i_k) = \frac{m_k^3(M + m_k)}{\mathbb{M}^3}\sin^3 i_k. \tag{14}$$

The cancellation of $\mathbb{M}$ and $M + m_k$ done to obtain Eq. (10) cannot be done here.

Another point to be taken into account is the fact known since Laplace (at least), that mutual perturbations affect the mean value of osculating elements. Two of these variations are particularly important: the mean perturbations in semi-major axis and the mean longitude at the epoch. The latter one is a drift of that value that directly affects the period of the motion. The classical first-order formulas giving these effects ([67]; see also [19]) are the following:

$$\widehat{a}_k = < r_k > = a_k\left(1 - \frac{1}{2}\sigma_k\right), \qquad \widehat{T}_k = T_k(1 - \sigma_k) \tag{15}$$

where, assuming that $a_1 < a_2$, we have

$$\sigma_1 = \frac{m_2}{M + m_1}\alpha^2\frac{db^0_{1/2}}{d\alpha}, \qquad \sigma_2 = -\frac{m_1}{M + m_2}\left(\alpha\frac{db^0_{1/2}}{d\alpha} + b^0_{1/2}\right); \tag{16}$$

$\alpha = \frac{a_1}{a_2}$ and $b^0_{1/2}(\alpha)$ is the lowest order Laplace coefficient. For $\alpha \approx 0.636$ (as in planets close to a 2/1 commensurability of periods), we have $b^0_{1/2} \approx 2.268$, $db^0_{1/2}/d\alpha \approx 1.132$ and

$$\sigma_1 = -0.46\frac{m_2}{M + m_1}, \quad \sigma_2 = 3.0\frac{m_1}{M + m_2}. \tag{17}$$

The 1-planet kinematical model assumes that the motion is Keplerian and lies on an ellipse of semi-major axis $\widehat{a}$ with period $\widehat{T}$. When two (or more)

planets are involved, this ellipse is no longer the osculating one. Therefore, the use of Kepler's third law is not correct.

Let us keep the notation a, T for the osculating elements and write, instead of Eq. (8),

$$\mathbb{K}_k = \frac{m_k}{\mathbb{M}} \frac{2\pi \widehat{a}_k}{\widehat{T}_k} \frac{\sin i_k}{\sqrt{1-e_k^2}}. \tag{18}$$

for each planet. From the given definitions of $\widehat{a}_k$ and $\widehat{T}_k$ and Kepler's law (for the osculating elements), we obtain

$$f(M, m_k, i_k)\left(1+\frac{1}{2}\sigma_k\right) = \frac{\mathbb{K}_k^3 \widehat{T}_k}{2\pi G}(1-e_k^2)^{3/2}. \tag{19}$$

One important result derived from the above equations is that the ratio

$$\frac{m_1 \sin i_1}{m_2 \sin i_2}$$

is not affected by the indetermination of the inclinations and star mass. Therefore, if the two planets are coplanar, the mass ratio may be known from the observations. This is a very important point, since adiabatic evolutions follow lines of constant mass ratio and are only weakly dependent on the actual masses of the planets. (see [3, 4, 23]; see Sect. 6). However, we have to keep in mind that, while plausible on the ground of planetary formation theories, coplanarity cannot be known from a kinematical orbit determination.

2.5 3-Body Fits

Let us start this section by comparing the predicted evolution of the orbits in two non-interacting Keplerian orbits and the actual evolution given by the numerical integration of a 3-body model. The example chosen is the radial velocity curve of the star HD 82943. The planetary system of this star is one of the two confirmed paradigms of systems near a 2/1 commensurability of periods. Fig. 9 shows the computed radial velocity curves in the two cases and, with a solid line, the difference between them (kinematical minus dynamical). This line shows that in the whole interval of the available observations, $1999.1 - 2003.4$, there is no significant difference between the results given by the two models. However, for a larger interval, important differences appear and a Keplerian fit can no longer be used. An orbit determination taking into account the gravitational interaction of the two bodies will be necessary.

One important point when comparing 3-body radial velocity curves is that they can give an independent way of assessing the planetary masses. In the 3-body model used to construct the curves of Fig. 9, the masses of the two planets are the published values of $m.\sin i$; that is, it is assumed that the planetary orbits are being seen exactly edge-on. If the observations

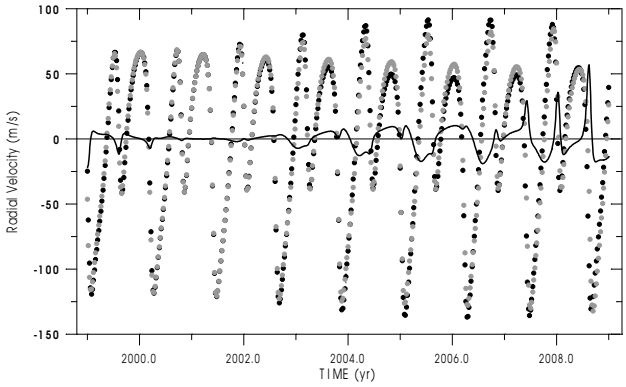

Fig. 9. Radial Velocity of the star HD 82943 following kinematical *(gray dots)* and dynamical *(black dots)* models corresponding to [48] elements and masses and coinciding at the date 2001.247. (*the solid line shows the difference kinematical−dynamical (*$\sin i = 1$*)*)

done in the next 4 years show a difference with the Keplerian model larger than that predicted in Fig. 9, it will mean that the actual masses of the planets are larger than $m. \sin i$, (that is, $\sin i < 1$, and $i < 90°$). Therefore, the extended observation of systems of planets with significant gravitational interaction may allow us to know their actual masses. However, if a new unknown (for instance, the inclination) is introduced in the fitting of a 3-body model to the observed radial velocity of that star, now, it will give rise to a correlation factor close to 1 between some derivatives making impossible to get the solution. A more extended observing time span is necessary to solve the indetermination.

The above discussion considered only the indetermination concerning the inclination. Another factor to be taken into account is the incertitude in the value of the star mass and its propagation to other quantities in the process. We adopted in the 3-body model used to construct Fig. 9 the mass $M_{\rm star} = 1.15 \pm 0.09 M_\odot$ as determined with the Geneva stellar evolutionary models [48]. However, other values of $M_{\rm star}$ may be found in the recent literature, for instance $0.93 \pm 0.09 M_\odot$ [1], $1.05 M_\odot$ (Geneva planet search web page, July 31th, 2002) and $1.11 M_\odot$ [13].

In order to avoid polluting the 3-body simulations with the inaccuracy of the star mass, it is convenient to use an adimensional formulation of the equations of motion. For that sake, we introduce a change in the units in which the star mass is equal to 1 and the distance unit is such that the value of G does not change (this is necessary to avoid transferring the inaccuracy on M to the gravitational constant G). We may adopt one of the usual time units and fix the value of G. For instance, we may adopt the day as time unit and take for G the standard value [17]

$$G = \left(\frac{2\pi}{365.2568983263}\right)^2 = (0.01720209895)^2. \tag{20}$$

The corresponding length unit is a modified astronomical unit (MAU) for the exoplanetary system considered. It is easy to see that

$$1 \text{ MAU} = 1 \text{ AU} \sqrt[3]{\frac{M_{\text{star}}}{M_\odot}}$$

The adimensional formulation depends on only two parameters: the masses of the planets in units of the star mass. The other external parameters are only introduced when a solution of the adimensional problem is compared to the radial velocity measurements. For that sake they are first projected on the line of sight and then the units are converted to our units by multiplying them first by the gauges

$$\gamma_k = \sqrt[3]{\frac{M_{\text{star}}}{M_\odot}} \sin i_k \qquad (k = 1, 2)$$

and then by the factor 1731.4568 to convert AU/d into km/s. In order to continue this discussion, we have to assume that during the period of the observations, the inclinations remained constant. Should a variation of the inclinations in the observational period to be taken into account, the procedure would become much more intricate even if simplified by the introduction of the invariable plane. The equations in the general case were not yet derived.

Therefore, the four undetermined parameters related to masses and inclinations in the dynamical orbit determination are $m_1/M, m_2/M, \gamma_1, \gamma_2$. If the two planet orbits may be assumed as coplanar, $\gamma_1 = \gamma_2 = \gamma$ and the number of unknowns is reduced to three: $m_1/M, m_2/M, \gamma_1$. However, these three involve four unknown parameters: $M, m_i, m_2, \sin i$ showing that we cannot produce an independent determination of the star mass M by radial velocities measurements only, and we are tied to use the values coming from astrophysical models. At least, in the above given procedure, the results are not affected by the inaccuracy in the value of the star mass.

As already discussed in Sect. 2.4, the ratio of the 2 planetary masses is not affected by the indetermination and may be determined from the observations.

3 The Known Planetary Systems

Currently (February 2005), 15 systems with 2 to 4 planets around main-sequence stars are known.[4]

[4] *cf.* "Extra-solar Planets Encyclopedia" `http://www.obspm.fr/planets` (©J.Schneider) ε Eri was not considered here since the second planet of this star is excessively uncertain.

At variance with the rest of this paper, this section is deeply rooted on the existing observational data. The data collected in the next tables can be substantially modified in a short time because of the new spectrographs in use since 2004, the increase of telescope time availability for transit detections and also the continuous disclosure of long-period planets due to the long span of observation accumulated during the past 10 years. For a frequent update, see `http://www.astro.iag.usp.br/~dinamica/exosys.htm`.

When the ratio of orbital periods in each pair of planets in consecutive orbits is plotted (Fig. 10), we see a more or less continuous distribution (at least in a logarithmic scale) between two extreme cases: In the lower end, we have planet pairs with $P_2/P_1 \approx 2$ and, at the upper end, $P_2/P_1 > 150$.

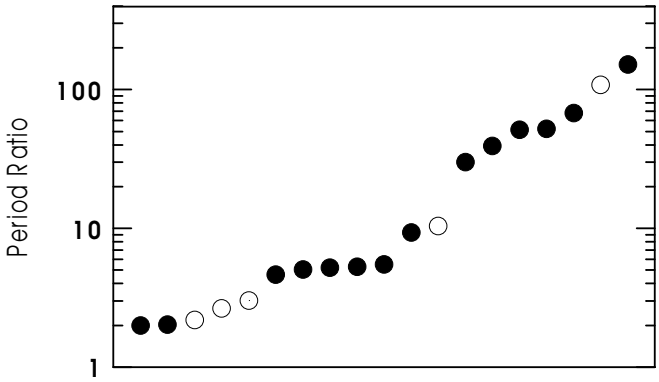

Fig. 10. Ratio of the orbital periods of pairs of planets in consecutive orbits in increasing order. Open circles indicate pairs for which the existence of one of the components is still under discussion

Class Ia. Planets in Mean-motion Resonance (MMR)

We put in Class I those planets at the lower end of the distribution shown in Fig. 10. Having large masses and eccentricities and orbiting in relatively close orbits, these planets are liable to strong gravitational interaction and are significantly perturbed in orbital timescales. They are unable to remain stable if not tied by a mean-motions resonance (MMR). They are among the more interesting extra-solar systems for Celestial Mechanics studies. The first two pairs are the two well-known pairs in 2:1 mean-motion resonance: HD 82943 and GJ 876 (= Gliese 876). The next three pairs appear in Fig. 10 with open circles because of doubts concerning one of the planets in the pair. They are HD 128311, 47 UMa and the planets b and c of 55 Cnc (= ρ^1 Cnc). The data on the outer planet of HD 128311 still are very uncertain; in what concerns the other two, there is no consensus among the observers about the

existence of the outer planet in the pair [56]. If the existence of these outer planets and the resonances are confirmed, they will be added to the three known resonant cases to form a set of examples whose study may contribute to our understanding of the physics underneath the capture of exoplanets into resonance. However, exactly in reason of this importance, we should be very careful and accept them only after the current doubts are dissipated. We note that the planets of 47 UMa have small eccentricities and may remain stable even in non-resonant orbits. For this reason, this system was not included in this class (Ia). It is being considered as paradigm of a separate sub-class (Ib), which will be discussed at the end of this section.

We also include in class Ia, the planet pairs of HD 128311 and HD 202206. The given elements of HD 128311 do not show a MMR, but simulations with the given data show instability in short time (even choosing the most favorable values for the unknown perihelia and relative positions). With the given large masses and eccentricities, they should not survive out of a resonance. The orbits of HD 202206 b,c are well known, but the mass of one of the companions allows to class it rather as a low-mass brown dwarf than as a planet. The large mass of HD 202206b is certainly responsible for the capture of HD 202206c in a high-order resonance (other planet pairs with the same period ratio are non-resonant and belong to Class II).

One may note that in the two first systems in Table 3, the outer planet was discovered before the inner one and the names b,c appear in inverse alphabetical order. This is a prosaic fact without a particular importance, but it deserves to be pointed because it is responsible for errors in some Internet

Table 3. Planets in mean-motion resonance

Star Star mass	Planet	mass $\times \sin i$ ($m_{\rm Jup}$)	Period (days)	Period ratio	Semi-major axis (AU)	Eccentricity
HD 82943([1])	c	1.7	219.5	1.99	0.75	0.39
$1.15M_\odot$	b	1.8	436.2		1.18	0.15
GJ 876([2])	c	0.597	30.38	2.01	0.13	0.218
$0.32M_\odot$	b	1.90	60.93		0.21	0.029
HD 128311([3])	b	2.58	420.5	~ 2	1.02	0.30
$0.80M_\odot$	c	3.24	919 (?)		1.74	0.29 (?)
55 Cnc([4])	b	0.784	14.67	2.99	0.115	0.02
$1.03M_\odot$	c (?)	0.217	43.93		0.24	0.44
HD 202206([5])	b	17.5	256.2	5.06	0.83	0.433
$1.15M_\odot$	c	2.41	1296.8		2.44	0.284

([1]) Ref: Ferraz-Mello et al. [24]
([2]) (= Gliese 876) Ref: Laughlin et al. [41]
([3]) Ref: California & Carnegie web page (`exoplanets.org`)
([4]) (=ρ^1 Cnc) Ref: McArthur et al. [49]
([5]) Ref: Correia et al. [15]

posted data where the reverse alphabetical order led to the interchange of some elements of the 2 planets.

