HIERARCHICAL STABILITY AND CHAOTIC MOTION OF GRAVITATIONAL FEW-BODY SYSTEMS

by

YAN CHAO GE B.Sc.

Thesis
submitted to the
University of Glasgow
for the degree of
Ph. D.

Department of Physics
and Astronomy,
The University,
Glasgow G12 8QQ.

April 1991

To Ying, Jin, Huizi, and Lin

CONTENTS

Page

PART 1

PART 2

ACKNOWLEDGEMENTS

The work in this thesis was carried out while the author was a research student in the Department of Physics and Astronomy, University of Glasgow. It is a pleasure to thank all the staff and fellow students of the department for making my stay here both enjoyable and profitable. In particular I would like to thank Professor John C Brown for providing the ideal facilities for carrying out this research within the Astronomy group.

My special thanks go to my supervisor, Professor Archie E Roy for many hours of fruitful discussion and guidance in the preparation of this thesis. Due to his consideration, I was able to participate in the meeting held in Cortina and I have benefited greatly from meeting all of the people working in the field. From him I have also learned much about psychic phenomena and corn circles, and have become attracted to these subjects.

I would like to thank my colleagues Dr David Alexander, Dr Vivette Carlaw, Dr Yasseen M Almleaky, Ms Amanda J Brace and Ms Xiaoling Leng for support, help, friendship and discussions on all sorts of matters. I am grateful to David for discussions on the work of chapters 6 and 7 and careful reading of part of my thesis; to Yasseen for teaching me the key commands of the computer system; to Mandy for telling me lots of things about the department and university; to Vivette for encouraging me to join the badminton group; to Xiaoling for discussions on chapter 4. Thanks also to Dr Robin Green for discussions on relativity; to Ms Daphne Davidson for her organisational skills and for supplying delicious tea and coffee every day; to Dr Alan Thompson for not blaming me for crashing a colour monitor while I was enjoying a game; to Ms Bonnie Steves for encouraging me climbing a cliff in Italy.

My thanks go to the British council and Chinese Educational Commission for their continuous financial support in the form of a Technical Cooperation award.

I am greatly obliged to Professor P A Sweet for numerous discussions on the theory of chaos, for encouragements and invaluable suggestions throughout the course of this work. Due to him my knowledge of mathematics was sharpened, and from him I learned

a lot about the philosophy of life.

Finally, I am indebted to Dr John F L Simmons. Over the years I was away from home he has not only been an effective teacher, teaching me the modern theory of mathematics, playing squash, but also a respected friend, encouraging me to explore the beautiful lochs and mountains of Scotland. I also learned from him how to master the machineries needed to make one's own home. I cannot thank him enough. Without his care and encouragement, the work of this thesis could not have been completed.

Yan Chao Ge

SUMMARY

In this thesis the hierarchical stability and chaotic motion of the classical few-body system are studied, and then extended into the framework of the relativistic theory of gravitation. Because of the importance of integrability to both hierarchical stability and Hamiltonian chaos, a general discussion is also given on integrals and symmetries using the modern language of differential geometry. The study of this thesis is closely related to the stability problem of our Solar System and the mass transfer process of compact binary star systems. The approach carried out is both computational and theoretical.

The computational part is a systematical investigation of the **hierarchical stability** (no drastic change in orbital elements or of the hierarchy) of the general 3-body problem, in comparison with the **Hill-type stability**. The importance of eccentricity in relation to stability is manifest, and the complexity of the phase space structure and fractal nature of the boundary between regular and chaotic regions are reflected in this study.

The theoretical work is a continuation of the investigations of the effects of integrals on possible motions. Using a canonical transformation method, a stronger inequality is found for the spatial 3-body problem, giving better estimation of the Hill-type stability regions. It is proved that a Hill-type stability guarantees one of the three hierarchical stability conditions. This classical study is then developed into an inequality method establishing restrictions of symmetries (integrals) on possible motions. The method is first applied to gravitational systems in general relativity and their post-Newtonian approximations.

The thesis is split into part I, a general introduction and discussion of the relevant methods, and part II, the original research and main body of the thesis.

In chapter 1 a general introduction to the problem of the Solar System's stability is given, with an emphasis on Roy's hierarchical stability and the divergence problem of classical perturbation theory due to chaos.

Chapter 2 is a review of the theory of Hamiltonian chaos, presented at a level of comprehending chaos mathematically. The importance of number theory, infinite series and integrability to chaos is emphasised. The geometrical method of studying nonlinear dynamical systems is introduced; classical perturbation theory is used to comprehend the KAM theorem. Particular attention is paid to coordinate-free interpretation of the

integrability and separability conditions. In this chapter, a collection of integrable and chaotic systems is given because of their conceptual value to later chapters. Based on the Toda and Henon-Heiles Hamiltonian systems, a discussion is given on the general relationship of a system to its truncated system. This suggests a similar situation for the geodesic motion in Kerr geometry.

Chapter 3 is the last chapter of part I on chaos. In this chapter we study the history of chaotic dynamics and its impact on science in general. Although it is standard to study quantization of regular and chaotic motions, the present author pays particular attention to a philosophical compatibility between the theory of chaotic attractors and quantum mechanics. Noting that the two revolutionary theories were born at almost the same time, and that Poincare was a contributor to both theories, the present author carries out a historical search for a possible mutual influence in the development of the theories. However, it is found that such a connection is surprisingly tenuous.

The original work is included in part II. The classical 3-body problem is studied in chapters 4 and 5; and the relativistic few-body problem is studied in chapters 6 and 7.

In chapter 4, we first review the previous approaches on the Hill-type stability of the general 3-body problem. It is found that all results of previous studies are equivalent and do not go beyond a direct use of Sundman's inequality. Zare's (1976) canonical transformation study on the coplanar 3-body problem is modified and applied to the spatial problem, thus obtaining inequalities stronger than Sundman's. These inequalities determine the best possible Hill-type stability regions for the general 3-body problem, although the critical configurations and the value of $(C^2H)_c$ cannot be improved. In this approach, it is found that the moment of inertia ellipse of the system may be used to simplify the calculation. Because of this, it is hoped that the same stronger inequalities may also apply to systems with more than three bodies.

On the other hand, Sundman's inequality is generalised in appendix B to facilitate a similar study of relativistic systems in chapters 6 and 7. It is also hoped that the stronger inequalities obtained in chapter 4 may be developed into an inequality approach so they can be applied to improve the results of chapters 6 and 7.

In chapter 5, we prove that a Hill-type stability guarantees hierarchical stability condition HS-(C). The general case of a result concerning the primary, secondary and tertiary bifurcation values of C^2H, which was proved in a limited case by Walker & Roy (1981), follows immediately from our proof of hierarchical stability. Based on analysis of the function C^2H, we were able to establish several upper bounds for the value of α_c, thus proving that no cross-over of orbit could occur if a system is inside the Hill-type region. The property of C^2H is also used to obtain a correlated variation in the semi-major axes and eccentricities of the two binary systems for coplanar hierarchical 3-body systems.

In the same chapter, a systematic numerical experiment is carried out to investigate the hierarchical stability of the coplanar 3-body systems with initially elliptic orbits. It is found that the eccentricity is the most important orbital parameter indicating the stability of the system, and the introduction of eccentricities into the initial orbits drastically complicates the behaviour of the 3-body problem. Stable systems (in the sense of all three conditions of hierarchical stability) have been found to exist outwith the Hill-type region, and unstable systems exist inside it. New complicated valley and plateau structures are observed in the lifetime vs. initial α plot. This is believed to be a reflection of the complicated island structures of a general nonlinear system. A failure of the elliptical C^2H stability criterion is concluded.

In chapter 6, we introduce the coordinate-free language of differential geometry and make a general discussion on symmetries and conserved quantities in general relativity. Because of the key role played by integrals in the study of both hierarchical stability and chaos, we go into some detail in this general investigation. New forms of integral conservation laws were found for general systems; and a relationship is found for geodesic motion between the Poisson bracket of a class of integrals and the Lie bracket of Killing vectors. The classical Sundman inequality is applied to geodesic motion in the Schwarzschild geometry to obtain the standard bounded motion results, thus providing the first successful example of generalising the inequality method to general relativity.

In chapter 7, we apply the generalised inequality method to investigate restrictions on possible motions by symmetries. Although the study in the general case is not complete, we were able to obtain some new relations and analyse the difficulties. An application to the post-Newtonian N-body problem yields useful results. In the 2-body case the result is satisfactory. In the 3-body case, the relations are good enough to show the existence of bounded motion, however, they are to be improved by future work.

The contents of chapters 6 and 7 have been accepted for publication by the journal *General Relativity and Gravitation* under the titles Symmetries of space-time, Conservation Laws and Forbidden Motion, Part I. General Discussion; Part II. Bounded

Motion of the Post-Newtonian N-Body Problem.

The computational results of chapter 5 have been accepted for publication in *Predictability, Stability and Chaos in the N-Body Dynamical System*, in the *NATO ASI Series*.

The results of chapter 4 have been submitted to the journal *Celestial Mechanics* with the title An Alternative Deduction of the Hill-Type Surfaces of the Spatial 3-Body Problem.

The results of sections 5.1 and 5.2 are currently being rewritten for publication.

We ought to regard the present state of the Universe as the effect of its preceding state and as the cause of its succeeding state. --- Laplace

It is the nature of mathematics to pose and to solve problems; there was no possibility of never knowing. In mathematics there can be no *ignorabimus*. --- Hilbert

CHAPTER 1
Introduction to Solar System Dynamics

Will the present configuration of the solar system be preserved for some long interval of time? Will the planets eventually fall into the Sun or will some of the planets recede gradually from the Sun so that they no longer belong to the Solar System? Will any planet approach another planet and form a binary system revolving around the Sun like the Earth-Moon system or become more eccentric or more inclined to the ecliptic, and break the present configuration of the solar system? How were the satellite systems in the Solar System formed in the past? Have the satellites been captured during the passage of the Sun through some aggregates of cosmic rocks? Are the meteoric swarms really the remnants of comets? Are the gaps in the distribution of the semi-major axes of the asteroids and the gaps in Saturn's rings actually caused by gravitational actions, so that it is impossible for any small mass of particles to stay for a relatively long interval of time with the corresponding value of the semi-major axes? (Hagihara, 1957).

These are the questions likely to be raised by anybody when considering the marvels of celestial phenomena. Put more technically, is our Solar System stable? Seemingly an easy question, this has long been one of the most acute problems in celestial mechanics. To many people this is a simple question, for it is but a system of 1+9 bodies interacting under the *well-known laws* of motion and gravitation, and it was Laplace (1749 - 1827) who said that given the present state of the Universe one can predict its past and future. Others, however, may regard the difficulty of the question as that no system found in reality is completely isolated from the uncontrolled influence of the environmental world. Who knows the stability or future of a *real* system?

It is now realised that the former group of people are wrong, because even the dynamics of a system with only three bodies is not soluble analytically (Poincare, 1892) and the existence of chaotic motions in principle implies unpredictability when employing numerical experiments. On the surface the second opinion goes to an opposite extreme (in fact Landau held the idea more or less like the former, while the young Fermi the latter), however, both are based on a single agreed intuition: theoretically speaking, a simple question such as the stability of the motion of a few bodies, if the bodies were well isolated or if the influence of the outside world was given, should have a simple answer. A simple system or a simple mathematical model should give simple phenomena and the comprehension of the complicated reality must be achieved by understanding enough simple specific cases (as was believed at the beginning of the century by eg. Klein and Sommerfeld). It turned out in the course of history that such intuitive beliefs were false either due to oversimplified understanding of the problem or to the limit imposed on men by their time, for at a certain epoch only part of a particular problem could be grasped and was often taken as the whole.

Such intuitions of a simple and comprehensible nature have appeared in various forms in history and have been held by many famous scientists as the back-bone of their life-long beliefs. For example, Fourier believed that every mathematical function, no matter how complex, could be expressed as the sum of the basic simple sinusoidal functions. The investigation of this idea lasted throughout most of the nineteenth century and involved many of the greatest mathematicians of that time, including Dirichlet, Riemann, Weierstrass and Cantor. These successors of Fourier discovered what make his methods work and what might cause them to fail. It turns out that, through the celebrated works of Poincare, these contributions are essential to an understanding of the stability of the Solar System and they lie at the heart of the theory of deterministic chaos (see later sections of this chapter and next chapter).

On the other hand, some beliefs of the simplicity and comprehensibility of Nature might still be justified by reality. A good example is Einstein's belief that Nature (in the sense of its laws) is simple, beautiful and symmetric. Simple natural laws could be compatible with a complicated reality if they have rich complicated solutions (see chapter 3).

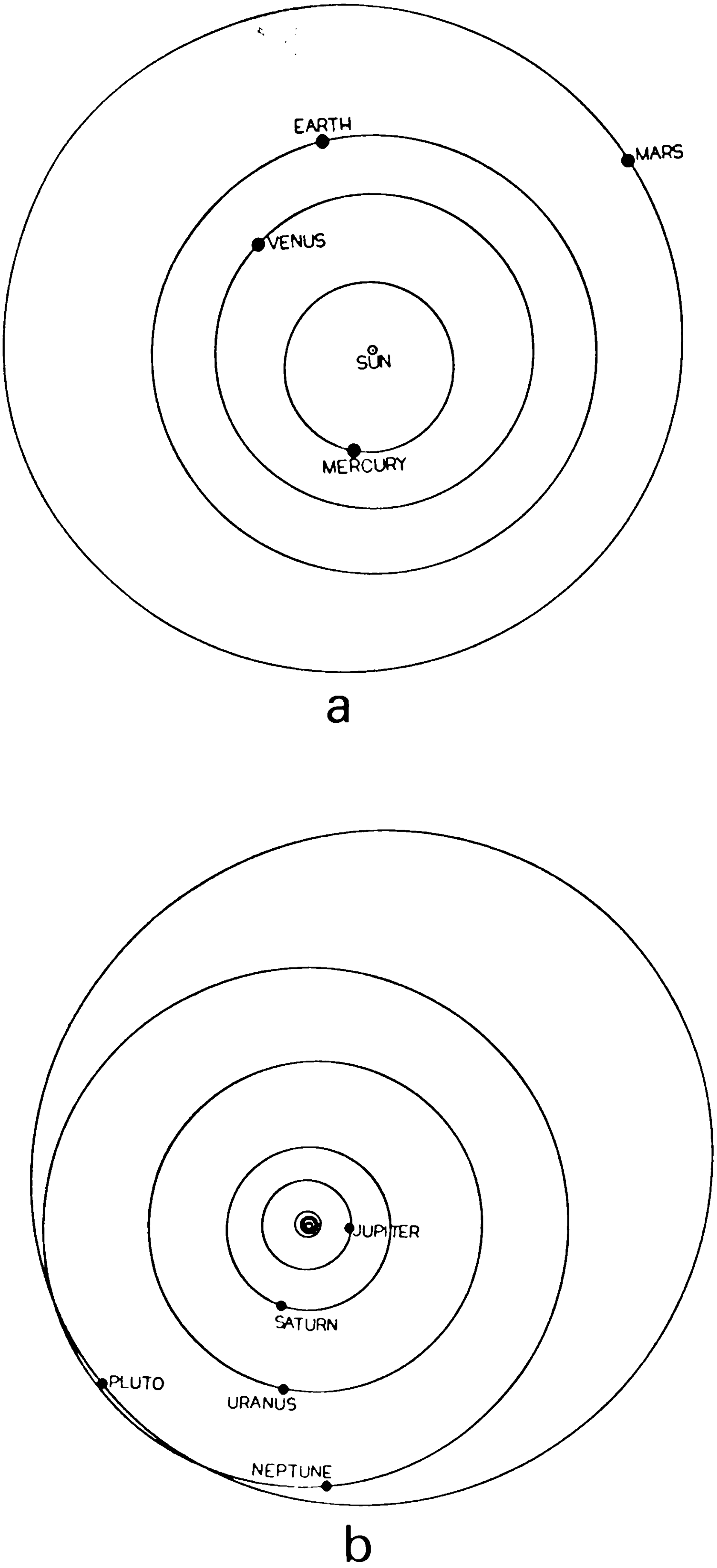
The main purpose of the present thesis is to study the stability of systems with a few gravitationally interacting bodies like the Solar System, whose formulation is very easy. The thesis is divided into two parts. In Part I, we review the recent progress made in understanding the stability of N-body systems and chaotic dynamics in general. Although most of the material in this part may be looked upon as standard, much original work and many ideas are included. For example, the present author attempts to clarify some points such as the relation between separability and integrability which are often confused. Also discussed in this part (chapter 3) are the possibility of interpreting quantum phenomena using the notion of strange attractors and the history of quantum mechanics in connection to that of chaos. Since this discussion deviates from the existing material, speculation does enter this part.

More original and conclusive work is included in Part II, in which a specific kind of stability, hierarchical stability, is investigated both theoretically and numerically. In chapter 5, we prove a relation between hierarchical stability and Hill-type stability of 3-body systems and investigate such relations in more detail by numerical experiments. In our experiments, phenomena not noticed before have been found. In chapter 4, stronger inequalities were established for the spatial 3-body problem and best possible Hill-type zero velocity surfaces are obtained. In chapters 6 and 7, the general relation between symmetry, conservation laws and forbidden motion is discussed, based on which a first effort is made to generalise the Hill-type surfaces found in nonrelativistic celestial mechanics to the framework of general relativity and the post-Newtonian approximation.

In this chapter we first describe in section 1.1 the reality and phenomena, namely, some of the relevant observations of the Solar System from ancient times up to the present epoch. In section 1.2 the three fundamental working theories (Newtonian mechanics, Einstein's relativity and quantum mechanics) on the origin, stability and future of the solar system are discussed. Finally in section 1.3, we summarise the successful explanations, open questions and make an overview of the relevant revolutionary concepts of the twentieth century mathematics such as fractal geometry and deterministic chaos.

1.1 The Phenomena - Observed Structures of the Solar System

The field of Solar System dynamics, namely celestial mechanics, studies the structures which exist in the Solar System, its possible dynamical origin, stability and future. In order to discuss the problem appropriately, it is useful to review some of the typical motions and basic structures observed in the Solar System. In addition to the fact that the motions of the planets are constrained to almost circular orbits on or close to a plane, the ecliptic, and that the motions are in the same direction, ie. prograde (direct, co-rotational), the characteristic phenomena also include the well-known Titius-Bode's law of the planetary orbits, commensurabilities in mean motion, the condensation and gap feature in the distribution of the asteroids, the ordered motion of the satellites and the


EARTH
MARS
VENUS
SUN
MERCURY
a
JUPITER
SATURN
PLUTO
URANUS
NEPTUNE
b

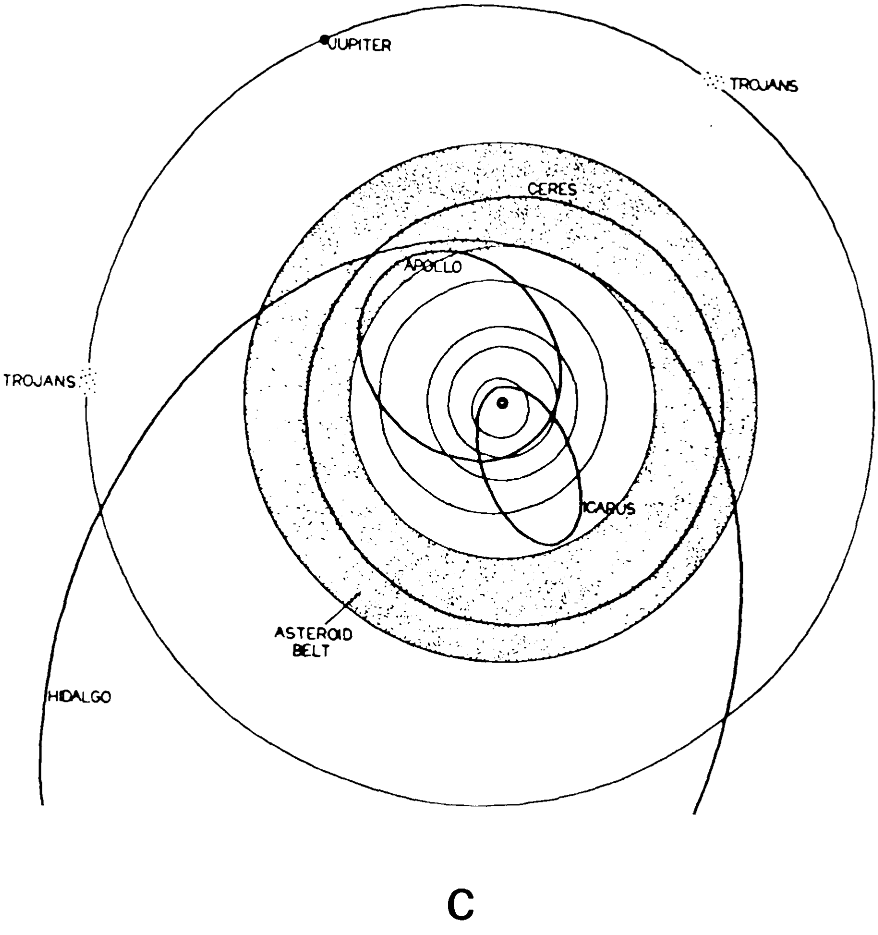


Figure 1.1 Orbits of the nine planets. (a). The orbits of the four inner planets. (b). The orbits of the outer five. (c). The orbits of the asteroids. (Taken from Baugher, 1988)

ring systems.

Kepler's Laws

Johannes Kepler (1571 1630), from a study of the mass of observational data on the planets' positions collected by Tycho Brahe (1546-1601), formulated the three laws of planetary motion forever associated with his name. They are:

1. The orbit of each planet is an ellipse with the Sun at one focus.
2. For any planet the rate of description of area by the radius vector joining planet to Sun is constant.
3. The cubes of the semimajor axes of the planetary orbits are proportional to the squares of the planets' periods of revolution.

Table 1.1 Planetary distance from the Sun (in AU)

Planet	n	Distance from Sun (AU)		Eccentricity	Inclination
		Bode's law	Actual		
Mercury	$-\infty$	0.4	0.387 099	0.205 627	7.003 99
Venus	0	0.7	0.723 332	0.006 793	3.394 23
Earth	1	1.0	1.000 000	0.016 726	0.0
Mars	2	1.6	1.523 691	0.093 368	1.849 91
asteroids ?	3	2.8	2.80		
Jupiter	4	5.2	5.202 803	0.048 435	1.305 36
Saturn	5	10.0	9.538 843	0.055 682	2.489 91
Uranus	6	19.6	19.181 951	0.047 209	0.773 06
Neptune	7	38.8	30.057 779	0.008 575	1.773 75
Pluto	8	77.2	39.438 71	0.250 236	17.169 9

Titius-Bode's Law

One of the most striking manifestations of order in the Solar System is found in the planetary distances from the Sun, the characteristics of which are shown in Table 1.1 and Figure 1.1. It is seen that the mean orbital radii, r_n, agree with Titius-Bode's law (found in 1766, Johann Titius 1729-1796, Johann Elert Bode 1747-1826) up to the orbit of Uranus, viz.

$$r_n = 0.3 \times 2^n + 0.4 \quad \Leftrightarrow \quad r_{n+1} - r_n = 0.3 \times 2^n$$
$$n = -\infty,\ 0,\ 1,\ 2,\ 3, \ldots\ldots$$

Although this law has not the same status as Kepler's laws, it was historically relevant particularly with the discovery of a number of minor planets (asteroids), and it is related to commensurable mean motions by Roy & Ovenden (1954). In addition similar laws can be found for the major satellite systems (Blagg, 1913; Roy, 1982, P5).

Commensurabilities in Mean Motion

There exists in the Solar System a remarkable number of approximate **commensurabilities (resonances)** in mean motion between two or more bodies in the planetary and satellite systems. If the mean angular velocities are denoted by $\omega = \{\omega_1, \ldots\ldots, \omega_n\}$ and a set of non-vanishing integers denoted by $\mathbf{k} = \{k_1, \ldots\ldots, k_n\}$, then the commensurability condition may be written as

$$< \mathbf{k}\ , \omega > = \Sigma\,(k_i\, \omega_i) = 0.$$

For example, if ω_J, ω_S, ω_N and ω_P are the mean motions in degrees per day of Jupiter, Saturn, Neptune and Pluto respectively, then

$$\omega_J = 0.0\underline{83\ 091},\quad \omega_S = 0.0\underline{33\ 460},\quad \omega_N = 0.00\underline{5\ 981},\quad \omega_P = 0.00\underline{3\ 979},$$

$$<\{2, -5\},\{\omega_J,\omega_S\}> = -\ 0.0\underline{01\ 118},\quad <\{3, -2\},\{\omega_N,\omega_P\}> = -\ 0.00\underline{0\ 025}.$$

One of the triple commensurabilities is among the mean motions of three satellites of Jupiter; Io, Europa and Ganymede. In the same units the mean motions and triple commensurability are

$$\omega_I = 203.488\ 992\ 435,\quad \omega_E = 101.374\ 761\ 672,\quad \omega_G = 50.317\ 646\ 290,$$

$$<\{1, -2\},\{\omega_I,\omega_E\}> = 0.739\ 469\ 091,\quad <\{1, -2\},\{\omega_E,\omega_G\}> = 0.739\ 469\ 092,$$

$$<\{1, -3,\ 2\},\{\omega_I,\omega_E,\omega_G\}> = 0.000\ 000\ 001.$$

Corresponding to this remarkable commensurability in the mean motions of the satellites, there is an equally exact one in their mean longitudes, viz.

$$<\{1, -3,\ 2\},\{\ell_I, \ell_E, \ell_G\}> = 180^0.$$

This is called a **critical argument**.

An example of resonances involving more elements is the well-known Saros found in the motion of the Moon, on which more information is given in Roy (1973, 1982).

In fact, for any set of given numbers (here, mean motions) there always exists a set of non-vanishing intergers which can satisfy the commensurability condition arbitrarily closely. However, it was shown by Roy and Ovenden (1954) that, if the integers are limited to small ones, the occurrence of approximated commensurable mean motions is higher than naturally possible (cf. KAM theory, small divisor).

Note. For any two numbers $\{a_1, a_2\}$, there always exist two rationally independent non-zero integers $\{k_1, k_2\}$ such that $(a_1/a_2 - k_2/k_1)$ and $< \mathbf{k} , \mathbf{a} >$ both arbitrarily tend to zero. However, care must be paid to the fact that the two expressions are not equivalent, since the value of the k's are allowed to go to infinity. In fact, the former is a necessary, but not sufficient, condition for the latter; thus there are more k's satisfying the former relation than the latter one. This is evident from a simple example: for the rational number 1/3 = 0.333..., the first relation can always be made arbitrarily small (below 0.0 ... 04) by the sequence of rationally independent integers {3 ... 3, 10 ... 0}, whereas for the same numbers the second relation equals to a constant, 1/3.

Distribution of Asteroids between Mars and Jupiter

The minor and major planets are divided by the asteroid belt centred on 2.8 AU from the Sun, and the majority of the small bodies are distributed in the range 2.2 - 3.2 AU. Shown in Table 1.2 are a few important ones out of the thousands of these small bodies. It is seen that the eccentricities and inclinations of the asteroids tend to be much higher than those of the planets but they are all in direct orbits.

However, let us pay more attention to a more interesting phenomenon in the distribution of orbital radii shown in Fig. 1.2. The structures existing in the distribution of the asteroids in relation to commensurabilities has caused much curiosity for a long time and is still attracting active research.

On the one hand there are the obvious breaks avoided by the asteroids, known as Kirkwood gaps after their discoverer. These distances correspond to mean motions that are commensurable with that of Jupiter, the main disturber of the asteroid orbits, namely, 1:2, 1:3, 2:5, 3:7 and so on. On the other hand, there is an accumulation of asteroid orbits near the commensurabilities of 2:3, 3:4 etc. Finally the Trojans, first discovered by Lagrange (1736 - 1813), may be said to be a special case of condensation close to commensurability 1:1.

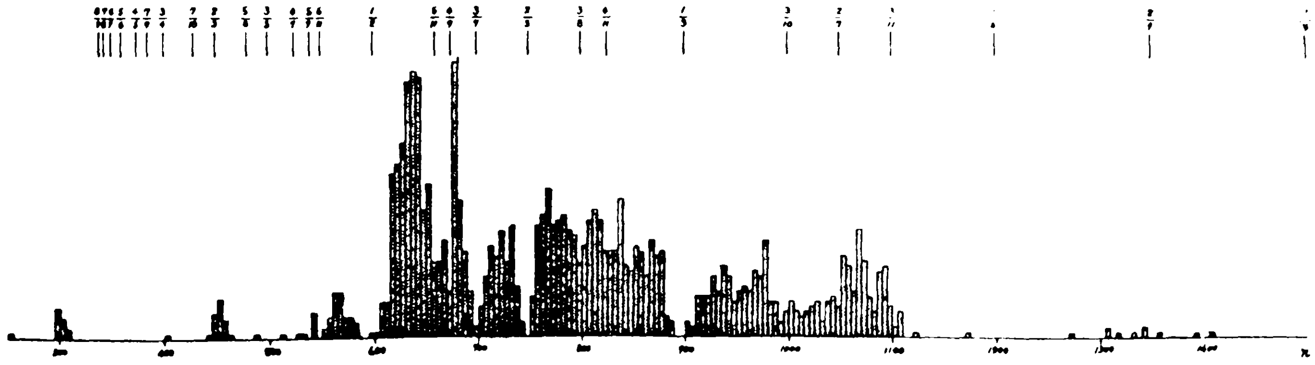


Figure 1.2 Distribution of asteroids with their mean motions about the Sun, in seconds of arc per day. Mean motions which are commensurable with that of Jupiter are indicated (from Hagihara, 1957).

Table 1.2 Some important asteroids

Asteroid	Year of discovery	Diameter (km)	Semi-major axis (AU)	Eccentricity	Inclination
1 Ceres	1801	946	2.77	0.08	10.6
2 Pallas	1802	583	2.77	0.23	34.8
3 Juno	1804	249	2.67	0.26	13.0
4 Vesta	1807	555	2.36	0.09	7.1
10 Hygiea	1849	443	3.14	0.12	3.8
433 Eros	1898	20	1.46	0.22	10.8
1566 Icarus	1949	2	1.08	0.83	22.9
1862 Apollo	1932	?	1.47	0.56	6.4
2102 1975YA	1975		1.29	0.30	64.0
2363			5.13	0.04	32.8
2146			5.22	0.10	38.1
1869 Philoctetes	1960		5.31	0.06	3.4

Motion of Other Smaller Bodies

In addition to the generalised Bode's law for the satellite systems and commensurabilities mentioned above, the mean motions of the satellites are found to be related to the spin periods of the planets. The rings of Saturn are easily observed, with the gaps showing a correspondence with distances at which the orbital periods around Saturn are some simple fraction of the periods of some of its inner satellites. While most objects found in the Solar System follow direct orbits, retrograde orbits are found in the satellite systems of Jupiter and Saturn.

The motion of comets and meteors is also of some interest; in particular, of dynamical interest are close approaches (encounters) of such smaller bodies with planets. For example, the orbit of Brook's comet was markedly changed by the action of Jupiter (see Roy, 1982). Before its encounter with the planet on July 20th 1886, its period of revolution about the Sun was 29.2 years, its orbit lying outside Jupiter's. After encounter, its period changed to 7.10 years, while its orbit shrank in size to lie completely inside Jupiter's orbit.

The study of such close encounters is important to the capture theory of the origin of the solar system, satellite systems and Pluto. For more detailed descriptions see Moore (1988), Baugher (1988) and Dormand & Woolfson (1989). For a discussion of dynamical capture theory see eg. Leimanis & Minorsky (1958) and Tanikawa (1983).

Binary and Multiple Hierarchical Systems

In contrast to motions found in the solar system, stellar motions may appear more random and uncorrelated. However, this is not the case. In addition to the large scale super-structures, more than half of the stars are found to be moving in binary systems, in which the members may be separated so far from each other that their orbital periods may be hundreds of years; in other cases the two stars are almost in contact, distorting each other's shape by tidal pull, sharing a common atmosphere or transferring material from one component to the other.

The proportion of triple and higher systems is also reasonably large, lying between one-quarter and one-third of all stars (see Roy, 1982). In studying the motion of such systems and the ordered motion in the solar system, Evans's (1968) hierarchical approach is found useful. In fact, the stability of such hierarchical systems has been studied by Walker (1980) and M^{c}Donald (1986), and is also the main subject of the present thesis.

In this connection the work of Heggie (1975) is worth mentioning. He found that binary and multiple systems can be formed dynamically in classical many-body systems. However, this mechanism does not produce such systems in sufficient numbers to match the observed proportion.

Summary

In this section some of the characteristic motions and most important features of the solar system have been summarised. The theories that may be used to supply satisfactory explanation of such features shall be mentioned in the following sections. However, it is important to note that no theory is an absolute reflection of truth, and its origin and development rely heavily on the observed phenomena.

In fact the origin and development of Newton's laws of motion and gravitation depended heavily upon the careful observations by Brahe and Kepler. The test of the theory needs more accurate observation over longer time periods and it is worth noting that an accurate test of the theory is not possible at present since many of the bodies have only been discovered and traced relatively recently; in fact Pluto has only covered half of its orbit since its discovery in 1930 (see eg. Walker et al, 1980).

1.2 The Theories - Newton's Laws of Motion and Gravitation

The problem of the motion of celestial bodies and objects on the Earth has stimulated much curiosity and speculation. As mankind is necessarily limited by personal activity and movements, the sensation of space came to man (either individual or society) earlier than time, and an Earth-centred universe was an almost obvious 'truth' held for a long time. While a normal man is born with sight, hearing and other sensation to feel the length, width and height of the 'universe', the realisation of time needs a conscious observation of recurrent (periodic, almost periodic) phenomena. In ancient times the observation of such recurrent phenomena was inevitably mixed with art, religion and superstition; for example, the rise and set of the Sun, periodic motion of the Moon, and more importantly, the relation of the motion of the Sun to the periodicity of seasons, flood and field work. Therefore, the character of the civilisation of ancient times was that everything was correlated and of a unified 'God'. This was continued until the time of Galileo (1564-1642), who was the main contributor to the modern scientific method of reasoning and experiment characterised by the strategy 'divide and rule'.

On the other hand the origin and development of a scientific theory or method is almost always characteristic of successfully formulation and abstraction, which usually requires men to be creative, to bring some apparently irrelevant experiences together and pursue the principle of beauty, simplicity and economics. The historical development of Newton's law of gravitation is a very good example of a successful formulation based on careful observations and creative intuition of seeing the common feature out of apparently irrelevant events.

Newton's Laws of Motion

At the time of Newton (1642-1727), the 'shoulders of the giants' were ready to be stood upon. In his celebrated work The *Principia*, Sir Isaac Newton proposed the three laws of motion by bringing together statics and kinetics:

(1). Every body continues in its state of rest or of uniform motion in a straight line except insofar as it is compelled to change that state by an external impressed force.

(2). The rate of change of momentum of the body is proportional to the impressed force and takes place in the direction in which the force acts.

(3). To every action there is an equal and opposite reaction.

While the third is independent, the first two can be formulated into a unified mathematical equation, namely,

$$\mathbf{F} = dP/dt = ma,$$

where the notation is standard. Based on this equation of motion, the theory of classical

Newtonian mechanics (both statics and dynamics) was established, and the origin and progress of the theory of calculus gradually followed.

Now let us note that the power of this formulation of motion cannot be justified only at this level or remain at a level of philosophy; its power lies in its capability to be tested. Most original theories share this property. However, the marvellous achievements of Newtonian mechanics did also lead to false generalisations and enthusiasm (eg. ideas of Laplace and Hook). It is quite common in the history of science and in fact inevitable in any personal activity to lay aside situations where a theory does not work and remain enthusiastic about where it works well. This is useful, but frequently, after many successes society and individuals tend to completely forget and remain blind to situations where the theory is not valid (see the quotations of Laplace at the beginning of this chapter).

In fact most of the problems of 'given force - find motion' are not completely solvable in closed form. Thus Newtonian mechanics is faced with the difficulty that it is purified out of a very small subset of simple facts (phenomena) and is solved by mental labour for a slightly larger subset, in which the theory is found to agree with some new facts, and thus the power is shown. However, the remaining large amount of phenomena are believed to be encompassed by the theory without any means of testing since the theory cannot be solved nor compared with the facts. Therefore the belief is faced with a serious drawback, but it is exactly due to this incompleteness that the theory is open to modification.

On a more technical level let us consider the difference between the configuration (physical, positional) space and state (phase) space. A configuration of a body is the position of it in space at a specific time, $\mathbf{r}(t_0)$; while a state of motion of a body is the position of it for some time interval, $\mathbf{r}(t)$, which is equivalent to knowing, for analytic motions, all derivatives of the position vector with respect to time at a specific time (according to Taylor expansion).

If a law of motion is, by assumption, an ordinary differential equation and capable of determining the state of motion, it could be of any order (zero, one, two, three or higher). However, a law of motion cannot be an ODE of order zero (ie. algebraic), since this is just a direct description of the state of motion, thus not a useful theory; it cannot be of order one either since absolute motion (velocity) is meaningless, which had been realised well before Newton. Therefore the simplest nontrivial ODE must be of second order, like Newton's second law of motion. The problem is then why it should be the second order derivative rather than the third one that is related to 'force'. This is justified because 'force', a concept important to everyday life, is in fact defined as a second order derivative; whereas only in the case of slow motion, **F** is identified as the 'force' in statics. So we see that Newton's laws of motion are purified from a rather limited class of motions analytic with time, with some implicit assumptions such as a continuous world and Galilean relativity; the approximate feature of the theory is also evident. Such methodology is also used in the foundation of Einstein's relativity and wave (quantum) mechanics.

The state of motion, **r**(t), during some time interval is determined by a Taylor expansion and the recurrence relations of the coefficients. It turns out that to find the state of motion defined by an nth order ODE, it suffices to know the time derivatives of **r** at a specific time up to the (n-1)th order. Therefore, the initial value problem of the Newtonian mechanics is determined by giving initial position and velocity. This is why the space of generalised coordinates and momenta is called the state (phase) space.

Because there are many a priori assumptions and simplifications in Newtonian mechanics, the theory was to be revolutionised by special relativity, general relativity and quantum mechanics, for which new fundamental concepts are needed. Because of the nonsolvability of the theory, a revolution completely within the framework of Newtonian mechanics, namely, deterministic chaos, was also of historical necessity.

Newton's Law of Universal Gravitation

Newton's law of universal gravitation is one of the most far-reaching laws ever formulated, and is the basis of the studies of celestial mechanics and astrodynamics. It is based on the work of Nicholas Copernicus (1473 - 1543), Tycho Brahe (1546 - 1601) and in particular Kepler's three laws of planetary motion. Newton was the first to realise the importance and study systematically the three Kepler's laws. By using his laws of motion, he was able to show that the inverse square law of gravitation is the only law of force compatible with the three empirical laws of Kepler regarding motions of planets around the the Sun. The law is stated as:

Every particle of matter in the universe attracts every other particle of matter with a force directly proportional to the product of the masses and inversely proportional to the square of the distance between them. In other words,

$$F = Gm_1m_2 / r^2,$$

where G is the universal constant of gravitation.

We note that the law of universal gravitation is a kind of two-body, linear law, since the interaction between any two of the many (or infinity) bodies is completely independent of the existence of other bodies and the forces can be added by the linear triangle principle to form the resultant forces. This is another example of many a priori assumptions of Newtonian mechanics which were to be abandoned in the theory of general relativity.

Summary

In this section we briefly discussed the main working theory in the studies of celestial mechanics, Newtonian mechanics, which enables the motion of celestial bodies to be studied mathematically. Although in certain circumstances, the theories of relativity and quantum mechanics are also relevant, in practice, it is Newton's universal gravitation that dominates the motion. Moreover the motion may often be modelled by that of some point-mass particles, with the influences of non-gravitational forces, the size and distribution of mass in a body considered as perturbations. Therefore in the following section we shall discuss the ideal N-body problem.

1.3 Stability of the Solar System as an N-Body Problem

With Newton's laws of motion and gravitation, the motion of planets and asteroids may be treated as point masses interacting under mutual attractions only, this is a specific example of the classical N-body problem with one dominating mass. The N-body problem may be defined by the following set of ordinary differential equations,

$$\ddot{\mathbf{R}}_i = \frac{d^2\mathbf{R}_i}{dt^2} = -\sum_{j \neq i} \frac{Gm_j}{R_{ij}^3}\mathbf{R}_{ij} \qquad (i,j = 1, 2, \ldots\ldots, N)$$

$$R_{ij} = \left|\mathbf{R}_{ij}\right|, \quad \mathbf{R}_{ij} = \mathbf{R}_i - \mathbf{R}_j,$$

where $\mathbf{R}_i$ and m_i are the position vector of the ith mass point in an inertial reference frame and its mass respectively, while t is the time and G the gravitational constant which in the present thesis is taken as unity.

In this formulation, the problem of celestial mechanics is changed to find the solutions of this set of nonlinear ordinary differential equations. Although the 2-body problem is solved, the 3-body problem poses a great difficulty. As a drawback even the restricted 3-body problem is not solvable in closed form, although the slightly different 2-centre problem is solvable. The solvability of the 2-body problem on the one hand and the nonsolvability of the 3-body problems on the other is due to the fact that there is only a limited number of symmetries in the underlying space-time background. Therefore only a limited number of global isolating (uniform) first integrals may be used to facilitate solving the problem. Related with the solvability of the equations are also the singularity problems caused by possible collisions, for which the Cauchy existence and uniqueness theorem does not apply.

In connection with both aspects, the series expansion method (essentially Taylor and Fourier expansion) may be invoked and this, in fact, has been the main tool exploited to produce ephemerides and regularising transformations. However, difficulties have been encountered regarding the convergence of the series. It was not realised until the works of Bruns, Poincare, Painleve, Sundman and Siegel that non-integrability and real singularity are intrinsic problems of the dynamics, for which divergence of infinite series is unavoidable rather than artificial (see eg. Siegel & Moser, 1971). It turns out that the previous belief of the integrability and existence of convergent series solution was incorrect.

Singularities and Regularization

One of the most obvious difficulties of the classical N-body problem may be the existence of singularities in the differential equations caused by collisions between two or more bodies. When this happens the general existence and uniqueness theorem (sufficient conditions) does not apply; thus whether a solution exists or not is not certain from a mathematical point of view. It may happen that a solution does not exist at such singularities or exists but is not unique, because careful investigation shows that a singularity of a differential equation does not necessarily imply singular solutions. For simple examples the classical book by Stiefel & Scheifele (1971) should be consulted.

The standard method of establishing solutions through singularities is by regularising transformations, whereby a change of variables transforms the original singular equations to regular ones for which the general existence and uniqueness theorem applies. It was shown by Sundman (1912) that collision with either primaries of the restricted 3-body problems can be regularised. Solution can also be continued through non-simultaneous binary collisions in the N-body problem (eg. Wintner, 1947); whereas not all collisions involving three (or more) bodies can be regularised (eg. Siegel and Moser, 1971) - they are real singularities in the sense that solutions at such collisions are necessarily singular in a topological sense. Furthermore, it may happen that a solution is singular but without any collision involved (eg. Leimanis & Minorsky, 1958, P97).

The singularity problems encountered in N-body problem and their regularization are not only of pure theoretical interest but also of computational value. For the references in this field Szebehely (1967), Stiefel & Scheifele (1971) and Heggie (1974) must be referred to.

General Perturbation theory

Under the limit that the many body problem, or even the two-body problem with at least one of the bodies of arbitrary shape and mass distribution, cannot in general be solved in closed form for all time, various perturbation methods have been used to infer the characteristic behaviour of such systems. For example, for the motions in the Solar System such as the motion of a planet or asteroid around the Sun which is perturbed by another planet, or the motion of natural and artificial satellites (treated as point particles) in the field of a planet (treated as an extended body), the general perturbation theory can present satisfactory predictions about the motion of the bodies for a finite time interval. In this theory, the motion of the body under study may be formulated as the motion in the potential field, U_0, of an integrable case and a perturbation potential, R, which is at least an order of magnitude smaller than U_0. Thus the equation of motion may be written as,

$$d^2\mathbf{R}/dt^2 = \nabla(U_0+R),$$

where U_0 is usually the potential function due to the point-mass 2-body problem.

The above equation may be equivalently written as the so called Lagrange planetary equations, which determine the variation of the osculating elements (eg. Roy, 1982). The importance of osculating elements and the Lagrange planetary equations are often explained as a result of the smallness of the changes of the orbital elements of the Kepler problem due to the small perturbation. However, this often causes the misunderstanding that the Lagrange equations are already approximate whereas they are rigorous. The more fundamental aspects lie in that from the study of two bodies the coordinates and velocity components at any instant permit the determination of a unique set of six orbital elements, and that the set of Delaunay elements, which are related to the classical elements by simple formulae, in fact forms a set of canonical variables. Therefore, the Lagrange equations are the equivalent laws of motion written in a different coordinate system (Brouwer & Clemence, 1961). For future reference, we write down the equations, the proof of them may be found from most standard textbooks (eg. Stiefel & Scheifele, 1971). These references must also be consulted for more technical treatments on practical problems.

Lagrange's planetary equations:

$$\begin{cases} \dot{a} = -\dfrac{2\sqrt{a}}{K}\dfrac{\partial R}{\partial M} \\ \dot{e} = -\dfrac{1-e^2}{Ke\sqrt{a}}\dfrac{\partial R}{\partial M} + \dfrac{1}{Ke}\sqrt{\dfrac{1-e^2}{a}}\dfrac{\partial R}{\partial \omega} \\ \dot{I} = -\dfrac{1}{K\sqrt{a(1-e^2)}}\left[\operatorname{ctg} I\dfrac{\partial R}{\partial \omega} - \dfrac{1}{\sin I}\dfrac{\partial R}{\partial \Omega}\right] \end{cases} \quad \begin{cases} \dot{M} = Ka^{-3/2} + \dfrac{2\sqrt{a}}{K}\dfrac{\partial R}{\partial a} + \dfrac{1-e^2}{Ke\sqrt{a}}\dfrac{\partial R}{\partial e} \\ \dot{\omega} = -\dfrac{1}{Ke}\sqrt{\dfrac{1-e^2}{a}}\dfrac{\partial R}{\partial e} + \dfrac{\operatorname{ctg} I}{K\sqrt{a(1-e^2)}}\dfrac{\partial R}{\partial I} \\ \dot{\Omega} = -\dfrac{1}{K\sqrt{a(1-e^2)}\,\sin I}\dfrac{\partial R}{\partial I} \end{cases}$$

where K^2 equals the sum of the two masses whose motion are under consideration, with the perturbation function R expressed in terms of the classical Keplerian elements (a, e, I, M, ω, Ω), namely, the semi-major axes, eccentricity, inclination to invariable plane, mean anomaly, argument of pericentre and the longitude of ascending node.

Canonical equations in Delaunay elements:

$$\mathcal{H}(L, G, H; l, g, h) = -\frac{K^4}{2L^2} + R$$

$$\begin{cases} dL/dt = -\partial R/\partial l \\ dG/dt = -\partial R/\partial g \\ dH/dt = -\partial R/\partial h \end{cases} \quad \begin{cases} dl/dt = K^4/L^3 + \partial R/\partial L \\ dg/dt = \partial R/\partial G \\ dh/dt = \partial R/\partial H \end{cases}$$

where the Delaunay elements (L, G, H; l, g, h) are related to the classical elements by

$$\begin{cases} L = K\sqrt{a}\ , \quad G = K\sqrt{a(1-e^2)}\ , \quad H = K\sqrt{a(1-e^2)}\cos I \\ l = M\ , \quad g = \omega\ , \quad h = \Omega \end{cases}$$

These two sets of equations describing the variation of arbitrary constants are in general nonlinear, nonintegrable ODEs, as are the equations in rectangular coordinates. Perturbation methods may be used to solve them because of the smallness of R in magnitude. Often such methods use successive approximations, such as the classical and secular perturbation theory and the averaging method. In fact, in transforming the equations from rectangular coordinates to the Delaunay canonical elements and equations, we have just performed the preparatory procedure of putting the equations into action-angle variables in the classical and secular perturbation theory discussed in

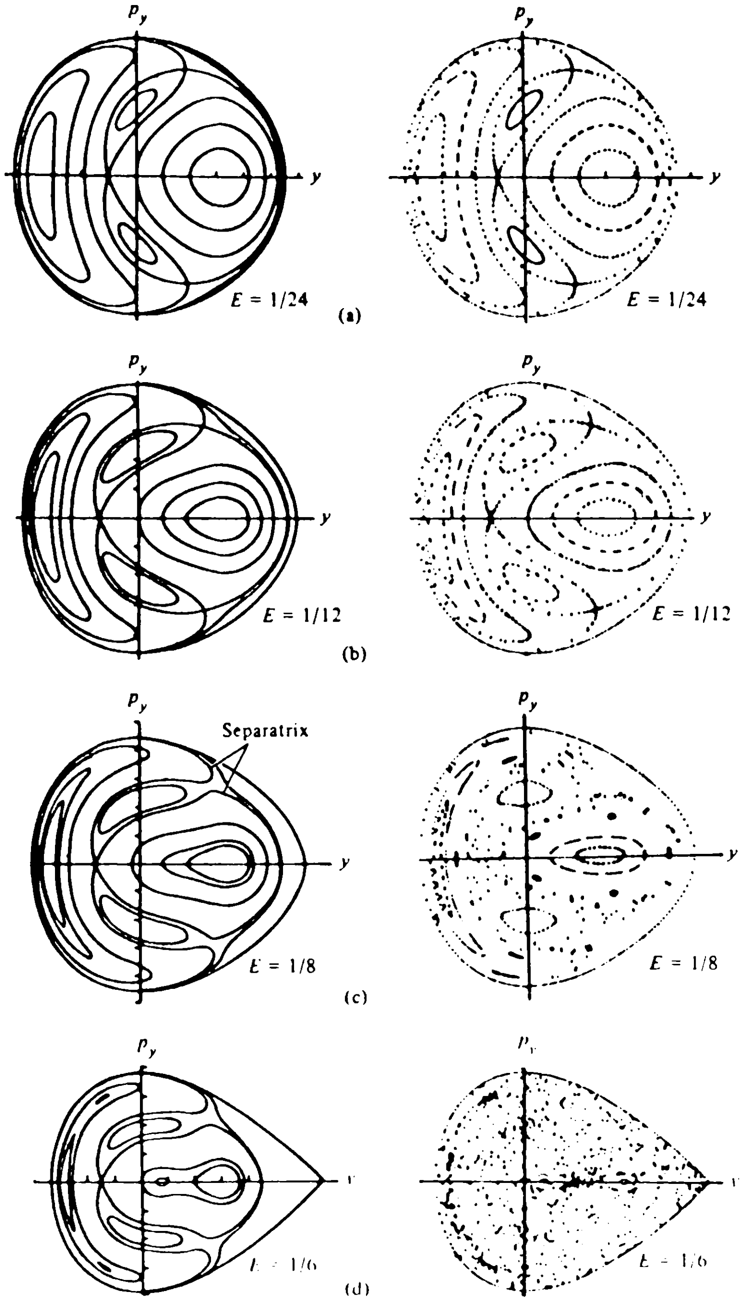


Figure 1.3 Comparison of surfaces of section of Henon-Heiles Hamiltonian system computed from perturbation theory with those computed numerically. The regular and chaotic regions are not separated by smooth boundaries (from Lichtenberg & Lieberman, 1983).

next chapter.

In the remaining part of this chapter the problems encountered in the perturbation methods are briefly discussed. More detailed discussions on the occurence of resonances, small divisors, quasi-periodic solutions and chaotic solutions are postponed to the next chapter in the much broader context of Hamiltonian dynami s.

Fractal Geometry and the KAM Theorem

Now it is well realised that the 'domain of a property' may not be defined by a simple set with smooth boundary manifold, but rather of fractal feature (Mandelbrot, 1977; Devaney, 1987; Feder, 1988). For example, let us take the simple mapping

$$Z_{n+1} = Z_n^2 - C \quad \text{with} \quad Z_0 = 0 \qquad (\text{ or equivalently} \quad Z_{n+1} = C'Z_n(1-Z_n\,)\,)$$

where all the quantities are complex with C as a complex parameter. The question is very simple: as n --> infinity, for what values of C does the mapping converge (respectively diverge)? It turns out that the range of the convergence (divergence) property of the mapping cannot be described by a smooth boundary curve, although there is nothing wrong with the continuity nor the differentiability in the above mapping. This is the well-known fractal Mandelbrot set (Mandelbrot, 1983). This is a typical example of simple questions with complicated answers. One can imagine the difficulties should one try to answer the question following a conventional method; one just lacks the notation to describe such a complicated boundary without the right notion for the solution to the problem.

In fact the above example is not artificial at all, it is found that the domain of a prescribed property is usually complicated in the parameter space. The property can be stability, equilibrium, convergence (divergence) and so on (see eg. Poston & Stewart, 1978; Lichtenberg & Lieberman, 1983). In the context of celestial mechanics, the topological methods and KAM theorem (named after Kolmogorov, Arnold and Moser) show that, in almost all nonintegrable systems, the properties of regular (periodic and almost periodic) and irregular (chaotic) motions are mixed in a very complicated way *similar* (not exactly) to that of the rational and irrational numbers. In phase space, the boundaries separating the two kinds of solutions are of fractal feature (see Fig. 1.3). Only for regular solutions can the expansion method be used rigorously; the expansion method is not compatible with chaotic trajectories (functions, motions). Therefore, the divergence of the series cannot always be justified by better expansions; this is an intrinsic difficulty.

The numerical integration method has been of great importance. However, one needs to be aware of the fact that the solution obtained on the computer may be quite different from the real solution if the problem is in the chaotic region. Again only regular single trajectories are computable with satisfactory precision.

Defining Chaos

Chaos is well recognised by scientific society as a rule for dynamical systems and in particular Newtonian mechanics. However, there is not an agreed definition for it. As is well known, modern science does not always follow the old fashioned axiomatic formulation, but rather 'to define is to understand' (Poincare). Since chaos is one of many nonlinear phenomena related to nonsolvability, nonpredictability and other limits, and is still a growing field not only digging in depth but also expanding in extent, its current status of not being universally defined should not be surprising. In the following we give some of the widely used definitions and point out their problem in order to comprehend the subject.

(1). Deterministic chaos is seemingly random and *apparently* irregular behaviour (solution, motion) of deterministic nonlinear dynamical systems, in contrast to smooth regular (eg. periodic) motions.

(2). Deterministic chaos is an *intrinsic* sensitive dependence on the initial conditions, exponential divergence of neighbouring trajectories (solutions), or occurence of positive Liapunov characteristic exponents of solutions to ordinary differential and difference equations.

(3). Deterministic chaos is an aperiodic solution to deterministic system, or solutions with continuous Fourier spectra.

(4). Deterministic chaos always exists in (bounded) nonintegrable, nonlinear ordinary differential equations and mappings.

(5). Deterministic chaos is due to the existence of hyperbolic fixed points (or unstable conditionally periodic orbits), whereby the adjacent trajectories, close to each other but on different sides of the stable and unstable manifolds, may approach the hyperbolic point (or unstable conditionally periodic orbits), and then depart quickly on receding from the hyperbolic point (or unstable conditionally periodic orbits).

(6). Deterministic chaos is defined by area-filling trajectories on a 2-dimensional Poincare surface of section.

(7). Deterministic chaos is homoclinic and heteroclinic motions in conservative systems, and strange (chaotic) attractors in dissipative systems.

(8). Deterministic chaos fills fractal regions in phase space.

(9). Deterministic chaos is what happens in a system with a large number of particles (eg. a box of ideal gas) where the motion of every particle is governed in the strict sense by the deterministic laws of Newtonian mechanics; however, because the number of particles involved and the frequency of collisions are so large that a dynamical description becomes practically impossible, thus a transition to statistical laws is needed.

(10). Deterministic chaos is defined as sensitive responses to errors, perturbations, and nonpredictability, incomputability etc thus caused.

A few incomplete comments on the above tentative definitions on deterministic chaos is in order. The most obvious characteristic of them is that by chaos we mean the behaviour completely intrinsic to *deterministic* dynamical systems such as the initial value problems of ordinary differential and difference equations, to which under very general conditions a unique solution exists for an *arbitrarily long but not infinite* time interval. It is commonly held that the behaviour of such deterministic systems is simpler than that of a completely random or stochastic system; and the prediction of its future behaviour based on the present state of motion is straightforward, if not trivial. However, the comprehension of generic chaotic solutions to such systems implies the futility of such expectation.

In fact the first four statements are generally true descriptions of various aspects of such deterministic chaos, with some underlying equivalence. However, none of them may be an ideal definition of chaos. The first is widely accepted by philosophers as a good definition because it uses the least exterior material and the most comprehensible language; the shortcoming of it is that the language is too descriptive, eg. a long periodic motion may seem irregular if observed in a relatively shorter time interval. The second needs more delicate specifications, although it does capture one of the most important points of chaos. For example, it is well-known that solutions to linear systems may diverge exponentially with time, but are not chaotic; moreover, the definition of the Liapunov characteristic exponent needs much more careful specification (eg. Lichtenberg & Lieberman, 1983, chapter 5). The third statement is probably a good definition for chaos in conservative systems, because periodic and aperiodic functions are very simple and theoretically accurate concepts; and in principle, it is the most directly comprehensible reason for the difficulties encountered in history related to chaos. However, in addition to its nonapplicability to linear systems, it may be blamed for the words being failing to convey all the beautiful aspects of the concepts. Furthermore, care must be paid not to confuse aperiodic functions with periodic functions of arbitrarily long but finite period. The fourth statement says where to look for chaos; but it uses the very delicate notions like linearity and integrability, the determination of which cannot be done in general.

The sixth and seventh are true; but as is specified in the statements, they cannot be general definitions for chaos in other systems. The fifth is correct in emphasising the significance of unstable periodic solutions. However, the mechanism of quick divergence of neighbouring orbits is not appropriate; it must be interpreted in the sense of the seventh statement where chaos is generated by hyperbolic points through homoclinic and heteroclinic points. The eighth is true only for strange attractors in dissipative systems, although chaos also causes regular and irregular motions to mix up in all scales and the boundaries separating the two kinds of motions to be fractalised.

The last two statements shall be considered to be erroneous. The tenth makes the occurrence of chaos a result of exterior influences, although the points stated are important outcomes of intrinsic chaos and structural instabilities. The error of the ninth statement needs particular attention, because it has been dominant in the ergodic theory of physical sciences. In statistical mechanics the H-theorem, which proves the non-decreasing feature of entropy, relies strongly on the collision process. In fact collisions are not responsible for the ergodicity because the measure of the collision manifold is zero in phase space (Siegel & Moser, 1971).

Therefore we will try to comprehend aspects of chaos without sticking to a particular definition. We would not even attempt to do that because we would like to leave the field open for new nonlinear behaviours to be included in the future, although not all nonlinear phenomena can be explained by the notion of chaos.

Planetary Motion by Large-Scale Numerical Integrations

Numerical experiments are essential in science nowadays due to the speed and accuracy of solution; it offers a very quick way of seeing the otherwise impossible results. Numerical results may be used to test theory. Moreover, some of the theoretical research must be guided by numerical results, especially when the behaviour of a system is too complicated to be achieved analytically. In celestial mechanics, the ODEs in rectangular coordinates or the Lagrange planetary equations may be integrated directly on the computer, because on the one hand the perturbation theory is not always applicable in the sense of rigorous mathematics, and on the other hand the actual calculation involved is too large, especially when the classical mixed variable transformation is used. In fact, analytical research and computational research progress in a parallel way. Moreover, when the perturbation is too large, as is often required in a theoretical approach, the perturbation method does not lead to very interesting results.

Thus the stability of the Solar System cannot be answered by perturbation theory whose convergence is questionable; the sufficient conditions for stability required by the

KAM theorem are also too restrictive. However, special perturbation theory may give some hints in such cases. Numerical experiments are now usually used in celestial mechanics for large scale prediction or systematic investigation. For example, in the work of Cohen et al (1972), the orbits of the outer planets are calculated up to 1,000,000 years centred at the epoch, January 6 of 1941. In the LONGSTOP consortium, the motion of the outer planets was computed forward and back in time over a total of 10^8 years. New results such as commensurable mean motions are still being confirmed and observed. For a more detailed account see Roy (1988).

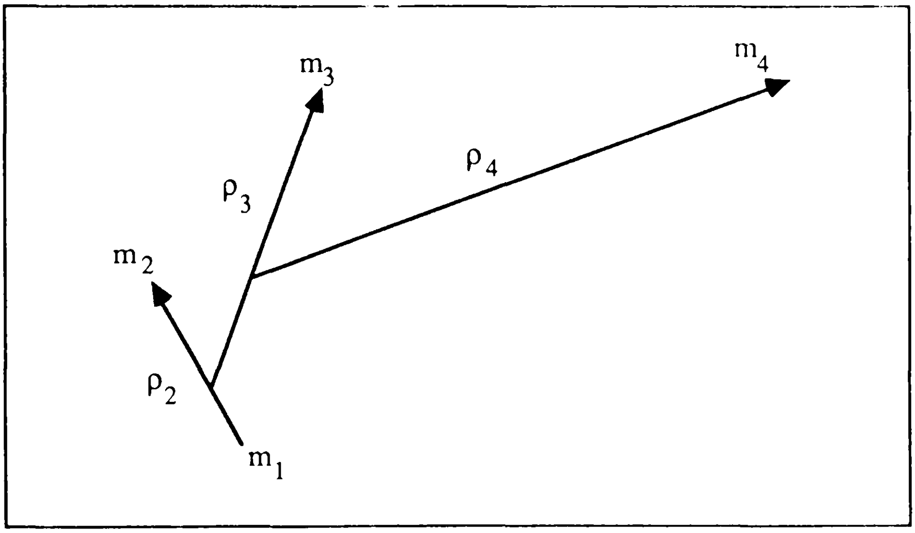


Fig. 1.4 The Jacobian vectors of the N-body problem

Roy's Hierarchical Stability and Hill-Type Stability

It is well known that whether the motion of a system is stable or not depends on how stability is defined. The very natural Liapunov stability is not useful for practical interests, because even the Kepler motion is not stable in this sense. Poincare's orbital stability is of great value in theory, but from the history of celestial mechanics and dynamics the condition of such stability is too hard to be established for a practical problem. Other stability worth mentioning in this connection is Poisson stability, Laplace stability and Hill stability, which are related with some of the geometrical studies of Poincare.

An N-body system is stable in the sense of Laplace if neither escape to infinity nor collision happens; whereas it is stable in the sense of Poisson if the system repasses to the initial situation infinitely often. Hill stability is defined for the circular restricted three-body problem if the zero-velocity curves close to trap the motion of the infinitesimal body. The Earth-Moon-Sun system is stable in this sense of Hill (Szebehely, 1967; Roy, 1982).

Very recently much work has been done investigating stability in the sense of Roy's hierarchical stability. This stability will be studied further in the second part of this thesis especially in the case of three-body problems. Like the Lagrange and Poisson stabilities, this stability again seems very simple; however, a complete answer is not so simple. The theoretical reason of choosing this stability to study is due to the successful generalisation of the Hill-stability to general three-body problem recently (see chapter 4).

Hierarchical stability (hereafter HS in short) was defined by Walker & Roy (1983) in connection with the so called Jacobian coordinate system. A dynamical N-body system is held to be HS if, during an interval of time substantially longer than the periods of revolution of the bodies in the system, the following conditions hold:

HS-(A). none of the bodies escapes to infinity from the system;

HS-(B). no dramatic changes occur in any orbit's size, shape or orientation to the invariable plane of the system.

HS-(C). $\rho_i < \rho_j$ for any i < j, where $\rho_i = |\rho_i|$ (i=2, 3, ... , n), being the Jacobian vectors which connect the barycentre of the first (i-1) masses and the i^{th} mass (see Fig. 1.4).

These conditions will be referred to as stability conditions HS-(A), HS-(B) and HS-(C) respectively. When anyone of them is not satisfied it will be referred to as instability condition A, B or C. This stability will be investigated in detail in chapter 5.

Summary

Although the motion of the bodies in the Solar System and its stability, as was quoted at the beginning from Hagihara, has not been answered by dynamical theory; much understanding and progress have been achieved by the N-body model. To sum up the success of the N-body model in answering the stability of the Solar System and questions remaining open, Roy's (Roy, 1982, chapters 1 and 8) formulation of stability of Solar System as an N-Body problem shall be quoted. Roy presents a list of questions which reasonably focuses the attempts made in the field of celestial mechanics, viz.

(1). How old is the solar system ?

(2). Does the distribution of planetary orbits alter appreciably in an astronomically long time?

(3). If so, do the orbits alter slowly; or can sudden far-reaching changes occur in one or more of the planetary orbits, even to the extent of planets changing their order from the Sun or colliding ?

(4). If the Solar System is stable and only slowly evolving, is this due to its present set-up with almost circular orbits, low inclinations, near-commensurabilities in mean motion and direct orbits ?

(5). Are the retrograde outermost satellites of Jupiter and Saturn captured asteroids?

(6). Are most of the other satellite orbits stable over astronomically long intervals of time, even if tidal action is taken into account ?

It appears that the most successful theory which has been used in answering the above questions is the theory of chaos. The advance made in this field will be reviewed in chapter 2, where we actually paid attention to chaotic dynamics in general. In chapter 3 we investigate the historical influence on each other in the development of chaotic dynamics and quantum mechanics.

In Part II of this thesis we will discuss in detail the hierarchical stability and Hill-type stability of the few-body problem. Compared with the theories reviewed in Part I, the attempts of Part II only has a limited power towards an answer to the above questions. However, many interesting results have been obtained in this field. In chapter 4 we have modified the previous approaches and obtained stronger inequalities governing Hill-type stability regions of spatial 3-body problem. Chapter 5 is a numerical exploration on the hierarchical stability of the coplanar 3-body problem. In chapters 6 and 7 we discuss the relationship between symmetries and conservation laws in general relativity, and make a first attempt to generalising the classical study into the framework of general relativity.

It may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an ***enormous*** error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon.

--- Poincare

No significant formal system can ever be strong enough to prove or to refute every statement it can formulate. --- Godel

---

CHAPTER 2
Stable and Chaotic Behaviour in Hamiltonian Dynamics

In the last chapter we briefly described the physical and astronomical phenomena of heavenly bodies and the fundamental theories relevant to their motion; stress is laid on particular cases in the Solar System. The successful explanations of such phenomena by Newtonian mechanics and the difficulties encountered in the classical N-body model are reviewed, with an emphasis on the generic behaviour of chaos and its effects in both continuous and discrete dynamical systems.

In the present chapter, we will give a deeper view of chaotic (or resonant, nonlinear) dynamics and in doing this we consider all three revolutionary physical sciences of the century (namely relativity, quantum theory and chaos). A discussion about the relationship to statistical mechanics is not included. This is not solely a review of the existing literature on the subjects which has received much popularity in the past several decades, but also a collection of the author's own opinions. Most of the material is not presented completely, nor is intended to be mathematically rigorous, but in a way to help comprehend the problem mathematically. Nevertheless, compared with the following chapter, the content of this chapter is closer to standard material; and many confusions often occurring in textbooks are clarified.

The chapter begins with a selected discussion on the theories of numbers, functions, differential equations and convergence of infinite series (section 2.1), followed by a summary of Lagrangian and Hamiltonian mechanics (section 2.2). Integrability and separability are discussed in more detail in section 2.3 because of their importance to chaos; emphasis is laid on the coordinate-free interpretation of such concepts. In section 2.4 perturbation theories are outlined to comprehend the problem of small divisors, the possibility of chaotic motion and convergent method to establish quasi-periodic motion. The geometrical method and KAM theorem are included in section 2.5; chaos in Hamiltonian systems is discussed using Poincare's surface of section. The chapter is concluded by a personal comment on the implication of the occurrence of commensurable mean motions in the solar system (section 2.8) suggested by Roy & Ovenden (1954).

In the discussion, effort is made to stress the importance of the few-body problem and the modern geometrical notion on manifolds. Although a detailed discussion on chaotic attractors in dissipative system is out of context, a collection of such mappings occupying some significance in history and still under active investigation is presented in section 2.7. Characteristic features of chaos in both area-preserving mappings and dissipative mappings can easily be observed by putting them onto a computer.

2.1 Introduction to Ordinary Differential Equations and Mappings

It is usually remarked that the nineteenth century was the century of linear dynamics, whereas the twentieth century that of nonlinear dynamics. As a result of historical inertia, even nowadays, scientific society living at the closing page of the century is still satisfied with the simple solutions inherent to linear systems, which are often easily distinguishable from stochastic phenomena. However, along the track of mathematical astronomy, it was already shown by the end of the last century by Poincare (1892), and early this century by Birkhoff (1917, 1920, 1927) followed by the work of Siegel and KAM, that nonlinear systems can produce in principle not only simple **regular** (ie. conditionally periodic, as is used by Wisdom, 1987, Binney et al, 1987) solutions but also very **irregular** (ie. aperiodic) solutions which appear to be random. This kind of chaotic solution, together with fractal geometry, has caused popular attention after being rediscovered from experiments (Lorenz, 1963) and made visualisable by the advent of the computer (eg. Henon and Heiles, 1964; Henon, 1965-70; Henon, 1969, 1976; Greene, 1979; Chirikov, 1979). Since then, chaotic dynamics has found its increasing application in various fields from engineering to biology and economics (Stewart, 1989).

Yet, chaotic dynamics and fractal mathematics are still growing subjects, and a great number of articles and literature have appeared to convey these new sciences to both the academic and public worlds. In this literature, much attention has been drawn to the fact that simple dynamical systems such as the **logistic mapping** can produce exceptionally complicated phenomena, implying the possibility that phenomena previously taken to be stochastic may in fact obey certain underlying deterministic laws. However, the more fundamental root of the 'simplicity producing complexity' rule is respected in this chapter. Indeed, what occurs in chaotic dynamics and fractal mathematics is very similar to the well-known chess-board game story. By simply placing one grain on the first square, two on the second, four on the third and so on, one finally finds oneself in an astonishing situation. This is just a result of iterating on the simple numbers. Chaos and fractals are the outcome of certain similar procedures: iteration of simple well-behaved functions. Repeating a few simple operations on simple elements would not cause a qualitative transition and usually results in something that is conceivable without detailed analysis; while increasing the number of operations can lead to results so complicated that is ultimately beyond any straightforward intuition.

In an iterative procedure, the nature of the final state depends on how the generating structures accumulate under iteration and how fast this process grows. Usually an infinite number of iterations leads to qualitatively different effects. This may be demonstrated by some very simple examples. If $\Sigma_n = x^{-1} + x^{-2} + ... + x^{-n}$, then it is easy to verify that the corresponding infinite series equals $1/(x-1)$. The singular point of the finite series is at $x=0$ no matter how large n is, whereas that of the infinite series is shifted to $x=1$. A second example is based on the observation of Fourier expansion: when the function is periodic, no matter how large the period is, it can always be expanded in convergent Fourier series; however, if the period is allowed to go to the limit of infinity, namely aperiodic limit, then the function usually cannot be developed by Fourier analysis.

Let us observe that a discrete mapping in fact creates an infinite sequence, the periodic points of the mapping being just the repeating elements in this sequence, whereas the irregular trajectories can be viewed as the irregularity of the sequence. Chaotic solutions are just complicated solutions; while fractals are a kind of order existing in the structure of chaos or in the domain where some specific properties appear for the solution when viewed in parameter space. Therefore chaos and fractals are just phenomena which exist in principle, with differential and difference equations being simply their generators; thus there must be a compatible way of describing them as functions.

Functions and Infinite Series

Continuous functions are fundamental in science. One of their most useful subsets is the set of **smooth** functions (continuous derivatives up to some finite or infinite order exist). In this thesis a finite number of bounded discontinuities are allowed; thus with slight modification, results on functions without discontinuities are also true for **piecewise** (or sectionally) continuous and smooth functions. Functions whose Taylor expansions converge in the neighbourhood of a point are called **analytic** at that point.

In physics it has been conventional to assume that functions are either (piecewise) smooth up to the required order or (piecewise) analytic. These assumptions are met by **elementary** algebraic (integral rational, fractional rational and irrational) and transcendental (trigonometric, exponential, hyperbolic and their inverses) functions and often by convergent infinite series built upon them. Non-elementary functions such as the Dirichlet function (that is, f(t) is 1 if t is rational and 0 if t is irrational) are usually taken as the exception since they are not well behaved. However, the recently recognised application of fractal geometry such as the Koch curve (eg. Feder, 1988) has shown the importance of continuous but not smooth functions; the continuous but nowhere differentiable Weierstrass function (infinite series) $\Sigma\ (k!)^{-1} \sin\ [(k!)^2x]$ is no longer regarded merely as a mathematician's abstract construction of minority counter-examples (Korner, 1988, P38). They are becoming increasingly important in application.

Even if only the class of smooth functions is concerned, its difference from analytic functions is not critically sharpened. In fact, the associated Taylor series of a smooth function may converge but to a different function, or it may diverge for all points in the neighbourhood of the expanding point except at it; well-known examples of these 'exceptional' cases are the functions, for the former, $\exp(-1/x^2)$, and for the latter, $\Sigma e^{-k}\cos(k^2x)$ defined in the domain [-1,1] (see Poston & Stewart, 1978, P44).

This point shows how narrow the class of analytic functions is, which is more obvious for functions of complex variables. This point should also be a warning towards solving ordinary differential equations by power series, in which method a power series is assumed to be the solution of an ODE and substituted into the equation to determine the coefficients. Since the formal solution always satisfies the equation (Poincare, 1892), a convergent series is always a smooth solution. However, if the series diverge, it may be that there is no solution, or the series expansion is inadequate (though useful solutions can often be achieved by such inadequate methods), or the solution is smooth but not analytic.

For initial value problems, the **existence** and **uniqueness** of smooth solutions are assured by theorems established in more powerful ways other than by series expansion (eg. Roxin, 1972); therefore the first case is discarded in regular regions. In particular, in occasions when the boundedness of a solution can be established (eg. Davies & James, 1966), a failure of the Taylor series method should be concluded from its divergence. It is also useful to note that the existence and uniqueness theorem is only a kind of local notion, since it is only established for infinitely small or finite time intervals in phase space. In general the interval is allowed to be arbitrarily large but not infinite. Moreover, the convergent successive procedures for constructing solutions is also limited to such a finite interval. This is not contradicted by the divergence of the successive procedures for solving the Lagrange planetary equations, because the method is applied on an infinite time interval. However, when the system is linear, the theorem becomes global in the sense that a unique solution exists and may be constructed by a suitable iterative method in the whole phase space for an infinite time interval.

Taylor and Fourier series are often used in solving initial value problems. However, neither is compatible with the existence and uniqueness theorem: smooth solutions are not always analytic; whereas a Fourier series imposes periodic restriction on solutions. It is usually remarked by celestial mechanists that Fourier analysis has an advantage over Taylor expansion in the sense that the former is valid over the whole real line, while the latter only in the vicinity of some points. However, this is not always true. Taylor expansion can sometimes also be valid over the whole real line (eg. e^x, sinx, cosx); whereas Fourier analysis is also only a local treatment, because periodic functions over the whole real line are in fact a repetition of a local property. When the period is taken to an infinitely large limit, the above local property in a period cannot be arbitrary; otherwise a divergence occurs in a Fourier analysis. Therefore there is no universal infinite series (nor procedures) valid for all smooth functions defined on $(-\infty, +\infty)$, not to mention all continuous functions.

This point is important in order to see why the Fourier series solution of the N-body problem should in general diverge as was studied in detail by Poincare (1892). To see this let us argue that a nonlinear ODE must in general have both quasi-periodic and aperiodic solutions. The case with all solutions quasi-periodic must be very atypical, because aperiodic solutions exist even in linear ODEs such as the well known *Mathieu equation* (eg. McLachlan, 1947; Jordan & Smith, 1977), although in general this aperiodicity does not mean chaos. Secondly, from intuitive observation, both (almost) periodic and aperiodic solutions exist in the real N-body problem; to be compatible with nature, the classical N-body problem must have both kinds of solutions. However, aperiodic solutions defined for all time cannot be expressed in Fourier series. Keeping in mind that the conventionally obtained formal Fourier series always satisfyies the ODE, one sees that the series should not converge in general. Therefore, the divergence of the formal series solution, which was proved by Poincare (1892), is not surprising based on the above heuristic argument, nor is this divergence literally related to chaos. What is really important is that he showed the detailed condition of convergence and its relation to the property of numbers, and that the series is asymptotic. Therefore we say that aperiodicity and divergence of Fourier series is only a signal of chaos. The more direct contribution of Poincare's relevant to chaos is his discovery of nonintegrability and the existence of homoclinic and heteroclinic points in the N-body problem. After a discussion of such problems in later sections, we will see that it is the existence of chaotic solutions that prevents Fourier series solutions from being convergent and the construction of solutions from being global.

For boundary value problems, usually there is neither existence nor uniqueness theorem; nevertheless, certain conclusions can usually be achieved by the shooting method in the light of initial value problem. The problem is simpler if the equation is linear so that a general solution may be established for the equation, then the constant of integration may be determined by the boundary conditions. When the boundary condition is imposed on a finite boundary and no singularity exists in the bounded domain, then the Fourier series should be applicable in general. Taylor series is also used, but it is not a complete method in principle.

For example, let us consider the Legendre linear ODE with boundary condition at x=-1 and 1. Although theorems exist to assure this is a Sturm-Liouville eigenvalue problem (Courant & Hilbert, 1953), they cannot be used to determine the eigenvalues, nor to obtain physically interesting smooth eigenfunctions. The introductory way of finding the eigenvalues and eigenfunctions is by power series, which is not a complete method - eigenvalues and smooth eigenfunctions may be missed by this method. Because of the linearity of the Legendre differential equation (using the parameter n(n+1) as usual), there is no doubt that when n takes non-negative integer values, the n^{th} order terminating polynomials (Legendre polynomials) are the unique solutions, nor on the method of abandoning the infinite part which diverges at either boundary points and retaining the terminating polynomials. However, doubt arises on the introductory argument that, since the power series diverges at either boundary point for any real n, and for non-integer values of n the series does not terminate, and hence no solution exists. Consequently, the problem is taken as an eigenvalue problem and all eigenvalues have been found.