Class II. Non-Resonant Planets with Significant Secular Dynamics

The period ratios of the planet pairs considered in this class lie above 4.6 making difficult a capture into a MMR. The gravitational interaction between these planets may be strong, but the angular momentum conservation limits the eccentricity variations allowing them to remain stable even if not in a MMR. They present long-term variations, primarily described by secular perturbations, large variation of the eccentricities and interesting dynamical effects such as the alignment and anti-alignment of the apsidal lines (see [54]).

Among the extra-solar planetary systems, the best-known example in this class is the system formed by the outer planets of υ And. They do not seem to be in MMR or close to one and are paradigms of systems showing apsidal lock due to a non-resonant secular dynamics.

The distinction between planets of this class and the next one is only circumstantial. One could be tempted to define the class by the presence of alignment or anti-alignment of the apsidal lines (the feature improperly called "secular resonance"). However, as discussed by Michtchenko and Malhotra [54], for every given mass and semi-axis ratios, one may have alignment, anti-alignment or circulation of $\Delta\varpi$ as well, just as a function of initial eccentricities and $\Delta\varpi$. This is a kinematical feature, not a dynamical one: all possible motions belong to a same continuous family of solutions (see the discussion in Sect. 5). Thus, we have used as a tracer, the variation of $\Delta\varpi$. If $\Delta\varpi$ is oscillating about 0 or π, or if it does not oscillate, but varies in a tortuous way, the system (or pair of planets) belongs to class II. If $\Delta\varpi$ varies almost uniformly, the system (or pair of planets) belongs to class III.

Another difference between classes II and III is the sensitivity to variations of the elements. The stability of the orbits of planet pairs in class II depends critically on the parameters and may become unstable if they are slightly changed. Because of this sensitivity, published orbits often lead to catastrophic events in short times. This, obviously, does not mean that a catastrophe will occur. It just means that the current accuracy of the orbit determinations is not good.

The first pair of planets in Table 4 is formed by the outermost planets of μ Ara (= HD 160691). The first orbit determination of these planets led to guess that they were in a 2:1 MMR, but recent determinations indicate a much higher orbit for μ Ara c. The orbit of the last pair included in Table 4, HD 37124, is poorly known. Different choices of elements may lead either to a very unstable orbit in which $\Delta\varpi$ oscillate about 180° or to a stable orbit with almost uniformly circulating $\Delta\varpi$. The current elements do not allow deciding between classes II and III.

Table 4. Non-resonant planet pairs with a significant secular dynamics

Star/ Star mass	Planet	mass $\times \sin i$ ($m_{\rm Jup}$)	Period (days)	Period ratio	Semi-major axis (AU)	Eccentricity
μAra([1])	b	1.67	645.5	4.63	1.50	0.20
$1.08 M_\odot$	c	3.1	2986		4.17	0.57
55 Cnc([2])	e	0.045	2.808	5.2	0.038	0.17
$1.03 M_\odot$	b	0.784	14.67		0.115	0.02
υAnd([3])	b	0.64	4.6171	52.2	0.058	0.019
$1.3 M_\odot$	c	1.79	241.18	5.32	0.805	0.26
	d	3.53	1282.6		2.543	0.25
HD 12661([3])	b	2.30	263.6	5.48	0.83	0.35
$1.07 M_\odot$	c	1.57	1444.5		2.56	0.20
HD 169830([4])	b	2.88	225.62	9.32	0.81	0.31
$1.4 M_\odot$	c	4.04	2102		3.6	0.33
HD 37124([5])	b	0.72	153	~ 10	0.54	0.1
$0.91 M_\odot$	c	1.3(?)	1595 (?)		2.5 (?)	0.7 (?)

([1]) (= HD 160691) Ref: McCarthy et al. [50]
([2]) (=ρ^1 Cnc) Ref: McArthur et al. [49]
([3]) Ref: Fischer et al. [25]
([4]) Ref: Mayor et al. [48]
([5]) Ref: California & Carnegie web page (`exoplanets.org`)

We have included in class II the whole υ And system, notwithstanding the high period ratio of the two first planets: 52.2. The planet υ And b has indeed many characteristics of one planet belonging to one class III pair: a large long-period variation of the eccentricity (the eccentricity grows from 0 to 0.39 in the half-period 150,000 years). However, the periastron of this planet oscillates with the mentioned long period around the periastron of υ And c (see Fig. 11). Even though the current longitude of the perihelion of υ And b is ill determined and a new determination may change the result presented here, this example shows the difficulty of finding a significant absolute criterion to classify non-resonant planet pairs in these two classes.

Class III. Hierarchical Planet Pairs

Lee and Peale [43] used the denomination "hierarchical" in the same sense as here, that is to indicate planets with large ratio P_2/P_1. The extreme example is HD 38529 where $P_2/P_1 \sim 150$.

However, they used a limit lower than here and used the word to designate several planetary system, including HD 12661 for which $P_2/P_1 = 5.48$. They have shown that the apsidal lines of the HD 12661 planets oscillate about anti-alignment and for this reason, following the criterion here introduced, it was included in class II. We included in class III only planets at the right of the jump seen in Fig. 10. The large period ratios (> 30) allow the

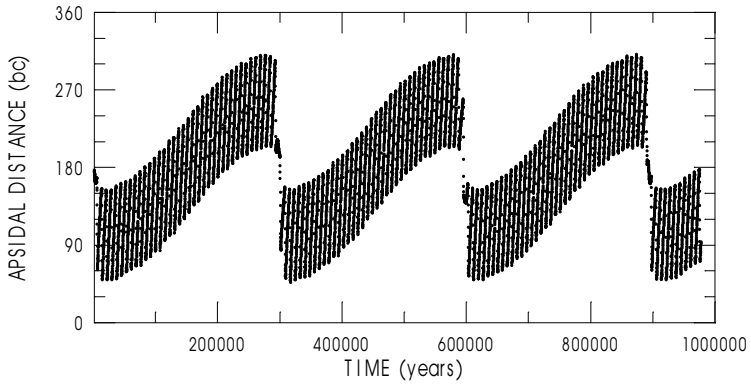

Fig. 11. Long-period oscillation of the difference between the periastra of the υ And b and c ($\varpi_c - \varpi_b$)

existing gravitational interaction between these planets to be weaker than in the previous classes and the probability of capture in a MMR is negligible. The interaction leads to long-period variations in the eccentricities, uniform or almost uniform variation of $\Delta\varpi$ and variations in the osculating elements associated with the wobbling of the system barycenter. One exception is the inner pair of planets of υ And that have been included in class II notwithstanding the high period ratio (52.2). In hierarchical pairs, the variation in the eccentricity of the inner planet may be large. For instance, the eccentricity of HD 74156 b has a total variation of 0.3.

It is worth noting that some planets in this class have large masses and one of the planets of HD 168443 is probably a low-mass brown dwarf rather than a planet.

One consequence of the weaker gravitational interaction is the low sensitivity of the published orbits to small variations. The orbits in this class are probably worse than are those of the previous classes (since they involve planets with long periods), but simulations always show stable motions.

Class Ib. Low-Eccentricity Near-Resonant Planet Pairs

This is a special class including systems with low-eccentricity planets in successive orbits with small period ratio but with circumstances making the gravitational interaction between the planets less important. No confirmed extra-solar system of this kind is yet known around MS stars. 47 UMa will be included here if forthcoming observations confirm that 47 UMa c indeed exists and that the planets of this system are not tied to a MMR. We put in this class the planets pairs of the pulsar PSR B1257+12. We may also include here the planets of the two sub-systems of the Solar System. In the planetary system of the pulsar PSR B1257+12 (see Table 7), as well as in the inner Solar System, the orbits are close one to another, but the masses

Table 5. Hierarchical planet pairs

Star/ Star mass	Planet	mass × sin i ($m_{\rm Jup}$)	Period (days)	Period ratio	Semi-major axis (AU)	Eccentricity
HD 168443([1])	b	7.7	58.116	29.9	0.29	0.529
$1.01 M_\odot$	c	16.9	1739.5		2.85	0.228
HD 74156([2])	c	1.86	51.643	39.2	0.294	0.636
$1.27 M_\odot$	b	6.17	2025		3.40	0.583
HD 11964([3])	b	0.11	37.82	51.3	0.229	0.15
$1.12 M_\odot$	c	0.70	1940		3.167	0.3
μAra([4])	d	0.044	9.55	67.6	0.09	–
$1.08 M_\odot$	b	1.67	645.5		1.50	0.20
55 Cnc([5])	c (?)	0.217	43.93	103	0.24	0.44
$1.03 M_\odot$	d	3.912	4517.4		5.257	0.33
HD 38529([6])	b	0.78	14.309	152	0.129	0.29
$1.39 M_\odot$	c	12.7	2174.3		3.68	0.36

([1]) Ref: Udry et al. [68]
([2]) Ref: Naef et al. [56]
([3]) Ref: California & Carnegie web page (`exoplanets.org`)
([4]) (= HD 160691) Refs: Santos et al [64], McCarthy et al. [50]
([5]) Ref: McArthur et al. [49]
([6]) Ref: Fischer et al. [25]

are relatively small. In the outer Solar System, the masses are larger, but the distances between the planets are always large allowing this system to show long-term stability notwithstanding the fact that the planets have low period ratios and are not in MMR.

In this class, one characteristic is the presence of a large number of pairs in near resonance. (But the sample is small and near-resonance may be just a coincidence). The most conspicuous examples are the pair Jupiter-Saturn with a period ratio ~ 2.5 (5:2 MMR) (see [53]) and the two outer planets of the pulsar PSR B1257 +12 with period ratio ~ 1.5 (3:2 MMR). The closeness of the pulsar outer planets to commensurability produces perturbations in longitude ("great inequality") large enough to be observed from Earth thus allowing the very existence of the planets to be confirmed in a few years of continuous observations ([62, 45]).

Table 6. Low-eccentricity Near-resonant Planet Pairs

Star/ Star mass	Planet	mass × sin i ($m_{\rm Jup}$)	Period (days)	Period ratio	Semi-major axis (AU)	Eccentricity
47 UMa([1])	b	2.9	1079.2	2.64	2.1	0.05
$1.03 M_\odot$	c (?)	1.1	2845.0		4.0	0

([1]) Ref: Fischer et al. [25]

Table 7. Planetary System of the pulsar PSR B1257+12([1])

Planet	Mass ($m_{\rm Earth}$)	Period (days)	Period ratio	Semi-major axis (AU)	Eccentricity	Inclination (degrees) ([2])
A	0.020([3])	25.262	2.63	0.19	0.	
B	4.3	66.5419	1.47	0.36	0.0186	53
C	3.9	98.2114		0.46	0.0252	47

([1]) Ref: Konacki and Wolszczan [36]. Adopted pulsar mass: $1.4 M_\odot$
([2]) over the tangent plane to the celestial sphere
([3]) adopting the inclination $i = 50^\circ$

Table 8. Outer Solar System

Star	Planet	Mass ($m_{\rm Jupiter}$)	Period (years)	Period ratio	Semi-major axis (AU)	Eccentricity
Sun	Jupiter	1.0	11.866	2.500	5.204	0.0489
	Saturn	0.30	29.668	2.831	9.584	0.0571
	Uranus	0.046	83.987	1.958	19.178	0.0468
	Neptune	0.054	164.493		30.004	0.0112

4 Chaos

Chaos is a common feature in systems with many degrees of freedom and is, nowadays, part of the standard knowledge in Dynamical Astronomy. A superb introduction to the aspects of the phenomenon was recently given by Contopoulos [14].

4.1 The Outer Solar System

The neighborhood of the outer Solar System planets is filled by mean-motion resonances. The actual motion of these planets occur very near to two-planet MMR giving raise to a dense set of three-planet resonances, which occurs when the periods corresponding to two two-planet MMR form critical linear combinations. Jupiter and Saturn lie very close to the 5:2 MMR, Uranus is confined between the domains of the overlap of the 7:1 MMR with Jupiter and the 2:1 MMR with Neptune in one side, and the 3:1 MMR with Saturn in the other; finally, Neptune is close to the 2:1 resonance with Uranus.

The resonant structure of the Solar System was studied by Michtchenko and Ferraz-Mello [52] by means of dynamical maps constructed in the neighborhood of the four planets whose equations of motion were integrated using the accurate Everhart's RA-15 integrator [18]. The initial values of the semi-major axis and eccentricity of one planet were uniformly distributed on a rectangular grid covering the vicinity of the actual position of the planet,

while the initial positions of the other planets were chosen to be the actual ones at epoch JD2451100.5. The initial inclination and angular orbital elements of the chosen planet were fixed at their present values. The short-term oscillations (of the order of the orbital periods) were eliminated by employing a low-pass filtering procedure *on-line* with the numerical integration. The resulting semi-major axes were used to construct dynamical maps as described in the Appendix. The spectral number N was defined as the number of spectral peaks above 5% of the largest peak. The counting is stopped at $N = 100$ and $\log N$ is plotted on the (a_0,e_0)-plane.

Uranus neighborhood

As an illustration, we present here the dynamical map of the neighborhood of Uranus (Fig. 12). This neighborhood is dominated by three two-planet resonances: 2:1 with Neptune, 3:1 with Saturn and 7:1 with Jupiter. There are also several narrow bands of chaotic motion associated with three-planet MMR (see Table 9). The small hatched domains in Fig. 12 are those in which collisions (i.e. disrupting close approaches) occur in the time span of the numerical integration (50 Myr).

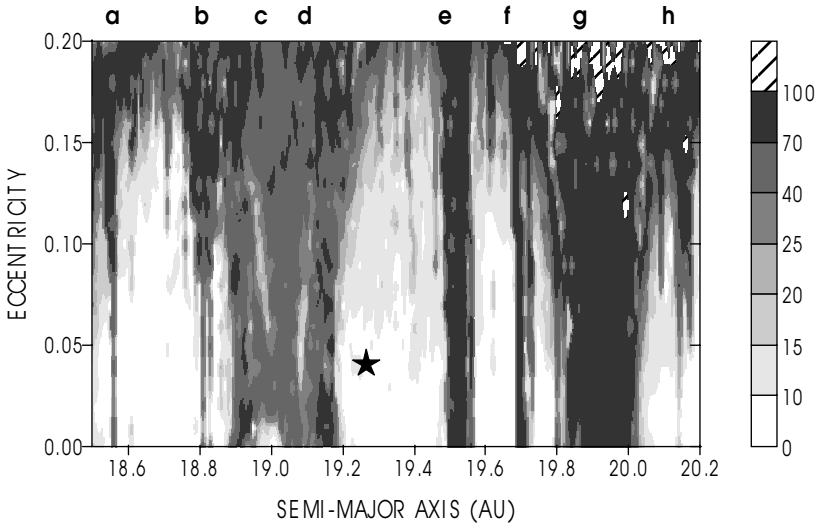

Fig. 12. Dynamical map of the neighborhood of Uranus. The main apparent MMR are indicated on top of the figure by the letters a-h and identified in Table 9. The mapped quantity is the spectral number N (see the Appendix). The star indicates the actual position of Uranus

Table 9. Main MMR in Uranus' neighborhood

Position (fig12)	MMR†	Remarks (beat)
a	J–S–4U	beat of J–7U and S–3U
b	3J–10S+7U	beat of 2(2J–5S) and J–7U
c	U–2N	
d	J–7U	
e	2J–6S+3U	beat of 2J–5S and S–3U
f	2J–6S+3U	beat of 2J–5S and 2(S–3U)
g	S–3U	
h	2J–3S+6U	beat of 2J–5S and –2(S–3U)

† k_1J $\pm$ k_2S $\pm$ k_3U $\pm$ k_4N means the MMR $k_1 n_{\rm J} \pm k_2 n_{\rm S} \pm k_3 n_{\rm U} \pm k_4 n_{\rm N} \approx 0$

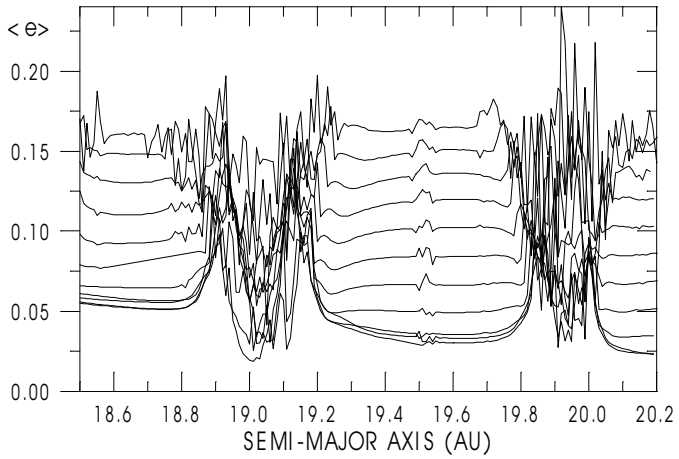

Fig. 13. Mean eccentricities over 50 Myr of solutions whose initial conditions are in the domain $a = 18.5 - 20.2$ AU, $e = 0 - 0.2$. Each line correspond to a value of the initial eccentricity, which was sampled at each $\Delta e = 0.02$

It is interesting to compare this dynamical map with a plot showing the amplitude of the perturbations of the motion for each initial condition in the grid. Fig. 13 shows the average eccentricities over 50 million years for solutions starting at each point of a grid similar to that used in Fig. 12, but with eccentricities sampled at each $\Delta e = 0.02$ to avoid a clumsy plot. This figures shows that the S-3U MMR is the one where the perturbations in eccentricity are the largest. If the initial eccentricity is larger than 0.15 the perturbation is large and many solutions in this domain were bound to a catastrophic event in less than 50 Myr (hatched areas in Fig. 12). In the neighborhood of the U-2N MMR, the perturbations are also large but less important than in the S-3U MMR. We may note that the more important perturbations accumulate in the V-shaped border of the U-2N MMR resonance, while the perturbations are very small in the central part of the resonance. This result is consistent

with the known dynamics of the U-2N MMR, which is characterized by thin chaotic regions along the seperatrices and almost regular motions far from them ([10]; see Sect. 5).

4.2 The Angular Momentum Constraint

One characteristic of chaotic domains is that two orbits with neighbor initial conditions may evolve in a completely different way. One common assumption is that one solution may evolve to reach any point in the chaotic domain and it is often true in the dynamics of asteroids. In that restricted case, the disturbing planet acts as an infinite source of energy and angular momentum and the asteroid may move in an almost arbitrary way in the chaotic domain as long as a topological barrier does not constrain its motion. In the case of a planetary system, the planets are themselves the only sources of energy and angular momentum and the evolution may obey at the conservation laws of these quantities. The domains of initial conditions shown in dynamical maps do not correspond to solutions of equal energy or angular momentum.

Both conservation laws are important, but the angular momentum plays a key role (see [39]). The z-component of the angular momentum of a system of N planets is

$$\mathcal{L}_z = \sum_{k=1}^{N} m_k n_k a_k^2 \sqrt{(1-e_k^2)} \cos i_k. \tag{21}$$

This equation is exact when Keplerian elements associated with Poincaré or Jacobi canonical variables are used (see [23]), but it also holds when using astrocentric (or barycentric) Keplerian elements with an error $\mathcal{O}(m_k^2)$. The angular momentum constrains the semi-major axes, eccentricities and inclinations of the planets of a system (The mean-motions n_k are linked to the semi-major axes through the third Kepler's law).

In the case of almost coplanar planets, taking a reference plane close to the planes of their motions, we have

$$\mathcal{L}_z \approx \sum_{k=1}^{N} m_k n_k a_k^2 \sqrt{(1-e_k^2)}. \tag{22}$$

The semi-major axes are inversely proportional to the energies of the planets and the variation of the energy of one planet is the work done by the forces disturbing its motion around the star. If no close approaches occur, the energy variations compensate themselves to give an average variation close to zero. This means that in absence of close approaches, the quantities a_k in the above equation may be regarded as constants. The other quantities in Eq. (22) are the functions $\sqrt{(1-e_k^2)}$ whose values vary monotonically in the interval [0,1]. When the eccentricity of the planet increases, $\sqrt{(1-e_k^2)}$ decreases. Then, when the eccentricity of one planet increases, to keep $\mathcal{L}_z$ constant,

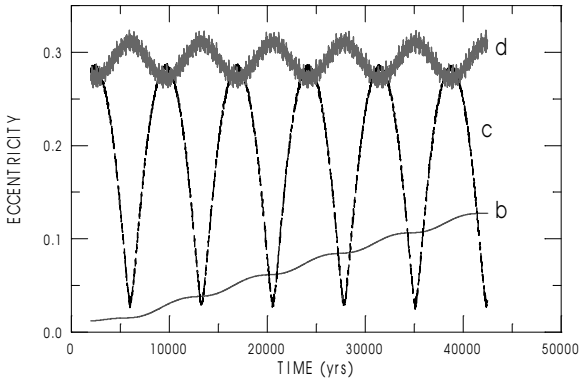

Fig. 14. Variation of the eccentricities of the υ And planets

the eccentricity of at least one of the other planets may decrease. However, they cannot decrease below zero, and this constrains the increasing of the eccentricity of the planets. Since the Angular Momentum of the system does not allow the planetary eccentricities to grow, close approaches are avoided, and the system may survive for a long span of time. The most notorious example in the outer solar system is the interaction between Jupiter and Saturn. Simulations lasting almost 1 billion years have shown that a fictitious outer solar system starting with the current eccentricities and inclinations and in which Jupiter and Saturn are inside the chaotic domain of the MMR $2\mathrm{J}-5\mathrm{S}$ evolves without dramatic eccentricity increases and close approaches [52].

The *e*-*e* coupling due to the constant angular momentum is more easily seen if we introduce, instead of $\mathcal{L}$ the so-called Angular Momentum Deficit:

$$\mathcal{A} \approx \sum_{k=1}^{N} m_k n_k a_k^2 \, (1 - \sqrt{(1 - e_k^2)}). \tag{23}$$

The Angular Momentum Deficit is related to the z-component of the Total Angular Momentum through $\mathcal{A} = \sum_{k=1}^{N} m_k n_k a_k^2 - \mathcal{L}_z$ and is thus a constant as long as the semi-major axes a_k may be regarded as constants. If we introduce the approximation $1 - \sqrt{(1 - e_k^2)} \approx \frac{1}{2} e_k^2$, Eq.(23) becomes

$$\mathcal{A} \approx \frac{1}{2} \sum_{k=1}^{N} m_k n_k a_k^2 e_k^2,$$

which is a weighted sum of squares. The *e*-*e* coupling is then a consequence of the fact that if one of the terms in this sum (that is, one eccentricity) increases, others must decrease to keep the sum unaltered. For instance, in the case of two planets, if one eccentricity increases, the other decreases; in addition, since one eccentricity cannot decrease below zero, the other cannot

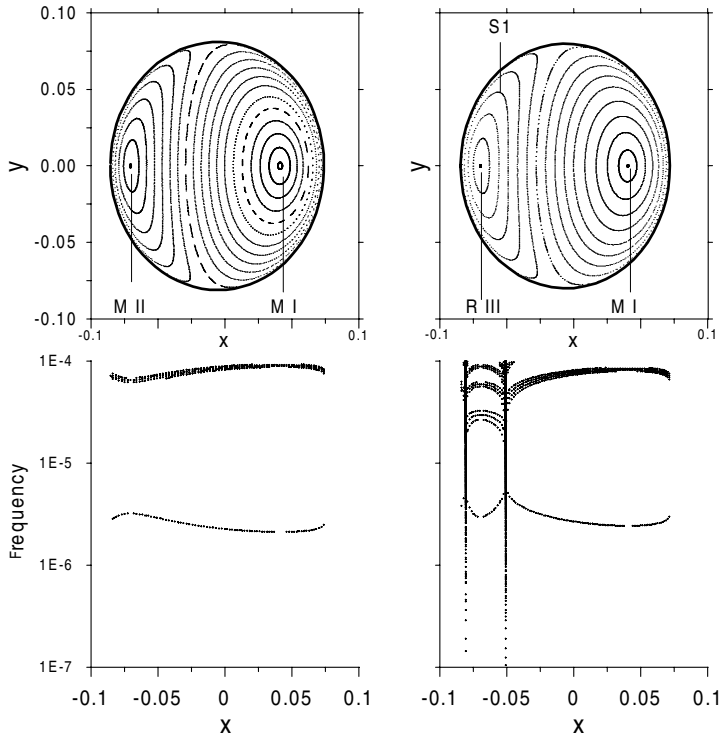

Fig. 15. *Top:* Surfaces of Section of the averaged Low-Eccentricity planetary system in the neighborhood of the 2:1 resonance for two energy levels. Axes $x = e_{\mathrm{U}} \cos \Delta\varpi$, $y = e_{\mathrm{U}} \sin \Delta\varpi$. *Bottom:* Dynamic power spectra of the solutions shown in the surfaces of section parameterized by the value of x corresponding to $y = 0$. Frequencies in yr^{-1} (see [10])

increase indefinitely, and if $\mathcal{A}$ is not large the eccentricities cannot grow to reach values allowing a close approach of the two planets.

5 Transition from Secular to Resonant Dynamics

We summarize in this section some results of the study of the low-eccentricity dynamics of the U–2N MMR by Callegari et al. [10] and show how the dynamics of a system of 2 planets evolves when it passes from outside to inside the MMR.

Figure 15 shows two surfaces of section of planetary motion defined by the condition $\theta_1 = 2\lambda_{\mathrm{N}} - \lambda_{\mathrm{U}} - \varpi_{\mathrm{U}} = 0$ and represented on the plane ($x = e_{\mathrm{U}} \cos \Delta\varpi$, $y = e_{\mathrm{U}} \sin \Delta\varpi$) with $\Delta\varpi = \varpi_{\mathrm{U}} - \varpi_{\mathrm{N}}$, for two different energy levels.

At the lower energy level, (Fig. 15, left) the system is outside the 2:1 MMR and its dynamics is dominated by secular interactions and characterized by two secular modes of motion known from the linear secular theories (see [58]). There are two periodic solutions: one where $\Delta\varpi = 0$ and another one where $\Delta\varpi = \pi$. In the surfaces of section, these solutions appear as fixed points. These periodic orbits and the domains around them are indicated as Mode I and Mode II. The Mode I of motion is always located on the right-hand side of the section, while Mode II is located on the left-hand side.

The smoothed curves around the fixed points are quasi-periodic solutions.

Around Mode I, the angle $\Delta\varpi$ oscillates about 0, while around Mode II, $\Delta\varpi$ oscillates about π; in both cases, the eccentricities of the planets vary regularly around their values in the periodic orbits. Between the two cases, the angle $\Delta\varpi$ is in *retrograde* circulation, and the motion is a composition of the two main oscillations. It can be shown that, in this near resonance zone, the critical angles

$$\theta_1 = 2\lambda_{\rm N} - \lambda_{\rm U} - \varpi_{\rm U}$$
$$\theta_2 = 2\lambda_{\rm N} - \lambda_{\rm U} - \varpi_{\rm N}$$

are in circulation. No infinite-period separatrix exists between the solutions around Mode I and Mode II and they form just one continuous family of solutions.