Although the conclusion is correct, the logic is false because smooth but not analytic eigenfunctions may exist which cannot be found in the above way. When the problem is defined on a bounded domain containing no singular points, the expansion method compatible with the physically interesting solutions, namely, smooth solutions, is a

Fourier series and the so called generalised Fourier series, such as series of Legendre polynomials, which can be introduced in a different way other than as a solution of the Legendre equation (see Courant & Hilbert, 1953). These can be shown to form a **complete** orthogonal system, namely, piecewise continuous functions satisfying the boundary conditions can be expanded in a unique generalised Fourier series which converges in the mean (Courant & Hilbert, 1953). Absolute and uniform convergence of the series can be established under stronger smoothness conditions; whereas a C^1 function satisfying some intermediate conditions and the boundary conditions can be expanded in a convergent series in terms of these orthogonal functions. Therefore, if the solution is assumed to be an infinite series in terms of Legendre polynomials other than a Taylor series, then it can be shown that no extra eigenvalues nor extra eigenfunctions exist for the Legendre differential equation with the boundary conditions.

In connection to the convergence of the perturbation theory and KAM theorem, it is also useful to note the speed of convergence of such series expansions; the smoother the function is, the more rapidly its Fourier series converges. In fact, if f(t) is C^n smooth and its associated Fourier series is $\Sigma C_k e^{ik\omega t}$, then the coefficients can be estimated by $|C_k| \leq c/k^n$, where c is a constant. This relation may be obtained immediately if the expression for the coefficients is integrated by parts k times. Similar results also exist for multiple periodic functions of many variables.

After pointing out the generality of (generalised) Fourier series, however, we shall mention a limitation of it. The sufficient smoothness condition for the function to be expandable cannot be relaxed too much; there always exist continuous periodic functions whose associated Fourier series diverge at a given set of measure zero points (Kolmogorov considered the problem in depth, see Korner, 1988, P75). Let us also note that Fourier series is only applicable to multiple periodic and **conditionally periodic** (or almost periodic, quasi-periodic; see eg. Szebehely, 1967; Arnold, 1978; Siegel & Moser, 1971; Berge et al, 1984) functions, which only have a finite number of extremes in a period; while aperiodic functions cannot be treated in this way except for a small class that may be developed by Fourier and Laplace transforms. Functions defined on a finite domain can be developed either by Fourier series or Fourier transform.

Although most aperiodic functions cannot be expanded in convergent series, divergent asymptotic expansion can often be developed for practical usage (Poincare, 1892; Whittaker & Watson, 1902). A series $\Sigma A_k x^{-k}$ is said to be an asymptotic expansion (either convergent or divergent) of a function f(x) if the sum over the first n terms $\Sigma A_k x^{-k} \to f(x)$ as $x \to \infty$ for fixed n. Such series are often the only possible means of obtaining the solution to some problems, although they are usually produced by nonrigorous expansion. The semi-convergent feature of the classical perturbation procedure shown by Poincare (1892) justifies its applicability in practice.

Historical interests in looking for convergent series for the N-body problem is in general related to asymptotic expansions. From the examples given in either Poincare (1892) or Whittaker & Watson (1902), we see that by rearranging the order of the terms, a convergent series may be made semi-convergent or even divergent. Therefore the divergence of one expansion procedure does not necessarily imply the divergence of a solution. In addition, the occurrence of secular or mixed-secular terms in celestial mechanics does not imply divergent solutions either. An example of the latter case is the function $\sin(1+\varepsilon)t = \sin t + \varepsilon t \cos t - 0.5\, \varepsilon^2 t^2 \sin t - \ldots$, which is convergent for all t in spite of the mixed-secular terms (Roy, 1988).

Finally it is important to keep in mind the richness of infinite series, and in particular their compatibility with chaos and fractals. Obviously, some discontinuous functions can be defined by different simple functions in different regions; but this is not a convenient expression for such functions. In order to see a possible alternative, let us recall that infinite series built up on well behaved functions can in fact produce discontinuous functions (eg. Whittaker & Watson, 1902, P44). The most apparent example is probably the infinite trigonometric series. Conventionally, when this happens, attention then is turned to finding conditions under which the infinite series converges to continuous functions, and the importance of converging to discontinuous functions is ignored. However, at the times of chaos and fractals, some emphasis must be paid to such so called exceptional cases.

Knowing how to generate chaos (chaotic attractors and fractals can all be studied utilising infinite sequences), it is useful to study whether infinite series (or more generally infinite sequence) may converge to functions with many or infinite extremes, or even discontinuities. In fact, Weierstrass constructed a continuous but nowhere differentiable function.

Ordinary Differential (Difference) Equations

Ordinary differential equations (**ODE** in this thesis) and discrete **mappings** (or transformations) are usually called (continuous and discrete) dynamical systems. Since most important theorems can be stated in a similar way, we shall concentrate on ODEs in general. This mathematical notion finds its wide usage in science because of the implicitly assumed continuity and smoothness of the world: an observable variable is assumed to be a function of position, time or other independent variables; by knowing the value of the observable variable at one point in the independent variable space, one hopes to grasp the behaviour in the neighbouring regions or future (or past), namely, making a prediction.

Algebraic equations were a great advance in the history of mathematics, because they treat unknowns as knowns to form the equalities, and then solve the equations by systematic routines, and thus offer a unified method to replace the previously scattered methods of solving these problems. Differential equations, as improved algebraic equations, take both the unknown functions and their derivatives as known material to formulate the equalities. The idea of equations is to find what is conserved in the case of a natural process and construct the equations, because the world is believed to be casually interrelated with cause and effect. Something, such as a combination of the variables, must be conserved, but whether it is conceivable or not depends on the creativity of human being (combination of variables, functions and their derivatives). The purpose of studying equations is to find their solutions. If they cannot be solved due to some principle limit, then they must be replaced by more appropriate laws. Therefore the relevance of differential equations is a result of regarding the world as variables, elementary functions and infinite series built on them. The discovery of chaos and fractals suggest a limit on the power of differential equations in general. Physically, this limit on dynamics is more fundamental than the two limits from relativity (speed of light limit) and quantum theory (uncertainty principle).

Differential equations are classified by type and order. Ordinary and partial differential equations are distinguished according to the types of derivatives involved; the order of a differential equation is the maximal order of the differentiation that appears in the differential equation. Only ordinary differential equations are of concern in the present study.

An ordinary differential equation is a functional relationship of the form $F(t, x, x', x'', \dots, x^{(n)}) = 0$ between an independent variable t, an unknown function x(t), and a finite number of its derivatives. Moreover, there may be systems of ODEs involving various unknown functions $\mathbf{x}(t)$. In general, it is always possible to reduce such a system of ODEs to that in which only derivatives of the first order appear. This can be done by introducing new unknown functions. Thus it suffices to consider first order systems of the form, $\mathbf{x}' = \mathbf{F}(t, \mathbf{x})$. This is an advantage for a unified theoretical study; nevertheless, second order ODEs are also of practical convenience.

In practice, autonomous systems and conservative systems are often encountered. A system is autonomous if F or $\mathbf{F}$ is independent of t; such a system is called conservative, if, furthermore, $\mathbf{F}(\mathbf{x})=\nabla_{\mathbf{x}}U(\mathbf{x})$. The class of systems with $\mathbf{F}(t, \mathbf{x})=\nabla_{\mathbf{x}}U(t, \mathbf{x})$ are also frequently used; but it is not conservative.

In the case where F (or **F**) is a linear polynomial in the unknown function and its derivatives then the differential equation is called linear; otherwise it is called nonlinear. Linear ODEs are further classified as homogeneous and nonhomogeneous linear ODEs; or linear ODEs with constant coefficients and time dependent coefficients.

Although the **local** existence and uniqueness of C^1 solution is usually established by a successive approximation method for the first order systems, it is applicable to higher order systems. It is also stated as satisfying initial conditions, and the solution is a continuous function of the initial conditions. The above existence and uniqueness theorem can be sharpened for linear systems (eg. Roxin, 1972) and, in fact, it becomes a global theorem. Moreover, the solutions have more simple properties which are not shared by that of nonlinear systems. Because of these properties, there are no chaotic solutions in linear systems. However, this does not mean that linear systems can always be solved easily; their behaviour is not always simple, as can be seen from Floquet theory on linear systems with periodic coefficients (Jordan & Smith, 1977).

There is no general existence nor uniqueness theorem for boundary value problems. Eigenvalue and eigenfunctions are a result of boundary conditions, which may happen in both classical mechanics and quantum mechanics. However, in quantum mechanics the most important condition leading to quantised states is that due to natural boundary conditions which are in fact symmetries of background space and periodicity. Thus quantization is a result of symmetry or periodicity.

In the general solution to a linear system of n first order ODEs, usually n arbitrary integration constants appear which must be determined by the n initial conditions in phase space; on the other hand a solution to a general nonlinear system of n first order ODEs is determined by n initial conditions. An **integration constant** is a function of the form C(t,**x**), which is constant on a trajectory. This should not be confused with a **first integral** (or, conserved quantity, constant of motion, invariant of motion, integral of motion) which is a function of the phase space variables, **I**(**x**), and is constant on a trajectory (but see Whittaker, 1904). A first integral is called **isolating** if it is single valued, or non-isolating if it is non-single valued. Since only isolating integrals are of importance, the word 'isolating' is usually dropped. The classical energy, momentum and angular momentum integrals are isolating; for examples of non-isolating integrals see, eg. Binney & Tremaine (1987). The existence and uniqueness theorem says that a unique solution exists in the neighbourhood of an ordinary point, $\mathbf{x}(\mathbf{x}_0, t_0, t)$, which is a continuous single-valued function. Therefore the n initial coordinates, $\mathbf{x}_0$, can always be solved for inversely as functions of t and **x**. Thus they are integration constants, but in general not first integrals.

A dynamical system of n first order ODEs (mappings, respectively) is **integrable** if it can be solved by quadrature, which usually requires the existence of n independent single-valued first integrals. We shall see in later sections that the class of Hamiltonian systems is a particularly interesting subset of ODEs, having more elegant properties and wide applications. For example, for a 2n dimensional Hamiltonian system, n first integrals suffice for its integrability. It must be noted that the notion of involution (defined by the Poisson bracked) is not defined in general dynamical systems; nor are the other conditions and outcomes (cf. section 2.3) of an integrable Hamiltonian applicable in the more general sense; because they may be a reflection of the particular property of canonical systems.

2.2 Standard Formulation of Lagrangian and Hamiltonian Systems

Since the variational principle is just a reflection of some invariant properties, we shall follow Abraham & Marsden (1978) and not include the variational principle in the following discussion. Complete classical discussions on such systems may be found in Whittaker (1904), Goldstein (1950) and Arnold (1983).

Lagrangian Systems

Lagrangian systems form a class of very important dynamical systems, which are defined in configuration space and have the following expression,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = 0, \qquad L \equiv L(\dot{q}, q, t), \qquad \dot{q} \equiv \frac{dq}{dt},$$

where L is the Lagrangian of the system and $q=\{q_i\}$ the generalised coordinates. For the motion of particles in a potential field, L=T-U, where T=T(q, dq/dt, t) is the kinetic energy and U=U(q,t) the potential of the field.

It is well known that when one of the generalised coordinates, q, is absent in L, then there is a corresponding a first integral of the system. But this is usually carelessly remarked as: if one of the q's is not involved in the potential, U, then a first integral results. This statement is generally false, although it is always true if Cartesian coordinates are used. The reason is that the kinetic energy is dependent on q unless Cartesian coordinates are exploited.

It is instructive to look at this problem using the language of modern differential geometry, as will be heavily relied on in the later chapters on relativity. To fix the idea, let us take the motion of a single particle in an exterior potential as an example. In fact, both T and U are coordinate independent scalar functions. Thus $T=mv^2/2=mg(v,v)/2$, (with v=d/dt) may be interpreted as a scalar of the contraction of the metric tensor with the tangent vectors of the solution curve q=q(t) in the standard Euclidean space. Although in the Cartesian coordinates, the components of the metric tensor are constants, they are usually functions of the coordinates in an arbitrarily chosen coordinate system. In order that a coordinate q is absent from L, that is, using geometric language, the Lie derivative of L along the vector field of the q-coordinate is vanishing, a sufficient condition is that the q-field is a Killing vector of the metric tensor and that U is independent of q. Therefore, the coordinate basis field along which U is invariant does not necessarily correspond to any first integral; to do so it must also be a Killing field (Schutz, 1980).

The above Lagrangian system may be put into a canonical form, to which the remaining part of the section is devoted, via the following Legendre transformation,

$$H(q, p, t) \equiv \Sigma(\dot{q}_i p_i) - L(\dot{q}, q, t), \quad \text{where } p \equiv \frac{\partial L}{\partial \dot{q}},$$

where the time derivative of H satisfies $dH/dt=\partial H/\partial t=-\partial L/\partial t$.

Hamiltonian System and Canonical Transformation

As is known, ODEs can be reduced to an equivalent first order system, $d\mathbf{x}/dt=\mathbf{F}(t, \mathbf{x})$, in phase space; whereas Hamiltonian systems form a special class of ODEs with even (eg. 2n) dimensional phase space, viz.

$$\frac{dq(t)}{dt} = \frac{\partial H(q,p,t)}{\partial p}, \quad \frac{dp(t)}{dt} = -\frac{\partial H(q,p,t)}{\partial q}, \quad \frac{dH(q,p,t)}{dt} = \frac{\partial H(q,p,t)}{\partial t},$$

which can be written in an equivalent form by a use of Poisson bracket, viz.

$$\dot{q} = \{q, H\}, \quad \dot{p} = \{p, H\}, \quad \dot{H} = \partial H/\partial t,$$

where H is the Hamiltonian, $q=\{q_i\}$ and $p=\{p_i\}$ are the generalised momenta and coordinates.

When H is independent of t, the canonical system is called an autonomous Hamiltonian system. Such a system is conservative, and H becomes the usual energy integral. Moreover, the conjugate momentum of any ignorable coordinate is a first integral of the system (not all integrals can be made conjugate to coordinates of the base space even if a transformation is allowed). The Poisson bracket is useful in finding first integrals because of the following relation for an arbitrary function F(p,q,t), viz.

$$\frac{dF}{dt} = \sum_i\left(\frac{\partial F}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial H}{\partial q_i}\right) + \frac{\partial F}{\partial t} = \{F, H\} + \frac{\partial F}{\partial t}.$$

In solving the above set of first order differential equations, it is useful to take the advantage of the ignorable coordinates so as to reduce the dimension of the problem, with the equations of the remaining variables being still in canonical form. If such coordinates do exist, they can usually be found by performing canonical transformations in the phase space. A canonical transformation is a transformation of the canonical variables (p,q) of the phase space to a set of new canonical variables $(\mathscr{p},\mathscr{q})=(\mathscr{p}(p,q,t), \mathscr{q}(p,q,t))$, whereas the canonical form of the differential equations is preserved. Such a transformation can usually be produced conveniently by a generating function. The four possible forms for the generating function, transformations and the new Hamiltonians are summarised in Table 2.1 (Whittaker, 1904; Szebehely, 1967; Stiefel & Scheifele, 1971).

Table 2.1

$S_1(q, \mathscr{q}, t)$:	$p=+\frac{\partial S_1}{\partial q}$,	$\mathscr{p}=-\frac{\partial S_1}{\partial \mathscr{q}}$;	$\mathscr{H}=H+\frac{\partial S_1}{\partial t}$
$S_2(q, \mathscr{p}, t)$:	$p=+\frac{\partial S_2}{\partial q}$,	$\mathscr{q}=+\frac{\partial S_2}{\partial \mathscr{p}}$;	$\mathscr{H}=H+\frac{\partial S_2}{\partial t}$
$S_3(p, \mathscr{q}, t)$:	$q=-\frac{\partial S_3}{\partial p}$,	$\mathscr{p}=-\frac{\partial S_3}{\partial \mathscr{q}}$;	$\mathscr{H}=H+\frac{\partial S_3}{\partial t}$
$S_4(p, \mathscr{p}, t)$:	$q=-\frac{\partial S_4}{\partial p}$,	$\mathscr{q}=+\frac{\partial S_4}{\partial \mathscr{p}}$;	$\mathscr{H}=H+\frac{\partial S_4}{\partial t}$

Extended and Reduced Phase Space

Since the solution of a canonical Hamiltonian system usually depends on the dimension of the phase space, in order for a unified understanding of the structure of the phase space, it is instructive to work in an extended phase space when the Hamiltonian is time dependent, or in a reduced phase space when the Hamiltonian is time independent, whereas the canonical feature of the system is preserved.

If the old Hamiltonian system in the 2n-dimensional phase space is,

$$H(q, p, t) \quad , \quad \frac{dq}{dt}=\frac{\partial H}{\partial p} \quad , \quad \frac{dp}{dt}=-\frac{\partial H}{\partial q} \quad ,$$

then by performing the following transformation,

$$\mathscr{q} = q \quad , \quad \mathscr{p} = p \quad ; \quad \mathscr{q}_{n+1} = t \quad , \quad \mathscr{p}_{n+1} = -H \quad ,$$

we obtain a new Hamiltonian system in the 2n+2 dimensional phase space, viz.

$$\mathcal{H}(\mathscr{q}, \mathscr{p}) = H(q, p, t) - H \quad , \quad \frac{d\mathscr{q}}{d\mu} = \frac{\partial \mathcal{H}}{\partial \mathscr{p}} \quad , \quad \frac{d\mathscr{p}}{d\mu} = -\frac{\partial \mathcal{H}}{\partial \mathscr{q}} \quad ,$$

where the new Hamiltonian, $\mathcal{H}$, is independent of the new arbitrary 'time' variable, μ.

Conversely, given an 2n-dimensional Hamiltonian system, viz.

$$H(q, p) \quad , \quad \frac{dq}{dt} = \frac{\partial H}{\partial p} \quad , \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q} \quad , \quad \frac{dH}{dt} = 0 \ ,$$

then by choosing any generalised coordinate as the new 'time' μ, and the conjugate generalised momentum as the new 'time' dependent Hamiltonian, we obtain a new Hamiltonian system in the 2n-2 dimensional reduced phase space (Whittaker, 1904; Lichtenberg & Lieberman, 1983).

Therefore, the motion of a system with a time dependent Hamiltonian is equivalent to that of a time independent Hamiltonian with an additional degree of freedom, and vice versa. In this way, the theory developed for a time independent Hamiltonian with n degrees of freedom also applies to a time dependent Hamiltonian with n-1 degrees of freedom. In particular, a time independent Hamiltonian with two degrees of freedom is dynamically equivalent to a time dependent Hamiltonian with one degree of freedom.

2.3 Solution Method I - First Integrals and Integrability

Although the successive approximation method exploited in establishing existence and uniqueness of solution may be used to construct solutions to an ODE, it is usually far from being applicable in practice, either because of the amount of calculation involved or the unsatisfactory speed of convergence. Nevertheless, such dynamical systems have a deterministic nature. This does not contradict the existence of chaotic trajectories in the same dynamical systems, nor does it contradict the new concepts like nonpredicatability and non-computability. The point is that such an iteration method is a local notion only, which is not useful in distinguishing various topologically different types of solutions. Even worse, most differential equations admit neither an exact analytic solution nor a complete qualitative description (Arnold, 1983). Therefore, various exact and approximation methods have been developed to solve the differential equations or to infer the qualitative feature of the solutions in phase space. The simplest cases of

Hamiltonian systems are discussed in this section, namely, systems reducible to quadrature. More complicated systems shall be considered in later sections.

Symmetries and Integrable/Nonintegrable Hamiltonian Systems

It is well known that isolating first integrals are related to ignorable coordinates; more geometrically they are related with invariant properties and symmetries through Noether's theorem (Noether, 1918; Abraham & Marsden, 1978; Olver, 1986). In the extreme cases, a system may possess so many symmetries that the system is completely integrable by quadratures. Integrability is a coordinate-free notion, and may be defined in the following equivalent ways for autonomous Hamiltonian systems with n degrees of freedom:

Integrability in general sense:

(1). A Hamiltonian is integrable if it possesses n independent global isolating first integrals in involution (Liouville's integrability).

(2). A Hamiltonian is integrable if it is independent of all generalised coordinates.

(3). A Hamiltonian is integrable if $\partial H/\partial p_i=f(q_i)$ for all i=1, ... , n, so that the n equations for the generalised coordinates can be integrated by quadrature, ie., $dt=dq_i/(\partial H/\partial p_i)$.

(4). A Hamiltonian is integrable if it is completely separable, namely, $H=\Sigma H_i(p_i,q_i)$. For example, dynamical systems of Liouville's type are integrable (Whittaker, 1904, P67).

(5). A Hamiltonian is integrable if canonical transformations (or generating functions) exist such that it can be reduced to one of the first four cases.

Integrability in restrictive sense:

(6). A Hamiltonian is integrable if all solutions are bounded and conditionally periodic.

(7). A Hamiltonian is integrable if it is equivalent (globally diffeomorphic) to a linear canonical system.

(8). A Hamiltonian is integrable if adjacent trajectories at worst diverge linearly.

It is useful to emphasise that we are restricted to Hamiltonian systems; otherwise the sixth and seventh statements would lead to contradictions: combining them will lead to the conclusion that linear systems can only have quasi-periodic solutions, which is obviously false. It is a classical result that (see Arnold, 1978)

$\{$Liouville's integrability$\} \subset \{$separability$\} \subset \{$integrability by quadratures$\}$.

However, the three sets may be identical in a practical (but coordinate-free) sense, because even the Toda Hamiltonian and geodesic motion in the Kerr geometry are easily verified to be integrable in the sense of Liouville. These two systems are integrable because of the existence of 'hidden' symmetries and extra integrals which are not conjugate to any generalised coordinates of the configurational space (cf. section 6.4).

Thus the first five statements but (2) (certainly there are more than listed here) may be regarded as alternative definitions for integrability, and there is nothing ambiguous in them (but see eg. Wintner, 1947, P144 for disagreement). On the other hand, integrability has always been a highly difficult problem (eg. Whittaker, 1904; Leimanis & Minorsky, 1958; Broucke, 1979; Lichtenberg & Lieberman, 1983); because it is hard to decide whether a specific Hamiltonian is integrable or non-integrable. The situation here is very like that for the prime numbers. The standard definition for a prime number is that it is number which is not divisible by any number except one and itself; which can also be equivalently stated as a number which is not divisible by any number not greater than the square root of the number. The principle for choosing one from many equivalent statements as the definition is that it involves as little conditions as possible, or it is conceptually as economical as possible; whereas the remaining ones are regarded as theorems useful for different purposes (eg. Steen, 1978).

From the sixth statement above, we see that a useful method of determining integrability is to assume that the solutions are all quasi-periodic, thus all solutions may be expanded as convergent Fourier series; then substitute such series solutions into the differential equations and investigate whether any contradiction arises.

The progress in chaotic dynamics shows that the very natural definitions of Liapunov and Poincare stability, integrability and periodicity are all related to resonance, and ultimately to number theory (Whittaker, 1964; Moser, 1973). These relations were embedded in the very foundation of mechanics, but they were not uncovered until the works characterised by Poincare (1892), Birkhoff (1927), Siegel (1941), and KAM. Therefore chaotic dynamics did not go by itself beyond the framework of Newtonian mechanics, it enriched the content and displayed the underlying relations, making them more apparent. In history, many different definitions for stability have been given for various theoretical and practical purposes. Of course, great progress has been made in this way by defining stability to be adapted to the physical problems (eg. Szebehely, 1984), but it was just because of this compromise that the recognition of the most fundamental chaotic behaviour of dynamical systems was delayed.

To see the possibility of relating periodicity to number-theoretic results, let us give a very simple example. Can a function of a single real variable have more than one finite principal period? By intuition, the answer is no; but the rigorous proof which was first given by Jacobi (see eg. Forsyth, 1893) relies heavily on the properties of rational, irrational numbers and continued fractions.

In order to fix our ideas and use the modern language of Riemannian manifolds to look at integrability, let us confine ourselves to the motion of a single particle in an exterior potential field; while the dimension of the configuration space is relaxed to any finite dimension.

In this way we can give a deeper view of the involution condition in Liouville's integrability (however, since hidden symmetries and Killing tensors are not well understood, we have to confine ourselves to integrals conjugate to coordinates, obvious symmetries, or Killing vectors). It is usually understood that if the n integrals are in involution (not in involution), then their conjugate coordinates exist (not exist). However, what might be less well known is the reason for the nonexistence of such conjugate coordinates if the integrals are not in involution. This becomes obvious by utilising the concepts of Lie derivatives and Killing vectors (cf. Schutz, 1980). In this context, first integrals are made correspondent to Killing vector fields (**isometries**, symmetries); moreover, the integrals being 'in involution' simply means that the corresponding Killing vector fields commute, hence form a set of coordinate bases (cf. section 6.4). When the integrals are not in involution, the Killing fields do not commute, and therefore they do not form coordinate bases. That such independent Killing fields are not in involution do not ensure integrability is because they do not offer a one to one mapping for the Riemannian manifold, hence they are not coordinate basis fields (Schutz, 1980). The advantage of working with the Tetrad formulation based on noncommutative independent fields is discussed in Chandrasekhar (1983).

In this way, first integrals are related to more fundamental and more apparent geometric concepts, namely, symmetries and Killing vector fields. This is a step forward, however, even the notion of symmetry is not always obvious (cf. section 6.4). In addition, there are symmetries (eg. reflection, Killing tensor) that cannot be included in such description. On the other hand, although Noether's theorem establishes stronger relations, it does not give any way of uncovering all invariant properties.

To see that the Killing vector version only uncovers a subset of all symmetries, let us note that it gives a sufficient but not necessary condition for the existence of (obvious) integrals. In the previous section, we required that both T and U were independent of the generalised coordinates. However, it may happen that an invariant property exists for L, with the coordinate dependent part in T and U cancelled out. In the Toda Hamiltonian, there is no ignorable coordinate exist in physical space accounting for the additional integral; it is due to a more subtle invariant property of L or H in phase space (Henon, 1974; also Lichtenberg & Lieberman, 1983). In the Kerr space-time, the extra integral is a result of a Killing tensor of the space-time.

The difficulties and efforts made in deciding integrability and finding integrable systems can be found in Whittaker (1904), Lynden-Bell (1962) and Lichtenberg & Lieberman (1983). More recent review on the advance in this subject may be found in Yoshida (1983), Hietarinta (1987) and Ramani et al (1989).

Finally we mention two important theorems considering 'how many' systems are integrable, and what occurs if a system is not integrable. Siegel's theorem considers the space of Hamiltonians analytic in their variables: non-integrable Hamiltonians are dense in this space, whereas integrable Hamiltonians are not. Nekhoroshev's theorem leads to the fact that all non-integrable systems have a phase space that contains chaotic regions (eg. Campbell, 1989).

Hamilton-Jacobi Equation and Action-Angle Variables

As is seen, integrability of a Hamiltonian is a coordinate-free property. One of the difficulties in determining whether a system is integrable or not is because a single Hamiltonian can show various forms in different coordinates. Thus one of the efforts in finding integrable Hamiltonians is to study their possible forms in some particular coordinate systems so they can be identified. The Hamilton-Jacobi equation is one such method which identifies integrable Hamiltonians in a class of coordinate systems; in a looser but practical sense, this is often said to identify a class of integrable Hamiltonians. In this method integrability of the set of first order ODEs is made equivalent to separability of a first order PDE.

A remark may be made here on the widely accepted comment that separability is only a sufficient condition for integrability. The confusion really depends on whether one is using coordinate-dependent or coordinate-free language. It is true that an integrable Hamiltonian may always be put into a coordinate system such that it is not separable; thus separability does not identify all integrable systems. However, there always exists at least one coordinate system in which any integrable Hamiltonian is separable. Therefore, separability is equivalent to integrability. It is in this coordinate-free sense that the equivalent definitions were given for integrability. From this we see why the action-angle variables are usually the most convenient coordinates to use in obtaining the approximate series solutions for **near-integrable** systems considered in more detail in the following sections.

Because of the equivalence of time-dependent and time-independent Hamiltonians, we shall only consider here the autonomous Hamiltonian systems. Moreover, only **Hamilton's characteristic function** is included. A more complete discussion on its relation to Hamilton's principle function is given in Goldstein (1980).

In attempting to obtain a closed-form solution, a given Hamiltonian may be transformed to a new Hamiltonian by a generating function, say of S_2 type, so that more ignorable coordinates are used. The relationship between the new and old systems may be found from Table 2.1; and in general, any function $S(q, \mathscr{p})$ would generate a new canonical system.

For a Hamiltonian $H(p, q)$ completely integrable in the sense of Liouville, a generating function S (say of S_2 type without loss of generality) exists to transform the system into a new Hamiltonian $\mathcal{H}(\mathscr{p})$, in which all the generalised coordinates $\mathscr{q}$ are absent and generalised momenta $\mathscr{p}=\alpha$ are integral constants of motion. The purpose of the Hamilton-Jacobi equation formulation is to find the generating function so the transformation can be carried out.

Suppose $H=\mathcal{H}=\mathscr{p}_1=\alpha_1$, then the generating function must satisfy the Hamilton-Jacobi equation, $H(\partial S(q, \mathscr{p}) /\partial q, q)=\alpha_1$, where the $\mathscr{p}$'s are to be regarded as parameters. In this way the effort of solving the original system is changed to finding a complete solution to the Hamilton-Jacobi equation, $S(q, \alpha, c)$, which is called the Hamilton characteristic function. The constant c is a pure additive constant, which is not important to the transformation.

In practice the PDE is equally hard to solve as the original ODE unless the H-J equation can be separated completely in the form, $H(p,q)=\Sigma H_i(p_i, q_i)$. The non-separable feature may be a reflection of nonintegrability or that an integrable Hamiltonian is put in a badly chosen set of coordinates. Therefore the advantage of the H-J equation is that integrable systems may be identified in a less restricted class of coordinates. In order that the system may be solved, the original system must be put into the appropriate coordinate system; and there is no general method with which to make the choice.

When the H-J equation is separable, the actual procedure of finding the transformation may be found from Goldstein (1950) or Lichtenberg & Lieberman (1983).

For integrable Hamiltonians, any function of the conserved momenta $\mathscr{p}$ may also be taken as the new generalised momenta. A particularly important class of Liouville integrable systems is one which possesses compact phase space. For such systems the action-angle variables are a very useful class of canonical variables which are defined by

$$J_i = \frac{1}{2\pi}\oint \mathcal{p}_i d\mathcal{q}_i \quad ; \quad \theta_i = \omega_i t + \beta_i \quad , \quad \omega_i = \frac{\partial \mathcal{H}}{\partial J_i} \, .$$

For nontrivial applications of such variables the books by Goldstein (1950) and Lichtenberg & Lieberman (1983) must be consulted.

Examples of Integrable and Nonintegrable Systems

It is instructive to summarise some of the integrable and nonintegrable systems which occupy some position in chaotic dynamics and have some significant implication to the later work of the present thesis.

(1). The Toda Lattice and Henon-Heiles System

The Toda lattice is a one dimensional lattice in which the repulsive force between neighbouring particles moving on a ring is an exponentially decreasing function of their angular distances. This is an integrable Hamiltonian system, for which Lichtenberg & Lieberman (1983) give more details and references. Here we only quote the relevant part of the problem.

After some transformations and use of the simple integrals, the original 3-particle Toda lattice problem is reduced to the **Toda** Hamiltonian $\mathcal{H}$ with two degrees of freedom, which possesses the first integral, $\mathcal{I}$, nonlinear in the momenta, viz.

$$\mathcal{H} = T(p_x, p_y) + U(x, y) = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{24}\{ e^{2y+2\sqrt{3}x} + e^{2y-2\sqrt{3}x} + e^{-4y} \} - \frac{1}{8}$$

$$\mathcal{I} = 8p_x(p_x^2 - 3p_y^2) + (p_x - \sqrt{3}\,p_y)e^{2y+2\sqrt{3}x} + (p_x + \sqrt{3}\,p_y)e^{2y-2\sqrt{3}x} - 2p_x e^{-4y}.$$

This Hamiltonian is integrable. However, there is no obvious and simple symmetry (in physical space) corresponding to the first integral $\mathcal{I}$.

If the above Toda Hamiltonian is expanded in a Taylor series with respect to x and y, and cubic terms are retained, we obtain the non-integrable **Henon & Heiles** (1963) Hamiltonian system,

$$H = T(p_x, p_y) + U(x, y) = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(x^2 + y^2) + x^2y - \frac{1}{3}y^3$$

$$\frac{d^2x}{dt^2} = -x - 2xy \quad , \quad \frac{d^2y}{dt^2} = -x^2 + y^2 - y \quad .$$

In fact, truncated systems of the Toda Hamiltonian to any order is not integrable. However, the following Hamiltonian system is only slightly different from the Henon & Heiles system, but it is integrable,

$$H = T(p_x, p_y) + U(x, y) = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(x^2 + y^2) + x^2y + \frac{1}{3}y^3$$

$$\frac{d^2x}{dt^2} = -x - 2xy \quad , \quad \frac{d^2y}{dt^2} = -x^2 - y^2 - y \quad .$$

The integrability of the last system is obvious by a change of variables, namely, X=x+y and Y=x-y (eg. Cooper, 1989, P244). In fact, this example belongs to a whole class of integrable systems which satisfy the Painleve property (eg. Lichtenberg & Lieberman, 1983, P40; Ramani et al, 1989).

At this point, we shall discuss some relationships between the integrability of a system and its truncated systems, which might be helpful in understanding some of the difficulties encountered in the later chapters on relativity and post-Newtonian approximations.

One of the reasons for studying truncated systems is because of the complexity of the original system and the belief that a truncated system is usually simpler, hence a nonsolvable system may be solved by such an approximation method. Examples are numerous; to list only two of immediate interest to the present thesis, stability of equilibrium points is usually studied to first or second order; Newtonian mechanics and post-Newtonian approximation are the lowest order truncation of the full relativistic theory.

However, what we learn from the the Toda lattice example is that the relation between original and truncated systems is not so simple. Firstly, a relation found from the original system may be lost in an approximation procedure; because in a complicated relation, the quantities on the two sides of '=' may ultimately be regarded as infinite series, which are not necessarily based on the same constructions (lim A = lim B does not guarantee that A = B).

Secondly, qualitative differences may be produced if a truncation is made with respect to more than one variable. To fix our ideas, let us again consider the n-dimensional motion of a single particle in a fixed exterior potential field. If the potential is expanded and kept up to quadratic terms in $q=(q_1, \dots , q_n)$, then it is always possible to find a transformation such that the truncated potential becomes a standard quadratic form which is separable. However, even for terms one order higher, the standard cubic forms are not always separable (Poston & Stewart, chapter 2); therefore there is no guarantee for the integrability of a truncated cubic potential, nor for higher order terms. This justifies the integrability of the Toda Hamiltonian and the non-integrability of the Henon-Heiles Hamiltonian.

We conclude from the above examples that although a truncated system may be simpler than the original system, this is not always true. However, this is not to deny the applicability of the approximation methods. These may be of some importance in two respects in the later chapters. Firstly, the difficulty encountered in constructing the best inequalities for the post-Newtonian many body problem may not be intrinsic to the full relativistic problem, but rather a feature of the particular truncation. Secondly, completely integrable systems in the full relativistic case, may become non-integrable and chaotic if the systems are approximated to some nonlinear orders. This is a method of studying relativistic chaos (see also Chandrasekhar, 1989; Contopoulos, 1990). This point shall be discussed further in connection with quantization in the next chapter.

(2). Few-Body Problems

Integrable and non-integrable few-body problems (Whittaker, 1904; Siegel & Moser, 1971) are discussed briefly here because of their importance in relation to quantum chaology and chaos in general relativity.

The motion of a single particle in a fixed central field is integrable in both Newtonian mechanics and general relativity (Schwarzschild geometry). It is not integrable if the field is not central (not static, spherically symmetric). The Newtonian motion is Keplerian if and only if the field is an inverse square law; any deviation from such law results in pericentre precession due to quasi-periodic motion or non-integrable motion (Goldstein, 1980).

The Newtonian motion of two bodies interacting with radial forces is reducible to the motion of a particle in a central field, hence integrable. Usually the problem is not integrable if the force is not radial. The motion of two bodies with at least one extended body of arbitrary shape is not integrable in Newtonian mechanics, nor in relativity.

The classical two-centre problem is integrable, whereas the nonaligned many-centre problem is not. None of such problems is integrable in general relativity. As an extreme case of the two-centre problem, the motion of one mass in the field of a point mass and a uniform field is integrable.

The motion of two bodies in an arbitrary exterior field is not integrable in Newtonian mechanics. Nevertheless, the problem is integrable if the exterior field is uniform; but care must be taken that this problem can be reduced to the two-body problem, hence it is simpler than when one of the masses is fixed.

The motion of two opposite charges in a uniform electric field is integrable and can be reduced to the corresponding gravitational problem with one body fixed (Stark effect). The motion of two bodies with opposite charges in a uniform magnetic field is integrable (Zeeman's effect). These two models are of importance in quantum mechanics (Born, 1927; Berry, 1978). Irregular spectra are observed corresponding to them (Hasegawa et al, 1989).

The restricted three-body problem is not integrable (Henon, 1965-1970); such a problem is not formulated in relativity. Hill's limiting problem is not integrable.

Many-body problems are not integrable in Newtonian mechanics, nor in general relativity.

(3). Others

The harmonic oscillator is often quoted as an example of an integrable system for which a closed-form solution may be obtained explicitly. The simple pendulum and the two-body problem are examples of integrable cases, but the solutions can only be obtained in an implicit infinite series of time (Lichtenberg & Lieberman, 1983; Stiefel & Scheifele, 1971).

The ideal resonance problem is an example of an integrable system with a small parameter, which may be written in action-angle variables as $H = H_0(I) - \varepsilon A(I) \cos\theta$. This has been used to investigate perturbation theory, resonance and small divisors (eg. Garfinkel, 1966; Ferraz-Mello, 1985).

The **Sitnikov** motion in the elliptic restricted 3-body problem is not integrable (eg. Moser, 1973). This is the motion of the infinitesimal mass on the line perpendicular to the plane of motion of the primaries and going through their centre of mass, where the primary masses are of the same size.

An interesting example of non-integrable systems is

$$H = \frac{1}{2}(p_x^2 + p_y^2) + \alpha x y^2$$

for which all solutions escape. Thus no quasi-periodic solution exists (Broucke, 1979).

2.4 Solution Method II - Perturbation Theory

As is shown in the preceding sections, in particular by Siegel's theorem, non-integrability is the generic case for Hamiltonian systems, for such systems it is impossible to obtain closed analytic solutions. This is not simply because ingenuity fails, but because the notion of closed-form functions is too limited to accommodate the solutions to the variety of differential solutions encountered in practice. Although, under such a situation, iterative methods can be invoked, the solutions so found are often too complicated to display clearly the principal features of the solution. Sometimes this is also true, even if an analytic solution can be found; this is particularly true of implicit solutions and of solutions which are in the form of integrals or infinite series. Therefore, qualitative study must be pursued, whereby important characteristics of the solutions can be deduced without actually solving the differential equations. However, one needs to keep in mind that some differential equations do not even admit a complete qualitative description.

In this section we discuss the classical and secular perturbation theories for solving Hamiltonian systems. They are among the most important methods, which do not only offer solutions valid for finite time scale but also reflect the qualitative features.

Classical Perturbation Theory and Lie Transformation Methods

It is instructive to first look at the effects of resonances and small divisors in simple linear and nonlinear systems of the form $x'' = f(x, x') + g(t)$, where $g(t)$ is a periodic function of t with principal frequency Ω. When there is no damping and $f(x)=\omega^2 x$, then the equation may be solved in Fourier series, which has a kind of blow-up oscillation whenever there is a resonance $\omega=k\Omega$. Thus linear response to a driving frequency simply results in divergence. However, when the function $f(x, x')$ is nonlinear the system is saved from an extreme blow-up instability, and a completely new regime of responses such as jump catastrophe and limit cycles is produced. This can be observed from the well-known **Duffing's** equation with $f(x, x')=kx'+\alpha x+\beta x^3$, and **van der Pol** equation with a small parameter $f(x, x')=\varepsilon(x^2 -1)x'+x$ (see Jordan & Smith, 1990).