This fact is clearly shown in the dynamic power spectrum given at the bottom left frame of Fig. 15 where only two fundamental frequencies and a few harmonics appear. The lower line (at $\sim 3 \times 10^{-6}$ yr^{-1}) corresponds to one of the fundamental frequencies. We can see that the line shows a discontinuity near the fixed points, indicating that the amplitude associated to the secular frequency tend to zero in these solutions. The secular period is $\sim 400,000$ years around Mode I and $\sim 300,000$ years around Mode II. The second fundamental frequency can also be seen in the dynamic power spectrum (at $\sim 10^{-4}$ yr^{-1}). It is the frequency associated with the circulation of the critical angles (transversal to the surface of section).

At the higher energy level, (Fig. 15 right), the resonance 2/1 is already visible. The surface of section seems to be the same as the previous one, but the dynamic power spectrum shows that an important difference exists. The two vertical broad lines seen in the dynamic power spectrum indicate that the curve labeled by S1 in the surface of section is indeed a separatrix (even if it does not show any visible feature of a separatrix). The domain inside S1 corresponds to a new regime of motion (called Regime III by Callegari et al. [10]) in which the critical angles θ_1 and θ_2 oscillate (librate) about 0 and π, respectively. Their difference $\Delta\varpi$ librates around π.

These surfaces of section show why the evolution of a system of two low-eccentricity planets towards resonance is always driven to the condition known as resonance corotation, with anti-aligned periastra (see Sect. 6).

The study of the surfaces of section with energies intermediary between the two shown in Fig. 15 shows that the separatrix S1 emanates from the very position of the periodic orbit labeled Mode II in the secular dynamics. It then grows to encompass the whole domain of solutions about $\Delta\varpi = \pi$ and even large amplitude oscillations around $\Delta\varpi = 0$. For higher energies, the domain inside S1 shows the rise of centers and saddles corresponding to true secular resonances and to new regimes of motion inside the MMR domain (Fig. 16). For a detailed discussion, see [10].

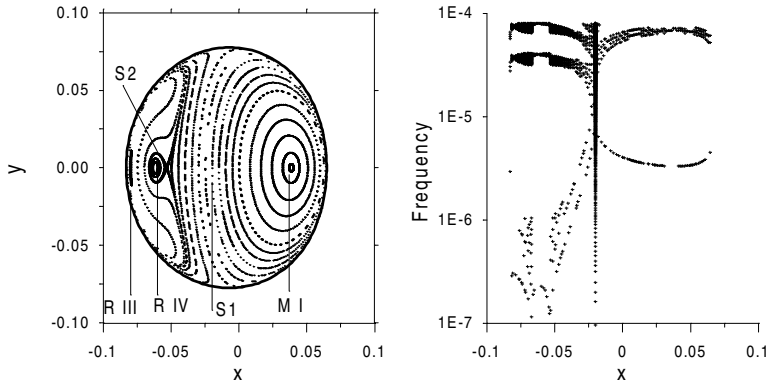

Fig. 16. Same as Fig. 15 for a higher energy level showing a secular resonance inside the MMR (see [10])

Secular Resonance

The denomination "secular resonance" is currently used with two different meanings and this is a nuisance. We will try to set the strong dynamical difference of the two situations in which this name is used.

Generally, it is said that a secular resonance occurs when the motion of the periastron (perihelion) or of the node is trapped by a forced frequency (for instance by another periastron or node). For instance, an asteroid is said to be in secular resonance when the perihelion of its orbit oscillates about the perihelion of Jupiter. The controversy comes from the fact that from the dynamical point of view, this oscillation is not a resonance at all. Looking at the equations of this oscillation, we see that the main perturbations of the eccentricity and perihelion are given by two periodic terms having as frequencies, the forced frequency of the motion of Jupiter's perihelion and a proper frequency associated with the asteroid's perihelion. The amplitude of the forced oscillation is a given function of the masses and elements while the amplitude of the proper oscillation is an arbitrary constant depending only on the initial conditions. In asteroids whose initial conditions are such that the amplitude of the proper oscillation is smaller than the amplitude

of the forced oscillation, the forced oscillation dominates; consequently, in a reference system rotating with the orbit of Jupiter, the motion of the perihelion of the asteroid is an oscillation. At variance, if the amplitude of the proper oscillation is larger than the amplitude of the forced oscillation, the perihelion of the asteroid circulates. In the first case, it is often said that a "secular resonance" occurs. However, from the dynamical point of view there are no differences between the two cases depicted. One solution may pass from one case to another by continuously varying the amplitude of the proper oscillation.

In the case of two planets, the phenomenon is the same. It was shown in the previous section that outside the MMR, the dynamics is characterized by two periodic solutions: the two centers seen in the surface of section. These periodic orbits and the domains around them are indicated as Mode I and Mode II. In Mode II, $\Delta\varpi$ always oscillates about π; In Mode I, we see both: solutions in which the angle $\Delta\varpi$ oscillates about 0 and solutions in which the angle $\Delta\varpi$ is in circulation. In the latter case, the torus section enclosing the Mode I center also encloses the origin and the angle $\Delta\varpi$ reaches all values between 0 and 2π.

From the dynamical or topological point of view, to have a "resonance" would mean to have the apsidal proper frequency vanishing. The dynamic power spectrum of Fig. 15 (*bottom*) shows that this does not occur. The apsidal proper frequency is always different of zero and the frequency variation from Mode II oscillations to circulations and to Mode I oscillations (from right to left in the surface of section of Fig. 15 (*top*)) is smooth (continuous).

Nevertheless, situations exist in which a true secular resonance exists.

In the asteroid case, true secular resonances were the subject of a great deal of results in the past 25 years (for a thorough discussion see [55]). In the planetary case, examples of true secular resonance were given [52, 54, 10]. The example given by Callegari et al. [10] is of a true secular resonance inside the U-2N MMR. It is shown in Fig. 16. The main qualitative difference between this figure and Fig. 15 (*right*) is the appearance inside the MMR domain (inside the separatrix S1), of a second separatrix (S2). The dynamic power spectrum shows that the lower frequency becomes equal to zero at the two points corresponding to the intersections of S2 with the x-axis. The direction of the phase flow inside the separatrix S2 is inverted with respect to what it is outside that separatrix. A true resonance happens (with one proper frequency passing through zero). The dynamic power spectrum also shows that the solutions in the immediate neighborhood of S2 are chaotic.

6 Apsidal Corotation Resonances (ACR)

The surfaces of section in Fig. 15 show that the anti-aligned periastra of the solutions in Mode II of the 2:1 MMR are preserved in the evolution of a system of two low-eccentricity planets towards resonance (notwithstanding

the transition of the critical angles from circulation to libration). This is a situation completely different of that occurring in the capture of a particle into a resonance with one planet (or satellite). In the latter case, the capture happens when the sidereal periods of particle and planet (or satellite) become approximately commensurable, the critical angle[5]

$$\theta_1 = (p+q)\lambda' - p\lambda - q\varpi$$

is trapped in the neighborhood of 0 or π, but the periastron of the particle orbit (whose longitude is ϖ) continues to rotate. That is, $\Delta\varpi = \varpi - \varpi'$ varies monotonically. However, for some well defined values of the eccentricity of the planet (satellite) orbit, it happens that not only the angle θ_1 but also the angle

$$\theta_2 = (p+q)\lambda' - p\lambda - q\varpi'$$

is trapped into a resonance as well. Consequently $\Delta\varpi$ is no longer circulating, but librating [20]. This phenomenon is known as "corotation resonance" in disc and ring dynamics (motion accompanying the resonance pattern speed), extending that denomination to beyond the narrow 1:1 resonance case of the epicyclical orbits theory. The characterization of corotation resonance is given by Greenberg and Brahic [29]: "resonance that depends on the eccentricity of the perturbing satellite, rather than on the eccentricity of the perturbed particle", means that we have a corotation resonance when θ_2 is in libration. However, in the restricted problem, θ_2 cannot be in libration if θ_1 is circulating. Therefore, the simultaneous libration of $\Delta\varpi$ and θ_1 is synonymous of corotation resonance.

The resonant planar planetary three-body problem (averaged over short-period terms) is a two-degrees of freedom system (see [3]). A stationary solution of it is a solution for which the angles

$$\begin{aligned} \theta_1 &= (p+q)\lambda_2 - p\lambda_1 - q\varpi_1 \\ \theta_2 &= (p+q)\lambda_2 - p\lambda_1 - q\varpi_2. \end{aligned} \tag{24}$$

and their conjugated momenta I_1, I_2 remain constant in time. It is important to notice that the stationary solutions of the averaged model correspond to periodic orbits of the non-averaged model. In general, corotation resonances are periodic solutions of the averaged equations (quasi-periodic solutions of the non-averaged equations): they are oscillations around the fixed point of the averaged system. One such solution with finite amplitude oscillations will be generically refereed to as "Apsidal Corotation Resonance" (ACR).

Although corotation resonances have gained certain notoriety in exoplanetary dynamics, they are not new and can be found in our own Solar System. It has long been known that the Io-Europa pair is trapped in a 2:1 MMR and

[5] the unprimed quantities refer to the trapped particle and the primed ones to the trapping planet (or satellite).

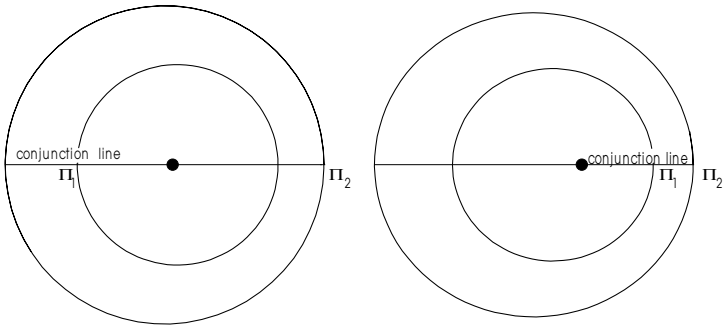

Fig. 17. Symmetric stationary solutions in the 2:1 MMR. *Left:* $(0,\pi)$-ACR − The conjunction occurs with the inner planet in the periastron and the outer planet in the apastron ($|\Delta\varpi| = 180^\circ$); *Right:* $(0,0)$-ACR − The conjunction occurs with both planets in the periastron ($\Delta\varpi = 0^\circ$).

is in corotation resonance (see [19]). Both θ_1 and θ_2, and consequently $\Delta\varpi$, oscillate (with very small amplitude) around fixed values.

The exact ACR is defined, in this case, by $\theta_1 = 0$ and $\Delta\varpi = \pi$. This case is referred to as a $(0,\pi)$-ACR, a denomination more accurate than just saying that the apsides are *anti-aligned*.

GJ 876, the first extra-solar resonant planetary system ever discovered, also exhibits corotation resonance, although in its case the angular variables oscillate around $\theta_1 = 0$ and $\Delta\varpi = 0$ (see [41]).

This case is referred to as a $(0,0)$-ACR (the apsides are *aligned*). The other extra-solar system discovered in the same commensurability (HD82943) also shows corotation resonance (see [24]). These examples seem to show that ACR constitutes a strong stabilizing mechanism for high-eccentricity resonant planets orbiting in nearby orbits.

Another important class of corotation resonance exists, in the 2:1 MMR, in which the apsidal lines are trapped in an asymmetric way, where the equilibrium values of the angles are not equal to zero or π, as shown in Fig. 18 [21, 42, 22, 3]. These "asymmetric ACR" were found in both the 2:1 and 3:1 MMR and the published orbits of the inner planets of 55 Cnc seem to correspond to such an asymmetric configuration [3, 71].

The diversity of apsidal corotations does not stop here. Numerical studies of the system HD82943 [31, 33] have shown a new type of stationary solutions at very high eccentricities. Although symmetric, they correspond to equilibrium values $\theta_1 = \pi$, $\Delta\varpi = \pi$. We have called them (π,π)-ACR. The orbital elements initially published for HD82943 (Geneva planet search web page, July 31th, 2002) seemed to correspond to such a configuration. In that case, the conjunction occurs with the inner planet in the apastron and the outer planet in the periastron. It is worth emphasizing that the apastron of the inner planet is more distant from the star than the periastron of the outer

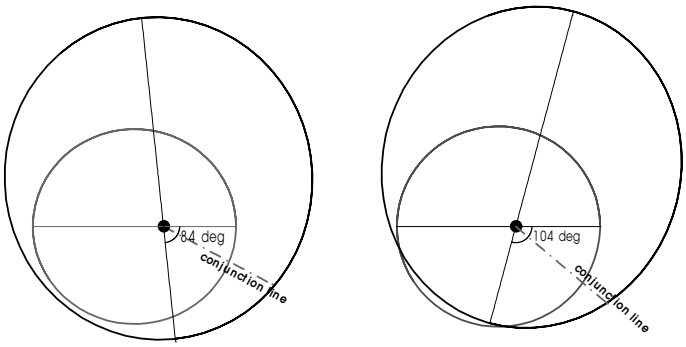

Fig. 18. Asymmetric stationary solutions in the 2:1 MMR. *Left:* Stationary solution with $|\Delta\varpi| = 84^\circ$, $e_1 = 0.286$ and $e_2 = 0.3$. *Right:* Stationary solution with $|\Delta\varpi| = 104^\circ$, $e_1 = 0.17$ and $e_2 = 0.38$. The conjunctions occur in a position between the two periastra

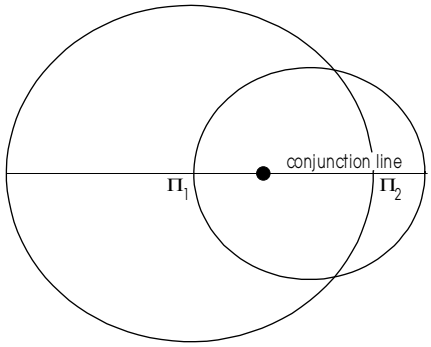

Fig. 19. Symmetric stationary solutions in the 2:1 MMR with $e_1 = e_2 = 0.4$: (π, π)-ACR − The conjunction occurs with the inner planet in the apastron and the outer planet in the periastron ($|\Delta\varpi| = 180^\circ$)

planet. Then, the orbits intersect themselves and, at conjunction, the inner planet is further away from the star than the outer planet. Lee [44] used numerical simulations with differential migration to map the extent of this new family finding that stable solutions are located beyond the line corresponding to two colliding anti-aligned orbits: $a_1(1 + e_1) = a_2(1 - e_2)$. It is obvious that solutions close to this line can only be stable if the masses of the planets are very small and that the true limit of the stable (π, π)-ACR should be a function of the planet masses. The used model is not valid if the two planets come very close one to another, the minimal distance allowed being proportional to the cube root of the planet masses (see [26]).