From these examples we see that linearity produces relatively uninteresting extreme responses to driving: either simple stability or simple divergence; while nonlinearity produces more complicated responses which are more interesting.

Now let us turn to the more relevant almost integrable Hamiltonian systems. Such Hamiltonians are usually produced in several ways: (1). the Hamiltonian is analytic in a small parameter involved, with respect to which the Hamiltonian may be expanded in a Taylor series; (2). the Hamiltonian is analytic in some variables, and may be expanded with respect to such variables in the neighbourhood of some point or trajectory; (3). the last two cases may be truncated to certain order; (4). a finite power series of a small parameter arising naturally or by assumption.

Attention must be paid to the fact that the usually phrased **order** of a perturbation method has nothing to do with the order where the truncation is made in the above four cases; it is the order of the truncation in the final solution. One is usually in a class of almost integrable Hamiltonians which may be written in the action-angle, $(\mathbf{J}, \theta)$, form of the integrable part, H_0 ,viz.

$$H(\mathbf{J}, \theta, \varepsilon) = H_0(\mathbf{J}) + \varepsilon H_1(\mathbf{J}, \theta),$$

where the perturbation $H_1(\mathbf{J}, \theta)$ is assumed to be a multiple periodic function in the angle variables; and the solution to H_0 is $\mathbf{J}$ = Const, $\theta = \omega\, t$ + Const, $\omega = \partial H_0 / \partial \mathbf{J}$.

The classical perturbation theory solves the complete Hamiltonian system by seeking a transformation to the new variables $(\mathfrak{p}, \mathfrak{q})$ for which the new Hamiltonian is a function of $\mathfrak{p}$ alone (certainly also of ε). Before we actually turn to details, let us observe what we may expect. If a well behaved transformation could be found for the above purpose, this would imply the integrability of the Hamiltonian $H(\mathbf{J}, \theta, \varepsilon)$. Because in general this is not the case, some contradiction must arise at the end. In fact the classical perturbation theory is based on formal series calculus, so that the transformation is found by truncating such **formal** power series in the small parameter ε. If convergence can be established for the infinite series, then the solution has been found; if, however, the series diverges, we must conclude that the assumption is probably false.

Conversely, if the Hamiltonian is not integrable, this formal procedure must produce some kind of divergence. In fact, it has been realised since Poincare that it is not simply a complete convergence nor complete divergence -- what happens depends much on initial conditions, and the results of number theory is of great relevance. Poincare's work was furthered by Birkhoff (1927); since then, great progress has been made in two apparently opposite directions: the conditions for divergence were sharpened by Siegel (1945), whereas the existence of quasi-periodic solution and conditions for convergent series were sharpened by Kolmogorov (1954), Arnold (1963) and Moser (1962). Now a comprehensive view has been built up for the topological structure of the phase space, in spite of its incompleteness.

Although it is now clear that the divergence of the formal transformation series is an intrinsic problem for aperiodic solutions, the classical perturbation theory is still of importance in practice, because it has been shown by Poincare that such a series is asymptotic (semi-convergent). Therefore the series truncated up to certain order can approximate the real solution. Another support to this theory is from the KAM theorem, which shows that the measure of quasi-periodic solutions is positive.

Now let us try to infer these points by a closer look at the classical perturbation theory. In the classical method of this theory the transformation is represented by a mixed variable generating function independent of time, $S(\boldsymbol{p}, \theta, \varepsilon)$, ($S_2$ type of Table 2.1) expanded in a power series of ε,

$$S = <\boldsymbol{p}, \theta > + \varepsilon\, S_1(\boldsymbol{p}, \theta) + \varepsilon^2\, S_2(\boldsymbol{p}, \theta) + \ldots$$

The transformation and new Hamiltonian are to be obtained according to Table 2.1. In order to express the new Hamiltonian in the new variables, half of the above transformation relations must be inverted to find the old variables explicitly in terms of the new ones. After these procedures we come to the following new Hamiltonian (eg. Lichtenberg & Lieberman, 1983),

$$\mathcal{K} = \mathcal{K}_0\,(\boldsymbol{p}) + \varepsilon\, \mathcal{K}_1(\boldsymbol{p}, \boldsymbol{q}) + \varepsilon^2\, \mathcal{K}_2(\boldsymbol{p}, \boldsymbol{q}) + \ldots$$

where

$$\mathcal{K}_0\,(\boldsymbol{p}) = H_0(J)\,,\;\; \mathcal{K}_1(\boldsymbol{p}, \boldsymbol{q}) = \omega\,(\boldsymbol{p})\, \partial S_1(\boldsymbol{p}, \boldsymbol{q})\,/\partial \boldsymbol{q} + H_1(\boldsymbol{p}, \boldsymbol{q})\,,\;\; \ldots$$

In order to obtain a new Hamiltonian of the form $\mathcal{H}(\boldsymbol{p}, \varepsilon)$, we simply average the S_i-independent terms of $\mathcal{K}_i$ over $\boldsymbol{q}$, and denote the averaged quantity by $\mathcal{H}_i(\boldsymbol{p})$. Then we sort for S_i to cancel all the remaining part of $\mathcal{K}_i$. Thus finally we come to the required new Hamiltonian,

$$\mathcal{H} = \mathcal{H}_0\,(\boldsymbol{p}) + \varepsilon\, \mathcal{H}_1(\boldsymbol{p}) + \varepsilon^2\, \mathcal{H}_2(\boldsymbol{p}) + \ldots$$

$$= H_0\,(\boldsymbol{p}) + \varepsilon < H_1(\,\boldsymbol{p}, \boldsymbol{q}) > + \ldots$$

where the explicit expression in terms of the old Hamiltonian is only given up to the first order. (In this section we denote the average of a function by <...>, which is similar to the inner product of two vectors but this may be understood from its content.)

The generating function may be obtained in multiple Fourier series, for example, we have the following for the first order perturbation,

$$\omega(\boldsymbol{p})\, \partial S_1(\boldsymbol{p}, \boldsymbol{q}) / \partial \boldsymbol{q} = H_1(\boldsymbol{p}, \boldsymbol{q}) - < H_1(\boldsymbol{p}, \boldsymbol{q}) > = \Sigma_{\mathbf{k} \neq 0} H_{1\mathbf{k}}(\boldsymbol{p})\, e^{\, i<k,q>}$$

$$=> S_1(\boldsymbol{p}, \theta) = \Sigma_{\mathbf{k} \neq 0} \{ H_{1\mathbf{k}}(\boldsymbol{p})\, e^{\, i<k,q>} / <\mathbf{k}, \omega(\boldsymbol{p})> \} \text{ , if } <\mathbf{k}, \omega(\boldsymbol{p})> \neq 0,$$

where $\mathbf{k} = (k_1, \dots , k_n)$ is a set of any integers.

From this formula let us observe the possibility of convergence and divergence of the series for S_1. The formula is only valid for nonresonant (incommensurable) frequency vectors; when this happens the so called secular perturbation theory is needed. By secular perturbation, one performs a separate transformation to eliminate one of the original actions of **J**, followed by the same procedures given here.

When the frequencies are not resonant, we have the small divisor problem, namely, the denominator $<\mathbf{k}, \omega(\boldsymbol{p})>$ can be an arbitrarily small number. Therefore one may expect a general divergence. This, however, is not the case; it turns out by a use of number theory that divergence is exceptional. To establish convergence one needs to estimate the size of the combination of the Fourier coefficients and small divisors. It can be shown that the coefficient of a multiple Fourier expansion has a property similar to that of a single variable, namely, it decreases exponentially with $k = \Sigma |k_i|$; moreover, it decreases more rapidly for smoother functions. On the other hand, some frequencies have a property similar to that of algebraic irrational numbers: their small divisors can be bounded from below, that is,

$<\mathbf{k}, \omega> \;\geq C(\omega)/k^{n+1}$, for positive constant $C(\omega)$ and **all** integer vector **k**.

In addition such frequencies form a set of positive measure, whereas only a zero measure set of frequencies does not satisfy this inequality (like transcendental numbers).

For sufficiently smooth perturbation H_1, convergence can be established for frequencies satisfying the above number-theoretic inequality, although divergence always occurs for some frequencies (Brouwer & Clemence, 1961).

When higher order perturbations are calculated, small divisors (and resonances) are also involved and a similar feature happens for divergence and convergence of each S_i. It was Siegel who proved that only for a class of measure zero analytic Hamiltonians, convergence can be established globally, which implies that divergence always exists for positive measure Hamiltonians. On the other hand, the KAM theorem (next section) shows that for an almost integrable Hamiltonian, convergence of the infinite power series for the generating function is the generic case.

Now let us see why the order of the original Hamiltonian is irrelevant to the order of the perturbation theory, which is the order of the power series in the final Hamiltonian.

Higher order terms in the original Hamiltonian may be included in H_1. On the other hand, the new Hamiltonian may be calculated up to any order in the small parameter even if the original Hamiltonian only involves the small parameter to the first order. The order of the perturbation arose completely from the assumption that the transformation and the new Hamiltonian is analytic in the small parameter.

It is in order to make a comment on the averaging method often used in solving ODEs. An averaging method simply averages the low order perturbations in the original Hamiltonian over fast variables (see Lidov, 1963; Aksenov, 1979; Arnold, 1983); however, there is no theorem to assure that the averaged system would agree with the original system. In fact the above classical perturbation method justifies the applicability of the averaging method, if the transformation can be developed in convergent series, and the averaging is performed on the new Hamiltonian system. The averaged part and the remaining fast part play different roles. Even if such a transformation does not exist, the method is still of practical value because of the semi-convergent feature of the procedure.

Now we turn to the secular perturbation theory dealing with commensurable cases. As was remarked by Poincare (1892), secular terms due to exact resonance are not an intrinsic problem of the dynamical system (in contrast the small divisor problem is intrinsic). Thus a secular perturbation procedure can always be carried out to avoid such a difficulty. In fact, a resonance appearing in a specific order leads to an extra (isolating) integral to that order of perturbation. The procedure needed for such case is to perform an additional transformation to eliminate the resonance by eliminating one of the actions (frequencies); then put the new system into action-angle form and carry out the above standard perturbation procedure. In principle the procedure is the same for resonances which appear in any order perturbation.

Let us take the simplest case as an example, namely, resonance in first order perturbation; in addition, the degree of freedom of the system is assumed to be two. If a resonance exists between the unperturbed frequencies, viz. $<(\omega_1,\omega_2),(k_1,-k_2)>=0$, we choose the generating function $S_2 = (k_1\theta_1 - k_2\theta_2)\mathscr{p}_1 + \theta_2\mathscr{p}_2$, which defines a canonical transformation

$$J_1 = k_1\mathscr{p}_1 \quad , \quad J_2 = \mathscr{p}_2 - k_2\mathscr{p}_1 \quad , \quad \mathscr{q}_1 = k_1\theta_1 - k_2\theta_2 \quad , \quad \mathscr{q}_2 = \theta_2 ,$$

where θ_2 is assumed to be the slower angle variable of the two. This transformation puts the observer in a rotating frame in which $\mathscr{q}_1$ measures the slow deviation from resonance due to the perturbation. The procedure followed is to average over $\mathscr{q}_2$ and obtain the corrected integral $\wp_2 = J_2 + k_2 J_1 / k_1$. For more detailed discussion on this, removal of higher order resonances, intrinsic and accidental degeneracy, generic separatrix motion, and islands see Lichtenberg & Lieberman (1983).

The above mixed-variable transformation method is called von Zeipel's method. However, there are serious inconveniences when higher order perturbations are actually calculated. The so called Lie transform method has been developed to offer an easier algorithm to carry out higher order calculations. For more details of this method see Deprit (1969a, 1969b), Kamel (1969), Campbell & Jefferys (1970), Mersman (1971) and Choi & Tapley (1973).

Finally, it is worth noting that the same convergence problem equally occurs in the Lie transform algorithm. Like the normal form notion, the perturbation method is a formal procedure, which works so long as there are no exact resonances; and an exact resonance can also be treated after a separate consideration. The questionable point of perturbation theory is the convergence of the series for incommensurable case.

2.5 Solution Method III - Geometrical Methods and KAM Theorem

The geometrical method stimulated by Poincare in solving dynamical systems has been very fruitful. The most important ones are the Poincare surface of section and Poincare-Birkhoff fixed point theorem. The KAM theorem is also included here because a combination of the KAM invariant tori and Poincare-Birkhoff fixed point theorem can give a very comprehensive description on the phase space structure of almost integrable systems; the importance of quasi-periodic solution to chaotic solution is also manifest from this. The converse KAM method is also briefly introduced. We shall begin the section by stating these important theorems.

Poincare's Recurrence Theorem

The theorem says that if T is a continuous 1-1 volume-preserving mapping which maps a bounded region onto itself, then any point in the region returns arbitrarily close to the point (namely, recurrence) for sufficiently many iteration of the mapping (Poincare, 1892). For an outline of the proof see eg. Arnold (1978).

This is actually the continuous version of the returning property of a finite denumerable cyclic group. On more practical grounds, the theorem is related to Poisson stability defined for an N-body system, which requires that, in addition to the distance between any two particles and its reciprocal being finite, the system repasses an infinite number of times to the initial situation.

Poincare Mapping and Surface of Section

This method was initiated by Poincare to determine periodic solutions. By this method, instead of the solution as a t-curve in phase space, one studies the relation (continuous Poincare mapping T) between the neighbouring points at t_1 and their corresponding points at t_2; and in doing this one can take advantage of first integrals (Poincare 1892; Siegel & Moser, 1971). When t_2-t_1 equals the period of a periodic solution, the periodic solution becomes a fixed point of the mapping in the neighbourhood of the point. Because of this relationship between ODEs and difference equations (transformations, mappings), the study of the behaviour of the latter becomes important even if it is not directly related to physically interesting ODEs.

The method is especially useful when the mapping is two dimensional (surface of section) and the mapping is bounded, whereby a quasi-periodic solution becomes a smooth invariant curve. In this way the search for a quasi-periodic solution is thereby reduced to determining the fixed point and invariant curve. The method is fruitful when at least one fixed point can be found followed by the determination of the higher order fixed points and invariant curves in the neighbourhood of it. This method was used by Poincare and furthered by Birkhoff (1917, 1922, 1927). The surface of section method is also the easiest way to display chaotic motion.

Because of the simplicity of the transformation method and since it can speed up the standard numerical integration of ODEs, it will be advantageous to obtain a mapping for a canonical system. This procedure and its conversion is described in Lichtenberg & Lieberman (1983).

Fixed Point Theorems of Poincare and Birkhoff

There are many theorems on the existence of periodic solutions and fixed points (see Szebehely, 1967), among these are a set of theorems conjectured by Poincare which were proved and extended by Birkhoff (1913, 1926). We shall only state them here; the original works may be referred to for the proof and more details. Alternatively, the proof can also be found in Birkhoff (1927) and Siegel & Moser (1971).

Poincare's fixed point theorem was conjectured by Poincare and is of interest for the restricted 3-body problem. Because this is the work Poincare regarded as of highest importance, although he was not able to prove it before his death in 1912, the theorem is also called Poincare's last geometric theorem. The characteristic style of Poincare's work is obviously seen again in this theorem: it is not so sharp for application, nevertheless, it shows the possibility of the existence of periodic solutions in the restricted 3-body problem. The theorem is geometrically beautiful although it may seem that the conditions are arbitrary.

Given a ring $0 < a < r < b$ in the (r, θ) plane (r, θ being polar coordinates) and a one-to-one continuous area-preserving mapping T of the ring onto itself, which maps the points on r=a and r=b in different directions, the theorem says that there exist at least two points of the ring invariant under T.

The desired proof for the above theorem was given shortly after Poincare's death by Birkhoff (1913). In fact the number of fixed points in Poincare's theorem was improved to even numbers greater than two. The number of fixed points is related to the rotational number of the mapping which we shall give in the following. This theorem was given the name the Poincare-Birkhoff fixed point theorem.

The fixed point theorem, extended by Birkhoff, establishes the existence of infinitely many higher order fixed points (or periodic solutions) in the vicinity of an elliptic fixed point (or stable periodic solution). Obviously the statement is stronger and the conditions are relaxed in favour of practical application.

It is useful to note that these theorems have not been successfully extended to higher dimensional mappings, as is remarked in Siegel & Moser (1971).

However, for systems reducible to a two dimensional mapping, a combination of these theorems with the KAM theorem presents a comprehensive qualitative description of all possible types of solutions and the interrelation of them. We shall see that the KAM invariant tori can in fact act as the boundary of the rings in Poincare's fixed point theorem.

KAM Theorem and Arnold Diffusion

The intrinsic small divisor problem in an almost integrable Hamiltonian system may cause divergence of various infinite series expression of solutions to such systems (Poincare, 1892). Moreover, Siegel's theorem says that most Hamiltonian systems are not integrable. These may be interpreted as that most quasi-periodic solutions are destroyed by small perturbations. This, however, is not the case. It was conjectured by Kolmogorov (1954), proved by Arnold (1963) and modified by Moser (1962) that for sufficiently small perturbations, almost all quasi-periodic solutions are preserved. Although the conditions are far from being of practical value, the theorem offers qualitatively significant guidance towards more realistic estimation of the existence of such solutions and stability of equilibrium and periodic solutions. The conditions to be satisfied are stated as follows (see Lichtenberg & Lieberman, 1983; Siegel & Moser, 1971; Moser, 1973), using the notations of the last section.

(1). the linear independence (sufficient nonlinearity) of the frequencies $\omega(\mathbf{J})$ over integers, namely, $< \omega, \mathbf{k} > \neq 0$, over some domain of $\mathbf{J}$;

(2). sufficient smoothness condition on the perturbation (sufficient number of continuous derivatives of H_1);

(3). initial conditions sufficiently far from resonance to satisfy $|< \omega, \mathbf{k} >| \geq c/k^{-\tau}$ for all $\mathbf{k}$, where τ is dependent on the number of degrees of freedom and smoothness of H_1, and c is dependent on ε, the magnitude of H_1, and the nonlinearity G of H_0.

Although the KAM theorem is proved by a different method, the importance of the above conditions can be observed from the perturbation theory of the last section. The second and the third conditions are important to establish convergence of the partial Fourier series and the complete power series; whereas the first condition precludes the resonant case which needs separate discussion. The first condition can also be interpreted as sufficient nonlinearity, as is evident if one observes that the frequencies are the derivatives of the integrable Hamiltonian. For further explanation of these conditions and the concepts of accidental and intrinsic degeneracy etc. the references should be consulted.

Arnold's smoothness condition required that the Hamiltonian should be analytic in a strip defined by the action-angle variables; whereas Moser's original condition required it to have 333 derivatives. Henon (1966) carried out the computations necessary to determine the size of the perturbation and showed that for a system of two degrees of freedom, Arnold's form of the theorem required it to be less than 10^{-333} and Moser's less than 10^{-48}. This estimation is practically useless. Moreover, the incommensurability condition was later weakened to the condition that there be no resonances of order k=3 and k=4 (Moser, 1973).

The structure of quasi-periodic and chaotic motion in the phase space of almost integrable systems is often made analogous to the structure of rational and irrational numbers on the real line. However, care must be paid not to read too much into such analogy. In fact, they are similar in the sense that the two kinds of properties are mixed in so complicated a way that neither covers a finite open domain no matter how small the domain is. The difference lies in that rational numbers have measure zero, whereas both chaotic and quasi-periodic solutions have positive measure in phase space.

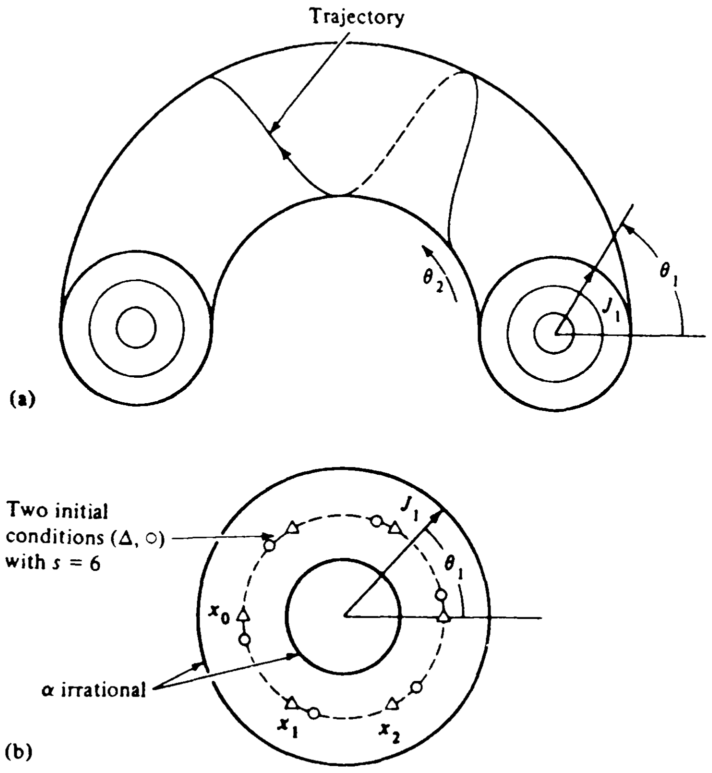


Figure 2.1 Motion of a phase space point for an integrable system with two degrees of freedom. (a). The motion lies on a torus J_1=const, J_2=const.

(b). Illustrating trajectory intersections with a surface of section θ_2=const after a large number of such intersections.

It is instructive to apply both the KAM theorem and Poincare's fixed point theorem to almost integrable autonomous Hamiltonians with two degrees of freedom and almost integrable area-preserving mappings to infer the generic behaviour of such systems, where in the former case the Hamiltonian is the only global integral and in the latter no global integral exists. The importance of such low dimensional systems lies in that an autonomous Hamiltonian of one degree of freedom is always integrable by quadrature; whereas for systems with degrees of freedom higher than two and high dimensional surface of section, fixed point theorems have not been established. Theorems like these may not exist for higher dimensional space, which may in turn be related with the so called Arnold diffusion in higher dimensional space.

For a Hamiltonian with n degrees of freedom, the KAM invariant hypersurfaces are n dimensional, whereas the solutions are all on a family of 2n-1 dimensional energy integral manifolds. From topological study it is evident that a set of n dimensional hypersurfaces can divide an n+1 dimensional space into separated bounded (n+1 dimensional) regions. Solving the simple equation n+1=2n-1, we see that complete isolation of the phase space happens only for a Hamiltonian with two degrees of freedom and two dimensional area-preserving mappings, which cannot be further separated. For higher dimensional systems, no matter how small the perturbation is, chaotic regions are not isolated by KAM tori, thus the so called **Arnold diffusion** throughout the phase space occurs.

Using the same notation as in the last section, and assuming the system's degree of freedom is two, we consider the system in action-angle space of the integrable part. The motion for a given energy of the unperturbed part is on a set of tori (J_1, θ_1; J_2, θ_2), where the two actions are related through the equation $H_0(J_1 , J_2)$ = const (Fig. 2.1.a). When the actions are such that the two frequencies are in resonance, the trajectories are closed, and the motion periodic; whereas when the trajectories are not in exact resonance the trajectories are not closed and repass the vicinity of every point on the torus infinitely many times.

If we choose the section of the torus after every 2π evolution in the slow angle θ_2, then the surface of section is characterised by the J_1-θ_1 (Fig. 2.1.b). If the unperturbed Hamiltonian is not intrinsically degenerate, that is, $H_0 \neq H_0(<\mathbf{C}, \mathbf{J}>)$, where **C** is a set of real constant, the surface of section of the unperturbed Hamiltonian is usually a set of smooth curves and a set of higher order fixed points, which are mixed in a way similar

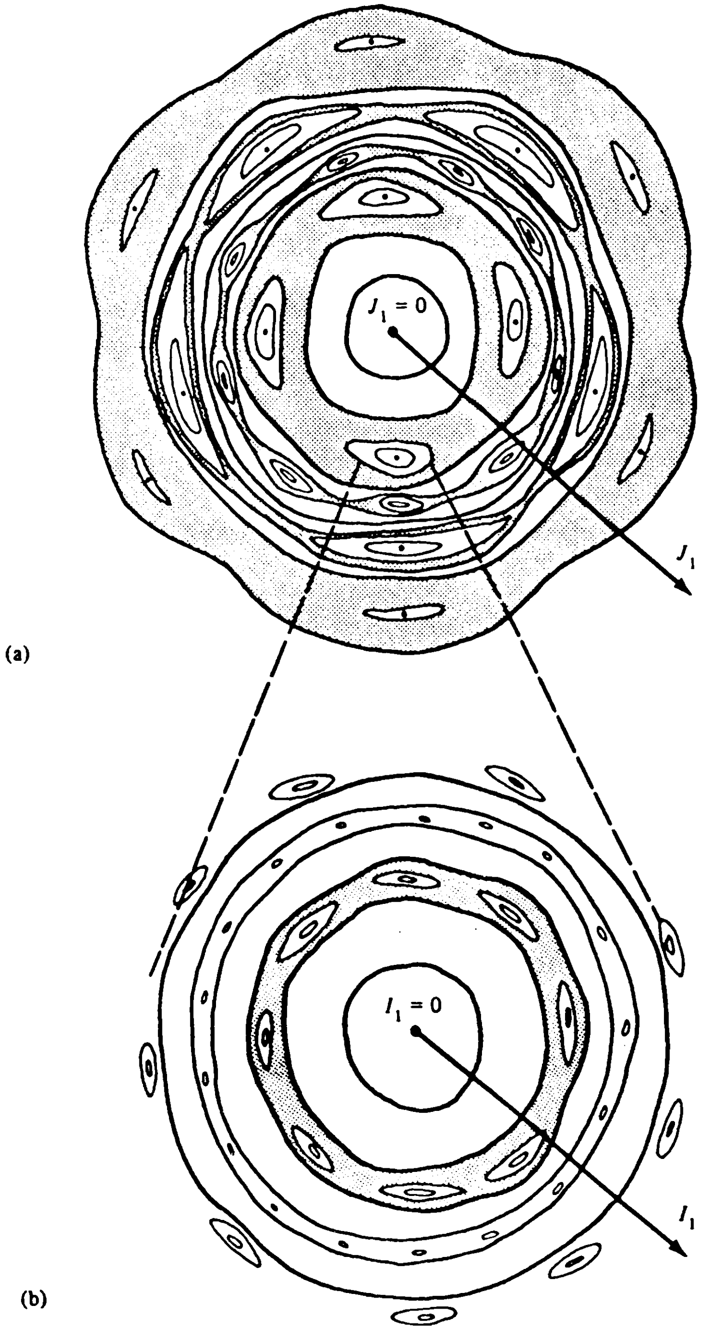


Figure 2.2 Regular and irregular trajectories for a Hamiltonian with relatively large perturbation (a). near the primary fixed point; (b). expanded (and circularized) scale near a second-order fixed point (from Lichtenberg & Lieberman, 1983).

to rational and irrational numbers. Defining the **rotation number** as $\alpha=\omega_1/\omega_2$, which has the meaning that corresponds to a 2π increase in θ_2 there is a $2\pi\alpha$ increase in θ_1, then the above process is described on the J_1-θ_1 surface of section by the following **twist mapping**,

$$\begin{cases} J_{n+1}=J_n \\ \theta_{n+1}=\theta_n+2\pi\alpha(J_{n+1}) \end{cases} \quad \text{or} \quad \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix}=\begin{bmatrix} \cos(2\pi\alpha) & -\sin(2\pi\alpha) \\ \sin(2\pi\alpha) & \cos(2\pi\alpha) \end{bmatrix}\begin{bmatrix} x_n \\ y_n \end{bmatrix},$$

which is area-preserving. The fixed points of the mapping correspond to rational α, and invariant curves correspond to irrational α. For example, if $\alpha=\omega_1/\omega_2=k_2/k_1$ with k_2 and k_1 relatively prime integers, the periodic trajectories close after k_2 revolutions in θ_1 and k_1 revolutions in θ_2. On the J_1-θ_1 surface of section, this produces a set of k_1 fixed points (with period k_1).

When the system is perturbed slightly, then according to the KAM theorem, most of the tori supporting incommensurable rotation numbers, ie. irrational α, are preserved as invariant tori but slightly distorted to support the slightly distorted quasi-periodic solutions (Fig. 2.2.a); whereas tori supporting the periodic solutions of rational α and those supporting quasi-periodic solution with not sufficiently irrational α break up into chains of islands and chaotic regions. The is why chaotic dynamics is also called **resonant dynamics**. The latter is not directly given by the KAM theorem, in fact this is also where the KAM theorem may be sharpened. This, however, may be explained by a combination of the KAM theorem and the fixed point theorems of Poincare and Birkhoff; or alternatively by the still growing converse KAM theory (eg. MacKay et al, 1989), which attempts to establish the nonexistence condition of invariant tori accommodating quasi-periodic solutions.

For such a slightly perturbed system, the surface of section may be represented by certain perturbed twist mapping, whose generic behaviour is not lost by the following simpler **radial twist mapping** and **standard mapping**, respectively written in the form

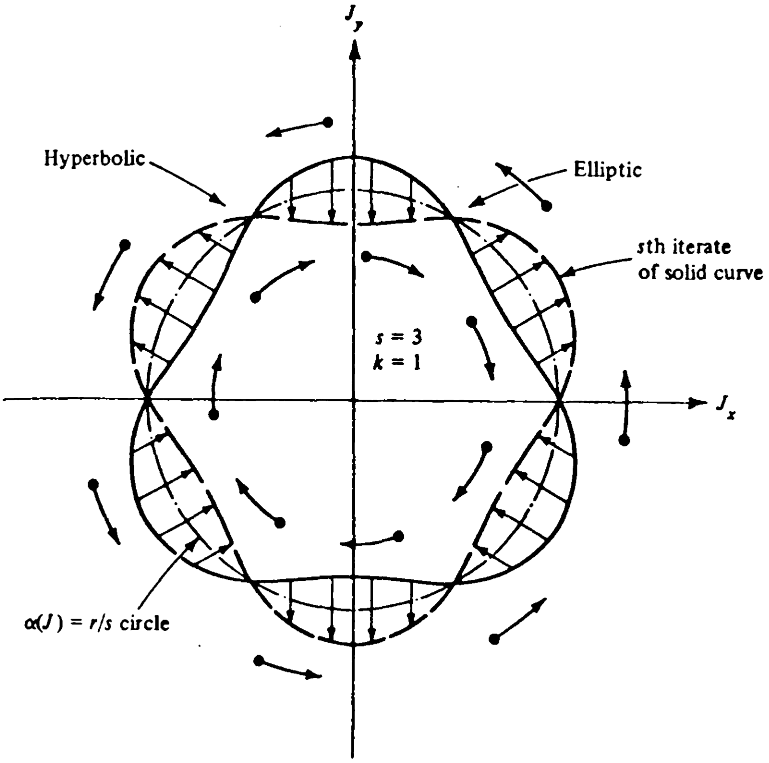


Figure 2.3 Illustrating the Poincare-Birkhoff theorem that some fixed points are preserved in a small perturbation. The intersections of the heavy solid and dashed curves are the preserved fixed points (from Lichtenberg & Lieberman, 1983).

$$\begin{cases} J_{n+1} = J_n + \varepsilon f(\theta_n) \\ \theta_{n+1} = \theta_n + 2\pi\alpha(J_{n+1}) \end{cases} , \qquad \begin{cases} J_{n+1} = J_n + \beta \sin\theta_n \\ \theta_{n+1} = \theta_n + 2\pi\,\alpha(J_{n+1}) \end{cases} .$$

which are both area-preserving.

From the KAM theorem, the initial conditions become most relevant to the conservation of torus as other conditions are the same. Particularly, in the case of two dimensional mapping, the torus most robust to perturbations (ie. the last KAM curve) is that with the rotation number equal to the **golden section** number $g=(\sqrt{5}-1)/2$, which is the number worst approximated by rationals according to number theory (see Baker, 1984). The next class of 'most irrational' numbers are those of the form $p+[1/(q+g)]$, where p and q are integers (eg. Greene, 1979; Greene & Percival, 1981; Contopoulos et al, 1987). In connection to a resonant rotation number $\alpha=k_2/k_1$, there are at least two KAM curves on each side close to it, which may often be constructed by the rotation numbers $\alpha\pm[1/(q+g)]$. Relative to the rotation number α, the points on these two KAM curves are generically mapped in opposite directions. Therefore, these KAM curves can be regarded as the boundary of the ring in the Poincare-Birkhoff fixed point theorem; using it to the mapping, we obtain $2kk_1$ fixed points, where k is a positive integer which is usually one, with half of them elliptic and half hyperbolic (Fig. 2.3). Therefore due to perturbation some of the fixed points are preserved, with the position usually shifted (eg. Berry, 1978; Lichtenberg & Lieberman, 1983).

Applying the Birkhoff fixed point theorem and the KAM theorem to higher order iteration of the mapping, we can establish the preservation of infinitely many higher order fixed points and higher order KAM curves. In the remaining part of this section we shall discuss what happens to the other fixed points of the unperturbed system and those invariant curves not sufficiently irrational. It is in fact the chaotic solutions asymptotic to hyperbolic points that occupy regions of the phase space where no invariant curves exist (Moser, 1973).

Homoclinic and Heteroclinic Points

We have seen, by a repeated use of the KAM and the fixed point theorems, the importance of quasi-periodic motions towards a comprehensive picture of the phase space structure. Its importance in relation to chaotic motions shall become obvious in the following discussion on elliptic and hyperbolic fixed points. Homoclinic and heteroclinic points are the very first examples of chaotic motion given by Poincare demonstrating the complexity of dynamical systems. In fact, their existence can be easily shown based on

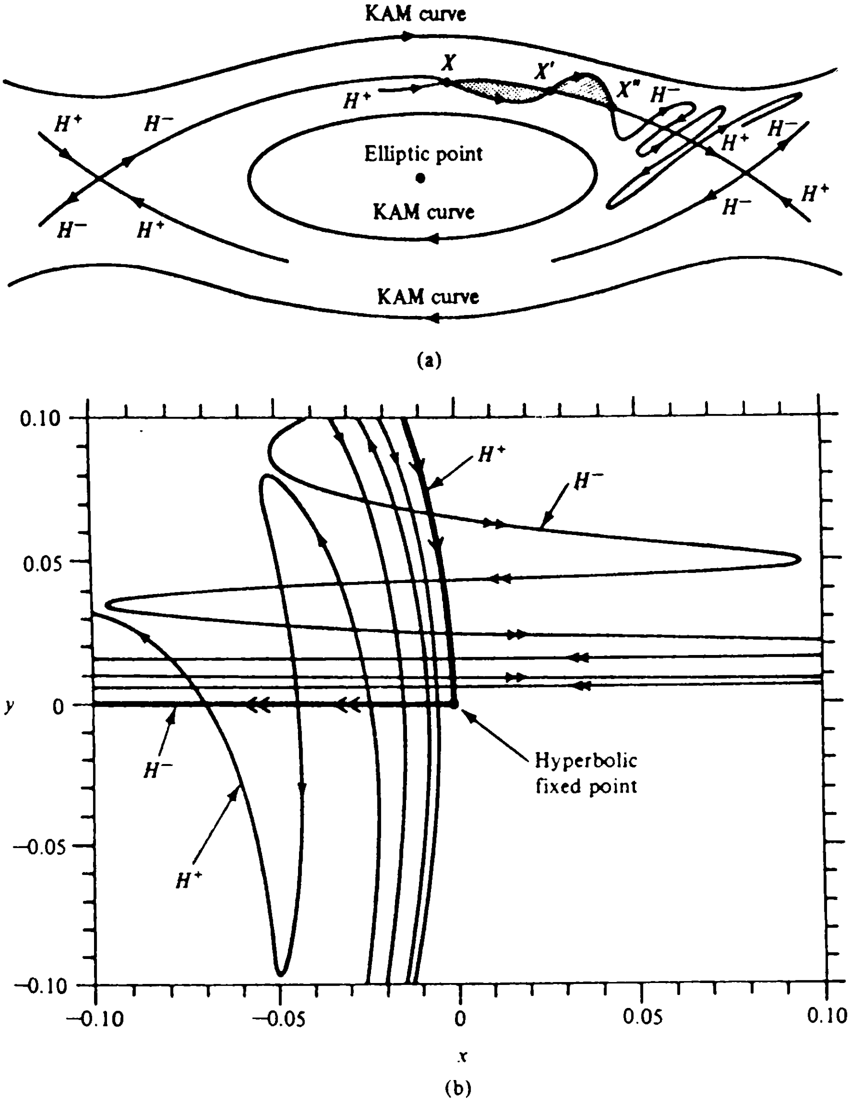


Figure 2.4 Illustrating the effect of a homoclinic point on the generation of chaos near a separatrix. (a). The stable (H$^+$) and unstable (H$^-$) branches of the separatrix intersect infinitely many times. (b). Details of the intersections near the hyperbolic fixed point (from Lichtenberg & Lieberman, 1983).

hyperbolic points of a continuous 2-dimensional area-preserving mapping.

When the system is integrable the stable and unstable manifolds of one (or many) hyperbolic fixed point(s) connects smoothly. When the system is not integrable, however, they **generically** intersect with one another. Since the mapping is continuous and area-preserving, one intersection implies infinitely many intersections (eg. Birkhoff, 1927; Berry, 1978); moreover, as the stable (unstable) manifolds comes closer to the hyperbolic point, it must be stretched to preserve the area. Since a stable (unstable) manifold cannot intersect with a manifold of the same type, it intersects with the elongated unstable (stable) manifold again and again, thus producing a wildly entangled net. In this way area-filling chaotic trajectories are generated (Fig. 2.4).

The intersections of stable and unstable manifolds of the same single hyperbolic point are called homoclinic points; whereas the intersections of such manifolds of different hyperbolic points are called heteroclinic points. Often both kinds of chaotic motions happen in a single system.

However, we must point out that these are only generic behaviour. Even in nonintegrable systems, it may happen that the stable and unstable manifolds of **some** hyperbolic points connect smoothly, and thus no homoclinic (or heteroclinic) points are produced by such hyperbolic points. For more detailed discussion on generic (typical) and non-generic (atypical) behaviour, see Poston & Stewart (1978).

Now we see another important aspect of quasi-periodic solutions in relation to chaos. The KAM curves are important to establish the existence of infinitely many hyperbolic points, which in turn generate homoclinic and heteroclinic points and chaos. Therefore, with the present knowledge of dynamical systems, quasi-periodic solutions and invariant tori are essential to present a comprehensive qualitative description of the generic behaviour for Hamiltonian systems and area-preserving mappings (Fig. 2.2.b).

It is useful here to make a comment on **Wintner**'s comments on Liapounov stability given on page 98, and integrability given on pages 142 to 145 of his 1947 book. He was quite right in saying that Liapounov stability is highly exceptional and that such stability is related to Diophantine problems. But he also remarked that this stability requires too much, thus is not important for practical purposes. Now it has been realised that it is exactly this stability together with Poincare's and concepts like Liapounov characteristic exponents that are most closely related to chaotic and quasi-periodic motions (eg. Moser, 1973), which are very useful practical concepts.

His comments on integrability in the text, are even more misleading (see eg. Broucke, 1979 for a comment). As a topologist, Wintner took a definitely more advantageous standpoint of looking at the qualitative feature of dynamical systems rather than a quantitative one; and realised that sometimes one is equally hopeless with a system integrable in the sense of Liouville. Because of the difficulty of inversion in analysis, a qualitative description is equally difficult to obtain (from the implicit solution in quadrature) for an integrable system and a nonintegrable system. This is true even in a slightly perturbed area-preserving mapping. For example the surface of section of the Henon-Heiles' system seems to be completely occupied by invariant curves when the energy is very low. This system is non-integrable; however, one could not easily distinguish this system from a completely integrable system with similar invariant curves. On the other hand, the invariant circles of the twist mapping may be distorted arbitrarily so as to represent another area-preserving mapping, whose solution would be complicated because of the complexity of the transformation. Yet, the system is integrable; it is topologically different from non-integrable systems.

Therefore, the present author would argue that integrability is a useful and theoretically rigorous concept, which is well defined at least in the present theory of dynamical systems. What is really uncertain is how to determine and distinguish whether a system is integrable, and how to search for integrable systems.

One should not conclude, from the difficulty of deciding convergence of infinite series that the notion of convergence is quite undefined. In a broader sense, the notions of nonpredictability and determinism, undecidability and definiteness etc. are not exclusive. The idea of 'no construction, no existence' (as was held by Kronecker) was already dismissed by the work of Cantor.

An example more likely to be accepted by Wintner may be that a good cup is diffeomorphic with a torus, where the global diffeomorphism is not trivial, and the analytic transformation between them is highly complicated in Euclidean space. Yet, the notion of diffeomorphism is very useful in topology.

2.6 Solution Method IV - Numerical Integration and Nonpredictability

Because of the existence of chaotic solutions, the reliability of conventional numerical routines for continuous dynamical systems needs a more careful discussion. The computability (predictability) of single trajectories becomes questionable especially for long term numerical integrations (eg. Heisenberg, 1967). Since this field is only at a beginning stage, we shall only mention the problem important to numerical routines of classical many-body systems.

The first integrals such as energy and angular momentum are often used as a check of the accuracy of the numerical routines (eg. Stiefel & Scheifele, 1971; Heggie, 1988). However, first integrals are usually 'adiabatic'; they are not sensitive to integration errors. Usually, no difference can be observed in the change of integrals between regular and irregular trajectories. Therefore they are very inadequate for this conventionally assumed role; at most they can be a necessary check that the routines are not too bad.

On the other hand, difference methods are usually used in numerical routines. It becomes questionable whether these integrals are still possessed by the numerical method. In general, integrals of continuous system may be lost by truncation, as is already shown in the Toda lattice problem. In fact, the extra integral of the Toda potential is lost in Henon-Heiles system, and any higher order truncation of the Taylor expansion (eg. Udry & Martinet, 1990). Although this has not been shown for the Taylor expansion with respect to time, the present author strongly feels that almost all integrals are lost in such a truncated system (eg. the numerical routines used in chapter 5).

A way of overcoming this shortcoming of standard numerical integration schemes is probably by the so-called simplectic integration algorithm (eg. Channell & Scovel, 1990). By this algorithm, the important global features of the systems are preserved. The method preserves all the Poincare integral invariants by seeking a generating function which produces canonical transformation between successive discrete times.

Integration errors are also estimated by integrating on initial conditions whose solutions are known. However, this is equally unreliable as the integral check. Because of the intrinsic difference of solutions, errors do not grow in the same way on different trajectories.

In fact, since adjacent trajectories depart exponentially with time in chaotic regions and linearly in regular regions, the long term computation of single trajectories in chaotic region becomes impossible. In this case, the increase in accuracy of the initial conditions and the process of computation grows linearly with the time interval of computation; this would push the power of modern machines to their limit.

2.7 A Collection of Mappings with Regular and Chaotic Solutions

In section 2.3 we have seen examples of chaos demonstrated on the surface of section of Hamiltonian systems. However, it is also evident that it is much easier to display chaos on computer by using mappings (eg. the standard mapping of section 2.5). In this section we give a collection of the important mappings which have attracted much active research interests, and have been successfully used to convey the meaning of chaos. They will be divided into two groups, namely, area-preserving mappings and mappings possesses attractors.

Area-Preserving Mappings

A 2-dimensional mapping is usually given the most general form

$$\begin{cases} X = f(x, y) \\ Y = g(x, y). \end{cases}$$

Such a mapping is integrable if it admits an integral I(X, Y)=I(x, y)=const. An area-preserving mapping is not necessarily integrable.

In constructing an area-preserving mapping in practice it is very useful to assume the following general forms

$$\text{I.} \begin{cases} X = \quad y + aF(x) \\ Y = -x + bG(X), \end{cases} \qquad \text{II.} \begin{cases} X = x + aF(y) \\ Y = y + bG(X), \end{cases}$$

where a and b are two arbitrary constants. It is easy to verify that the Jacobians of both are equal to one, hence area-preserving. In fact, the constants a and b are included here to indicate that the signs preceding the functions F and G are not important to the area-preserving property of the mappings (although they can be used to indicate chaos); it is the signs of x and y that matter.

(1). Sweet's Mapping

Sweet constructed a class of nonlinear mappings of type-I with F=G (in the 1980's, but not published), which is included here because of the richness of their structure. The very general mapping is

$$\begin{cases} X = \quad y + \dfrac{ax + bx^2 + kx^3}{1 - 2bx + cx^2} \\ Y = -x + \dfrac{aX + bX^2 + kX^3}{1 - 2bX + cX^2}, \end{cases}$$

where a, b, c and k are arbitrary constants.

When k=0 the mapping becomes integrable, although it is still nonlinear. In order to obtain the integral, we rewrite the first equation and square it to give

$$y^2 = X^2 - 2xX\frac{a + bx}{1 - 2bx + cx^2} + \frac{x^2(a + bx)^2}{(1 - 2bx + cx^2)^2} .$$

Thus we have

$$y^2(1 - 2bx + cx^2) + x^2\left[1 - \frac{(a + bx)^2}{1 - 2bx + cx^2}\right]$$

$$= x^2 + X^2 - 2axX + cx^2X^2 - 2bxX(x + X) .$$

Note that the RHS is a symmetrical function of (x, X). On squaring the second equation of the mapping one arrives at the same function on the RHS while on the LHS is the same function of (X, Y). Hence we obtain the quantity invariant under the transformation, viz.

$$I(x, y) = y^2(1 - 2bx + cx^2) + x^2\left[1 - \frac{(a + bx)^2}{1 - 2bx + cx^2}\right] = \text{const.}$$

It is also interesting to note that when b=0 the mapping is simplified, but all the important qualitative features of the original system are preserved. For example, the system is integrable if k=0. As the value of k increases more and more of the invariant curves break up into islands and chaotic sea.

(2). Henon-Heiles' Mapping

Henon & Heiles (1964) introduced an nonlinear mapping of type II, which displays the characteristic feature of their nonintegrable Hamiltonian system. The mapping is

$$\begin{cases} X = x + a(y - y^3) \\ Y = y - a(X - X^3), \end{cases}$$

where a is a parameter. Chaos occurs when a≠0 (eg. a=1.6).

(3). Henon's Quadratic Mapping

In his 1969 paper, Henon introduced and studied in detail his nowadays classical quadratic mapping

$$\begin{cases} X = x\cos\alpha - (y - x^2)\sin\alpha \\ Y = x\sin\alpha + (y - x^2)\cos\alpha, \end{cases}$$

where α is a parameter. Typical structures may be observed in the (-1, 1)×(-1, 1) square domain with parameter values $\cos\alpha$=0.8, 0.4, 0, -0.01 etc.

It is also interesting to note that the linear mapping

$$\begin{cases} bX = bx\cos\alpha - ay\sin\alpha \\ aY = bx\sin\alpha + ay\cos\alpha, \end{cases} \qquad I = \frac{x^2}{a^2} + \frac{y^2}{b^2} = \text{const.}$$

is integrable, where a, b, and α are parameters and I is an isolating integral. The mapping

$$\begin{cases} bX = bx\cos\sqrt{(x^2 + y^2)} - ay\sin\sqrt{(x^2 + y^2)} \\ aY = bx\sin\sqrt{(x^2 + y^2)} + ay\cos\sqrt{(x^2 + y^2)} \end{cases}$$

is apparently nonlinear, but it is easy to show that it possesses the following integral

$$I = \frac{x^2}{a^2} + \frac{y^2}{b^2} = \mathrm{const.}$$

hence it is integrable.

(4). Froeschle's Mapping

Froeschle (1971, 1973) studied the following area-preserving mapping of angular variables

$$\begin{cases} X = x + a\sin(x + y) \\ Y = x + y, \end{cases} \qquad (\mathrm{mod}\ 2\pi)$$

where typical values of the parameter giving interesting features are a=-0.3, -1.3.

(5). Rannou's Mapping Operating on Integers

To avoid the round-off errors usually involved in any computation, Rannou (1974) studied Froeschle's mapping by limiting the variables (x, y) in the field of integers. The mapping is

$$\begin{cases} P = p + \left[\frac{\lambda k}{2\pi}\sin\frac{2\pi}{k}(p + q)\right] \\ Q = p + q \end{cases} \qquad (\mathrm{mod}\ k)$$

where (p, q), (P, Q) and k are integers; a [...] represents the integral part of the number. This mapping displays similar invariant curves, islands and chaotic seas as Froeschle's mapping; so it confirms that the disappearance of isolating integral is not due to computing errors. See Appendix A for a similar example well studied in number theory.

Mappings with Attractors

When a dynamical system has both driving and damping mechanisms, it is very likely to produce the phenomena of attractors. Attractors can either be a smooth manifold having integral dimensions (eg. point attractor, limit cycles), or they can have fractal dimensions; in the latter case, they are called strange (or chaotic) attractors. While the phenomena of limit cycles was known at the turn of the century, the first fractal attractor was not discovered until 1963 when Lorenz observed his strange attractor from a set of three nonlinear differential equations. Later more and more such attractors was observed from both experiment and theoretical computation (see Holden, 1986). Today numerous efforts have been directed towards finding attractors even in biological science and economics (see Stewart, 1989). However, the easiest way of obtaining attractors on computer is to use mappings which are not area-preserving. The following is a brief list of the important ones. They may find some use for the proposal of the next chapter.

(1). Logistic Mapping

The logistic mapping is of some importance in biological science, which is usually written as

$$X=\lambda x(1-x) \qquad \text{or} \qquad X=x^2-c$$

where λ and c are parameters.

(2). Mandelbrot Set

The mapping giving the Mandelbrot set in the parameter space is the complex version of the logistic mapping, viz.

$$Z=z^2-C$$

where z is a complex variable and C is a complex parameter. This mapping may be expressed by a 2-dimensional mapping using real variables and parameters

$$\begin{cases} X=x^2-y^2-a \\ Y=2xy-b\ , \end{cases} \qquad \text{where} \begin{cases} z\equiv x+iy \\ C\equiv a+ib. \end{cases}$$

(3). Cantor Set

A simple Cantor set can be produced as the attractor of the following mapping

$$X=3\left(\frac{1}{2}-\left|x-\frac{1}{2}\right|\right).$$

(4). Henon's Attractor

Henon (1976) constructed his well-known 2-dimensional mapping with a strange attractor, which is

$$\begin{cases} X=y+1-ax^2 \\ Y=bx, \end{cases}$$

where a and b are two parameters. The typical values of the parameters are $a=1.4$, $b=0.3$, and the attractor lies in the domain $(-1.5, 1.5)\times(-0.5, 0.5)$. The Jacobian of the mapping is a constant, -b.

This is a very good example to demonstrate how the embedding theorem works (see Holden, 1986; Stewart, 1989). To explain the meaning of this theorem, let us suppose that the above mapping defines a complete hidden dynamics (say that of economics) with two coupled variables (x, y), but only x is observable; then the theorem says that one can recover the attractor by looking at the x sequence (or time series). The technique is to construct a 2-dimensional sequence (x_i, x_{i+n}) for a fixed n, and regard it as the sequence of a 2-dimensional mapping; then the same attractor (may be unfolded) appears in such a **phase portrait**. Most of the current research on chaos in economics is based on this theorem.

(5). The following mappings produce attractors with symmetry

$$\begin{cases} X = (2x^2 + 2y^2 - p)x - \frac{1}{2}(x^2 - y^2) \\ Y = (2x^2 + 2y^2 - p)y + xy \end{cases}, \qquad \begin{cases} X = (2x^2 + 2y^2 - p)x - \frac{1}{2}(x^2 - y^2) \\ Y = (2X^2 + 2y^2 - p)y + Xy, \end{cases}$$

where p is a parameter. Interesting values of the parameter are p=1.5, 1.8, 1.9.

(6). The following mappings have point attractors, limiting cycles and other 1-dimensional attractors, and have some connection to one Diophantine equation given in Appendix A. We will give the equations of the mappings and the local invariants defining the limit cycles, they are

$$\begin{cases} bX = b^2x^2 - a^2y^2 \\ aY = 2abxy, \end{cases} \quad b^2X^2 + a^2Y^2 = (b^2x^2 + a^2y^2)^2 \Rightarrow I = b^2x^2 + a^2y^2 = 0,1$$

$$\begin{cases} bX = b^2x^2 + a^2y^2 \\ aY = 2abxy, \end{cases} \quad b^2X^2 - a^2Y^2 = (b^2x^2 - a^2y^2)^2 \Rightarrow I = b^2x^2 - a^2y^2 = 0,1$$

where a and b are two arbitrary constants. For example, a=1, √2 and b=1 are interesting parameter values.

2.8 On the Occurrence of Commensurabilities in the Solar System

To end this review chapter, we shall make a few comments, base on the KAM theorem, on the occurrence of commensurable mean motions in the Solar System, a problem which was studied by Roy & Ovenden (1954, 1955).

The existence of the Titius-Bode law in the distance distribution of planets from the Sun has always been a curious problem. Is the law fortuitous, or is it a result of dynamical evolution? On attacking this problem, Blagg (1913) proposed more complicated relations which are applicable to satellite systems of Jupiter, Saturn and Uranus. Very recently, the problem has been studied numerically by Conway (1988); in his work the significance of the Titius-Bode law to stability is shown.

On a different track, Roy & Ovenden (1954, 1955) related the problem to the occurrence of commensurable mean motions approximated by small integers; because orbital motions approximated by low order resonances can display simple geometrical distribution in distance. It was found by Roy & Ovenden (1954) that the occurrence of commensurable mean motions in the Solar System is more frequent than by a chance distribution. From this they concluded that commensurable mean motions may be preferred by a dynamical mechanism. This point was shown to be supported by the Mirror theorem in their second paper; but as was pointed out in the paper, the explanation is not rigorous. Later, by introducing the influence of tidal forces this preference of commensurable mean motions was explained by Goldreich (1965).

In fact, the dynamical significance of resonances to stability is made more obvious in chaotic dynamics; which has been applied to the distribution of asteroids (eg. Siegel & Moser, 1971). However, in chaotic dynamics (in particular the KAM theorem) a resonance is often used to explain instability rather than stability; thus the above interpretation of Roy et al is not compatible with the Hamiltonian chaos theory.

According to chaotic dynamics of Hamiltonian systems, related to a resonant frequency, usually there are both elliptic (stable) and hyperbolic (unstable) fixed points (periodic motions); thus in general both stable quasi-periodic solutions and chaotic solutions exist in connection with a resonance. Although the occurrence of chaos at a resonance cannot be used to argue against the preference of resonant motions, it is shown not only by the KAM theorem but also in numerical work that initial conditions most further away from resonances are more stable. Therefore, if the planetary motions can be approximated by the N-point-mass model, then conceptually commensurable mean motions must be the **least** favoured for stability by dynamics.

However, it must be pointed out that chaotic dynamics of Hamilton system is not against the statistical analysis given by Roy & Ovenden, but only against the resonance preference interpretation of such an analysis. Moreover, controversy arises only when exact resonance is referred to in their interpretation; chaotic dynamics is not yet strong enough to dismiss **near** resonances, which might be meant by Roy & Ovenden.

It is also useful to note the following fact considering the distribution of numbers. It is readily seen that if x is uniformly distributed in the domain (0, 1), then in general $f(x) \in (0, 1)$ is not uniformly distributed. This is true, for example, if x and f(x) are defined by a finite number of digits 0.A...B.

If the ratio of two frequencies is assumed to be uniformly distributed in (0, 1), then the analysis of Roy & Ovenden may suggest that the occurrence of resonances is greater than natural. However, there is no reason why such an assumption must be true. In fact, if we assume that the frequencies are distributed uniformly in the sense that all numbers of the form 0.A...B are equally possible, then the ratio of frequencies would display certain resonance preferences.

A final answer to this question is far from been possible at present because chaotic dynamics of dissipative systems can produce resonance locking mechanisms (Berge, et al, 1984). Therefore a more complete investigation on the N-body model and the effect of tidal forces using the modern chaotic dynamics is desirable. Before this is possible, both the statistical analysis and its interpretation remain open.

I believe that God does not play dice. --- Einstein
He doesn't need ***notations***, he needs ***notions***. --- Gauss
The essence of mathematics resides in its freedom. --- Cantor

CHAPTER 3
Deterministic Chaos and Quantization
- a Heuristic Discussion

In the last two chapters we have briefly discussed the phenomena and fundamental theories relevant to the motion of heavenly bodies, in particular those in the solar system. The successful explanations by Newtonian mechanics and difficulties encountered in the classical N-body model are reviewed. Emphasis was made on the generic chaotic behaviour found in both continuous and discrete dynamical systems. In the present chapter, we will deviate from the standard material and give a rather speculative and personal view of chaotic dynamics, paying some attention to its impact on science in general.

In section 3.1, the author proposes a possible alternative interpretation to wave mechanics or quantum mechanics (QM, hereafter). This idea came to the author's mind based on the observation of certain formal similarities between the chaotic attractor and quantum state. In the spirit of looking for a deterministic interpretation to quantum mechanics, we shall speculate the mechanism required for attractors in atomic level, and demonstrate that the key characteristics of quantum mechanics can be well accommodated in the notion of a chaotic attractor. In section 3.2, the historical relation between deterministic chaos and quantum mechanics is discussed. An attempt is made to explain in a broader sense the nonpredictable feature of the deterministic history, which is not only manifest in the history of quantum mechanics but also history in general. In section 3.3, we will look at quantization in a mathematical way, and make some comments on the revolutionary differences between the microscopic and macroscopic worlds which seem to have been overemphasised based on the conventional interpretation to QM. The possible routes leading to quantization in the framework of classical physics (CP) are summarised. Finally, the impact of chaos on science, particularly on the significance of the concepts such as equilibrium and linearity to natural phenomena, is discussed in section 3.4.

3.1 Chaotic Attractors and Quantization - Formal Compatibility

It is well known that quantum mechanics is one of the three revolutionary physical theories born at the turn of the century, because it resolved the classical physics difficulties in explaining a number of well-established experimental and observational phenomena such as the black-body spectrum (ie. the ultraviolet catastrophe), absorption spectra and stability of atom (the classical radiation catastrophe). It has always been stressed that classical physics produces a continuous and deterministic macro-world, whereas quantum physics creates a discrete and probabilistic micro-world. Due to the revolutionary theory of quantum mechanics we now have built up a quantised and probabilistic picture for the atomic and sub-atomic physical world. However, from mathematical physics we know that quantization is not unique to quantum (or wave) mechanics, nor is the probabilistic view of nature solely due to quantum mechanics. However, compared to the discreteness, uncertainty and indeterminism familiar in classical physics, those arising from quantum mechanics have been given completely different significance in physics.

At almost the same time when the old quantum theory was born, there also occurred the revolution in relativity and chaos, where the latter's importance was recognised completely within the framework of deterministic Newtonian mechanics (Poincare, 1892). History has witnessed an almost parallel progress and mutual influence on each other in the development of quantum mechanics and relativity; however, because of the delayed recognition of the revolutionary theory of chaos, the interpretation of QM may have been accidentally made probabilistic. Here we shall propose a possible alternative deterministic interpretation for quantum mechanics using chaotic attractors discussed in previous chapters. According to this interpretation, QM may be put into a position like that of statistical mechanics relative to Newtonian mechanics; it would become a theory with underlying determinism, and the revolutionary aspects of QM would only lie at the transition from Newtonian mechanics to QM. Thus nature may deviate from some QM descriptions due to effects of classical mechanics; similar deviation from statistical mechanics (eg. Fermi & Pasta & Ulam, 1955) has already been recognised because of the work of Kolmogorov, Arnold and Moser.

No attempt is assumed (in fact one could not, at the present stage) to argue for the point that the conventional interpretation to QM is false; however, we would at least argue against the general opinion on the failure of classical physics and need for the revolutionary quantum theory contained in almost all standard textbooks. Also discussed is the problem of continuity and determinism of classical mechanics for the macro-world and the discreteness and indeterminism of quantum mechanics governing the micro-world.

Escape from Classical Catastrophes

Firstly, the oftentimes remarked continuity of classical physics cannot be taken literally. From chaotic dynamics (eg. the KAM theorem), motions in nonintegrable conservative systems cannot be stable (here, quasi-periodic) for all initial conditions; in phase space, stability does not change continuously with initial conditions. Thus nonglobal first integrals appear which exist only on invariant tori. If such integrals are formally generalised to the whole phase space, then one finds that only for some discrete set of values of them stable orbits exist; otherwise motions are not stable (ie. chaotic).

In this direction much investigation has been carried out; a comprehensive review may be found from Berry (1978, 1987). For example, semi-classical quantization of adiabatic invariants has been studied by Gutzwiller (1971, 1980), Berry & Tabor (1976) etc., which can be traced back to the very first effort of Bohr. Efforts on quantization of quasi-periodic, homoclinic and heteroclinic motions can be found in Ozorio de Almeida (1989) and references therein.

Even in linear systems such as Mathieu's and Hill's equations, where chaos does not occur, the stability condition leads to quantization of the parameters of the ODEs (see Jordan & Smith, 1990). The importance of such equations will be stressed in section 3.3 in guiding the construction of ODEs simulating simple models treated in QM.

Secondly, the deterministic view of classical physics is not true either. Before the theory of chaos, the two words 'deterministic' and 'predictable' were not critically distinguished in physics, nor in philosophy. Therefore, in the pre-chaos view, the behaviour of classical physics systems is not only taken as deterministic but also predictable as a result of the general existence and uniqueness theorem of ODEs. However, classical physics is now realised as deterministic because of the local existence and uniqueness of solution to ODEs; the behaviour of a dynamical system is not predictable for an arbitrarily long time scale. Because of this nonpredictability feature of classical systems, a probabilistic and statistical approach must be adopted. This is the indeterministic aspect of classical physics.

Thirdly, the classical radiation catastrophe may not be so catastrophic to classical physics as was always thought. It is in fact only catastrophic to linearised classical physics, because the old quantum theory was developed unfortunately under the shadow of linearity and integrability assumed implicitly for classical physics. If nonlinear terms are not omitted in the analysis, it may happen that the electron in the hydrogen atom would not collapse into the nuclei. The way out is easily seen with the notion of chaotic attractors in dissipative systems with driving terms, since chaotic attractors supply not only chaotic motions but also a new kind of equilibrium state.

In Bohr's treatment of the hydrogen atom, the classical radiation was artificially discarded by imposing a rather artificial quantization condition. In contrast, chaotic attractors may offer a more natural and comfortable explanation for the stability of quantum states and quantum jumps; because all the classical effects may be preserved. At first glance, one may find that the model lacks a driving mechanism. This, however, is not a serious problem to classical physics, because we see that all atoms and molecules are staying in the atmosphere of a radiation sea made up of electromagnetic waves (at least the cosmological background radiation is always there, though this may not be the correct one). Therefore both driving and dissipation mechanisms are inherent to classical physics; the hard problem is in fact how to construct or choose some models and study their chaotic dynamics in more detail.

In doing this there are two points that need particular care. One is that the problem may have to be formulated in the framework of special relativity because of the high velocity involved in the micro-world; general relativity is also important when elementary particles are studied. These suggest the importance of studying relativistic chaos. The other point is that in such an approach, the radiation sea may have to be quantised before the behaviour of the radiator, as historically Planck's work also preceded Bohr's. However, Planck's black body radiation may be used in studying quantization of an atom, since we are not challenging the result of QM at the present stage. It may happen that quantization of the black body radiation is easier to study first. In fact, such a view is held by many scientists (eg. Galgani, 1985 and references therein). Deductions of Planck's law have been carried out in various different ways, and a possible deviation from Planck's law in the Rayleigh-Jeans region is theoretically conceived.

In addition, the situation of this proposal may be compared to some problems related to relativity. For example, Einstein's second assumption on special relativity, namely, the constancy of light speed, is not needed since it is consistent with classical physics, although it is of fundamental significance to special relativity. The interpretation to QM by Heisenberg, Born and Bohr may not be needed because of the discovery of chaotic attractors, although it is fundamental to QM in withstanding the historical questioning by

Einstein. A second example suggesting a possible significance of relativity is the spin degree of freedom familiar in QM. Spin is usually introduced in QM as a new degree of freedom, which is totally exterior to the nonrelativistic quantum theory. It, however, may be explained in the relativistic quantum theory (eg. Schweber, 1961; Streater & Wightman, 1964).

Quantum States and Quantum Jumps

We have pointed out that the historical catastrophes were not so fatal to classical physics as previously thought. The mechanism needed in producing chaotic attractors can be found within classical physics to supply possibly a deterministic interpretation for QM phenomena. However, it remains to investigate the more detailed compatibility between chaotic attractors and the quantum state and quantum jump process.

Historically, the quanta phenomena would not be catastrophic to classical physics if attractors (not necessarily chaotic, eg. limit circles found in the van der Pol equation) were known to physicists explaining the stability of atoms; neither would the line spectra due to quantum jump be so difficult to classical physics, supposing the jump phenomena in systems such as Duffing's equation were familiar to physicists at the turn of the century. If these phenomena completely within classical physics were admitted to the science community, then at least conceptually classical physics would not have had to face a fatal challenge.

However, from the results of QM, we see that the structure of manifold attractors (eg. points, circles) is too simple to accommodate quantum states. Chaotic attractors are usually fractals, in which the motion of the system is chaotic and the state distribution of the system becomes ultimately probabilistic. Quantum jumps then may be interpreted as jumps between attractors either due to a disturbing excitation or even a statistical fluctuation. In this way the width of line spectra may be related to the size of attractors; while fine structures of spectra may be a reflection of the fractal feature of the attractors.

Here many problems cannot be made clear without actually constructing models and studying their detailed behaviour, and in fact this is the key difficulty of the proposal. For example, can one construct attractors simulating the quantised energy states, say in the hydrogen atom? Will the time-averaged distribution of states in attractors agree to the distribution determined by the wave function? Another problem is, supposing jumps between attractors produce the line spectra, then what observational effects would the classical radiation due to the motion inside each attractor produce? Although these points are not yet clarified, they cannot yet be used to argue against the proposed interpretation.

Uncertainty Principle and Probability Interpretation

As previously stated, the key point of the proposed deterministic interpretation to QM is to actually construct some nonlinear models and study their behaviour. However, before any construction is possible it is useful to investigate whether such an interpretation is consistent with the main indeterministic and probabilistic characteristics of QM, namely, Heisenberg's Uncertainty Principle and Born's probability interpretation to the wave function.

If the detailed information of the uncertainty principle and probability interpretation is laid aside, it is very easy to see the agreement between chaotic attractors and QM. Firstly, the indeterminism of QM must not be made in contradiction to the determinism of Newtonian mechanics; because the indeterministic aspect of Newtonian mechanics, eg. nonpredictability, has been uncovered by chaotic dynamics. Here, care must be paid to that in today's theory of chaos, the two words 'deterministic' and 'predictable' carry different meaning. In the conventional discussion of indeterminism of QM and determinism of Newtonian mechanics, however, they are not sharply distinguished. Secondly, the uncertainty feature of motion in the chaotic attractor is evident, since one can only tell which attractor the system is in, but not exactly where it is inside the attractor based on prior knowledge of its motion. Because of the high speed of motion and high dimension of the attractors, it may happen that a macroscopic short time scale involved in any ideal measurements is too long to locate a single position of the system in the attractor; thus the determination of all generalised coordinates and momenta becomes impossible in principle. However, it remains to study in detail why the uncertainty should be between conjugate quantities. In this connection it is worthwhile mentioning that uncertainty phenomena between conjugate quantities have been found numerically in the restricted three-body problem (see Szebehely, 1984).

The conceptual compatibility between attractors and Born's probability interpretation was already evident from the discussion of the previous subsection. In fact, many approaches have been carried out in studying the probability problem in chaotic attractors (see eg. Lichtenberg & Lieberman, 1983). What is still not clear is whether a detailed agreement can be made to the wave function.

3.2 History of Determinism (of CP) and Indeterminism (of QM)

One obvious question to ask in connection with the deterministic interpretation of QM proposed here is, if it were true, then why did history choose the false one and stay with it for almost a century? In this section we turn to a historical consideration because of the difficulty involved in a mathematical approach on the proposal of the last section. Such a discussion is admittedly difficult because of the enormous amount of available material and the fact that the answer to the above question really depends on personal view of history.

The author believes that natural and social courses are mutually related and have their deterministic feature locally (in space-time and other degrees of freedom); the history flow on some isolated degrees of freedom is deterministic in a short time scale. Thus history is deterministic, but it is not necessarily predictable; some accidental selection and bifurcation may happen. In fact, irrational numbers were already conceived by the ancient Greek, but a complete theory for it was not developed until only the last century by Cantor (see Bell, 1937). In Newton's time, his particle interpretation for light dominated over Huygens' wave interpretation more than a century, although the latter is closer to truth (see any textbook on light). More recently, the three classical tests of relativistic gravity computed by Einstein were the gravitational red-shift experiment, the deflection of light, and the perihelion shift of Mercury. However, the fourth test proposed by Shapiro (1964), namely, the time delay of light, which in principle is closely related to the deflection of light, is one of the most precise tests of general relativity to date. It remains a mystery why Einstein did not discover this effect (Will, 1981). These are examples of defects in science due to personal influence.

On the other hand, examples exist showing the influence of technology and human comprehensibility. Let us look at the fate of the three revolutionary physical sciences, relativity, quantum mechanics and chaos. History has witnessed a continuous growth of quantum mechanics in spite of its conceptually abstract features; this is actually because QM has always been related to experiments and productive results. In contrast, the great interest of the physical community in general relativity lasted only until the 1930's, a couple of decades since its foundation, then this interest had almost lapsed during the next twenty years. The situation was not changed until the late 50's and early 60's when great experimental progress was made, and for example the 3K microwave background radiation was discovered, the Kerr metric was found and a break-through was made on the mathematical structure of general relativity (Infeld, 1964).

Deterministic chaos was already well understood by Poincare at the end of last century, but it was not accepted by the scientific community until the 1960's when chaos was observed from experiments and easily demonstrated due to the advent of computing facilities. Although the approach along this line did not stop in the first half of this century, neither Birkhoff nor Siegel popularised the notion. In fact, even the development of fractals also followed a similar course. Thus it may be concluded that history does not necessarily choose the truth and make it flourish for ever.

Since it is almost certain that the notion of chaos did not influence much of the development of quantum mechanics, an obvious alternative way of answering the question stated at the beginning of the section is probably to find out whether the development of chaos has influenced the main contributors to quantum theory, wave mechanics and quantum mechanics. In this investigation, the main references used are Mehra (1975, 1982), Born (1971), Wheeler et al (1983), the materials on the Solvay Conferences, in particular those on the first and the fifth and those on the Bohr-Einstein debate. The recognised figures involved in the debate have also written many books reflecting their opinion.

In talking about the history of QM, it is conventional to distinguish the period of the old quantum theory from that of the quantum (and wave) mechanics. It is useful to note that the old quantum theory concentrated on the question of 'quanta', or the discreteness of the microscopic world; whereas the later quantum mechanics paid more attention to the uncertainty and probabilistic questions of the theory. As a historical coincidence, one can consider the death of Poincare as a useful mark between the old and new quantum theory. As is already shown, chaotic dynamics can produce very naturally both discrete and indeterministic (in the pre-chaos sense) phenomena. The indeterministic and unpredictable feature of classical mechanics was already clear to Poincare. However, considering how mathematically involved Poincare's works are and the fact that the old quantum theory was developed in the years immediately following the publication of his works, one must not expect much influence on the contributors to QM when they were founding the old quantum theory, for example, Planck's work on the black body radiation, Einstein's work on the photo-electric effect and Bohr's quantisation of the hydrogen atom. Moreover, Poincare probably did not conceive of the discreteness of classical mechanics; this had to wait until the establishment of the KAM theorem and the discovery of Lorenz attractor in the 1960's. Thus one should not be surprised that in the early development of the quantum theory, it was Poincare who dismissed on the first Solvay Conference any attempt (say Jeans) of producing discreteness in the framework of classical physics, and later 'proved' that an essential discreteness was needed to produce quanta.

Here a few questions may be appropriately asked. What would happen to quantum mechanics and modern physics in general, had Poincare lived at least another fifteen years so that he could contribute to the development of the new quantum theory and its interpretation? Would he still insist on the need for a discrete mechanism for the quanta. If he had joined the Bohr-Einstein debate, would he agree that classical mechanics was deterministic and quantum mechanics is probabilistic? Since he had always shown interests towards the physical sciences, as can be seen from his contribution to special relativity and his efforts showed on quantum theory, would not his work at least have influenced more deeply those figures involved in the historical Bohr-Einstein debate?

Among the contributors to the new quantum mechanics, de Broglie showed great interest in interpreting wave mechanics in a manner different from that of Bohr's school. However, his ideas must be considered almost irrelevant to the indeterminism in classical physics, and they are often confusing. It is more hopeful to look at the figures involved in the interpretation of quantum and wave mechanics in the mid-1920's, and the historically heated Bohr-Einstein debate on determinism or indeterminism of QM in 1930's. Here Einstein showed his strong feeling against the QM probabilistic view of the natural world and the belief of a deterministic interpretation to quantum theory. Moreover he took an opinion opposite to Poincare's; as can be seen from his correspondence to Born, Einstein mentioned many times that quanta must be a result of continuous differential equations. Although his specific suggestion of 'redundancy of variables' is obviously false, clearly he did not mean anything like the well-known eigenvalue problem of partial differential equations, but something dynamical.

On the other hand, although the debate mainly concentrated on the interpretation of QM, both Bohr and Einstein, as well as Heisenberg, Born and Jordan are good masters of Poincare's work. Among them both Bohr (1932) and Einstein (1917) studied deeply almost periodic functions, which are of key importance to the divergence caused by chaos; while Born and Heisenberg mastered Poincare's work on the divergence of the general perturbation theory, as may be seen from Born's (1924) classical book which was completed with some help by Heisenberg. However, it seems that none of them understood the chaotic notion in Poincare's work; in fact they only adopted Poincare's perturbation theory as a mathematical tool.

In spite of the similarity between the indeterminism in classical physics and the uncertainty and probabilistic feature of quantum mechanics it seems that no evidence supports the notion that Poincare's work contributed to the studies of Born and Heisenberg when they discovered the probability interpretation and the uncertainty principle in the years 1926 and 1927.

It seems that when the debate was relatively heated in the 1930's, none of them, nor anybody else quoted the indeterministic feature of Newtonian mechanics made clear by Poincare (in both academic and popular writings) to argue against Einstein. This is really a mystery, for at least Einstein met Poincare in 1911 on the first Solvay Conference; and in the same year both Poincare and Madame Curie highly commended Einstein (see Mehra, 1975), although Einstein was not impressed by Poincare at that time. From the debate and Einstein's famous say 'I believe God does not play dice' one usually comes to the unfortunate conclusion that Einstein, an important contributor to the old quantum theory, although open minded elsewhere, became very stubborn towards the new QM. However, there might be some justice to look at Einstein's 'conservative' attitude in a positive way. It might be Poincare's assumed 'progressive' attitude towards the need for a new discrete theory that made his work on chaos and indeterminism leave no mark on the history of quantum mechanics.

It is also interesting to look at the writings of Born and Heisenberg after the heated debate. In 1955, Born wrote an essay 'Is Classical Mechanics Deterministic ?', where his opinion was that it was not; but his argument still relied heavily on the collisional phenomena (see his 1956 Book). One thus has to assume that he did not understand chaos even in the 1950's. In contrast, Heisenberg presented a review paper on nonlinear physics in 1967, in which he clearly stressed the nonpredictability of the classical three-body problem and showed his proper understanding of chaos. These are the articles relevant to the proposal of the present chapter, however, neither of the authors mentioned the relation to their articles on quantum mechanics. Here one is inclined to ask the following questions.

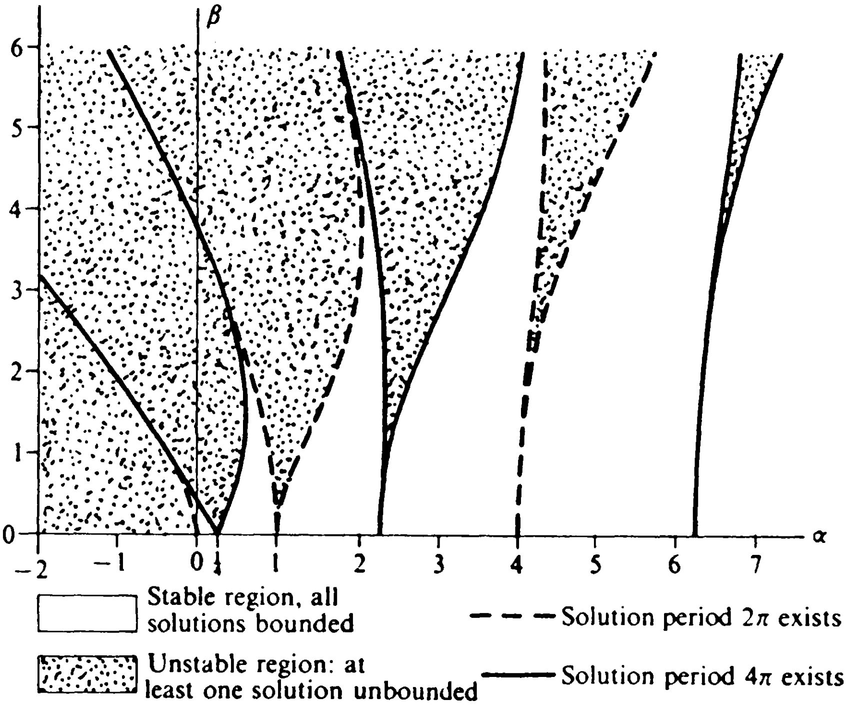
Why was Born interested in indeterminism of classical mechanics in 1950's? Was he trying to find indeterminism from classical mechanics to argue against Einstein's questioning? If yes, then why did he turn to this direction so late? One must note that although this was the time when Einstein mentioned more often to Born his 'quanta by redundancy of equations', this is to be considered as a different matter. Because in the context of quantum mechanics, the notions of probability and quanta are usually regarded as two independent fundamental rules.

When did Heisenberg become interested in nonlinear dynamics? Was his 1967 review only a result of his work in nonlinear quantum theory, or was it that he was also looking for indeterminism from classical mechanics and found the correct one? If it were the latter, then why did he not mention anything towards the historical debate in this review?

Considering the good collaboration between Heisenberg and Born, did the former communicate to the latter about the nonpredictability of classical mechanics. It is would be interesting to find out whether the figures mentioned wrote any further articles discussing the problem related to the debate on determinism or indeterminism.