Systematic searches using averaged analytical models and numerical simulations with adiabatic migration, [3, 22, 44, 4] made it possible to map the domains of existence of each type of ACR in the $(e_1, e_2)-$plane, in the 2:1

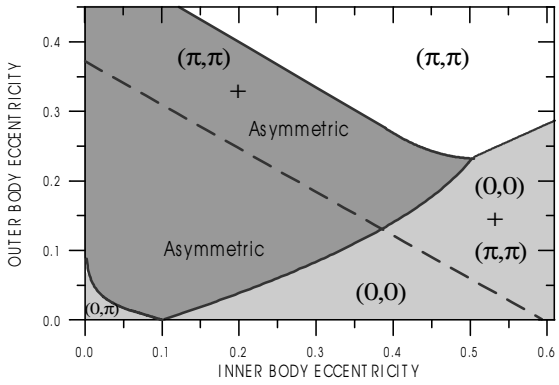

Fig. 20. Domains of existence of the 2:1 MMR stationary solutions, in the $(e_1, e_2)-$plane. The domains of existence of the main ACR are shown in light and dark gray. The domain of existence of the (π, π)-ACR is not colored since it overlaps with the others. The broken line is the planet collision limit for (π, π) stationary solutions. The lower limit of the domain of existence of (π, π)-ACR for real planets, is situated somewhere above that line

and 3:1 MMR. The main results for the 2:1 resonance are shown in Fig. 20. It is worth noting that the domain of existence of the (π, π)-ACR overlaps with the domains of existence of apsidal corotations of other types. The figure also shows the planet collision limit for (π, π) stationary solutions. The lower limit of the domain of existence of these solutions, for real planets, depends on their masses and is situated somewhere above that line.

Stationary solutions are equilibrium points of the averaged equations [3, 4]. It can be easily shown that, on a given resonance, these solutions depend almost only on the ratio of the masses of the two planets [23]. They depend only weakly on the actual masses of the planets as long as these masses are not large enough to impair the stability of the solution. For example, if both eccentricities are smaller than 0.3, the given solutions are stable, provided the masses of the planets are smaller than $1 - 2$ percent of the star mass [3]. In the domain of the (π, π)-ACR, this limit may be more stringent, mainly in the neighborhood of the inner boundary of the domain because this boundary corresponds to a true collision of the two planets.

Theoretical models and numerical simulations have shown that the stationary solutions inside each family are linked by isopleths of equal mass ratio m_2/m_1. In Fig. 21 every line corresponds to a well-defined mass ratio and a given two-planet system may exhibit only those types of apsidal corotation whose domains of existence are crossed by the corresponding isopleth.

For instance, if in a 2:1 resonant system the mass of the outer planet is significantly larger than the mass of the inner planet ($m_2/m_1 >> 1$), all possible ACR are symmetric.

Solutions found for given values of the six-parameter set

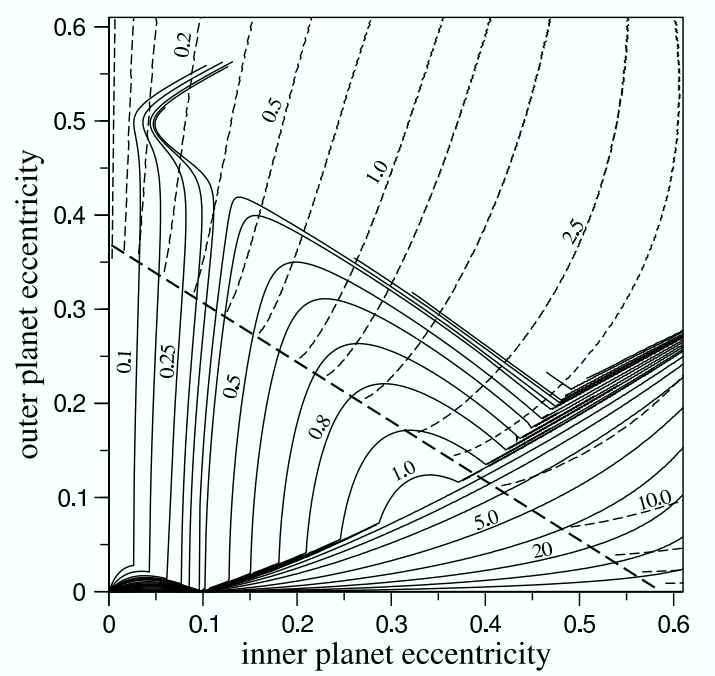

Fig. 21. Stationary solutions in the 2:1 MMR: Isopleths $m_2/m_1 = \text{const}$.

$$(m_2/m_1, \alpha, e_1, e_2, \theta_1, \Delta\varpi)$$

are valid for any exoplanetary system with these parameters, independent of their proximity to the star or of the individual planet masses.

7 Planetary Migration

ACR is a necessary condition for the survival of massive planets in nearby high-eccentricity orbits. The fact that several extra-solar planetary systems of this kind exist raises some questions about their origin: Were the planets formed in MMR? Or did they evolve towards a MMR?

In the outer Solar System, the existence of several pairs of resonant satellites is explained by aeons of smooth variation of their semi-major axes due to tidal interactions: the current configuration is due to a "migration" of primordial non-resonant orbits towards a commensurability of periods allowing the system to be trapped into a MMR. After the trapping, the satellites remained locked in the stable MMR. If the agent responsible for the migration continues to act, the orbits continue to migrate but in such a way that the resonance relationship is preserved. To say that a similar scenario occurred

in extrasolar planetary systems depends on two things: *(i)* to find plausible driving mechanisms for planetary migration, compatible with the formation process of the system, and *(ii)* to prove that the orbital evolution allows a capture into ACR.

It seems that the most probable migration process stems from the interaction between the planets and the gaseous primordial disk. Hydrodynamic simulations ([34, 35, 65, 57] and others) have shown that an adequate choice of the parameters of the gaseous disk can favor both, a large-scale inward migration and a MMR trapping. Even though there are significant uncertainties in this model, particularly with respect to the timescale of this process and the stopping mechanism for the planets at their present semi-major axes, in general this scenario seems to work very well. Fundamentally, it can explain the current orbits of resonant exoplanets starting from primordial quasi-circular orbits far from the star.

However, not all ACR can result from this scenario. In the case of the 2:1 MMR, Fig. 21 shows that the different families of stationary solutions can be of two distinct types:

- Type I. Families that can be obtained through analytical continuation from low-eccentricity orbits with $e_1 \sim 0$ and $e_2 \sim 0$. They are the $(0, \pi)$, $(0, 0)$ and the asymmetric ACR;
- Type II. Families that do not appear to be reachable from initial low-eccentricity orbits through a smooth variation. They include the (π, π)-ACR.

From the point of view of planetary migration, only Type I corotation resonances can be attained through a smooth orbital evolution starting from quasi-circular orbits. Thus, if the planetary migration hypothesis is correct, and if all exoplanets entered the mean-motion resonance in quasi-circular orbits, then we should only expect to observe Type I solutions in real systems. The extra-solar planetary systems presently confirmed in the 2/1 resonance: GJ 876 and HD82843 show Type I ACR. In the 3/1 resonance, the published orbit of 55 Cnc also corresponds to a Type I ACR ([3], [71]).

This is not the case of the (π, π)-ACR, which do not appear to be reachable via a smooth variation of the parameters from a path starting from the low-eccentricity domains. Even in the domain where $(0, 0)$ and (π, π) overlap (in the e_1, e_2-plane), it does not seem possible to have a smooth change of one into another. In fact, there is only one system for which a Type II ACR has been proposed: μ Ara (=HD 160691) [6]. However, the more recent observations have shown that the second planet has a larger orbital period [28, 49]. Therefore, even if in a 2:1 MMR, it would be unlikely that this system shows a (π, π)-ACR, since no adiabatic evolutionary process starting from low eccentricity orbits able to drive the system to such a situation is known. It could only appear in the event of an impulse perturbation able to impart a huge modification in one planet's orbit.

7.1 Capture into Corotation Resonance

In this section, we present the results of a series of numerical simulations of the dynamical evolution of fictitious pairs of planets under the action of a non-conservative perturbation that adds angular momentum and energy to the orbit of the innermost planet. The planets are small (some 10^{-5} of the central body mass) and the mass ratio is $m_2/m_1 = 0.538$ [22].

The initial distances of the planets to the star are just behind the 2:1 resonance: $\alpha = a_1/a_2 = 0.612$. When the semi-major axis of m_1 increases, a_1 increases and the mean-motion resonance ($\alpha = 0.63$) between m_1 and m_2 is reached. Capture then can take place. The probability of capture depends on the rate of variation of a_1: if the rate is high, the orbit crosses the resonance without capture, one phenomenon very well studied in the case of one massless particle. Other factors influencing the probability of capture are the orbital eccentricities: Capture is more probable when orbital eccentricities are small [16, 27]. In our calculations, initial eccentricities were lower than 0.001 and the physical parameters were adjusted to have an adiabatic approximation to resonance. Fig. 22 shows the evolution of the semi-major axes.

In this simulation, the system evolves with the innermost orbit receding from the central body (instead of having the outermost orbit approaching to the planet as in the case of interaction with an outer gas disk) up to the moment where the system is captured into a MMR, a_2 is almost constant. When the 2:1 MMR is reached, the system is trapped by the resonance. As known since Laplace, after the capture, m_1 continuously transfers one fraction of the energy (and angular momentum) that it is getting from the non-conservative source to m_2, so that a_2 also increases. One may note in Fig. 22 that, after the capture into the resonance, a_1 increases at a smaller pace than before the capture. The increase of the semi-major axes is such that the ratio a_1/a_2 remains constant.

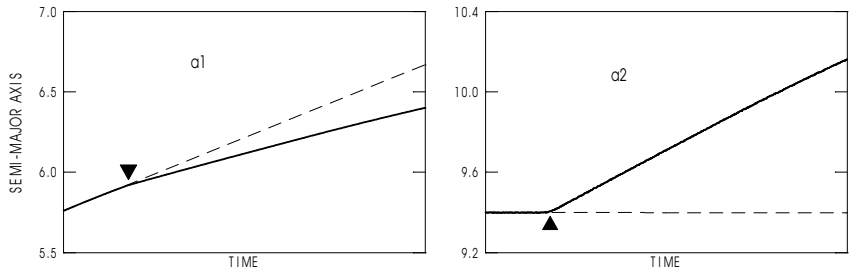

Fig. 22. Evolution of the semi-major axes before and after the capture into resonance. Triangles mark the moment of the capture. Dashed lines extrapolate the evolution before the capture and show the change in slope of the evolution lines *(Arbitrary units)*

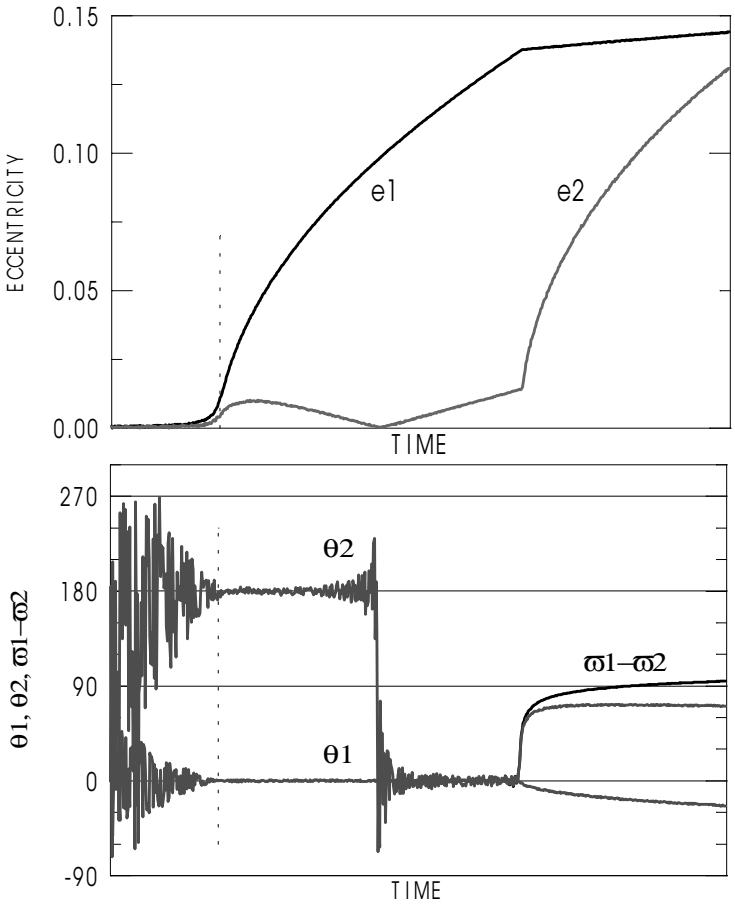

Fig. 23. Variation of the eccentricities, critical angles (θ_i) and $\Delta\varpi$ in the same time interval as Fig. 22. The vertical dotted lines show the moment of the resonance capture. $\Delta\varpi$ is only shown in the final part since it does not differ significantly from θ_2 in the time interval between the capture into resonance and the bifurcation

Figure 23 shows the variation of the eccentricities, critical angles θ_1, θ_2 and $\Delta\varpi$ in the same time interval as the previous figures. They show that, after capture, the two critical angles become trapped in the neighborhood of 0 and π, respectively and, consequently, the angle $\Delta\varpi$ is trapped in the neighborhood of π. The capture into a $(0,\pi)$-ACR (anti-aligned periastra) is thus simultaneous with the capture into the MMR.