3.3 Mathematical Aspect of Quantisation

In the previous sections we have shown the possibility of producing quantization by the



least one solution unbounded
Solution period 4π exists

Figure 3.1 Stability diagram for Mathieu's equation $d^2x/dt^2 + (\alpha + \beta \text{ cost})x=0$ (taken from Jordan & Smith, 1977).

deterministic chaotic attractor; also discussed was the historical relationship between the theories of quantum mechanics and chaos. In this section we will discuss briefly the mathematical aspects rather than the physical aspects of quantization, so as to provide further support for the proposal.

In physical sciences, the revolutionary aspects of the quantum theory have always been emphasised. Today even school students know that the microscopic world is dramatically different from the familiar macroscopic world; they are governed by completely different laws, although there is a certain correspondence principle connecting the two worlds. However, based on the progress made in deterministic chaos, the author has come to a different opinion.

In the context of mathematical physics, it is well known that quantization is not unique to quantum mechanics. A large class of classical problems related to wave phenomena share the same type of quantization with Schrodinger's wave mechanics. In both classical and quantum wave mechanics, quantization is usually a result of boundary conditions and natural boundary conditions. If we regard quantum (wave) mechanics as phenomenological, then what has been tested by experiment is that the micro-world does obey the theory. However, the interpretation of a new theory usually depends on what the community of the time has achieved based on a previous understanding of nature. This is a different matter and probably where a mistake is unavoidable.

It is worth noting that quantization is usually due to boundary conditions of differential equations, which , in quantum mechanics, is in turn a manifestation of the symmetry of the space background. Now we see that in chaotic dynamics, symmetry, periodicity and stability restrictions can also result in discreteness. Therefore there is a possibility of producing quantization phenomena by imposing symmetry, periodicity and stability conditions, which are more natural than the conventional quantization conditions. In fact, the simple Bohr-Sommerfeld quantization condition was only successful in classically integrable problems (neglect radiation). It is a mystery why there should be such a relationship between classical mechanics and quantum mechanics, and why the integrability of an ODE should matter to what is governed by PDEs.

In fact, deviation from QM has been observed in experiment; and the current opinion of a way out is to go back to classical chaotic dynamics and study how classical mechanics transits to QM. For example, irregular spectra have been observed in the Zeeman and Stark effects (Hasegawa et al, 1989; Friedrich & Wintgen, 1989).

In looking at quantization as chaotic attractors, we shall specially emphasise two points. The first one is Mathieu's equation, whose solution is surprisingly similar to what happens in quantum mechanics problems (say the simple harmonic oscillator), as can be seen in Fig. 3.1. The parameters of the equation (one is the linear frequency, which is always related to energy in QM) are quantised to produce stable solutions; moreover, in the parameter space the boundaries separating the stable and unstable regions correspond to periodic solutions to the differential equation. In this classical problem we see that discreteness can be produced by stability and periodicity. This type of equations may be very useful in guiding the search for the appropriate problem to study. It may even happen that a set of chaotic attractors may be found by adding nonlinear driving and damping terms into Mathieu's equation.

A second point is that an investigation into relativistic chaos may provide some hints on how to construct the attractors. In fact, relativity is a more suitable theory for the microscopic world than classical nonrelativistic physics. Based on this, one may learn what type of terms must be added to nonrelativistic problems, hence form some semi-relativistic models. It may even happen that classically integrable systems can become nonintegrable in relativity and produce the attractors we are looking for.

3.4 Impact of Chaos on Scientific Methodology

In this section we will discuss some philosophical meanings of chaos and its impact on science in general. Many articles exist on this subject, so we shall concentrate on the questions which the author has been considering.

To physical scientists and mathematicians, both Hamiltonian chaos and attractors are important; whereas attractors and fractals are more familiar to the public and nonphysical scientists. One of the important impacts on scientific thinking is that very irregular behaviour may be produced by simple deterministic dynamics; therefore some phenomena previously regarded as random may in fact be governed by simple deterministic laws. Today, based on the embedding theorem, much effort has been directed to finding fractals and attractors from statistic data of economics, biology and other sciences; for if one can find an attractor then one probably has found a new natural law. However, the author's personal opinion on such researches is more conservative. Firstly, the behaviour of attractors is very sensitive to noise; it is not clear what will happen when attractors are modulated by noise. Therefore the present author favours a mathematical construction of attractors rather than looking for them from data. Secondly, random fractals also exist, which cannot be distinguished from an attractor by the currently available methods. In fact, the so-called phase portrait analysis on, say, the stock market data, almost always shows a structure like that of attractors; but a simple random time series can produce similar structures as well. Therefore the current research in these fields is questionable.

A second impact of chaos is that the discovery of chaotic attractors has changed the idea of stability and equilibrium. Conventionally, stability has always been defined in connection to equilibrium. Chaotic attractors provide a new type of stable behaviour and equilibrium. Since chaotic attractors are not smooth manifolds, the behaviour inside them can show much irregularity and divergence. Based on the conventional notion of stability and equilibrium, one is very likely to conclude an instability; but the system is at a stable equilibrium state, which cannot escape from the attractor. When the motion of the system in the attractor is very fast, one can observe a kind of statistical stable equilibrium.

Thirdly, we come to one of the most challenging questions, is nature deterministic or probabilistic? This question has been discussed not only in connection to statistical mechanics but to quantum mechanics. We do not attempt to give an answer here; but try to add some of the author's personal understanding of the problem which is related to chaos. From the progress on deterministic chaos, today 'deterministic' and 'predictable' have been distinguished from one another. The complete classical physics description of nature is that we can know everything exactly at least in principle, although a practical prediction is always limited. On the other hand, quantum mechanics implies that even nature itself does not know exactly its classical physics variables. If we combine the post-chaos classical physics and QM together to look at the macroscopic world then we see that the QM uncertainty can always act as the error of the system's states. Therefore, nature itself is ultimately probabilistic. What chaotic dynamics adds to the comprehension of nature is that even if without QM, the natural world is also practically probabilistic. Since such practical indeterminism is closer to the conventional notion of indeterminism other than conventional determinism, one may have to adopt the former notion as a principle.

It is also in order to make a comment on Einstein's belief that natural laws are simple. By saying this he does not mean that the solutions to the laws are simple, nor are the natural phenomena. Then the question arises how can simple laws be compatible with complicated phenomena. The present author thinks that the notion of nonlinearity, nonintegrability and chaos provides a possible answer. By a complicated world we mean that phenomenologically the variation of the variables describing the world are a mixture of simple quasiperiodic and complicated chaotic behaviours. If we stick to the historical linearity and integrability rule of natural laws, then we see that they cannot produce wild properties. Natural laws must be nonlinear to be simple; one cannot have both simple and beautiful laws, and at the same time simple solutions as well.

However, there are several points that need particular mention. Firstly, one can argue that complicated behaviour may still be produced by laws equivalent to linear laws, if the transformation is complicated. This is not really true; because however complicated a transformation is, quasiperiodic motion can only be transformed to quasiperiodic motion, which is unlikely to be a good description of natural phenomena. The second point is that it remains to verify that nature does admit laws. This is a very fundamental assumption of science which is not appropriate to discuss here.

Finally, if laws are nonlinear, nonsolvable, then there is a question on how to test the laws. Laws are usually purified and abstracted from the observation on a very small set of natural phenomena; then they are solved for a larger set of cases and compared with reality, thus the laws are conventionally regarded as having been tested. However, if most cases of the laws cannot be solved (eg. Newtonian mechanics, relativity, quantum mechanics), then can one still believe that the laws are truth? In particular, for laws in the form of ODEs, how can we decide whether the laws are true in chaotic region? It seems to me that all these questions need further investigation.

Table 3.1 A Selected Chronology of the Three Revolutionary Physical Sciences

Year	**Persons and Events**
1638	Galileo: *Two New Sciences*
1675	Newton: corpuscular theory of light
1678	Huygens: wave theory of light
1687	Newton: *Principia*
1850	Clausius: second law of thermodynamics
1860	Maxwell: speed distribution law
1865	Clausius: entropy
1864	Maxwell: *Dynamical Theory of Electromagnetic Field*
1877	Boltzmann: entropy <=> thermodynamic probability
1892	Poincare: *New Method of Celestial Mechanics*
1896	Zeeman (& Lorentz): Zeeman effect
1901	Planck: quantum theory of radiation and h.
1904	Whittaker: *Analytical Dynamics*
1905	Einstein: photo-electric effect, Brownian motion Einstein: special relativity

1908	Minkowski: geometrical interpretation of special relativity
1911	Rutherford: atomic nucleus First Solvay Conference on physics
1912	Poincare: died Bohr: quantization of hydrogen atom
1913	Stark: Stark effect Second Solvay Conference on physics
1916	Einstein: general relativity Schwarzschild: Schwarzschild metric
1918 1922	Duffing: Duffing's equation van der Pol: van der Pol equation
1924	Compton: Compton effect Born: *The Mechanics of the Atom,* German edition Fourth Solvay Conference on physics
1925	de Broglie: matter wave Born & Jordan: quantum mechanics
1926	Schrodinger: wave mechanics Born: probability interpretation of wave function
1927	Heisenberg: uncertainty principle Born: *The Mechanics of the Atom,* English translation Birkhoff: *Dynamical Systems* Fifth Solvay Conference on physics
1929	Hubble: Hubble's law
1954	Kolmogorov: conservation of invariant tori Siegel: generic divergence of transformation
1955 1956	Born: *Is Classical Mechanics Deterministic ?* Born: *Physics of My Generation*
1962	Arnold: conservation of invariant tori
1963	Lorenz: strange attractor Moser: conservation of invariant tori
1964	Henon & Heiles: Chaos of Henon-Heiles Hamiltonian Penzias & Wilson: 3K microwave background Shapiro: fourth test of general relativity Kerr: Kerr metric
1967	Heisenberg: *Nonlinear Problems in Physics*

CHAPTER 4

First Integrals, Possible Motion and Hierarchical Stability of the Classical N-Body Problem

In this and the following chapters we will concentrate on gravitational few- and many-body problems and study a specific type of stability of their motion, namely, hierarchical stability (shorthand as HS) defined by Walker & Roy (1983). The definition of this concept is already given in chapter 1.

There are several reasons for studying hierarchical stability. The first one has already been discussed in chapter 1, which is based on the practical observation that most motions of the heavenly bodies are found to be ordered. The second reason is due to the progress made in the understanding of the general 3-body problem since the 1970's in Europe and America. Articles are numerous in this direction, including for example Easton (1971), Smale (1970), Marchal & Saari (1975), Bozis (1976) and Zare (1976, 1977).

It is well known that the circular restricted 3-body problem possesses the well-known Jacobian integral (Szebehely, 1967; Roy, 1982), which is positive definite in the velocities relative to the rotating frame. This integral thus leads to the Hill **zero-velocity surfaces**, which divide the physical space into possible and impossible regions of motion, such that when the Jacobian integral constant is sufficiently large the motion of the infinitesimal mass is restricted to lie within some disjoint Hill regions. This is called the **Hill stability**, and the Earth-Moon-Sun system is stable in this sense. These Hill surfaces are also called the **Roche lobe** in the field of close binary stars, where one is interested in the motion of fluid in the field of two stars (modelled as point masses) circulating each other (Pringle & Wade, 1985).

However, even in the slightly more complicated elliptical restricted 3-body problem, no integral exists, thus Hill stability cannot be strictly established in this case, although the 'Jacobian integral' is shown to be only slowly changing with time. Based on this 'adiabatic' feature, Hill stability has been generalised to a finite time scale (Roy & Steves, 1988).

A real break-through was not made in Europe and America until the work of Easton (1971), Smale (1970) and Marchal & Saari (1975). Geometrical results having dynamical value similar to that of the circular restricted problem were found to exist in the general 3-body problem, where the integral playing the role of the Jacobian integral is a combination of the total energy and angular momentum integrals; moreover, the determination of the Hill-type surfaces is essentially based on Sundman's inequality.

The purpose of this chapter is to review this advance in the field of 3-body problem and to generalise and improve the current results; in doing this the approaches that could be applied to relativistic problems are discussed in more detail. In section 4.1, the relevant formulation of the classical N-body problem is summarised in a form favourable to the problems to be studied in the following sections and chapters. In section 4.2, in order to facilitate the present study we choose, among the existing literature, to review in more detail Sundman's inequality method used by Marchal & Saari (1975), later simplified by Saari (1976), and the canonical transformation method used by Zare (1976). The formally different results of the two methods are shown to be equivalent in this section. Attention is also paid to a related open question, namely, the central configuration. In section 4.3, we shall deduce some stronger inequalities for the 3-dimensional motion of the general 3-body problem. This work was originally carried out by the present author in 1987 by modifying Zare's transformation method; however, when equation (4.13) was arrived at, Saari's (1987) work appeared in the journal Celestial Mechanics, a work which dealt with a broader class of problems, namely, the flat N-body problems. Therefore, his approach shall be adopted in the deduction following this equation, but his deduction is shortened and made more apparent by noting its relation to the well-known moment of **inertia ellipsoid**.

Finally, we must mention that the so-called Hill-type stability was actually obtained earlier in Russia by Golubev (1968) following the most efficient Sundman's inequality method (see also Saari, 1976). However, this work was almost isolated and the works produced in the West were completely independent until the first mention of it by Szebehely & Zare (1976). It seems that Golubev's work did not even lead to any further work in Russia. In fact, efforts towards a generalised Hill-type trapped motion were already clearly made by Poincare (1892, Vol. 3, Chap. 26, Sec. 301). He was obviously on the right track not only by using the complete Sundman's inequality but also dropping terms as Golubev (1968) and Saari (1976) did, that is, using the inequality $H-U \geq C^2/2I$. However, since he pursued too much mathematical rigour rather than a pictorial approach, he did not obtain the inequality $IU^2 \geq -2C^2H$, which is the key equation for the break-through and is almost an automatic outcome of the former inequality.

The inequality $IU^2 \geq -2C^2H$ has been called Easton's inequality by some celestial mechanists, however, this inequality already appeared earlier in Golubev's work. It is also important to keep in mind that a Hill-type stability does not establish any trapped motion in the original sense of the Hill stability, since escape is not precluded.

4.1 Standard Results in the N-Body System

The N-body problem is usually studied in different coordinates depending on ones objectives. Here we shall summarise the formulation of the problem in an inertial frame and the barycentre frame (barycentre not necessarily at coordinate origin), using Cartesian coordinates, Jacobian coordinates and relative coordinates as well. The main references are Wintner (1947) and Szebehely (1973). In either the inertial or barycentre frame we have the following coordinate-free equations,

The Virial Theorem $\ddot{I} = 4T + 2U$

Sundman's Inequality $C^2 + \frac{1}{4}(\dot{I})^2 \leq 2IT$

where I is the system's moment of inertia, U and T being the potential energy and kinetic energy respectively, and C the norm of the system's total angular momentum **C**. I and **C** must be calculated with respect to the same origin. A dot denotes the time derivative.

It is important to realise that the Virial theorem is only valid to a limited class of potentials of the specific problem. Whereas Sundman's inequality is completely independent of the actual system; it is solely a result of Newton's laws of motion. The term $(dI/dt)^2$ in Sundman's inequality is irrelevant, thus it will be dropped in the current study.

We shall always call the inequality without the $(dI/dt)^2$ term Sundman's inequality. Later an independent proof will be given and the inequality will be generalised in Appendix B so that it may be used in relativity.

Though the two equations are coordinate-free, the quantities I, T and **C** have different expressions in different coordinate systems. In what follows we shall give their formulae in the barycentre frame using Jacobian coordinates and relative coordinates

Jacobian Coordinates

The Jacobian vector coordinates are defined in chapter 1 for the classical N-body problem (see Fig. 1.4). Using these vectors we have, in a barycentric frame, the following expressions for the system's moment of inertia I, kinetic energy T, and angular momentum **C**, viz.

$$I = \sum_{i=2}^{N} \mu_i \rho_i^2 \ , \quad T = \frac{1}{2} \sum_{i=2}^{N} \mu_i \mathbf{v}_i^2 \ , \quad \mathbf{C} = \sum_{i=2}^{N} \mu_i \,\boldsymbol{\rho}_i \times \mathbf{v}_i \,. \qquad (4.1a)$$

where $\mu_i = M_{i-1}\, m_i / (M_{i-1} + m_i)$, with M_{i-1} being the sum of the first (i-1) masses. Moreover, in these equations, I and **C** must be calculated with respect to the system's centre of mass.

Relative Coordinates

The expressions for I, T and C in a barycentric frame may be formulated such that no absolute position and velocity vectors are involved, namely, only intrinsic relative position and velocity vectors appear, viz.

$$\left.\begin{array}{l} \sum m_i \mathbf{R}_i = 0 \\ \sum m_i x_i = 0, \ \sum m_i y_i = 0 \\ \sum m_i \mathbf{v}_i = 0 \\ (\sum m_i \mathbf{R}_i) \times (\sum m_i \mathbf{v}_i) = 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} I = \frac{1}{M} \sum\limits_{(i,j)} m_i m_j (\mathbf{R}_i - \mathbf{R}_j)^2 \\ I_z = \frac{1}{M} \sum\limits_{(i,j)} m_i m_j (x_{ij}^2 + y_{ij}^2) \\ T = \frac{1}{2M} \sum\limits_{(i,j)} m_i m_j (\mathbf{v}_i - \mathbf{v}_j)^2 \\ C = \frac{1}{M} \sum\limits_{(i,j)} m_i m_j (\mathbf{R}_i - \mathbf{R}_j) \times (\mathbf{v}_i - \mathbf{v}_j) \end{array}\right. \qquad (4.1b)$$

where (i, j) means all possible pairs without repetition. The proof of these equations is straightforward.

Moment of Inertia Ellipsoid

It is well-known that the Euler angles, inertia tensor and inertia ellipsoid are very useful in studying the motion of rigid bodies (Goldstein, 1980). We shall deduce some equations involving the elements of the moment of inertia tensor, which rely heavily on the inertia ellipsoid and will be used in later investigation of the spatial 3-body problem.

Saari (1987) obtained the best possible restrictions on the possible motion of the flat N-body problem by defining 'reference positions' and 'principal reference positions', whose physical meaning are clearer if they are related to the inertia ellipsoid (Saari did not point out this link). Saari's work will be interpreted here in connection to the inertia ellipsoid.

Consider N point masses which are distributed on a single O-xy plane. There is certainly a spatial inertia ellipsoid associated with this system, however, we are only interested in the **inertia ellipse** in the O-xy plane.

(1). Inertia Ellipse

The moment of inertia of the above system about an axis defined by the unit vector $\mathbf{n}=\{\cos\alpha,\cos\beta\}=\{\cos\alpha,\sin\alpha\}$ in the the O-xy plane is easily shown to be

$$I_n = I_{xx}\cos^2\alpha + I_{yy}\cos^2\beta - 2I_{xy}\cos\alpha\cos\beta \ .$$

If one defines a vector $\mathbf{R} = \{R_x, R_y\} = \mathbf{n} / \sqrt{I_n}$, then one sees that $\mathbf{R}$ defines an ellipse, ie. the inertia ellipse, in the O-xy plane by the equation

$$I_{xx} R_x^2 + I_{yy} R_y^2 - 2 I_{xy} R_x R_y = 1 \ ,$$

since the inertia tensor is positive definite.

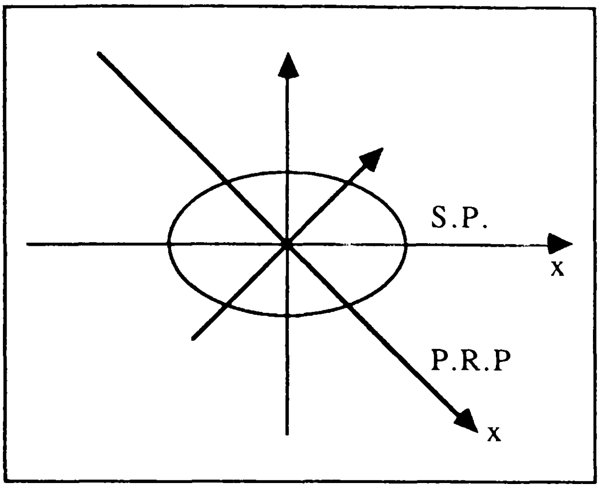


Fig. 4.1. Inertia ellipse, examples of P.R.P. and S.P.

According to Saari (1987), the system is at a **reference position** (R.P.) when the system's position in the coordinate system O-xy is such that $I_x = I_y$; whereas a reference position with $I_{xy} \geq 0$ is called a **principal reference position** (P.R.P.). When the system is rotated in the plane by -45^0 or -225^0 from a principal reference position, the system is said to be at a **standard position** (S.P.). The meaning of these concepts is made transparent if we look at them relative to the inertia ellipse. We have shown an example for each of these positions in Fig. 4.1. Obviously, there are usually four reference positions with $k\pi/2$ angle differences, out of which two are principal reference positions with $k\pi$ angle differences; moreover there are usually two standard positions with $k\pi$ angle differences, which are the positions when the long major axis of the inertia ellipse coincides with the Ox axis.

For such a planar mass distribution, a **configurational angle** can be defined at a principal reference position by $\alpha = \arccos (I_{xy} / I_x) \in [0, \pi/2]$. This is justified because by Cauchy's inequality (see Appendix B) we always have $(I_{xy})^2 \leq I_x I_y$, which at a principal reference position implies $I_{xy} \leq I_x$.

(2). Transformation of the Inertia Tensor

It is useful to investigate how the inertia tensor changes when a system is rotated in the plane by an angle ϕ with the coordinate system fixed. If the elements of the inertia tensor are denoted by I_x, I_y, and I_{xy} before the rotation, and by I'_x, I'_y, and I'_{xy} after the rotation of the system with respect to the coordinate system, then simple calculation shows

$$\left\{\begin{array}{l} I'_x = \sum m_i(x_i \sin\phi + y_i \cos\phi)^2 = I_x \cos^2\phi + I_y \sin^2\phi + I_{xy} \sin 2\phi \\ I'_y = \sum m_i(x_i \cos\phi - y_i \sin\phi)^2 = I_x \sin^2\phi + I_y \cos^2\phi - I_{xy} \sin 2\phi \\ I'_{xy} = \sum m_i(x_i \cos\phi - y_i \sin\phi)\,(x_i \sin\phi + y_i \cos\phi) \\ \quad = \frac{1}{2}\sin(2\phi)(I_y - I_x) + I_{xy}\cos 2\phi \\ I' = \quad I'_x + I'_y = I_x + I_y = I. \end{array}\right.$$

The advantage of defining principal reference positions is that any position of a system can be obtained by rotating the system an angle ϕ from a principal reference position, and the elements of the inertia tensor can be expressed as very simple functions of the rotation free quantity, $I_r \equiv I/2$, and the configurational angle α, viz.

$$\left\{\begin{array}{ll} I'_x = I_x + I_{xy} \sin 2\phi = I_r[1 + \cos\alpha \sin 2\phi] & \\ I'_y = I_x - I_{xy} \sin 2\phi = I_r[1 - \cos\alpha \sin 2\phi] & \\ I'_{xy} = I_{xy} \cos 2\phi \quad = I_r \cos\alpha \cos 2\phi & \\ I' = I = 2I_r\,. & \end{array}\right. \qquad (4.2)$$

(3). Equations Involving the Inertia Tensor

In the later development of the spatial 3-body problem, we will encounter the determinant of the inertia tensor with respect to the centre of mass, $I_x I_y - (I_{xy})^2$, which is the key link between our result and Saari's (1987). What we want to show here is that this function may be related to the area of the triangle formed by the three point masses. In general we have

$$I_xI_y - I_{xy}^2 = (\sum m_i x_i^2)(\sum m_i y_i^2) - (\sum m_i x_i y_i)^2$$

$$= \sum\sum m_i m_j x_i^2 y_j^2 - \sum\sum m_i m_j x_i y_i x_j y_j = \sum_{\substack{i,j \\ i \neq j}} m_i m_j x_i y_j (x_i y_j - x_j y_i)$$

$$= \sum_{(i,j)} m_i m_j (x_i y_j - x_j y_i)^2 = \sum_{(i,j)} m_i m_j (\mathbf{R}_i \times \mathbf{R}_j)^2 = 4 \sum_{(i,j)} m_i m_j S_{ij}^2$$

for any N point masses lying in a single plane, where S_{ij} is the area of the triangle formed by Om_im_j, and the summation notation

$$\sum_{(i,j)} = \frac{1}{2} \sum_{i} \sum_{j \neq i}$$

means that the sum is taken over all possible pairs without repetition of indices.

If only three bodies are involved and the origin is set at the centre of mass, then we have $m_im_jS_{ij}=m_1m_2S_{12}$ for any pair (i j | i≠j), and $S_{ij}=m_kS/M$ for any triplet (i j k | i≠j≠k≠i). These are easily shown by taking the cross product of the equation $\Sigma m_i\mathbf{R}_i=0$ with every $m_i\mathbf{R}_i$. Substituting these into the above result we obtain

$$I_xI_y - I_{xy}^2 = 4S^2 m_1 m_2 m_3 / M \geq 0 \qquad (4.3)$$

for three point masses lying in the plane O-xy, with the origin at the barycentre of the system.

4.2 Sundman's Inequality, Possible Motion and Central Configuration

There are many approaches used to establish the restriction on possible motion of dynamical systems by first integrals. A well-known method is the so-called effective potential method, which was generalised by Zare (1976) to a class of Hamiltonian systems with a positive definite property and applied to the coplanar 3-body problem. In principle, his method is complete and widely applicable. However, in this section we will concentrate on a simpler inequality method based on a direct use of Sundman's inequality to demonstrate the meaning of the Hill-type stability of the general 3-body problem obtained in this way (Golubev, 1968; Marchal & Saari, 1975; Saari, 1976). This method can establish equally good results, and shows its advantage in dealing with more complicated systems.

It turns out that the critical configurations for the Hill-type stability is exactly the same as the so called central configurations (Wintner, 1947), the determination of which is still an open question for systems with more than three bodies. The recent advances in this direction are also reviewed.

Finally we will compare the bounded motions of some classical systems and show their relation to Sundman's inequality. In doing this we try to show the generality of the inequality method, supporting its applications to relativistic problems in later chapters.

Sundman's Inequality and Possible Motion

Sundman's inequality was introduced to study collisions in the classical N-body problem (see Wintner, 1947). For the purpose of studying forbidden motion the term involving dI/dt can be dropped from the original expression, giving a weaker but still useful inequality. Hence we write,

$$C^2 \leq 2\,IT \qquad (4.4\,a)$$

where C is the value of the total angular momentum vector C, T the kinetic energy and I the moment of inertia. This equation will be called **Sundman's inequality** in the present approach. The proof of it is straightforward by utilising Cauchy's inequality, equation (B2) in Appendix B, viz.

$$\begin{aligned} C^2 &= \left|\sum(m_i \mathbf{R}_i \times \mathbf{V}_i)\right|^2 \leq \left(\sum\left|m_i \mathbf{R}_i \times \mathbf{V}_i\right|\right)^2 = \left(\sum m_i R_i V_i \sin\theta\right)^2 \\ &\leq \left\{\sum_i\left(\sqrt{m_i R_i^2}\sqrt{m_i V_i^2}\right)\right\}^2 \leq \left\{\sum_i m_i R_i^2\right\}\left\{\sum_i m_i V_i^2\right\} = 2IT. \end{aligned}$$

Since the total energy, H, of the N-body system is the sum of the potential energy, U, and the kinetic energy, the Sundman's inequality (4.4a) can be written as

$$C^2 \leq 2\,I(H-U) \quad \text{or} \quad -H+U+\frac{C^2}{2I} \leq 0. \qquad (4.4\,b)$$

If the total energy of the system is negative, then we can further perform the following deduction almost automatically,

$$\begin{aligned} -H+U+\frac{C^2}{2I} \leq 0 \quad &\Leftrightarrow \quad \sqrt{IU^2} \geq \frac{C^2}{2\sqrt{I}} - H\sqrt{I} \\ &\Rightarrow \quad Z(\Delta) \equiv IU^2 \geq -2\,C^2H \qquad (4.4c) \end{aligned}$$

where $Z(\Delta)$ is a scale-free function of positions only, C^2H is scale-free as well. For a 3-body system $Z(\Delta)$ is a function of the shape of the triangle formed by the three mass points. From the proof, one sees that these equations are very general; they are not only valid to an isolated system, but also to a subsystem. Moreover, the systems are not

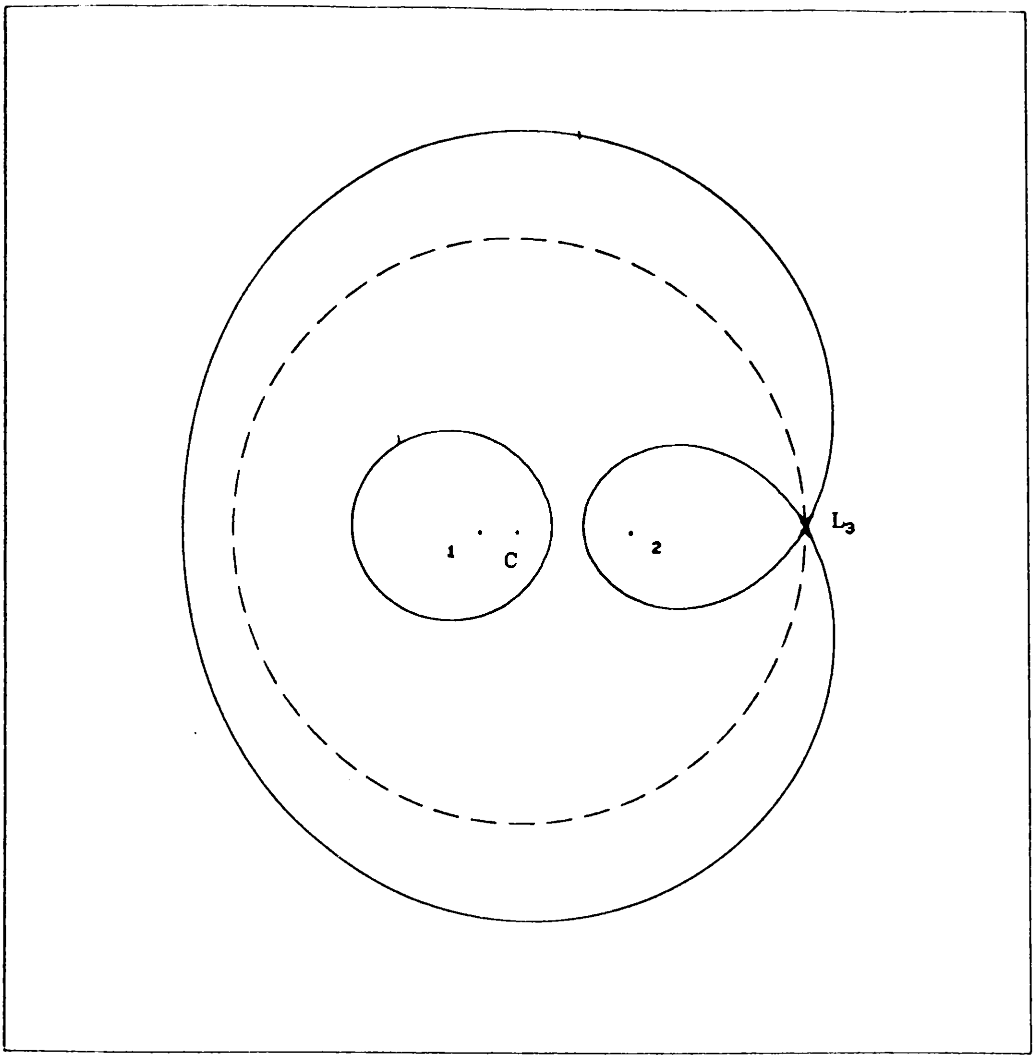


Figure 4.2 The critical contour of Z for m_1=0.75, m_2=0.25 and m_3=1.

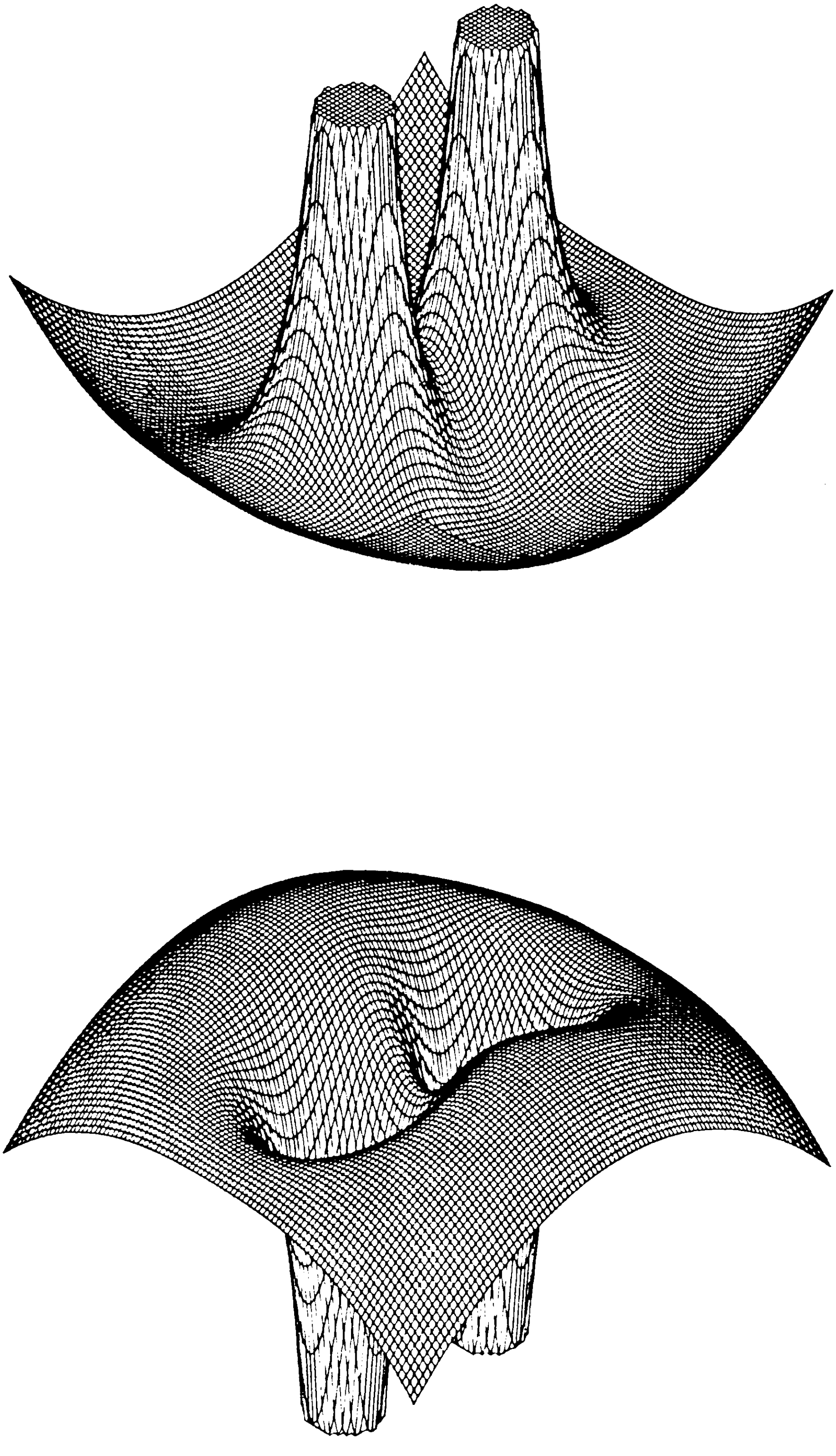


Figure 4.3 The surface defined by the function $Z(X, Y)=IU^2$ for $m_1=0.505$, $m_2=0.495$ and $m_3=0.1$. The mirror image of the surface is shown to display the detail of the critical points.

necessarily point mass N-body systems, they can be rigid body systems or a system of continuous medium. There are no requirements on the actual action force or potential. The quantities H and C are not necessarily conserved.

Equations (4.4b) and (4.4c) can be used to obtain some interesting and useful forbidden motions valid for all time if H and C are conserved quantities, or for some finite time scale if H and C are changing slowly with time.The basic idea is that the combination of the conserved (or slowly changing) quantities on the right side of equation (4.4c), C^2H, imposes restrictions on the function, $Z(\Delta)$, of the positions of the bodies on the left hand side, hence on the possible configurations the bodies could form.

We also note that inequality (4.4b) (alternatively (4.4a)) is the relation utilising both energy and angular momentum integrals; it is stronger than the relation $H\text{-}U\geq 0$, which only uses the energy integral. However, without more information about the specific problem it is difficult to determine which of the two inequalities, $H\text{-}U\geq 0$ or (4.4c), is the stronger and therefore the better one to adopt.

Applying equation (4.4c) to the 3-body system in its barycentre system, we obtain the Hill-type regions shown in Fig. 4.2. More specifically, the restrictions on possible configuration by equation (4.4c) can be studied by selecting two of the particles to define a reference line and unit of distance, then seeking the possible relative positions of the remaining body. In this way, one finds that, based on the level sets of $Z(\Delta)$ (see Fig. 4.3), there exist triply connected forbidden regions in the scaled physical space if the values of C^2H satisfy the relation $C^2H \leq (C^2H)_c$. The critical value of $(C^2H)_c$ is equal to that of $-Z/2$ estimated at one of the critical points of Z. These critical points are called **critical configurations**, while this kind of stability is termed the **Hill-type stability**.

However, in order to show the relation to the so-called central configurations, we will call all critical points of the function Z critical configurations of the N-body problem.

Obviously, the result is not only true for coplanar 3-body problems but also for spatial 3-body problems. Thus, the **Hill curves** are on the invariable plane for the coplanar problem; and the same Hill curves are on the instantaneous plane defined by the three mass points for the spatial problem. Moreover, if the mass parameters are fixed, the critical configurations and critical values of $(C^2H)_c$ are the same for the spatial and coplanar problems. Thus finally one obtains the **Hill surfaces** for the spatial 3-body problem by rotating the Hill curve found on the invariable plane in the coplanar case around the reference line determined by two of the mass points. This result will be improved in the next section.

It is worth pointing out that while the critical configurations of the coplanar 3-body problem correspond to the important particular solutions of the system (called homographic solutions), namely, Lagrange's equilateral triangle solutions and Euler's collinear solutions, the same critical configurations in the spatial problem are not necessarily related to any real motion; they are merely some special geometrical arrangements of the positions of the masses. Thus, in order to avoid confusion, when necessary the critical configurations will be called Lagrange's or Euler's configurations (or points) instead of solutions. Similarly, the central configuration is also just a concept of statics. However, these configurations are all related to each other and have dynamical value.

Marchal & Saari (1975) studied the Hill-type curves by a direct use of Sundman's inequality. However, they used the mean quadratic distance σ and the mean harmonic distance ν, which are respectively defined by $M^*\sigma^2=MI$ and $M^*\nu^{-1}=-U/G$, where $M^*=\Sigma m_i m_j$. Obviously their essential equation $\sigma/\nu \geq (-2C^2HM/G^2M^{*3})^{1/2}$ is equivalent to ours, ie. $I^{1/2}(-U) \geq [-2C^2H]^{1/2}$. Here the gravitational constant G is retained.

The notion of Hill-type stability of the general 3-body problem has been regarded as a break-through in celestial mechanics, because it guarantees the hierarchical stability condition HS-(C) defined by Roy & Walker (1983) (cf. chapters 1 and 5). However, its theoretical value must be balanced by the fact that unlike the Hill stability of the restricted problem, a Hill-type stability does not preclude escape of one body nor binary collisions. Next, similar geometrical results have been established for a 3-body system with non-negative total energy (Marchal & Bozis, 1982). Moreover, Hill-type stability cannot be established in the same way for systems with more than three bodies (Milani & Nobili, 1983, 1985); nor for the problem of the circular restricted 2+2 bodies (Milani & Nobili, 1988).

Nevertheless, the notion is important in application because real systems are usually approximated by a 2-body model, and a 3-body model will be a better approximation. If the total energy is negative, then at most one body can escape to infinity. As a result of this, the inner binary cannot escape if the system is Hill-type stable. Moreover, if the system is Hill-type stable, and the total angular momentum is not zero, then the outer mass cannot collide with either of the inner masses. We will see in the next chapter that the outer mass is always bounded out of the circle spanned by the line joining the two inner masses.

Critical Configuration and Central Configuration

The significance of central configurations is that a very important class of particular solutions, namely, the homographic solutions, to the classical gravitational N-body problem is related to it. We will show that the central configurations are exactly the critical configurations of interest here. Moreover, the determination, or even the counting, of such configurations is still a fascinating but unsolved question. We will review briefly the recent advances made in this field and obtain two simple theorems as an outcome of defining central configurations in different coordinates. The dynamical meaning of one of the theorems still needs further exploration.

(1). Central Configuration in Barycentre System

Following Wintner (1947), a central configuration may be defined, in a barycentre system, as:

The N position vectors $\mathbf{R}_i$ of the N bodies m_i will be said to form a central configuration with respect to the N fixed positive mass parameters, if the force of gravitation acting on m_i at the moment of the given configuration is proportional to the mass m_i and to the barycentric position vector $\mathbf{R}_i$, ie. if the set of equations

$$\mathbf{F}_i = -\nabla_{\mathbf{R}_i} U = -\sigma m_i \mathbf{R}_i \qquad (i=1, \ldots, N) \qquad (4.5a)$$

hold for some scalar σ which is independent of i.

In fact, since $\sigma \Sigma m_i (R_i)^2 = \Sigma < \mathbf{R}_i , \nabla_{\mathbf{R}i} U > = -U$, the value of σ is uniquely determined by $\sigma = -U/I$, where U is the potential energy of the system, and I is the system's moment of inertia with respect to the centre of mass. Moreover, equation (4.5a) implies that $\Sigma m_i \mathbf{R}_i = 0$, if we notice that for a self gravitational N-body system, $\Sigma \nabla_{\mathbf{R}i} U = \Sigma \mathbf{F}_i = 0$. This implies that not all of the equations are independent.

By the Lagrange multiplier theorem, it is evident that equation (4.5a) may be interpreted as the critical points of U with the constraint I=const. Thus equation (4.5a) is equivalent to

$$\nabla_{\mathbf{R}_i} U = 0 \quad , \quad I = \text{const.} \qquad (i=1, \ldots, N) \qquad (4.5b)$$

The central configurations may be further equivalently interpreted as the critical points of the function $Z = IU^2$, namely, the critical configurations of the Hill-type stability, viz.

$$\nabla_{\mathbf{R}_i} IU^2 = 0 \qquad (i = 1, \dots, N) \qquad (4.5c)$$

if one notices the following relations

$$\nabla_{\mathbf{R}_i} IU^2 = U^2 \nabla_{\mathbf{R}_i} I + 2IU\nabla_{\mathbf{R}_i} U = U^2 2m_i\mathbf{R}_i + 2IU\nabla_{\mathbf{R}_i} U \ .$$

It is clear that the notion of a central configuration determined by the three equivalent equations (4.5a, b, c) is independent of the orientation of the barycentric coordinate system and the unit of length. Correspondingly, the class of central configurations which can go into each other through a rotation or scaling will be considered as identical. Because of this, one can always find the central configurations of the N-body system by solving only (N-1) equations for the (N-1) unknown **R**'s, if one chooses the origin of the coordinates at the barycentre and arbitrarily fix one of the masses, say m_N, at the position (1, 0, 0). Therefore, only (N-1) of the N equations are independent. Thus central configurations can be equivalently defined by

$$\mathbf{F}_i = -\nabla_{\mathbf{R}_i} U = -\sigma m_i \mathbf{R}_i, \quad \sigma = -U/I \quad (i = 1, \dots, N-1) \qquad (4.5d)$$

$$\nabla_{\mathbf{R}_i} IU^2 = 0 \qquad (i = 1, \dots, N-1) \qquad (4.5e)$$

These two equations are a result of the above symmetry and scaling argument. In addition, one can deduce equations (4.5a) and (4.5c) from these two equations by a use of the equation $\Sigma m_i\mathbf{R}_i = 0$. The set of equations (4.5d) allows a simple and interesting interpretation stated in the following theorem:

Theorem 4.1. For an N-body system, if the resultant gravitational forces on (N-1) of the masses satisfy equations of the form (4.5d) or (4.5e), then so does the resultant force on the remaining mass.

(2). Central Configuration in an Arbitrary Coordinate System

Similarly one can define the central configurations in an arbitrary coordinate system by the set of equations

$$\mathbf{F}_i = -\nabla_{\mathbf{R}_i} U = -\sigma m_i(\mathbf{R}_i - \mathbf{A}) \qquad (i = 1, \dots, N) \qquad (4.6a)$$

Based on the same arguments as before, σ is uniquely determined by σ=-U/I, where U is the potential energy of the system, and I again is the system's moment of inertia with respect to the centre of mass. Equation (4.6a) also implies that $\mathbf{A} = \Sigma m_i\mathbf{R}_i / M$, thus **A** is necessarily the system's centre of mass and not all of the equations are independent. Similarly, equation (4.6a) is equivalent to the following equations

$$\nabla_{\mathbf{R}_i} U = 0 \quad , \qquad I = \text{const.} \qquad (i = 1, \ldots, N) \qquad (4.6b)$$

$$\nabla_{\mathbf{R}_i} I U^2 = 0 \qquad (i = 1, \ldots, N) \qquad (4.6c)$$

where I is the system's moment of inertia relative to the barycentre rather than the coordinate origin.

However, based on the translational, rotational and scaling symmetries of the system, one sees that only (N-2) of the equations are independent. Thus one can find the central configurations by selecting a coordinate system such that two of the masses, say m_N and m_{N-1}, are located at the positions (0, 0, 0) and (1, 0, 0) respectively, then solving the following set of (N-2) equations

$$\mathbf{F}_i = -\nabla_{\mathbf{R}_i} U = -\sigma m_i (\mathbf{R}_i - \mathbf{A}) \ , \qquad (i = 1, \ldots, N-2) \qquad (4.6d)$$

$$\text{where} \quad \sigma = -U/I \ , \quad \mathbf{A} = \sum m_i \mathbf{R}_i / M$$

$$\nabla_{\mathbf{R}_i} I U^2 = 0 \qquad (i = 1, \ldots, N-2) \qquad (4.6e)$$

Equation (4.6d) now allows the following interpretation,

Theorem 4.2. For an N-body system, if the resultant gravitational forces on (N-2) of the masses satisfy equations of the form (4.6d) or (4.6e), then so do the resultant forces on the remaining two masses.

This theorem was obtained by the symmetry arguments, thus one must be able to deduce equation (4.6a) or (4.6c) from equation (4.6d) or (4.6e) by a use of the equations $\Sigma \mathbf{F}_i = 0$ (or equivalently, $\Sigma m_i \mathbf{R}_i = M\mathbf{A}$) and $\Sigma(\mathbf{R}_i \times \mathbf{F}_i) = 0$. However, the author has not been able to verify this point. This is desirable for a future work, because if it were true, then it reveals the dynamical meaning of the theorem.

For the 3-body problem, one can verify the theorem after the central configurations are all found. However, for systems with more than three bodies, the central configurations has not been solved; in fact even the question of counting the number of central configurations is still open. Therefore, if the above suggestion cannot be proved, then the theorem raises a problem related to the more general open question.

(3). Recent Advances

Here we will summarise very briefly several points of the recent progress made in counting the number of central configurations. But first let us clarify several concepts needed in studying critical points of functions in general (definitions may be found from Poston & Stewart, 1978).

A critical point is isolated if there is no other critical point in a sufficiently small neighbourhood of the point. A critical point of a function is nondegenerate (degenerate) if the Hessian matrix of the function is nondegenerate (degenerate) at the point. While nondegenerate critical points are always isolated, the converse is not true. Moreover, nondegenerate critical points are structurally stable, and degenerate critical points are not structurally unstable. One of the difficulties met in counting the number of critical points is that degeneracy can produce infinitely many critical points, thus the more mathematically involved Morse theory of critical points and measure theory must be invoked.

It was already shown by Moulton (1910) that for an N-body system, there are $N!/2$ collinear central configurations. It is suggested by Wintner (1947) that the largest contribution to the number of central configurations is due to the collinear configurations. This is true for $N=3$ but false for $N\geq 4$, because Palmore (1973) showed that when the mass parameters are such that all central configurations are nondegenerate, then the minimum number of planar central configurations equals $(3N-4)[(N-1)!/2]$. In the nondegenerate 4-body cases, this minimum estimation gives 24 planar central configurations, among them 12 are collinear. However, it is shown in the same article that the system of four equal masses has a total number of 120 nondegenerate central configurations. Therefore, the contribution of noncollinear planar central configurations spectacularly exceeds that of the collinear ones.

It is a classical fact that the 3-body problem has only five central configurations; all of them are isolated and nondegenerate critical points (see Wintner, 1947). Degeneracy already happens in the case of 4-body problems, and an example is given by Palmore (1975). It is easy to show that the configuration with three unit masses at the vertices of an equilateral triangle and the fourth mass at the centre of mass of the first three masses is a central configuration; by a direct calculation one can show that the Hessian matrix is degenerate when the value of the fourth mass is $(2+3\sqrt{3})/(18-5\sqrt{3})<1$. Examples of degeneracy can be constructed for any $N>4$ in a similar way.

Palmore (1975) also proved the following conjecture of Smale (1970): for almost all (in the sense of Lebesgue measure) mass parameters $m=\{m_1 ,..., m_N\}$ there are only a finite number of planar central configurations; and they are all nondegenerate. Moreover, Palmore (1976) proved that in the N-dimensional mass parameter space ($N\geq 4$), the set of masses giving degenerate central configurations has Lebesgue measure zero, but positive k-dimensional Hausdorff measure, with $0 \leq k \leq N-1$. These results partially answered a question raised by Wintner (1947), namely, is the number of central configurations finite or infinite?

Trapped Motion of Various Few-Body Problems

It is well-known that bounded motions can be found by the effective potential method, or after the problem has been solved in closed form. Examples of the first case may be found from standard textbooks (eg. Fetter & Walecka, 1980; Goldstein, 1980); in fact, the circular restricted problem is a well-known nontrivial example. Well-known examples of the second case are the 2-body problem and the 2-centre problem (Born, 1927).

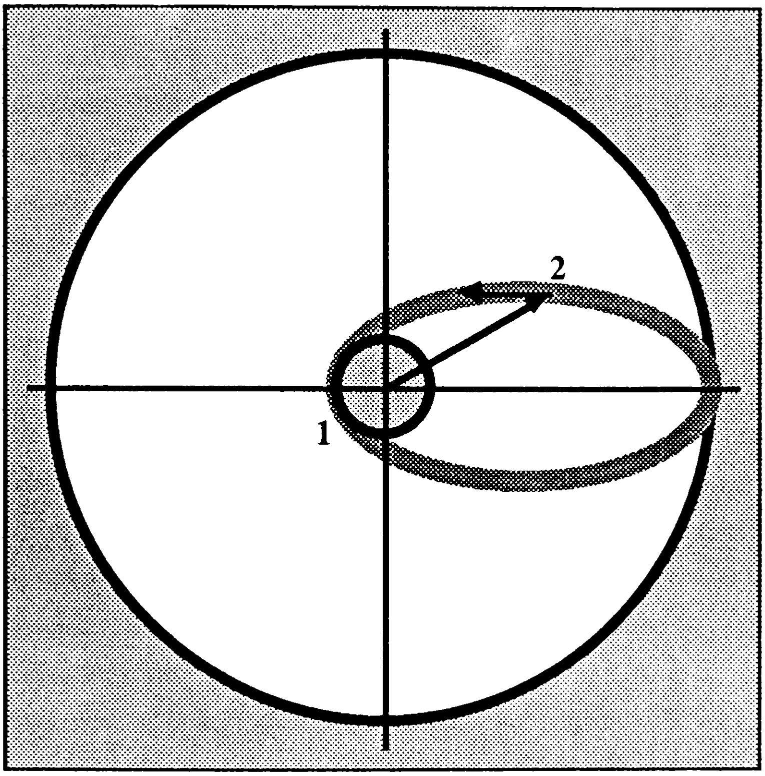


Fig. 4.4 The Kepler motion is bounded between the pericentre and apocentre distance by Sundman's inequality, equation (4.4b).

It is interesting to note that most of these standard results may be obtained as direct outcomes of Sundman's inequality by a use of equations (4.4b) or (4.4c); we favour this inequality method because it is more general. For example, it is well known that the motion of the two-body problem (with negative total energy) is limited between apocentre and pericentre distances and this is the best possible result since one can find this result after the problem is solved. Exactly the same result may be obtained by applying equation (4.4b) in the barycentre system. One can easily verify that the motion is possible only if the distance between the two bodies is limited in the region a(1-e) and a(1+e), with $a=-m_1m_2/2H$ and $(1-e^2)= -2C^2H(m_1+m_2)/(m_1m_2)^3$ (see Fig. 4.4).

For more applications of Sundman's inequality in studying bounded motion, see for example, Chapsiadis et al (1988), Sergysels (1988) and Veres (1989).

4.3 A Stronger Inequality for the Spatial 3-Body Problem

As has already been mentioned, many authors have made the approach of establishing Hill-type stability for the general 3-body problem. Although their methods of approach differ, their results did not go beyond that which can be obtained by a direct use of Sundman's inequality until the work of Saari (1987). Moreover, some of the approaches merely reproduced the very general Sundman's inequality in certain rather limited cases. For example, Zare (1976) only applied the (extended point) transformation method to the coplanar 3-body system, and obtained equation (4.8a) which will be shown in this section to be equivalent to Sundman's inequality. Therefore there is a need to extend some of the approaches to more general cases, and to work for stronger results.

Among the methods used so far, two of them need particular mention for the interests of this thesis, namely, direct Sundman's inequality method used by Golubev (1968) and Saari (1976) and the canonical transformation method used by Zare (1976). The first method was reviewed in the last section. Because of its simplicity and general applicability, it will be generalised in Appendix B and used later to study relativistic problems. However, Sundman's inequality and those established in Appendix B are not the best possible ones, they still allow improvements.

The second method is a modified version of Zare's approach, which was originally limited to extended point transformations so as to preserve the positive definite property of the Hamiltonian with respect to all generalised momenta. The modified method will be used here to study the spatial 3-body problem. In principle, the canonical transformation method is a complete method since no information is lost in such reductions. However, when the number of bodies involved increases, the amount of algebra needed for carrying out the calculation would become too large. For example, it would be rather hopeless should one try to deduce Sundman's inequality for the N-body problem following Zare's procedure.

In this section the modified transformation method will be applied to study the spatial motion of the 3-body problem. In doing this it is found that a similarity exists in the form of the Hamiltonians of the spatial and coplanar problems, which is a reflection of the fact that three points always lie in a plane. It is exactly due to this property that Zare's results of the coplanar problem can be used to simplify the investigation here. Inequalities stronger than Sundman's have been found.

The results were also obtained before the present approach by Saari (1987), who studied a more general class of problems, ie. the flat N-body problem (3-body is always flat). His work also shows some connection between the Virial theorem and Hill-type surfaces through the so-called rigid motion (see also Palmore, 1979). However, the present approach was independent of Saari's. The objective of the present author was limited in developing a more general inequality method to study bounded motion in both classical and relativistic gravitational systems, with the transformation method as an aid (since it is in principle complete). Saari's (1987) work was published when equation (4.13) was obtained, so the development after that equation is mainly due to him.

Equivalence of Zare's Result with Sundman's Inequality

Zare (1976) established restrictions of first integrals on possible motion for dynamical systems possessing time-independent Hamiltonians or systems reducible to that form using only extended point transformations. The method depends on the positive definiteness of the Hamiltonian in all generalised momenta, and that this property is preserved by extended point transformations. The study is in fact a generalisation of the so-called effective potential method often encountered when a rotating frame is used.

The method was applied to the coplanar general 3-body problem by Zare and derived the Hill-type curves which may be used to make statements concerning possible configurations of the three bodies applicable for all time. He obtained a reduced Hamiltonian by using the extended point transformations discussed in detail by Whittaker (1904). In spite of a slight formal difference, the Hamiltonian appearing in Zare's paper is equivalent to that of Whittaker (1904, section 161). In order to facilitate our study on the 3-dimensional problem (Whittaker also gives the transformations, which are all extended point transformations), we shall write down the final Hamiltonian, taken from Whittaker, viz.

$$\begin{aligned} H= & \left[\frac{1}{2m_1}+\frac{1}{2m_3}\right]\left[P_1^2+\frac{1}{Q_1^2}(P_3Q_2-P_2Q_3-C)^2\right] \\ & +\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](P_2^2+P_3^2) \\ & +\frac{1}{m_3}\left[P_1P_2-\frac{P_3}{Q_1}(P_3Q_2-P_2Q_3-C)\right]+U \qquad (4.7) \end{aligned}$$

where m_1, m_2 and m_3 are the masses, H and C are the total energy and angular momentum integrals with respect to the system's centre of mass. Q_1 is the distance of m_3m_1, Q_2 and Q_3 are the projections of m_3m_2 on and perpendicular to m_3m_1, Q_4 the angle between m_3m_1 and the x-axis fixed in space through the centre of mass. Furthermore, P_1 is the component of the momentum of m_1 along m_3m_1, P_2 and P_3 are the components of the momentum of m_2 parallel and perpendicular to m_3m_1, and P_4 is the total angular momentum of the system with respect to the centre of mass. U is the potential energy, which is a function of the generalised coordinates Q's only.

By introducing the variables X, Y, Z and η, where

$$\eta \equiv X + iY \equiv \frac{Q_2}{Q_1} + i\,\frac{Q_3}{Q_1} \quad , \quad Z = Q_1 \quad , \quad i = \sqrt{1}$$

and solving the set of linear equations $\partial H/\partial P_i = 0$ (i=1, 2, 3) he obtained the unique solution for the P's expressed in the above variables, viz.

$$P_1 = \frac{C\Psi Y}{Z} \; , \quad P_2 = -\left[1 + \frac{m_3}{m_1}\right]\frac{C\Psi Y}{Z} \; , \quad P_3 = \left[1 - \left(1 + \frac{m_3}{m_1}\right)X\right]\frac{C\Psi}{Z}$$

where $\Psi = \Psi(m, X, Y)$.

Finally, by using the positive definite feature of the Hamiltonian with respect to the momenta P's and substituting the above equations of P's into equation (4.7), he obtained the inequality governing the regions of possible motions, namely,

$$F(\eta, Z) = HZ^2 + Gb(\eta)Z - \frac{C^2M}{2}c(\eta) \geq 0 \qquad (4.8a)$$

where

$$M = m_1 + m_2 + m_3$$

$$b(\eta) = m_1m_2|\eta - 1|^{-1} + m_2m_3|\eta|^{-1} + m_3m_1$$

$$c(\eta) = \left[m_1m_2|\eta - 1|^2 + m_2m_3|\eta|^2 + m_3m_1\right]^{-1} .$$

This equation defines regions of possible and impossible motions in the 3-dimensional space O-XYZ whose boundaries are given by a quadratic equation in Z. If the total energy is negative, the projection of the regions of possible motions on the complex η-plane establishes all possible configurations independent of the scale and is given by

$$\Delta(\eta) = G^2 b^2(\eta) + 2\,MC^2 H c(\eta) \;\geq 0. \qquad (4.8b)$$

The critical configurations are defined as the singularities of the manifold F(η, Z)=0 in the (η, Z) space, namely,

$$F(\eta, Z) = 0 \quad \text{and} \quad \partial F/\partial Z = 0 \quad \text{and} \quad \nabla_\eta F \equiv \partial F/\partial X + i\partial F/\partial Y = 0$$

or equivalently, the projection of the singularities on the η-plane determined by

$$\Delta(\eta) = 0 \quad \text{and} \quad d\Delta/d\eta = \partial\Delta/\partial X + i\partial\Delta/\partial Y = 0 \;\; .$$

The equivalence of equations (4.8a, b) to equations (4.4a, b, c) in the case of three bodies is evident if we notice the expression of the U and I in the barycentre system and using the relative distances, namely, equation (4.1b). Obviously, this transformation deduction is only valid for coplanar 3-body problem, whereas the deduction of last section is valid for any N-body system, which is not necessarily planar.

Reduction of the Spatial 3-Body Problem and Stronger Inequalities

To describe the motion of the general 3-body problem, we follow Whittaker (1904) and use the following notations defined in a rectangular coordinate system

mass	general coordinate			general momentum		
m_1	q_1	q_2	q_3	p_1	p_2	p_3
m_2	q_4	q_5	q_6	p_4	p_5	p_6
m_3	q_7	q_8	q_9	p_7	p_8	p_9

In an inertial frame the Hamiltonian is

$$H = \; T + U = \sum_{i=1}^{9} \frac{p_i^2}{2m_k} + U$$

$$= \frac{p_1^2 + p_2^2 + p_3^2}{2m_1} + \frac{p_4^2 + p_5^2 + p_6^2}{2m_2} + \frac{p_7^2 + p_8^2 + p_9^2}{2m_1}$$

$$- \frac{m_1 m_2}{R_{12}} - \frac{m_2 m_3}{R_{23}} - \frac{m_3 m_1}{R_{31}}$$

where k is equal to the integer part of (i+2)/3 and

$$R_{12}^2=\sum_{i=1}^{3}(q_{i+3}-q_i)^2,\quad R_{23}^2=\sum_{i=1}^{3}(q_{i+6}-q_{i+3})^2,\quad R_{31}^2=\sum_{i=1}^{3}(q_i-q_{i+6})^2\ .$$

This system has nine pairs of canonical equations of motion

$$\dot{q}_i=\partial H/\partial p_i\quad,\quad \dot{p}_i=-\partial H/\partial q_i\quad(i=1,\ldots,9)$$

and possesses seven independent integrals corresponding to one time translation and three spatial translation symmetries, and three rotation symmetries, namely,

$$\begin{cases} H=\text{const.}\\ p_1+p_4+p_7=\text{const.}\\ p_2+p_5+p_8=\text{const.}\\ p_3+p_6+p_9=\text{const.}\\ q_1p_2-q_2p_1+q_4p_5-q_5p_4+q_7p_8-q_8p_7=\text{const.}\\ q_2p_3-q_3p_2+q_5p_6-q_6p_5+q_8p_9-q_9p_8=\text{const.}\\ q_3p_1-q_1p_3+q_6p_4-q_4p_6+q_9p_7-q_7p_9=\text{const.}\end{cases}$$

This Hamiltonian is positive definite in all generalised momenta p's. It may be reduced by performing a series of canonical transformations from (q, p) space to (Q, P) space, which are given by Whittaker (1904) and are all extended point transformations of the form Q=Q(q). In fact Whittaker only gives the generating functions (all are of S_3-type); the transformations may be found easily using Table 2.1. Zare (1976) proved that extended transformations preserve the positive definite property of a Hamiltonian.

Several technical points need special mention. If the integrals of a Hamiltonian are not conjugate to ignorable generalised coordinates of the working coordinate system, then in general they cannot be put into H before the derivatives of H are formed in the canonical equations; nor is H usually positive definite in the p's if such integrals are substituted.

In the following calculation, we always assume that a transformation is made from (q, p) variables to (Q, P) variables. After the transformation, we change the Q's and P's back to be denoted by q's and p's so as to perform the next transformation.

(1). Reduction by Means of Linear Momentum Integrals

The explicit transformation may be found using Table 2.1 using the following generating function

$$
\begin{aligned}
W_1 = {} & \sum_{i=1}^{6} p_i Q_i + \sum_{i=1}^{3} (p_i + p_{i+3} + p_{i+6}) Q_{i+6} \\
= {} & p_1 Q_1 + p_2 Q_2 + p_3 Q_3 + p_4 Q_4 + p_5 Q_5 + p_6 Q_6 \\
& + (p_1 + p_4 + p_7) Q_7 + (p_2 + p_5 + p_8) Q_8 + (p_3 + p_6 + p_9) Q_9
\end{aligned}
$$

where (Q_1, Q_2, Q_3) and (Q_4, Q_5, Q_6) are the relative coordinates of m_1 and m_2 with respect to m_3 respectively, with (Q_7, Q_8, Q_9) being the coordinates of m_3 in the previous rectangular coordinates. Furthermore, (P_1, P_2, P_3) and (P_4, P_5, P_6) are the momentum components of m_1 and m_2 respectively, with (P_7, P_8, P_9) being those of the centre of mass (i.e. total momentum of the system).

On substitution of the new variables for the old the new Hamiltonian may be obtained; it is found that (Q_7, Q_8, Q_9) are ignorable coordinates. Hence without loss of generality one can choose that (P_7, P_8, P_9)=0, as this only means that the centre of mass is taken to be at rest. Finally the Hamiltonian is simplified to

$$
\begin{aligned}
H = {} & \frac{1}{m_3}\left[\sum_{i=1}^{6} \frac{\mu_k}{2} p_i^2 + \sum_{i=1}^{3} p_i p_{i+3}\right] + U \\
= {} & \left[\frac{1}{2m_1} + \frac{1}{2m_3}\right](p_1^2 + p_2^2 + p_3^2) + \frac{1}{m_3}(p_1 p_4 + p_2 p_5 + p_3 p_6) \\
& + \left[\frac{1}{2m_2} + \frac{1}{2m_3}\right](p_4^2 + p_5^2 + p_6^2) + U
\end{aligned}
$$

where $\mu_k = (m_k + m_3)/m_k$, and for the sake of further reductions the symbols (p, q) are used to denote the new variables instead of (P, Q). Now the system is defined by six pairs of canonical equations

$$
\dot{q}_i = \partial H / \partial p_i \quad , \quad \dot{p}_i = - \partial H / \partial q_i \quad (i = 1, \ldots, 6)
$$

which possesses four independent integrals, namely,

$$
\begin{cases}
H = \text{const} \\
q_1 p_2 - q_2 p_1 + q_4 p_5 - q_5 p_4 + q_7 p_8 - q_8 p_7 = \text{const.} \\
q_2 p_3 - q_3 p_2 + q_5 p_6 - q_6 p_5 + q_8 p_9 - q_9 p_8 = \text{const.} \\
q_3 p_1 - q_1 p_3 + q_6 p_4 - q_4 p_6 + q_9 p_7 - q_7 p_9 = \text{const.}
\end{cases}
$$

(2). Further Reduction by Means of the Angular Momentum Integral

We perform another canonical transformation defined by the generating function,

$$\begin{aligned} W_2 = \ & p_1(Q_1\cos Q_5 - Q_2\cos Q_6\sin Q_5) + p_4(Q_3\cos Q_5 - Q_4\cos Q_6\sin Q_5) \\ & + p_2(Q_1\sin Q_5 + Q_2\cos Q_6\cos Q_5) + p_5(Q_3\sin Q_5 + Q_4\cos Q_6\cos Q_5) \\ & + p_3 Q_2\sin Q_6 \qquad\qquad + p_6 Q_4\sin Q_6 \ . \end{aligned}$$

In addition to the old rectangular coordinates O-x'y'z' fixed in space, we take a new set of moving coordinates O-xyz, where O is at the centre of mass and Ox is the intersection (or node) of the plane O-xy with the plane of the three bodies; Oy is perpendicular to Ox and lying in the plane of the three bodies, while Oz is normal to the plane of the three bodies and forms a right hand coordinates with Ox and Oy. Then the new variables may be interpreted as follows:

(Q_1, Q_2) and (Q_3, Q_4) are the coordinates of m_1 and m_2 respectively, relative to the axes drawn through m_3 and parallel to Ox and Oy; Q_5 is the angle between Ox and Ox'; Q_6 is the angle between Oz and Oz'. Furthermore, (P_1, P_2) and (P_3, P_4) are the momentum components of m_1 and m_2 respectively, relative to Ox and Oy; P_5 and P_6 are the angular momentum components of the system along Oz' and Ox axes respectively.

On substitution we obtain the new Hamiltonian, in which Q_5 does not occur, thus P_5 is a first integral of the system, which we shall denote as $P_5 \equiv C$. Again we shall use (p, q) to denote the new variables of the system. Then the new Hamiltonian system is

$$\begin{cases} H = H(p_1, p_2, p_3, p_4, p_6; \ q_1, q_2, q_3, q_4, q_6; \ C) \\ \text{Five pairs of canonical equations} \\ H = \text{const} \\ \text{Two angular momentum integrals} \end{cases}$$

For more details see section 158 of Whittaker (1904).

The expression of the above system can be greatly simplified if the invariable plane is set to coincide with the old O-x'y' plane without loss of generality. Because of this choice of the coordinate system O-x'y'z' relative to the invariable plane of the system (with the coordinate origin at barycentre), C becomes the norm of the total angular momentum vector with respect to the system's barycentre; moreover, the equations of the other two angular momentum integrals become simpler, viz.

$$\begin{cases} p_6 = 0 & (4.9a) \\ C\cos q_6 = p_2 q_1 - p_1 q_2 + p_4 q_3 - p_3 q_4 & (4.9b) \end{cases}$$

So far the invariable plane has only been used to simplify the above two angular momentum integrals; the Hamiltonian and (five pairs) canonical equations are still the same as before. Although in general equations (4.9a, b) cannot be put into the Hamiltonian before forming derivatives and determining the positive definiteness of the Hamiltonian; equation (4.9a) can be substituted because if the invariable plane was used at the very beginning, then after the same series of canonical transformations one obtains the same Hamiltonian. Such a Hamiltonian will be positive definite in the p's no matter whether p_6 is included; the canonical equations can be formed either before or after the substitution of equation (4.9a).

It is proved in Whittaker (1904) that equation (4.9b) can also be substituted into the Hamiltonian replacing q_6 without influencing the calculation of the derivatives of H with respect to the first four p's.

Therefore, equations (4.9a, b) can be regarded as replacing the pair of canonical equations of p_6 and q_6; moreover they can be substituted to obtain the Hamiltonian as a function of (p_i, q_i), with i=1, ..., 4. This Hamiltonian has four degrees of freedom, and possesses only one integral H=const. However, this Hamiltonian is no longer positive definite in the four p's because of the substitution of equation (4.9b).

It must be noted that these two equations are extra restrictions on the system, thus the change of variables according to them does not produce a canonical transformation. Therefore the positive definite feature of the Hamiltonian preserved by extended point canonical transformations cannot be preserved in the Hamiltonian following the above replacements. In fact the first replacement $p_6=0$ does not change this feature of the function H, it is the second replacement that makes the new Hamiltonian not positive definite in the variables (p_1, p_2, p_3, p_4). So Zare's theory cannot be directly applied to the spatial 3-body problem.

From here on, if the invariable plane is always used, then at least three types of further reductions may be followed for different purposes: (a). retain the occurrence of q_6 in H, and the positive definite property of H without using equation (4.9b); (b). keep the occurrence of q_6 in H, but abandon the positive definite property of H by a use of equation (4.9b) so as to simplify the expression of H; (c). abandon both the occurrence of q_6 in H and the positive definite property of H by using equation (4.9b) so as to simplify the expression of H.

Whittaker chose (c), because his interest was to simplify the Hamiltonian and canonical equations. Thus for our purpose, a choice must be made between (a) and (b) so as to keep the variable q_6, because it carries significant physical meaning, namely, the inclination of the plane of motion relative to the invariable plane. However, (a) follows exactly Zare's approach and will lead to a lengthy calculation, since the invariable plane is not used to simplify H. Therefore we will choose to follow routine (b) by noting a similarity between the spatial and planar Hamiltonians. In this way the complete advantage of the invariable plane is taken to simplify H, so long as q_6 is not lost from the final expression; although one has to abandon the positive definite property as a price. Nevertheless, this property is not so important as stressed in Zare (1976).

In what follows we shall give the Hamiltonian following all three routines so as to keep the physical meaning clear; but the study of possible and forbidden motions will be carried out along routine (b) only.

In order to make the explicit expression of the Hamiltonian more compact we introduce a function of the generalised coordinates, viz.

$$F(q) \equiv \frac{1}{(q_2q_3-q_1q_4)^2}\left[\left(\frac{1}{2m_1}+\frac{1}{2m_3}\right)q_4^2+\left(\frac{1}{2m_2}+\frac{1}{2m_3}\right)q_2^2-\frac{q_2q_4}{m_3}\right]$$
$$=\frac{I_n}{4S^2}\frac{m_1+m_2+m_3}{2m_1m_2m_3}>0 \qquad (4.10)$$

where S is the area of the triangle formed by the three bodies, I_n is the moment of inertia of the 3-body system about the node, that is, the line through the system's barycentre in which the plane of the three bodies meets the invariable plane.

(Routine a). If equation (4.9b) is not used, the new Hamiltonian becomes

$$H=\left[\frac{1}{2m_1}+\frac{1}{2m_3}\right](p_1^2+p_2^2)+\frac{1}{m_3}(p_1p_3+p_2p_4)$$
$$+\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](p_3^2+p_4^2)+U$$
$$+F(q)[C\operatorname{cosec}q_6-(p_2q_1-p_1q_2+p_4q_3-p_3q_4)\operatorname{ctg}q_6]^2 \qquad (4.11a)$$

(Routine b). If equation (4.9b) is used, the last term of the above Hamiltonian may be further simplified, thus obtaining the Hamiltonian

$$H=\left[\frac{1}{2m_1}+\frac{1}{2m_3}\right](p_1^2+p_2^2)+\frac{1}{m_3}(p_1p_3+p_2p_4)$$
$$+\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](p_3^2+p_4^2)+U$$
$$+F(q)C^2\sin^2q_6 \qquad (4.11b)$$

(Routine c). When equation (4.9b) is used, one can also obtain a Hamiltonian of the form

$$H= \left[\frac{1}{2m_1}+\frac{1}{2m_3}\right](p_1^2+p_2^2)+\frac{1}{m_3}(p_1p_3+p_2p_4)$$
$$+\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](p_3^2+p_4^2)+U$$
$$+F(q)[C^2-(p_2q_1-p_1q_2+p_4q_3-p_3q_4)^2] \qquad (4.11\,c)$$

The dynamics of the system is governed by four pairs of canonical equations

$\dot{q}_i=\partial H/\partial p_i \quad , \quad \dot{p}_i=-\partial H/\partial q_i \quad (i=1,\ldots,4)$

and the extra equations replacing the pair of equations of (q_6, p_6), viz.

$p_6=0 \quad \text{and} \quad C\cos q_6=p_2q_1-p_1q_2+p_4q_3-p_3q_4$

where the partial derivatives of H must be calculated from either equation (4.11a) or (4.11c). The system still possesses the energy integral. The canonical equations of motion cannot be formed from equation (4.11b). However, we will see that it is this equation that is important for our purpose here.

(3). Further Reduction in the Plane of the Three Bodies

Perform another canonical transformation defined by the generating function

$$W_3=p_1Q_1\cos Q_4+p_2Q_1\sin Q_4+p_3(Q_2\cos Q_4-Q_3\sin Q_4)$$
$$+p_6Q_6 \qquad +p_4(Q_2\sin Q_4+Q_3\cos Q_4)$$

which is constructed by observing the similarity of the spatial and planar problems. In fact the term p_6Q_6 is the only extra term compared to the generating function of the planar problem (Whittaker, 1904, section 161). Since the actual calculation is quite involved and this transformation is not given by Whittaker (1904), here we will give the detailed calculation.

The explicit transformations may be found from

$q_i=\partial W_3/\partial p_i \quad , \quad P_i=\partial W_3/\partial Q_i \quad (i=1, 2, 3, 4, 6)$

that is

$$\begin{cases} q_1 = Q_1\cos Q_4 \\ q_2 = Q_1\sin Q_4 \\ q_3 = Q_2\cos Q_4 - Q_3\sin Q_4 \\ q_4 = Q_2\sin Q_4 + Q_3\cos Q_4 \\ q_6 = Q_6 \end{cases} \qquad \begin{cases} P_1 = \ p_1\cos Q_4 + p_2\sin Q_4 \\ P_2 = \ p_3\cos Q_4 + p_4\sin Q_4 \\ P_3 = -p_3\sin Q_4 + p_4\cos Q_4 \\ P_4 = -p_1Q_1\sin Q_4 + p_2Q_1\cos Q_4 \\ \qquad -p_3(Q_2\sin Q_4 + Q_3\cos Q_4) \\ \qquad + p_4(Q_2\cos Q_4 - Q_3\sin Q_4) \\ P_6 = p_6 = 0 \end{cases}$$

where Q_1 is the distance of m_3m_1, Q_2 and Q_3 are the projections of m_3m_2 on and perpendicular to m_3m_1, Q_4 is the angle between m_3m_1 and the x-axis, (ie. the node through the centre of mass), Q_6 is the inclination of the plane of motion relative to the invariable plane. Furthermore, P_1 is the component of the momentum of m_1 along m_3m_1, P_2 and P_3 are the components of the momentum of m_2 parallel and perpendicular to m_3m_1, and P_4 is the component of total angular momentum of the system on Oz axis.

Since the equation for P_4 may be written as

$$P_4 = -Q_1(p_1\sin Q_4 - p_2\cos Q_4) + P_3Q_2 - P_2Q_3$$

we can introduce an auxiliary variable Ω to replace P_4 in the course of calculation, via.

$$\Omega \equiv \frac{1}{Q_1}(P_3Q_2 - P_2Q_3 - P_4) = p_1\sin Q_4 - p_2\cos Q_4$$

Then the useful transformation relations, with P_4 replaced by Ω, become

$$\begin{cases} P_1 = \ p_1\cos Q_4 + p_2\sin Q_4 \\ \Omega = \ p_1\sin Q_4 - p_2\cos Q_4 \\ P_2 = \ p_3\cos Q_4 + p_4\sin Q_4 \\ P_3 = -p_3\sin Q_4 + p_4\cos Q_4 \end{cases}$$

From these equations we may construct the following equations involving compound terms which appear in the old Hamiltonian,

$$\begin{cases} P_1^2 + \Omega^2 = p_1^2 + p_2^2 \\ P_2^2 + P_3^2 = p_3^2 + p_4^2 \\ P_1P_2 - P_3\Omega = p_1p_3 + p_2p_4 \end{cases} \qquad \begin{cases} -(P_3Q_2 - P_2Q_3 - P_4) = p_2q_1 - p_1q_2 \\ \quad P_3Q_2 - P_2Q_3 = p_4q_3 - p_3q_4 \\ \qquad\quad -Q_1Q_3 = q_2q_3 - q_1q_4 \end{cases}$$

$$P_4 = p_2q_1 - p_1q_2 + p_4q_3 - p_3q_4 = C\cos Q_6 \ \in [-C, C] \ .$$

Substituting these equations into the old Hamiltonian we finally obtain the explicit expressions for the new Hamiltonian:

(Routine a). If equation (4.9b) is not used, we have

$$
\begin{aligned}
H= & \left[\frac{1}{2m_1}+\frac{1}{2m_3}\right]\left[P_1^2+\frac{1}{Q_1^2}(P_3Q_2-P_2Q_3-P_4)^2\right] \\
& +\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](P_2^2+P_3^2)+\frac{1}{m_3}\left[P_1P_2-\frac{P_3}{Q_1}(P_3Q_2-P_2Q_3-P_4)\right] \\
& + \ U \ + F^*(Q)[C\operatorname{cosec}Q_6-P_4\operatorname{ctg}Q_6]^2 \qquad (4.12a)
\end{aligned}
$$

where F*(Q)=F[q(Q)], C is the norm of the total angular momentum.