7.2 Evolution after Capture

Figure 23 also shows that, after some time, θ_2 jumps from π to 0 and the $(0,\pi)$-ACR becomes a $(0,0)$-ACR (with aligned periastra). This change is

not the result of a discontinuous process. The change happens when the eccentricity e_2 is zero. Thus, we may describe the process by a momentary circularization of the orbit such that, when it becomes an ellipse again, the periastron is not at the same side as before. The large transients shown by the variation of the angle θ_2 are just due to the sensitivity of this angle to small changes when $e_2 \sim 0$.

The $(0,0)$-ACR also does not last forever. Fig. 23 (*bottom*) shows that the angles depart from zero and the ACR becomes asymmetric. At this moment, there is a discontinuity in the rates of variation of the eccentricities (the elbows seen in the curves $e_i(t)$).

It is interesting to note that this evolution has a counterpart in the study of periodic orbits of the 3-body problem. The study of symmetric periodic solutions shows the existence of two separated stable branches with aligned periastra; these two branches are tied with continuity by a branch of unstable periodic orbits [30]. In Fig. 21, this branch would appear along the line separating symmetric and asymmetric ACR as a shortcut tying the initial and final segments of stable $(0,0)$-ACR.

7.3 The 3:1 MMR

The previous sections considered almost only the 2:1 MMR. This choice is justified by the fact that the best-known resonant pairs are in 2:1 MMR. However, the innermost pair of 55 Cnc (if the existence of planet 55 Cnc c is confirmed) is in 3:1 resonance. The available results on this resonance [22] show two main ACR families: One family of symmetric (π,π)-ACR, situated below $e_2 \sim 0.11$ (*i.e.* below the elbows line in Fig. 24), and a family of asymmetric ACR for e_2 above this limit.[6]

The mass ratios m_2/m_1 of the solutions shown in Fig. 24 are in the range $0.1 - 1.2$. This is a new fact in what concerns asymmetric ACR. Asymmetric librations are known to happen in the asteroids in p:1 resonance with Jupiter only when the asteroid is external to the orbit of Jupiter [2]. With the convention that m_1, m_2 are the mass of the bodies, respectively, in the inner and outer orbits, this means $m_2/m_1 < 1$. In the 2:1 MMR studied in Sect. 6, this limit was slightly above 1 (~ 1.03), repeating more or less the result known for asteroids. However, in the 3:1 resonance, at least in the range studied, the behavior is independent of the mass ratio, up to the limit corresponding to a curve whose elbow occurs for $e_1 \sim 1$. Beyond this case, only symmetric ACR are expected to exist what would be consistent with the fact that we do not know asymmetric librations in the restricted case, when the asteroid orbit is internal to the orbit of Jupiter.

[6] cf. Eqs. (24), in this MMR we have $\theta_i = 3\lambda_2 - \lambda_1 - 2\varpi_i$ (The angles $\sigma_i = \theta_i/2$ are also often used). (π,π)-ACR means that the stationary solutions corresponding to the lower part of Fig. 24 are such that $\theta_1 = \pi$ and $\Delta\varpi = \pi$.

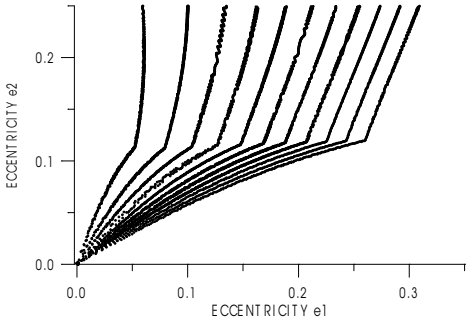

Fig. 24. ACR families in the 3:1 MMR. The line formed by the elbows of the curves, at $e_2 \sim 0.11$, separates the symmetric (π, π)-ACR (below) and asymmetric ACR (above). The mass ratios m_2/m_1 are in the range $0.11 - 1.22$ (from left to right)

Figure 24 was obtained through a simulation of planets suffering an adiabatic migration. As in the case of the 2:1 MMR, as soon as the system is captured into the MMR, it is in corotation resonance and $\Delta\varpi$ reaches 180°. The system remains in a (π, π)-ACR up to reach a bifurcation where the corotation resonance becomes asymmetric. Stationary solutions with aligned periastra were not found, although, in the very beginning of the simulations, $\Delta\varpi$ appeared temporarily oscillating about 0° with a large amplitude.

Figure 25 shows the evolution of the angle $\Delta\varpi$ during the migration, for the same solutions shown in Fig. 24. It is worth stressing the fact that for all of them the change from symmetric to asymmetric ACR occurs at almost the same value of e_2. It shows a pitchfork bifurcation. The system evolves in a stable (π, π)-ACR, with increasing e_2, up to reach the critical value. At that point, the corotation resonance becomes asymmetric with equal probability for the two possible branches ($\Delta\varpi$ less or greater than 180°). We may guess that for $e_2 >\sim 0.11$ symmetric stationary solutions continue to exist, but are unstable. In the 2:1 MMR, if the migration is continued for long time, the two branches reunite in a new bifurcation point and become symmetric again. This was not observed in the 3:1 MMR notwithstanding the fact that in one of the simulations e_2 was pushed up to values larger than 0.5.

8 Conclusion

Up to a very recent date, the dynamics of planetary systems was restricted to study the solar system: one system of 8 planets and several systems of inner satellites. These systems were characterized by almost coplanar and generally low-eccentricity orbits. In addition, in the case of inner planetary satellites, the dominant perturbation was due to the oblateness of the central

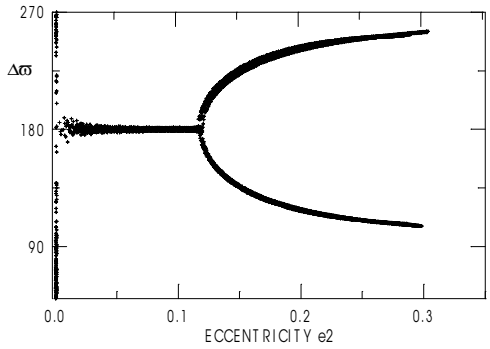

Fig. 25. Variation of $\Delta\varpi$ in the same solutions shown in Fig. 24, showing a pitchfork bifurcation at $e_2 \sim 0.11$

body. The discovery of planets around the PSR B1257+12 added one new example of planetary system, but it was very similar to the solar one: low eccentricities and near resonant orbits (see Sect. 3). The discovery of a dozen multi-planet systems around main-sequence stars opened a new front in planetary dynamics. The new planetary systems, even if they are yet only a few, already present an unexpected diversity. In the discussion done in Sect. 3, we have introduced three main classes founded on the strength of the gravitational interaction between the planets. As a guiding parameter, we have used the period ratio of their motions. For low period ratios (say, $P_2/P_1 < 3$), we had to consider two sub-classes: Class Ia with planet pairs in mean-motion resonance and class Ib with planets in low-eccentricity orbits. One additional pair in class Ia, HD 202206, has a larger period ratio (~ 5) but the relatively large P_2/P_1 is compensated by the large masses of the planets (one of them is a low-mass brown-dwarf). In the next classes, the period ratios are larger. In class II, we included planet pairs showing a significant secular dynamics and, in class III, those planets so away one to another that the gravitational interaction only introduces long-period perturbations that cannot impair the stability of the system.

Other conclusions included in this review concern the orbit determination from radial velocity data. The basic equations of the process were established one century ago to study spectroscopic binary stars. However, their application to exoplanets needs some precautions. One of them is the careful reduction to the barycenter of the solar system of both radial velocities and timings, because of the much low radial velocity variation due to exoplanets. The other concerns the critical dependence of the results on the adopted star mass: the inaccuracy in the star mass is almost reproduced in the planet masses (one factor almost never considered when transforming the raw data into elements and planet masses). Even N-body fits will leave some gauge factor involving the inclinations and the star mass undetermined.

The sections on chaos and mean-motion resonances review mostly already published results. However, some results on the several kinds of apsidal corotation resonance, in the 2:1 MMR [44, 4] deserve some comments. These results are summed up in Figs. 20 and 21. The first shows the domains of existence of stable ACR in the 2:1 MMR and the second the isopleths $m_2/m_1 = \text{const}$. From these two figures one may see that symmetric $(0, 0)$, $(0, \pi)$ and asymmetric ACR may be obtained by continuous transformation of almost circular motions after they are trapped into the 2:1 resonance. At variance, symmetric (π, π) ACR cannot be obtained in such a way and we should not expect this kind of ACR to be found among extrasolar systems if the evolutionary processes are adiabatic.

A Appendix

A.1 FFT Techniques to Study Chaos

A review of the many techniques used to map chaos is found in Contopoulos ([14]; Sect. 2.10). We do not review the whole set of techniques here, but just explain the ones used in Sect. 4. Each technique has its own virtues and limitations and often their combination is convenient to get a better understanding of the dynamics of one system. Their common shortcoming is the nonexistence of an exact correlation between their results and the macroscopic instability of the motion. It is our understanding that it is impossible to reduce the qualitative analysis of a planetary system to the blind application of one particular map technique.

The spectral analysis method used in Sect. 4 to detect the chaotic domains in a planetary system is based on the well-known features of power spectra [61]. It involves two main steps. The first is the numerical integration of the chosen model, with on-line filtering of the short-periodic terms. The choice of the model requires a careful previous analysis of the system, since computing time in the construction of dynamical maps is large. The application of the digital filtering procedure is an essential step in the construction of dynamical maps. The typical output of a long numerical integration consists of time series of osculating orbital elements that include both short- and long-periodic terms. Since we are interested only in the long-term features of the system behavior, the information about the short-term oscillations is unnecessary. Moreover, these short-period terms generate too large data output and make the identification of the long-term oscillations inefficient. For this reason, the time series of the planets osculating elements should be smoothed by digital filtering, to remove the short-period oscillations (those of the order of the orbital periods). The filtering procedure should be implemented on-line with the numerical integration, as described in detail in [51].

The second step of the technique is the spectral analysis of the output of the numerical integrations. The series giving the variation of chosen planetary

orbital elements are Fourier-transformed using a standard FFT algorithm. The Fourier transforms of the output allow one to distinguish between regular and chaotic motion.

Regular orbits

Regular motions are conditionally periodic and any orbital element ele(t) depends on time as a function:

$$\text{ele}(t) = \sum_{\mathbf{k}} A_{\mathbf{k}} \exp(2\pi i \mathbf{k}\mathbf{f}t) \tag{25}$$

where $\mathbf{f}$ is a frequency vector whose components are the fundamental frequencies of motion and $\mathbf{k} \in \mathbf{Z}^N$. When the independent frequencies are constant in time, the spectral decomposition of the motion may be obtained from its Fourier transform.

For a smooth function ele(t), the amplitudes $A_{\mathbf{k}}$ decrease rapidly with $|\mathbf{k}|$, so that the sum in Eq. (25) is dominated by a few terms. Therefore, the spectrum of regular motion is characterized by a countable (and generally small) number of frequency components. It consists of the lines associated with the independent frequencies, whose number is equal to the number of degrees of freedom of the dynamical system, as well as those corresponding to higher harmonics and to linear combinations of the independent frequencies. The half-width of the each line is of the order of $\Delta f = 1/T$, where T is the time span of the integration. T defines the transforms resolution: the longer the time T, the smaller is Δf, and the finer are the details in the Fourier spectrum that can be distinguished. For sufficient large T, each spectral peak may be approximated by a Dirac δ-function.

We illustrate a power spectrum of regular motion in the left panel of Fig. 26, which shows the spectrum of the semi-major axis of a regular resonant asteroid orbit obtained with the current initial configuration of the outer Solar System. In this example, the number of significant lines in the power spectrum is equal to 3.

Chaotic orbits

Chaotic motions are no longer conditionally periodic and the fundamental frequencies of the system vary in time. The Fourier transform of the orbital elements is not a sum of Dirac δ-functions: the power spectrum of chaotic motion is not discrete, showing broadband components. If the fundamental frequencies variations are enough large and rapid to allow a diffusion over the chosen time span T to occur, the power spectrum yields a large quantity of peaks. If the variation of the main frequencies is large and fast enough to be detected on the chosen time span T, the power spectrum yields a large amount of spectral peaks.

We illustrate a power spectrum of a chaotic motion in the right panel of Fig. 26, which shows the spectrum of the semi-major axis of the same asteroid

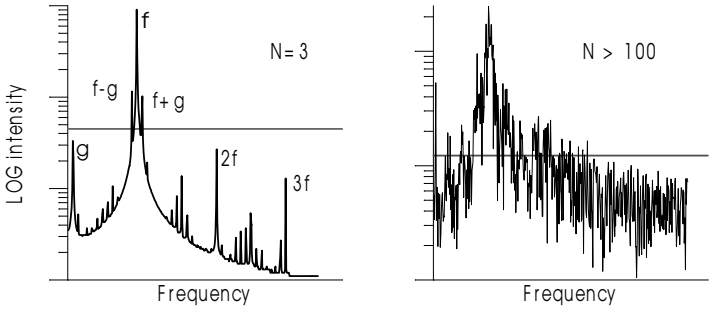

Fig. 26. *Left:* Power spectrum of the semi-major axis of a regular orbit. *Right:* Power spectrum of the semi-major axis of a chaotic orbit. In each case, N is the associated spectral number

as before when Saturn's semi-major axis is incremented by 0.03 AU to put the Jupiter-Saturn system inside the domain of the 5:2 mean-motions resonance. The number of significant lines in the power spectrum is very large.

A.2 The Spectral Number

In each power spectrum, we can determine the number of peaks that are above an arbitrarily defined "noise" level. The number thus obtained is called spectral number. In other words, the spectral number N is the number of significant peaks in the power spectrum of the chosen variable. In general, we consider in this reckoning those peaks with amplitude larger than 5% of the largest peak amplitude. To the case shown in the left panel of Fig.. 26, we associate the spectral number $N = 3$, while to the case shown in the right panel, where the number of peaks is so large that an exact reckoning is meaningless, we assign the value of an arbitrary upper limit. (In the cases studied in this paper, we have used $N = 100$.)

The spectral number N can be used to qualify the chaoticity of the system in the following way: small values of N correspond to regular motion, large values of N indicate the onset of chaos. It should be noted, however, that an orbit classified as regular can appear as chaotic if a larger time span is used in the integrations. Indeed, if the diffusion rate of the main frequencies is below the Fourier Transform resolution (defined by the time span), the spectral analysis methods are unable to detect chaos. The total integration time should be chosen large enough to allow one to distinguish chaos generated by mean motion resonances. Higher order resonances should appear in the dynamical map just by extending the integration time.

A.3 Dynamical Maps

Dynamical maps allowing chaotic domains to be identified are useful tools in the study of the stability of a planetary system. The dynamical maps

shown in this paper map the spectral numbers N. Once N is determined for all initial conditions on a grid, we plot it on the plane of initial conditions using a gray level scale that varies (logarithmically) from white ($N = 1$) to black (N maximum). Figure 27 shows the dynamical map of the well-known Taylor-Chirikov standard map

$$\begin{aligned} x_{i+1} &= x_i + \epsilon \sin(x_i + y_i) \qquad (\mathrm{mod}\, 2\pi) \\ y_{i+1} &= x_i + y_i \qquad\qquad\qquad\quad (\mathrm{mod}\, 2\pi) \end{aligned}$$

for $\epsilon = -1.3$, on a grid of 401×401 initial conditions.

Since large values of N indicate the onset of chaos, the gray tones indicate degrees of stochasticity of solutions with initial conditions starting at the map points: lighter regions correspond to regular motion, darker tones indicate chaotic motion (see the discussion in next paragraph). One may appreciate the finesse of details shown by the dynamical map inside and around the main regularity islands of the map.

A.4 Dynamic Power Spectrum

Power spectra, as shown in Fig. 26, are plots of the amplitude of the Fourier Transform against frequency. In order to see how the spectra change when

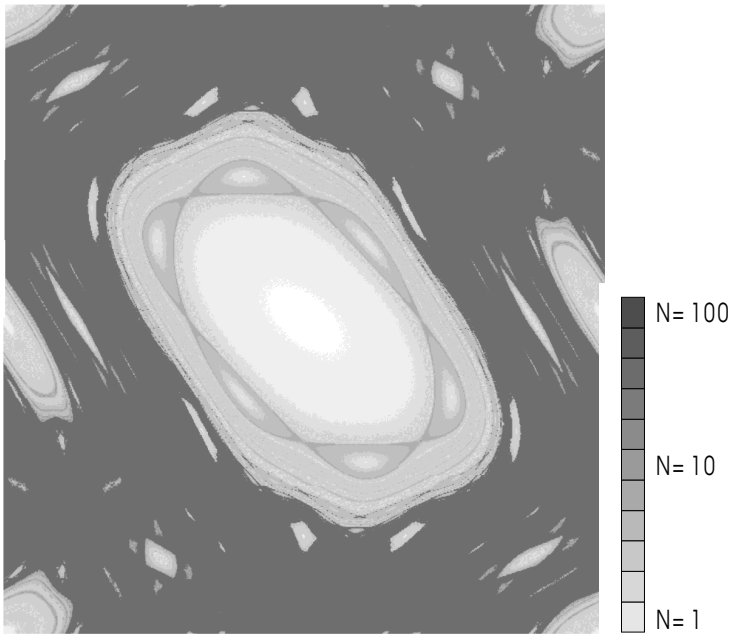

Fig. 27. Dynamical Map of the standard map for $\epsilon = -1.3$ on a grid of 401×401 initial conditions. For each initial condition, the degree of chaoticity is given by the spectral number using a uniform logarithmic scale

initial conditions vary, we should plot and compare a large number of spectra. The joint representation of the results thus obtained is cumbersome. In some analyses done in this paper, we have adopted a dynamic power spectrum (or Frequency map) obtained by marking the points where the frequency is larger than a chosen limit above noise level in a plane whose axes are one parameter describing a family of solutions (as abscissas) and the frequency of the peaks. For instance, Fig. 28 shows the dynamic power spectrum of a set of solutions of the Taylor-Chirikov standard map with initial condition $y = 0$ (the line shown in the map section in the upper part of Fig. 28). The family is parameterized by the initial x. Obviously, one chaotic solution has a huge number of significant amplitude peaks and appear in the dynamic power spectrum as a dark vertical line, and a chaotic region as a dark vertical band. The periodic orbit inside an island appears as one point at the ordinate corresponding to the periodic orbit frequency. Neighbor solutions still show only this frequency but as the initial condition goes away from this value, new frequencies appear: the arcs at several levels seen around $x = -2.4$. The selected island is a thin one and the transition to chaos is abrupt. A

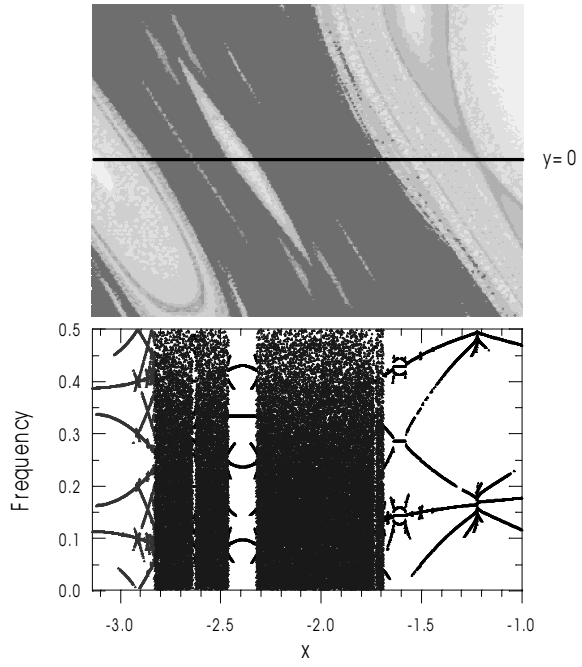

Fig. 28. Dynamic power spectrum of the solutions of the standard map. Each feature in the spectrum corresponds to the initial condition with the same abscissa on the line $y = 0$ (shown by a horizontal line on the map section in top of the figure). The abscissas in the dynamic power spectrum and in the dynamical map section are the same.

more continuous picture is seen at the left of $x = -2.82$ where the number of points increases continuously up to reach the maximum at the border of the chaotic region. Another interesting feature may be seen at $x = -1.22$ (which corresponds to a saddle point in the standard map). One dark vertical segment is shown in this position and we may see the confluence of many lines showing the increasing number of frequencies whose amplitude becomes larger than the noise level as the saddle point is approached. The remaining parts are repetitions of these basic behaviors.

In dynamic power spectra, the lines have the same behavior found in frequency analysis: frequencies remain almost constant inside resonance islands, have a vertical displacement when crossing a saddle point and become erratic when a chaotic layer is reached [38].

Dynamic power spectra are not only important complements when studying chaoticity through dynamical maps. In systems with two degrees of freedom, in which the chaoticity may be studied with the help of Poincaré maps (surfaces of section), dynamic power spectra allow us to understand the dynamics of the systems in areas where the maps show intricate features or in which the features are too thin to be visible. They were used in the discussion of the transition from secular dynamics to the 2:1 MMR (Sect. 5; [10]) and in the study of the low-eccentricity dynamics of the 5:2 MMR [53].

Acknowledgments

We acknowledge the support of CAPES, CNPq, CONICET, FAPESP, IGCE-UNESP and LCCA-USP to the investigations reported in this paper. SFM thanks the Helmholtz-Institut for Supercomputational Physics (Potsdam) and Profs. R. Dvorak (University of Vienna) and J. Kurths (University of Potsdam) for the invitation to give these lectures at the Helmholtz Summer School on Chaos and Stability in Extrasolar Planetary Systems.

References

1. C. Allende Prieto and D.L. Lambert: Fundamental parameters of nearby stars from the comparison with evolutionary calculations: masses, radii and effective temperatures. *A&A* **352**, 555-562 (1999) 230, 234
2. C. Beaugé: Asymmetric librations in exterior resonances. CMDA **60**, 225-248 (1994) 259
3. C. Beaugé, S. Ferraz-Mello and T.A. Michtchenko: Extra-solar planets in mean-motion resonance: Apses alignment and asymmetric stationary solutions. *ApJ* **593**, 1124-1133 (2003) 233, 251, 252, 253, 254, 256
4. C. Beaugé, S. Ferraz-Mello and T.A. Michtchenko: Planetary migration and extrasolar planets in the 2/1 mean-motion resonance. *MNRAS* submitted (2005), astro-ph/0404166 233, 253, 254, 262

5. G.F. Benedict, B.E. McArtur, T. Forveille, X. Delfosse, E. Nelan, R.P. Butler, W. Spiesman, G. Marcy, B. Goldman, C. Perrier, W.H. Jefferys and M. Mayor: A mass for the extrasolar planet Gliese 876b determined from Hubble Space Telescope Fine Guidance Sensor 3 astrometry and high-precision radial velocities. *ApJ* **581**, L115-L118 (2002) 221
6. E. Bois, L. Kiseleva-Eggleton, N. Rambaux and E. Pilat-Lohinger: Conditions of dynamical stability for the HD160691 planetary system. *ApJ* **598**, 1312-1320 (2003) 256
7. P. Bretagnon: Theory for the motion of all planets – The VSOP82 solution. *A&A* **114**, 278-288 (1982) 225
8. P. Bretagnon and G. Francou: 'Planetary theories in rectangular and spherical variables. VSOP87 solutions. *A&A* **202**, 309-315 (1988) 225
9. P. Butler, G. Marcy, D. Fischer, T. Brown, A. Contos, S. Korzennik, P. Nisenson and R. Noyes: Evidence for multiple companions to upsilon Andromedae. *ApJ* **526**, 916 (1999) 219
10. N. Callegari Jr., T.A. Michtchenko and S. Ferraz-Mello: Dynamics of two planets in 2:1 mean-motion resonance. *CMDA* **89**, 201-234 (2004) 245, 247, 248, 249, 250, 267
11. J.F. Chambers: Pulsars and solar-system ephemerides. *IAU Symposium* **172**, 105-112 (1996) 226
12. M. Chapront-Touzé and J. Chapront: The lunar ephemeris ELP 2000. *A&A* **124**, 50-62 (1983) 225
13. Y.Q. Chen and G. Zhao: *Chin. J. Astron. Astrophys.* **2**, 151-160 (2002) 234
14. G. Contopoulos: *Order and Chaos in Dynamical Astronomy*, (Springer 2002) 242, 262
15. A.C.M. Correia, S. Udry, M. Mayor, J. Laskar, D. Naef, F. Pepe, D. Queloz and N.C. Santos: The CORALIE survey for southern extra-solar planets. XIII. A pair of planets around HD 202206 or a circumbinary planet. *A&A*, in press (2005), astro-ph/0411512 237
16. S. Dermott, R. Malhotra and C.D.Murray: Dynamics of the Uranian and Saturnian satellite systems: A chaotic route to melting Miranda. *Icarus* **76**, 295-334 (1988) 257
17. *Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac*, H.M.Stationery Office *London 1961* 234
18. E. Everhart: An efficient integrator that uses Gauss-Radau spacings. In: *Dynamics of Comets: Their Origin and Evolution* ed by A. Carusi and G.B. Valsecchi (Reidel, Dordrecht 1985), pp. 185 242
19. S. Ferraz-Mello: *Dynamics of the Galilean Satellites*, (IAG-USP, São Paulo 1979) 232, 252
20. S. Ferraz Mello, M. Tsuchida and J.C. Klafke: On symmetric planetary corotations. *CMDA* **55**, 25-45 (1993) 251
21. S. Ferraz-Mello: Tidal Acceleration, Rotation and Apses Alignment in Resonant Extra-Solar Planetary Systems. *Bull. Amer. Astron. Soc.* **34**, 894 (2002) 252
22. S. Ferraz-Mello, C. Beaugé and T.A. Michtchenko: Evolution of migrating planet pairs in resonance. *CMDA* **87**, 99-112 (2003), astro-ph/0402335 252, 253, 257, 259
23. S. Ferraz-Mello, T.A. Michtchenko and C. Beaugé: Regular motions in extra-solar planetary systems. In: *Chaotic worlds* ed by B.A.Steves et al. (Kluwer, Dordrecht 2004) (in press), astro-ph/0402335 233, 245, 254
24. S. Ferraz-Mello, T.A. Michtchenko and C. Beaugé: The orbits of the extra-solar planets HD 82943 c,b. *ApJ* **621**, 473-481 (2005), astro-ph/0411734 237, 252

25. D.A. Fischer, G.W. Marcy, R.P. Butler, S.S. Vogt, G.W. Henry, D. Pourbaix, B. Walp, A.A. Misch and J.T. Wright: A planetary companion to HD 40979 and additional planets orbiting HD 12661 and HD 38529. *ApJ* **586**, 1394-1408 (2003) 239, 241
26. B. Gladman: Dynamics of Systems of Two Close Planets. *Icarus* **106**, 247-263 (1993) 253
27. R.S. Gomes: Resonance trapping and evolution of particles subject to Poyinting-Robertson drag: Adiabatic and non-adiabatic approaches. *CMDA* **61**, 97-113 (1995) 257
28. K. Goździewski, M. Konacki and A.J. Maciejewski: Where is the second planet in the HD 106691 planetary system?. *ApJ* **594**, 1019-1032 (2003) 256
29. R. Greenberg and A. Brahic: *Planetary Rings*, (University of Arizona Press, 1984) 251
30. J. Hadjidemetriou: Resonant periodic motion and the stability of extrasolar planetary systems. *CMDA* **83**, 141-154 (2002) 259
31. J. Hadjidemetriou and D. Psychoyos: Dynamics of extrasolar planetary systems: 2/1 resonant motion. In: *Galaxies and Chaos* ed by G. Contopoulos and N. Voglis, (Lectures Notes in Physics, Springer 2003), 412-432 252
32. F.C. Henroteau: Double and multiple stars. In: *Handbuch der Astrophysik*, (Springer, Berlin 1928), 299-468. 228
33. J. Ji, H. Kinoshita, L. Liu, G. Li and H. Nakai: The apsidal antialignment of the HD 82943 system. *CMDA* **87**, 113-120 (2003) 252
34. W. Kley: On the migration of a system of protoplanets. *MNRAS* **313**, L47-L51 (2001) 256
35. W. Kley: Dynamical evolution of planets in disks. *CMDA* **87**, 85-97 (2003) 256
36. M. Konacki and A. Wolszczan: Masses and orbital inclinations of planets in the PSR B1257+12 system. *ApJ* **591**, L147-150 (2003) 242
37. M. Konacki, G. Torres, D.D. Sasselov, G. Pietrzyński, A. Udalski, S. Jha, M.T. Ruiz, W. Gieren and D. Miniti: The transiting extrasolar giant planet around the star OGLE-TR-113. *ApJ* **609**, L37-L40 (2004) 224
38. J. Laskar: Frequency analysis for multi-dimensional systems. Global dynamics and diffusion. *Physica D* **67**, 257-281 (1993) 267
39. J. Laskar: On the spacing of planetary systems, *Phys. Rev. Let.* **84**, 3240-3243 (2000) 245
40. G. Laughlin and J.E. Chambers: Short-term dynamical interactions among extrasolar planets. *ApJ* **551**, L109-L113 (2001)
41. G. Laughlin, R.P. Butler, D.A. Fischer, G.W. Marcy, S.S. Vogt and A. Wolf: The GJ 876 Planetary System – A progress report. *ApJ*, in press (2005), astro-ph/0407441 237, 252
42. M.H. Lee and S.J. Peale: Dynamics and origin of the 2:1 orbital resonances of the GJ 876 planets" *ApJ* **567**, 596-609 (2002) 252
43. M.H. Lee and S.J. Peale: Secular Evolution of hierarchical planetary systems. *ApJ* **592**, 1201-1216 (2003) 239
44. M.H. Lee: Diversity and origin of 2:1 orbital resonances in extra-solar planetary systems. *ApJ* **611**, 517-527 (2004) 253, 262
45. R. Malhotra, D. Black, A. Eck and A. Jackson: Constraints on the putative companions to PSR 1257+12. *Nature* **355**, 583-585 (1992) 241
46. G.W. Marcy, R.P. Butler, D. Fischer, S.S. Vogt, J.J. Lissauer and E.J. Rivera: A pair of resonant planets orbiting GJ 876. *ApJ* **556**, 296-301 (2001) 219

47. G.W. Marcy, R.P. Butler, D. Fischer, G. Laughlin, S.S. Vogt, G.W. Henry and D. Pourbaix: A planet at 5 AU around 55 Cancri. *ApJ* **581**, 1375-1388 (2002) 219
48. M. Mayor, S. Udry, D. Naef, F. Pepe, D. Queloz, N.C. Santos and M. Burnet: The CORALIE survey for southern extra-solar planets. XII. Orbital solutions for 16 extra-solar planets discovered with CORALIE. *A&A* **415**, 391-402 (2004) 222, 230, 231, 234, 239
49. B.E. McArthur, M. Endl, W.D. Cochran, G.F. Benedict, D.A. Fischer, G.W. Marcy, R.P. Butler, D. Naef, M. Mayor, D. Queloz, S. Udry and T.E. Harrison: Detection of a Neptune-mass planet in the ρ^1 Cancri system using the Hobby-Eberly telescope. *ApJ* **614**, L81-l84 (2004) 237, 239, 241, 256
50. C. McCarthy, R.P. Butler, C.G. Tinney, H.R.A. Jones, G.W. Marcy, B. Carter, A.J. Penny and D.A. Fischer: Multiple companions to HD 154857 and HD 160691. *ApJ* **617**, 575-579 (2004) 239, 241
51. T.A. Michtchenko and S. Ferraz-Mello: Comparative study of the asteroidal motion in the 3:2 and 2:1 resonances with Jupiter. I. Planar model. *A&A* **303**, 945-963 (1995) 262
52. T.A. Michtchenko and S. Ferraz-Mello: Resonant structure of the Outer Solar System in the neighborhood of the Planets. *AJ* **122**, 474-481 (2001) 242, 246, 250
53. T.A. Michtchenko and S. Ferraz-Mello: Modeling the 5:2 mean-motion resonance in the Jupiter-Saturn planetary system. *Icarus* **149**, 357-374 (2001) 241, 267
54. T.A. Michtchenko and R. Malhotra: Secular dynamics of the three-body problem: Application to the υ Andromedae planetary system. *Icarus* **168**, 237-248 (2004) 238, 250
55. A. Morbidelli: *Modern Celestial Mechanics: Aspects of Solar System Dynamics*, (Taylor and Francis, London 2002) 250
56. D. Naef, M. Mayor, J.L. Beuzit, C. Perrier, D. Queloz, J.P. Sivan and S. Udry: 'The ELODIE survey for northern extra-solar planets. III. Three planetary candidates detected with ELODIE. *A&A* **414**, 351-359 (2004) 237, 241
57. J.C.B. Papaloizou: Disc-planet interactions: Migration and resonances in extrasolar planetary systems. *CMDA* **87**, 53-83 (2003) 256
58. T. Pauwels: Secular orbit-orbit resonance between two satellites with non-zero masses. *CMDA* **30**, 229-247 (1983) 248
59. M.A.C. Perryman: Extra-solar Planets. *Rep. Progr. Phys.* **63**, 1209-1272 (2000) 224
60. H.C. Plummer: *An Introductory Treatise on Dynamical Astronomy*, (Cambridge Univ. Press 1918 - reprint Dover 1960) 224, 229
61. G.E. Powell and I.C. Percival: Spectral entropy method for distinguishing regular and irregular motion of Hamiltonian systems. *JPhA* **12**, 2053-2071 (1979) 262
62. F.A. Rasio, P.D. Nicholson, S.L. Shapiro and S.A. Teukolsky: Planetary system in PSR 1257+12: A crucial test. *Nature* **355**, 325-356 (1992) 241
63. C. Ron and J. Vondrák: Expansion of annual aberration into trigonometric series. *Bull. Astron. Inst. Czech.* **37** 96-103 (1986) 225
64. N. Santos, F. Bouchy, M. Mayor, F. Pepe, D. Queloz, S. Udry, C. Lovis, M. Bazot,W. Benz, J.L. Bertaux, G. LoCurto, X. Delfosse, C. Mordasini, D. Naef, J.P. Sivan and S. Vauclair: The HARPS survey for southern extra-solar planets. II. A 14 Earth-masses exoplanet around μ Arae. *A&A* **426**, L19-L23 (2004) 241
65. M.D. Snellgrove, J.C.B. Papaloizou and R.P. Nelson: 'On disk driven inward migration of resonantly coupled planets with application to the system around GJ876. *A&A* **374**, 1092-1099 (2001) 256
66. P. Stumpff: Two self-consistent Fortran subroutines for the computation of the Earth's motion. *A&AS* **41** 1-8 (1980) 225

67. F. Tisserand: *Traité de Mécanique Céleste vol. I*, (Gauthier-Villars, Paris 1889 - reprint 1960) 232
68. S. Udry, M. Mayor, D. Naef, F. Pepe, D. Queloz, N.C. Santos and M. Burnet: The CORALIE survey for southern extra-solar planets. VIII. The very low-mass companions of HD 141937, HD 162020, HD 168443 and HD 202206. Brown dwarfs or superplanets?. *A&A* **390**, 267-279 (2002) 241
69. P. van de Kamp: Astrometry with long-focus telescopes. In: *Astronomical Techniques* ed by W.A.Hiltner, (Chicago Univ. Press 1962), pp. 487-536. 221
70. A. Wolszczan and D.A. Frail: A planetary system around the millisecond pulsar PSR 1257+12. *Nature* **355**, 145-147 (1992) 220
71. L.Y. Zhou, H.J. Lehto, Y.S. Sun and J.Q. Zheng: The apsidal corotation in mean-motion resonance: the 55 Cancri as an example. *MNRAS* **350**, 1495-1502 (2004) 252, 256

Index

Lecture Notes in Physics

For information about earlier volumes
please contact your bookseller or Springer
LNP Online archive: springerlink.com

Vol.635: D. Alloin, W. Gieren (Eds.), Stellar Candles for the Extragalactic Distance Scale

Vol.636: R. Livi, A. Vulpiani (Eds.), The Kolmogorov Legacy in Physics, A Century of Turbulence and Complexity

Vol.637: I. Müller, P. Strehlow, Rubber and Rubber Balloons, Paradigms of Thermodynamics

Vol.638: Y. Kosmann-Schwarzbach, B. Grammaticos, K.M. Tamizhmani (Eds.), Integrability of Nonlinear Systems

Vol.639: G. Ripka, Dual Superconductor Models of Color Confinement

Vol.640: M. Karttunen, I. Vattulainen, A. Lukkarinen (Eds.), Novel Methods in Soft Matter Simulations

Vol.641: A. Lalazissis, P. Ring, D. Vretenar (Eds.), Extended Density Functionals in Nuclear Structure Physics

Vol.642: W. Hergert, A. Ernst, M. Däne (Eds.), Computational Materials Science

Vol.643: F. Strocchi, Symmetry Breaking

Vol.644: B. Grammaticos, Y. Kosmann-Schwarzbach, T. Tamizhmani (Eds.) Discrete Integrable Systems

Vol.645: U. Schollwöck, J. Richter, D.J.J. Farnell, R.F. Bishop (Eds.), Quantum Magnetism

Vol.646: N. Bretón, J. L. Cervantes-Cota, M. Salgado (Eds.), The Early Universe and Observational Cosmology

Vol.647: D. Blaschke, M. A. Ivanov, T. Mannel (Eds.), Heavy Quark Physics

Vol.648: S. G. Karshenboim, E. Peik (Eds.), Astrophysics, Clocks and Fundamental Constants

Vol.649: M. Paris, J. Rehacek (Eds.), Quantum State Estimation

Vol.650: E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks

Vol.651: J.S. Al-Khalili, E. Roeckl (Eds.), The Euroschool Lectures of Physics with Exotic Beams, Vol.I

Vol.652: J. Arias, M. Lozano (Eds.), Exotic Nuclear Physics

Vol.653: E. Papantonoupoulos (Ed.), The Physics of the Early Universe

Vol.654: G. Cassinelli, A. Levrero, E. de Vito, P. J. Lahti (Eds.), Theory and Appplication to the Galileo Group

Vol.655: M. Shillor, M. Sofonea, J.J. Telega, Models and Analysis of Quasistatic Contact

Vol.656: K. Scherer, H. Fichtner, B. Heber, U. Mall (Eds.), Space Weather

Vol.657: J. Gemmer, M. Michel, G. Mahler (Eds.), Quantum Thermodynamics

Vol.658: K. Busch, A. Powell, C. Röthig, G. Schön, J. Weissmüller (Eds.), Functional Nanostructures

Vol.659: E. Bick, F.D. Steffen (Eds.), Topology and Geometry in Physics

Vol.660: A.N. Gorban, I.V. Karlin, Invariant Manifolds for Physical and Chemical Kinetics

Vol.661: N. Akhmediev, A. Ankiewicz (Eds.) Dissipative Solitons

Vol.662: U. Carow-Watamura, Y. Maeda, S. Watamura (Eds.), Quantum Field Theory and Noncommutative Geometry

Vol.663: A. Kalloniatis, D. Leinweber, A. Williams (Eds.), Lattice Hadron Physics

Vol.664: R. Wielebinski, R. Beck (Eds.), Cosmic Magnetic Fields

Vol.665: V. Martinez (Ed.), Data Analysis in Cosmology

Vol.666: D. Britz, Digital Simulation in Electrochemistry

Vol.667: W. Dieter Heiss (Ed.), Quantum Dots: a Doorway to Nanoscale Physics

Vol.668: H. Ocampo, S. Paycha, A. Vargas (Eds.), Geometric and Topological Methods for Quantum Field Theory

Vol.669: G. Amelino-Camelia, J. Kowalski-Glikman (Eds.), Planck Scale Effects in Astrophysics and Cosmology

Vol.670: A. Dinklage, G. Marx, T. Klinger, L. Schweikhard (Eds.), Plasma Physics

Vol.671: J.-R. Chazottes, B. Fernandez (Eds.), Dynamics of Coupled Map Lattices and of Related Spatially Extended Systems

Vol.672: Radyadour Kh. Zeytounian, Topics in Hyposonic Flow Theory

Vol.673: Carles Bona, Carlos Palenzula-Luque, Elements of Numerical Relativity

Vol.674: Allen G. Hunt, Percolation Theory for Flow in Porous Media

Vol.675: Martin Kröger, Models for Polymeric and Anisotropic Liquids

Vol.676: I. Galanakis, P.H. Dederichs (Eds.), Half-metallic Alloys

Vol.678: M. Donath, W. Nolting (Eds.), Local-Moment Ferromagnets

Vol.679: A. Das, B.K. Chakrabarti (Eds.), Quantum Annealing and Other Optimization Methods

Vol.680: G. Cuniberti, G. Fagas, K. Richter (Eds.), Introducing Molecular Electronics

Vol.681: A. Llor, Statistical Hydrodynamic Models for Developed Mixing Instability Flows

Vol.682: J. J. Souchay (Ed.), Dynamics of Extended Celestial Bodies and Rings

Vol.683: R. Dvorak, F. Freistetter, J. Kurths (Eds.), Chaos and Stability in Planetary Systems