(Routine b). If equation (4.9b) is used, we have

$$
\begin{aligned}
H= & \left[\frac{1}{2m_1}+\frac{1}{2m_3}\right]\left[P_1^2+\frac{1}{Q_1^2}(P_3Q_2-P_2Q_3-P_4)^2\right] \\
& +\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](P_2^2+P_3^2)+\frac{1}{m_3}\left[P_1P_2-\frac{P_3}{Q_1}(P_3Q_2-P_2Q_3-P_4)\right] \\
& + \ U \ + F^*(Q)C^2\sin^2 i \qquad (4.12b)
\end{aligned}
$$

where P_4 = C cos i, with i being the inclination.

(Routine c). If equation (4.9b) is used, we can also obtain

$$
\begin{aligned}
H= & \left[\frac{1}{2m_1}+\frac{1}{2m_3}\right]\left[P_1^2+\frac{1}{Q_1^2}(P_3Q_2-P_2Q_3-P_4)^2\right] \\
& +\left[\frac{1}{2m_2}+\frac{1}{2m_3}\right](P_2^2+P_3^2)+\frac{1}{m_3}\left[P_1P_2-\frac{P_3}{Q_1}(P_3Q_2-P_2Q_3-P_4)\right] \\
& + \ U \ + F^*(Q)(C^2-P_4^2). \qquad (4.12c)
\end{aligned}
$$

These three equations are those governing the possible and forbidden motion of the problem. Since all three canonical transformations generated by W_1, W_2 and W_3 are extended point transformations (i.e. Q=Q(q)), then according to the theorem given by Zare (1976), the positive definite property of the Hamiltonian in the generalised momenta is preserved in the Hamiltonian, equation (4.12a). So one can follow Zare's method to find the restriction on the possible motion by solving a set of linear equations

$\partial H/\partial P_i=0$ (i=1, ..., 4) based on equation (4.12a).

However, we will try to avoid such a tedious calculation by noticing a formal similarity between the above Hamiltonian and the final Hamiltonian of the coplanar problem, equation (4.7), and taking advantage of Zare's result.

(4). Restrictions on Possible Motions by C and H

Let us observe that there are only two differences between the Hamiltonians of the spatial and coplanar problems. The Hamiltonian of the spatial problem, equation (4.12b), may be obtained by replacing C of the Hamiltonian of the planar problem, equation (4.7), by $P_4 \equiv C \cos i$, and adding the extra term, $F^*(Q)C^2\sin^2 i$. Since the Hamiltonian defined by equation (4.7) is positive definite with respect to the three P's (also in C, though this is irrelevant), the part of the Hamiltonian defined by equation (4.12b) not including the last term is also positive definite with respect to the first three P's. Therefore, if P_4 is regarded as a parameter replacing C of the planar problem and the extra term is left aside, then the result of Zare (1976), ie. equation (4.8a), or equivalently equation (4.4b), can be applied directly to the remaining terms of the above Hamiltonian of the spatial 3-body problem. Written out explicitly, we have the following inequality governing the possible motions,

$$\begin{cases} H - U \geq \dfrac{C^2\cos^2 i}{2I} + F^*(Q)C^2\sin^2 i \\ F^*(Q) = F[q(Q)] = \dfrac{I_n M}{8S^2 m_1 m_2 m_3} \end{cases} \qquad (4.13)$$

where H and C are the total energy and angular momentum integrals in the barycentre system respectively, U is the potential energy of the system, with I being the system's moment of inertia with respect to the barycentre; M is the total mass, S is the area of the triangle formed by the three mass-points, with I_n being the moment of inertia of the system with respect to the intersection line of the invariable plane and the plane of the three bodies (ie. node); this line necessarily passes through the centre of mass. The variable i is equal to Q_6, the inclination of the plane defined by the three bodies with respect to the invariable plane.

The function $F^*(Q)$ is always greater than (or equal to) a function of the shape of the triangle formed by the three masses; because from the property of the inertia ellipse in the plane defined by the three mass points, one sees that I_n is always greater than (or equal to) the moment of inertia of the system with respect to the long major axis of the inertia ellipse. This is exactly the standard position defined by Saari (1987). However, we will not carry out the calculation directly from the above expression of F*(Q), but instead, we will prove the equivalence of this expression to that given by Saari (1987) and adopt his development.

(5). Saari's Expression of F*(Q)

Saari (1987) obtained the following key inequality governing the possible and forbidden motions of the flat N-body problem (for definition of flat problem, see also Wintner, 1947),

$$\begin{cases} H - U \geq \dfrac{C^2 \cos^2 i}{2I} + \dfrac{I_x C^2 \sin^2 i}{2\,(I_x I_y - I_{xy}^2)} = \dfrac{C^2 D(\mathbf{R})}{2I} \\ D(\mathbf{R}) \equiv \cos^2 i + \sin^2 i \dfrac{I I_x}{I_x I_y - I_{xy}^2} \end{cases} \qquad (4.14a)$$

where the O-xyz coordinate system is defined relative to the system's invariable plane in the same way as ours (eg. Ox is the node through the system's barycentre). The equation gives automatically the inequality

$$IU^2 \geq -2D(\mathbf{R})C^2H \quad \text{or} \quad IU^2/D(\mathbf{R}) \geq -2C^2H \qquad (4.14b)$$

where D is a function of the position vectors **R**'s only.

From equation (4.3) we see that for the three-body problem we have

$$F^*(Q) = \frac{I_n M}{8S^2 m_1 m_2 m_3} = \frac{I_x}{2(I_x I_y - I_{xy}^2)} .$$

Therefore, in the case of 3-body problem, our result deduced from canonical transformation, equation (4.13), is equivalent to Saari's, ie. equation (4.14a). The following development of the function D(**R**) is due to Saari (1987); but we shall simplify his original deduction to obtain the best possible Hill-type surfaces.

Using equation (4.2) we obtain the following expression for the function D(**R**), viz.

$$\begin{aligned} D(\mathbf{R}) &= \cos^2 i + \sin^2 i \ (I I_x)/\ (I_x I_y - I_{xy}^2) \\ &= \cos^2 i + 2\sin^2 i \,[1 + \cos\alpha \ \sin(2\phi)]/\ (1 - \cos^2\alpha) \\ &= 1 + \sin^2 i \ [1 + 2\cos\alpha \ \sin(2\phi) + \cos^2\alpha]/\ (1 - \cos^2\alpha) \\ &\geq 1 + \sin^2 i \ (1 - \cos\alpha)\ /\ (1 + \cos\alpha) \quad (\text{'=' if } f - \phi = \pi/4,\, 5\pi/4) \\ &\geq 1 \ . \end{aligned}$$

From this we obtain the following inequality which is weaker than equation (4.14b)

$$E(\Delta, i) \equiv IU^2/[1 + \sin^2 i\ (1 - \cos\alpha)\ /\ (1 + \cos\alpha)] \geq -2C^2H \qquad (4.15)$$

which is the version of equation (4.14b) at the two standard positions.

However, since E(Δ, i) is a function of the shape of the triangle and the inclination of the plane of motion with respect to the invariable plane, by regarding i as a parameter, this inequality determines the best possible Hill-type curves at all inclinations, they form better Hill-type surfaces in the 3-dimensional space (Fig. 4.5).

(6). Critical Configuration due to E(Δ, i)

The critical points of the function E(Δ, i), where i is to be treated as a parameter, are important in determining hierarchical motion, as those of the function Z (Δ) in the planar case. Such critical points of E(Δ, i) will be called critical configurations at inclination i. This point was studied by Saari (1987), who obtained the following results for the 3-body problem:

For N=3 and i≠0, there is one and only one noncollinear critical configuration (with respect to reflection) at each inclination. This configuration is a (equilateral triangle) central configuration iff all three masses are equal. In general, a noncollinear critical configuration of E(Δ, i) is not a central configuration (namely, critical configuration of Z(Δ)).

A collinear configuration is a critical configuration of E(Δ, i) iff it is a central configuration. These critical configurations all lie in the invariable plane.

Let us note that at a collinear configuration, $\cos\alpha=1$, and thus E(Δ,i) reaches its maximum value Z(Δ) at a collinear configuration. Therefore the critical value $(C^2H)_c$ must be evaluated at a collinear critical configuration. Combining this point with the above statement it is clear that although stronger restrictions and larger forbidden regions exist for spatial motion, the critical value of $(C^2H)_c$ is not improved by these stronger inequalities. Moreover it is the same for the spatial and planar motions of the three bodies if the mass parameters are kept unchanged.

Saari (1984) has proved that, for the 3-body problem, if i≠0 at one moment then i≠0 for all time. Thus motions with and without inclination cannot pass into each other. On the other hand, at the date of syzygy all three masses must lie in the invariable plane (Wintner, 1947). Thus should the three masses form a collinear configuration, they necessarily lie along the line of nodes. However, according to Saari (1984), this is

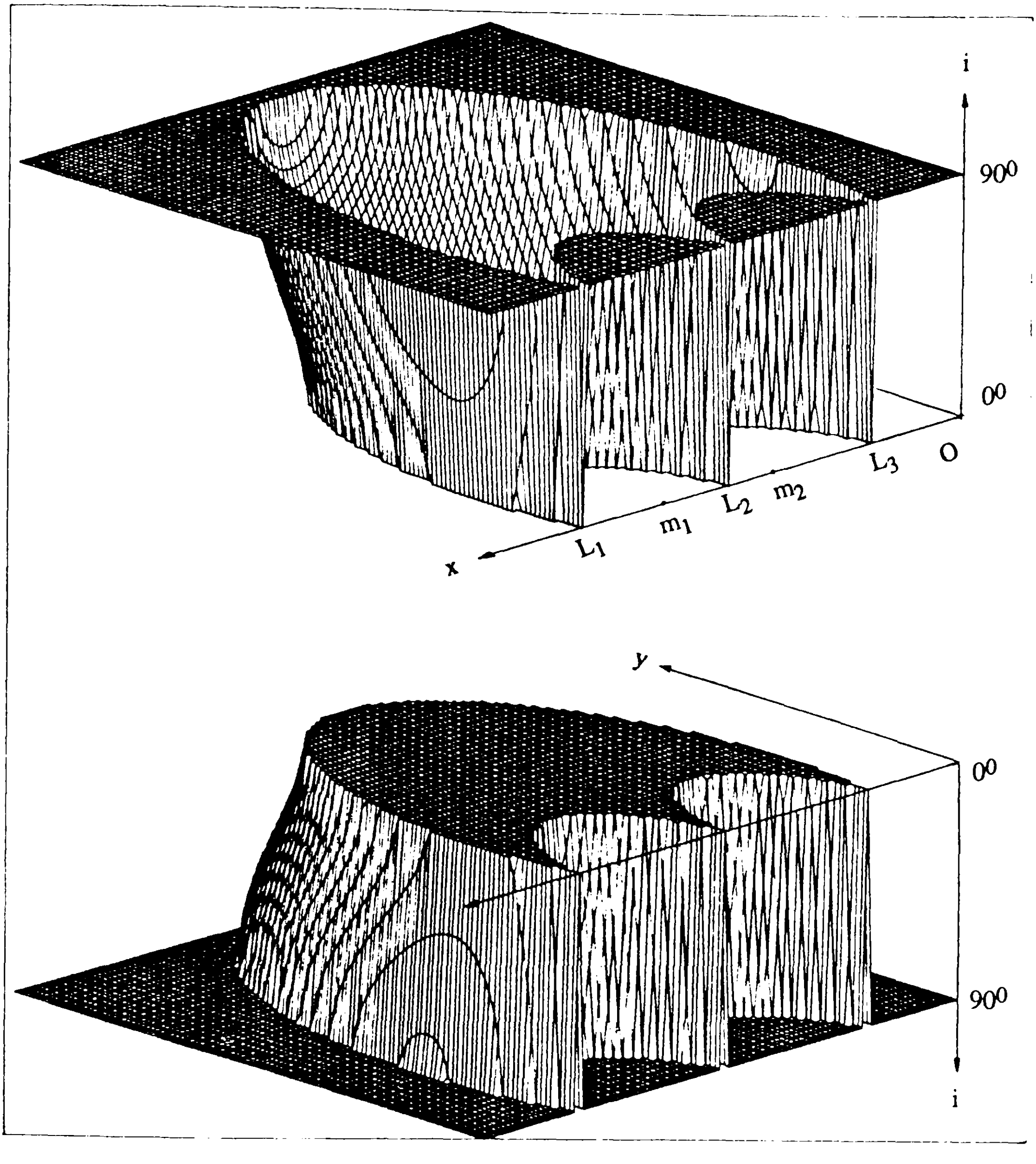
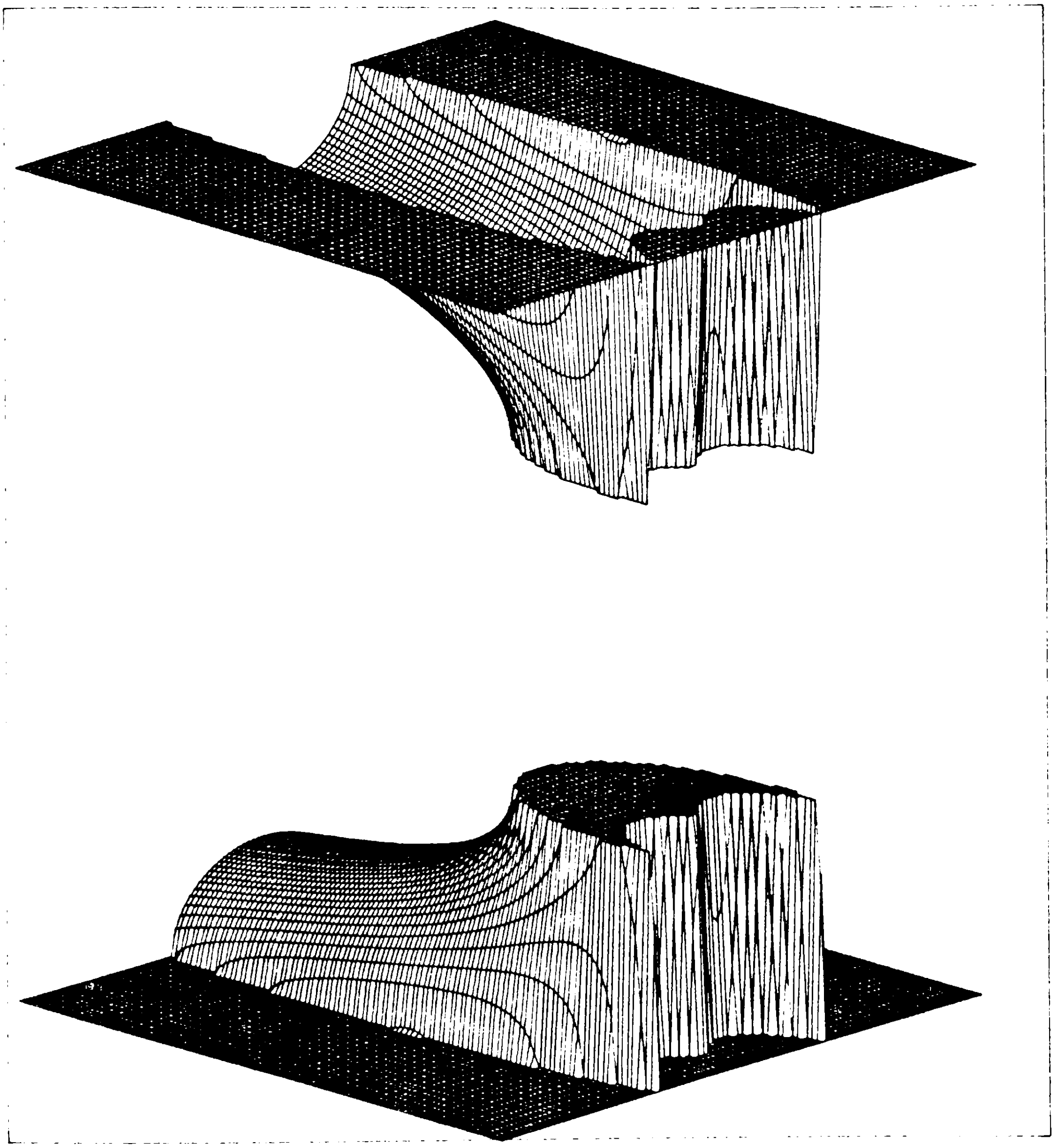


Figure 4.5 The change of the Hill-type stability regions with inclination for (a). $m_1=m_2=m_3=0.5$; (b). $m_1=m_2=0.5$ and $m_3=0.01$. The masses m_1 and m_2 lie on the Ox-axis, and the mirror image is given to show the detail of the Hill-type regions. It is clear that as the value of the outer mass m_3 decreases the outer surface tends to that of the restricted problem.



b

improbable. Therefore, syzygy is a rare configuration in the case of spatial motion of three bodies, and the collinear critical configuration is very unlikely to be achieved. This is different from coplanar motion of three bodies.

(7). Conclusion

In this section we have obtained stronger inequalities for the spatial 3-body problems, which determine larger forbidden regions of motion. However, although these inequalities are stronger than Sundman's inequality, they do not give a better estimation for the critical values of $(C^2H)_c$. The same results were obtained earlier by Saari (1987), but the deduction given here is independent. The concept of the inertia ellipsoid is used to interpret and simplify Saari's deduction, and to show the equivalence of the present author's results with Saari's.

In Fig. 4.5, the Hill-type curves are plotted against inclination for two sets of mass parameters. They may be more intuitively visualised in a 3-dimensional physical space as follows:

Consider a coordinate system O-xyz, with the two masses m_1 and m_2 lying on Ox. Put the Hill-type curves at inclination i on the plane passing through Ox and having inclination i with respect to O-xy. Then these curves with a continuous parameter i form a surface in the above coordinate system. This surface divides the space into possible and forbidden regions of motion. We will call such surfaces and regions **Hill-type surfaces** and regions respectively.

Comparing these Hill-type surfaces with those obtained for the circular restricted 3-body problem (see Lundberg & Szebehely et al, 1985), one finds that the inner surfaces are quite similar, but the outer surfaces differ: it is closed and sphere-like in the general problem, while in the restricted problem it is open and cylinder-like.

4.4 Summary

Instead of summarising what we have obtained in this chapter, we shall mention in particular some possible future work, which will certainly benefit the improvement of the inequality method emphasised in this thesis, the result of Appendix B, and the approach to relativistic problems in chapters 6 and 7.

The important point is that equation (4.14a) found by Saari (1984, 1987) for flat N-body problems ($N \geq 3$) using a rigid motion method must admit an alternative inequality deduction. In fact, this result implies the existence of general mathematical inequalities stronger than Sundman's and those collected in Appendix B. Moreover, it is very likely that the result may be generalised for spatial N-body problems.

The value of an alternative proof is obviously seen from Zare's (1976) work. A proof of Sundman's inequality for the N-body problem would be likely to become very lengthy should one attempt to follow the transformation method, although such an effort would be important if one could succeed.

CHAPTER 5
Hierarchical Stability and Hill-Type Stability of the General 3-Body Problem

In the last chapter we discussed the restrictions on possible motions of the general 3-body problem by the energy and angular momentum integrals. The basic results are obtained by a direct use of Sundman's inequality. It is found that a Hill-type stability exists when the value of C^2H is below a critical value $(C^2H)_c$, which is determined solely by the size of the three masses. Because of the key role the function C^2H plays in the problem, the above stability criterion is called a C^2H stability criterion as well. In connection with this analytical stability, Roy & Walker (1983) defined hierarchical stability (see chapter 1) and studied it both theoretically and numerically. The purposes of the present chapter is to prove a relation between the Hill-type stability and hierarchical stability condition HS-(C), and to continue the numerical investigation on hierarchical stability of initially elliptic coplanar 3-body systems. Throughout the chapter, the notion of hierarchical stability (shortened as **HS**) will be used as the main concept for stability. However, before we go into any detail, let us first make clear the relevance and limitation of such an approach.

The concept of hierarchical stability is relevant because hierarchical arrangements are widely observed in the universe. For example, multiple-star systems and our planetary system are found to move in well ordered orbits such that their orbital elements are closely approximated to by two-body motions. The significance of it has been strengthened since the discovery of Hill-type stability (**HTS** hereafter) in the general 3-body problem. It is widely held that this analytical stability assures condition HS-(C) although there is no simple analytical criterion to guarantee all three conditions (cf. section 5.1). A further contribution to the hierarchical stability approach arose from the numerical integration experiments on initially circular **CHT** (shorthand for Coplanar Hierarchical Three-body) systems by Walker and Roy (1983). This work positively demonstrated the good agreement between the above two types of stability (hierarchical stability and Hill-type stability) and the existence of empirical (hierarchical) stability regions outside the analytical Hill-type stability region.

These observations and numerical experiments may suggest that hierarchical stability could be a generally applicable concept of stability, and supports the following attractive picture: although instabilities (A), (B) and collision are not precluded in Hill-type stability regions, they were not observed in such regions; thus Hill-type stability analytically defines an significant hierarchical stability region, out of which empirical stability regions were found by fitting empirical stability curves to the expected life-time versus ratio of major axes plots of the numerically integrated fictitious systems (Walker and Roy, 1983; M^{c}Donald, 1986). It would be satisfactory if this were true for all 3-body systems. Unfortunately, Hill-type stability does not exist for problems with more than three participating masses. Moreover, the numerical experiments of the present chapter suggests that even for the coplanar 3-body problem, the previous picture appears to be false if the initial orbits are elliptic.

The stability condition HS-(C) is probably the most distinguished and attractive one among the three conditions; but there are certain limits on this geometrical condition. As is well known, Jacobian coordinates can be applied to any N-body system. However, not all configurations can be described by hierarchies. Contrariwise some configurations can be defined by more than one hierarchical structure. Consequently, it is possible that hierarchical stability may not include all important stable motions. One example is the well-known Lagrange (equilateral triangle) solution of the 3-body problem, where the motion is periodic, and stable (at least linearly) if the masses satisfy some conditions (Danby, 1964; Siegel & Moser, 1971). It is obvious that the Jacobian vectors define no ordering, thus this motion cannot be covered by our approach. Nevertheless, numerical experiments on the 3-body problem suggest that once the motion is such that the system exhibits no hierarchy, instabilities usually set in very quickly - stability is atypical for the motion of non-hierarchical systems. A second point is that we only need to study hierarchical stability for one of the possible hierarchies, since all possibilities are covered by varying the mass parameters. For example, a system may possess a hierarchy for a while, and destroy it to achieve another stable hierarchy after a time. In the present study we terminate the programme at the break-up of the first hierarchy and conclude that the first is not stable. We do not find and study the new hierarchy, because a new hierarchy simply means a new set of masses and orbital parameters - our systematic investigation covers, in principle, all possible combinations of these parameters and all initial conditions so long as the system is hierarchically positioned.

Limits also exist concerning conditions HS-(A) and HS-(B). Motions may still be quasi-periodic even if the orbits suffer drastic changes (Ferraz-Mello, 1990). In fact even collisional quasi-periodic motions exists (eg. Henon, 1976; Henon & Petit, 1986).

In section 5.1 we will prove that Hill-type stability guarantees hierarchical stability condition HS-(C). Section 5.2 summarise some of the results that may be obtained based on the properties of C^2H; the proof of them is outlined. The numerical experiments are presented in sections 5.3, 5.4 and 5.5.

5.1 Hill-Type Stability and Hierarchical Stability Condition HS-(C)

The aim of this section is to prove analytically that in the case of the general 3-body problem, either coplanar or spatial, a Hill-type stability guarantees the hierarchical stability condition HS-(C), although it does not preclude 'escape' nor 'collision' instabilities.

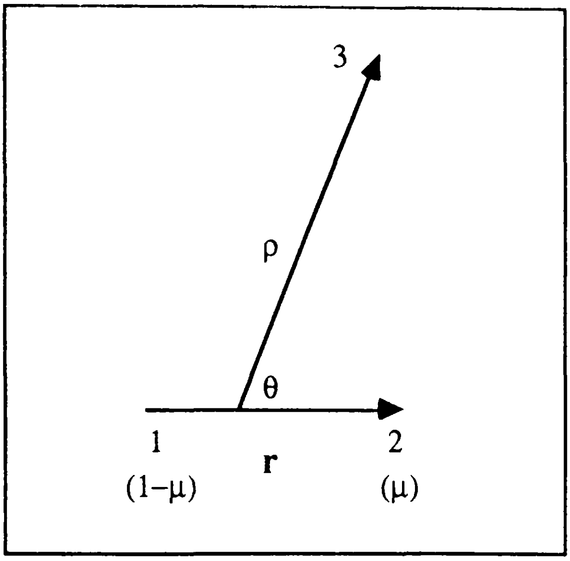


Fig. 5.1 The Jacobian vectors

To do this let us consider the 3-body problem in a Jacobian coordinate system, and denote the Jacobian vector connecting m_1 and m_2 by **r**, and that connecting the centre of mass of the first two masses, C, and m_3 by ρ; the angle between these two Jacobian vectors is denoted as θ (see Fig. 5.1). Let the unit of mass be such that $m_1+m_2=1$ ($m_1 \geq m_2$), hence $(m_1, m_2, m_3)=(1-\mu, \mu, \mu_3)$, with $\mu \in [0, 0.5]$. We also choose **r** as the reference line and r as the variable unit of length without loss of generality. Then the functions I (with respect to the system's centre of mass), U (in this section we use U to denote $|U|$) and Z may be written out explicitly in the Jacobian coordinate according to equation (4.1), viz.

$$\begin{cases} I = I(\rho) = \mu(1-\mu) + \dfrac{\mu_3}{1+\mu_3}\rho^2 \\ U = U(\rho,\theta) = \mu(1-\mu) + \dfrac{(1-\mu)\mu_3}{R_{13}} + \dfrac{\mu\mu_3}{R_{23}} \end{cases} \tag{5.1}$$

where

$$R_{13} = \sqrt{\mu^2 + \rho^2 + 2\mu\rho\cos\theta}$$

$$R_{23} = \sqrt{(1-\mu)^2 + \rho^2 - 2(1-\mu)\rho\cos\theta}$$

Suppose that m_1 and m_2 form the inner binary, with m_3 being the outer mass. As the value of C^2H is increased from $-\infty$, the forbidden regions first appear around the equilateral triangular points if $C^2H=-(\Sigma m_i m_j)^3/2M$; then these regions expand until they become triply connected, hence the system is stable in the sense of Hill. It is clear that in our case L_2 is not the critical configuration at which we are to estimate $(C^2H)_c$. Therefore, it must be either L_1 or L_3 according as which one has a greater $Z=IU^2$ (see Fig. 5.2).

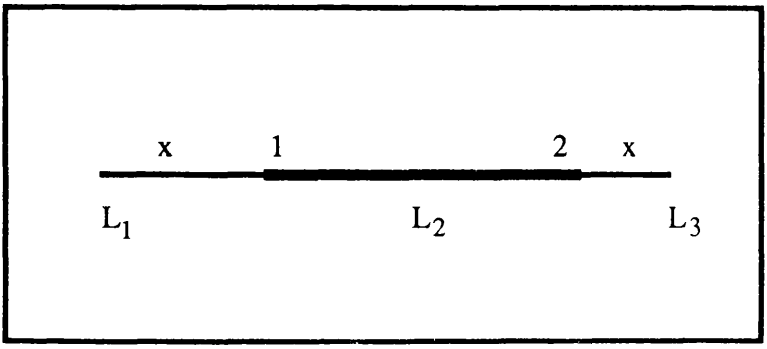


Fig. 5.2 The collinear critical configurations

Now let us prove that if $C^2H \leq -\max\{Z_1, Z_3\}/2$, then $\rho > 1$ (ie. $\rho > r$) in the following steps.

(1). $\rho > 1$ at L_1 and L_3

It is a classical result that the position of the two collinear critical configurations may be determined by the equation (see Roy, 1982)

$$F(x, \mu, \mu_3) = x^5 + x^4(3-\mu) + x^3(3-2\mu) - x^2(3\mu_3 + \mu) - x(3\mu_3 + 2\mu) - (\mu_3 + \mu) = 0$$

which gives

ρ at L_3 by $\rho = 1-\mu+x$, if $\mu = m_2/(m_1+m_2) \in [0, 0.5]$,

ρ at L_1 by $\rho = 1-\mu+x$, if $\mu = m_1/(m_1+m_2) \in [0.5, 1]$.

It is well known that there is one and only one positive real solution x to the above polynomial equation, if μ and μ_3 are considered as two parameters. Moreover, it is easy to verify that $F(0) < 0$, and $F(+\infty) = +\infty > 0$. On the other hand

$$F(x=\mu) \le x^5 + (3-\mu)x^4 + (3-2\mu)x^3 - \mu x^2 - 2\mu x - \mu \Big|_{x=\mu} \qquad (\text{'=' iff } \mu_3 = 0)$$

$$= -\mu(1-\mu)(\mu^2 + 3\mu + 1) \le 0 \qquad (\mu \in [0,1], \text{ '=' iff } \mu = 0 \text{ or } 1).$$

It follows that $x > \mu$ for $\mu \in [0, 1]$ and $\mu_3 \in [0, +\infty)$ if at most one mass is allowed to vanish. Thus under the same condition we obtain $\rho > 1$ at L_1 and L_3.

Furthermore, if $F(x, \mu, \mu_3)$ is considered as a function of all three variables, then according to the implicit function theorem, we may solve the equation $F(x, \mu, \mu_3) = 0$ for $x=x(\mu, \mu_3)$, hence write $f(\mu, \mu_3)=F(x(\mu, \mu_3), \mu, \mu_3)=0$. It is straightforward to verify that $\partial F/\partial\mu < 0$, $\partial F/\partial\mu_3 < 0$ and $\partial F/\partial x \,|_{F=0} > 0$. Thus according to the chain rule we have $\partial x/\partial\mu \,|_{F=0} \ge 0$, $\partial F/\partial\mu_3 \,|_{F=0} \ge 0$. Since $\mu(L_3) \le 0.5 \le \mu(L_1)$, we have $x(L_3) \le x(L_1)$.

It will be also useful if we can show that *the primary bifurcation value of C^2H requires the smallest mass to be the central mass in the collinear configuration, whereas the tertiary requires the largest mass and the secondary the intermediate mass to be so positioned.* This statement was proved by Walker & Roy (1981) near the limit of three equal masses, and shown to be always true from numerical calculation. However, it has not been proved analytically in the general sense, mainly due to a complicated relation between $\rho(L_3)$ and $\rho(L_1)$ (but see Golubev, 1968). This point can be used to shorten the following proof; however, we will proceed without it. In fact it may be obtained as an immediate result of the following proof.

(2). L_1, L_2 and L_3 are saddle points; L_4 and L_5 are minima of Z (see Fig. 4.3)

Consider the function $Z=IU^2$ as a function of θ with ρ as a parameter, we study the behaviour of the function with respect to θ. Using the fact that I is independent of θ, it is straightforward to show that the value of $\partial Z/\partial\theta$ is proportional to $U\partial U/\partial\theta$, namely,

$$U\left[\frac{\rho\sin\theta}{R_{13}^3}-\frac{\rho\sin\theta}{R_{23}^3}\right]$$

and that of $\partial^2 Z/\partial\theta^2$ is proportional to $U\partial^2 U/\partial\theta^2+(\partial U/\partial\theta)^2$, namely,

$$U\left[\frac{\rho\cos\theta}{R_{13}^3}-\frac{\rho\cos\theta}{R_{23}^3}+\frac{3\mu\rho^2\sin^2\theta}{R_{13}^5}+\frac{3(1-\mu)\rho^2\sin^2\theta}{R_{23}^5}\right]$$
$$+\ \mu(1-\mu)\left[\frac{\rho\sin\theta}{R_{13}^3}-\frac{\rho\sin\theta}{R_{23}^3}\right]^2 .$$

From these we obtain

$$\begin{cases}\partial Z/\partial\theta=0\ , & \partial^2 Z/\partial\theta^2<0\ , \quad \text{at}\ \theta=0,\pi\\ \partial Z/\partial\theta=0\ , & \partial^2 Z/\partial\theta^2>0\ , \quad \text{at}\ \theta=\pm\theta_0\end{cases}$$

where θ_0 is the angle corresponding to $R_{13}=R_{23}$. Therefore, the function Z has two local maxima at $\theta=0$ and π, and two local minima at $\theta=\pm\theta_0$ with respect to the θ variable. These are the only critical points of the function with respect to θ.

A tedious but straightforward calculation will show that at all five critical points, we have $\partial Z/\partial\rho=0$ and $\partial^2 Z/\partial\rho^2 > 0$. Thus the result concerning the property of the critical points is proved. In addition it is also straightforward to show that all five critical points are nondegenerate (cf. section 4.2).

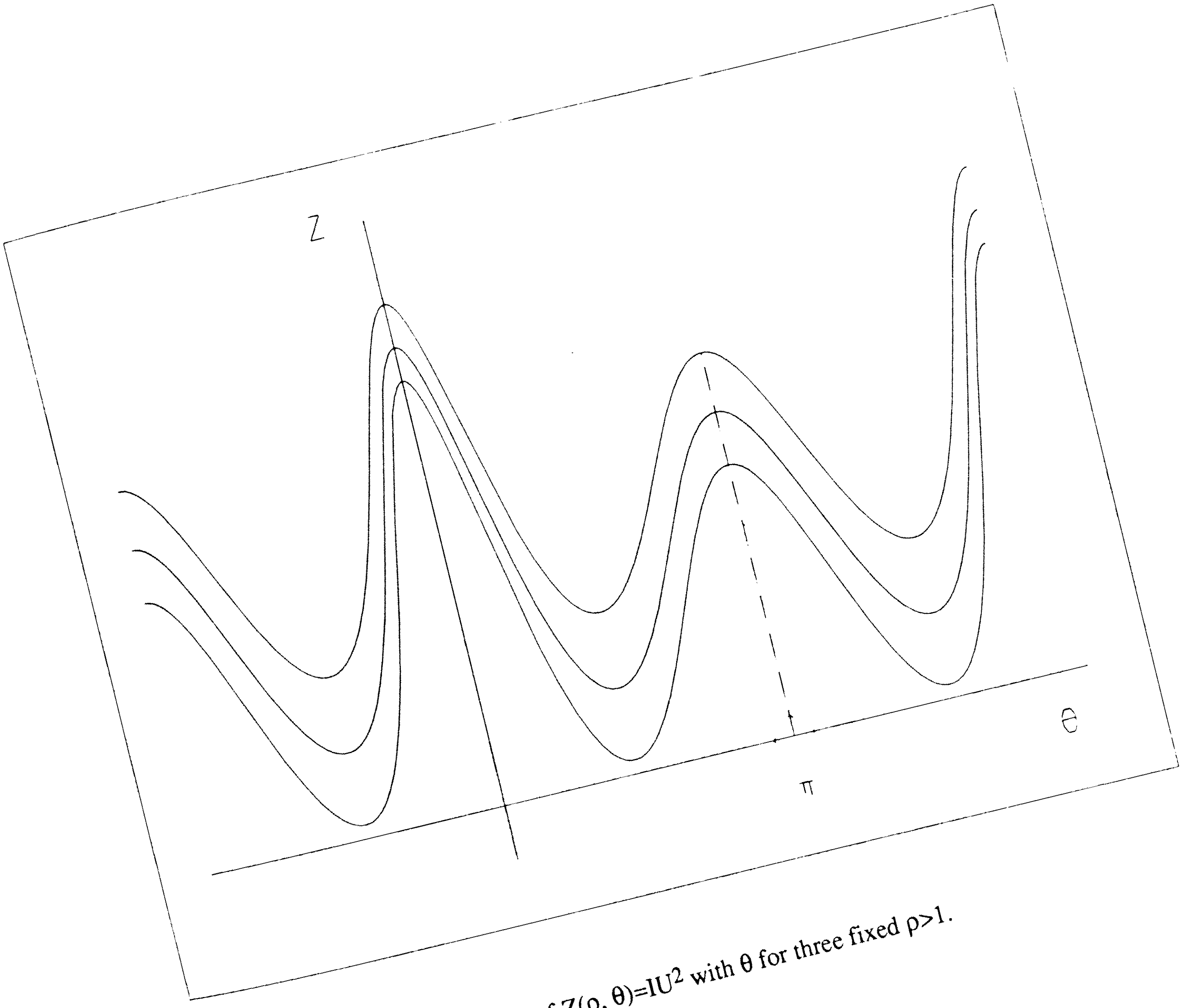


Figure 5.3 The change of $Z(\rho, \theta)=IU^2$ with θ for three fixed $\rho>1$.

(3). $Z(\theta) \leq Z(\theta=\pi) \leq Z(\theta=0)$ if ρ=const.≥ 1

In order to compare the values of the function Z at the two collinear cases, namely, $\theta=0$ and π, we again use the fact that I is independent of θ; thus only the values of $U(\rho, \theta)$ need to be compared. Obviously the difference of U at the two collinear cases, $U(\theta=0)-U(\theta=\pi)$, is proportional to the following quantity,

$$\left[\frac{1-\mu}{\rho+\mu}+\frac{\mu}{\rho-(1-\mu)}-\frac{1}{\rho}\right]-\left[\frac{1-\mu}{\rho-\mu}+\frac{\mu}{\rho+(1-\mu)}-\frac{1}{\rho}\right]$$
$$=\frac{\mu(1-\mu)}{\rho(\rho+\mu)[\rho-(1-\mu)]}-\frac{\mu(1-\mu)}{\rho(\rho-\mu)[\rho+(1-\mu)]} \geq 0$$

which is valid whenever $\rho \geq 1$ and '=0' is true iff $\mu=0.5$.

Therefore the value of Z on a circle with a radius $\rho \geq 1$ and the centre at C, the centre of mass of the first two masses, has its greatest value in the direction $\theta=0$. As the value of θ is increased from zero, the value of the function Z decreases until it reaches its minimum at θ_0; thereafter it increases until it reaches the secondary local maximum at $\theta=\pi$. This is shown in Fig. 5.3.

(4). $\rho>1$ in general

Consider the circle with its centre at C and of radius $\rho(L_3)$, then the greatest value of Z is achieved at $\theta=0$, ie. $Z(\theta) \leq Z(\theta=0)$. Because L_3 is a nondegenerate saddle point, the contour curve passing through L_3 bifurcates into two branches, one lies inside the circle, the other outside.

Because the two major directions of this saddle point are tangent and perpendicular to the circle at this point (eg. $Z \rightarrow +\infty$ as $\rho \rightarrow +\infty$), and that there is no singular point out of this circle, with a possible exception at $\theta=\pi$ (nothing is known about its position relative the circle; this uncertainty does not influence the following points), we may conclude that within some neighbourhood of L_3 the outside branch of the contour curves passing through L_3 lies completely outside the circle.

Because the value of Z on the circle cannot be the same as that at L_3 except at $\theta=0$ or

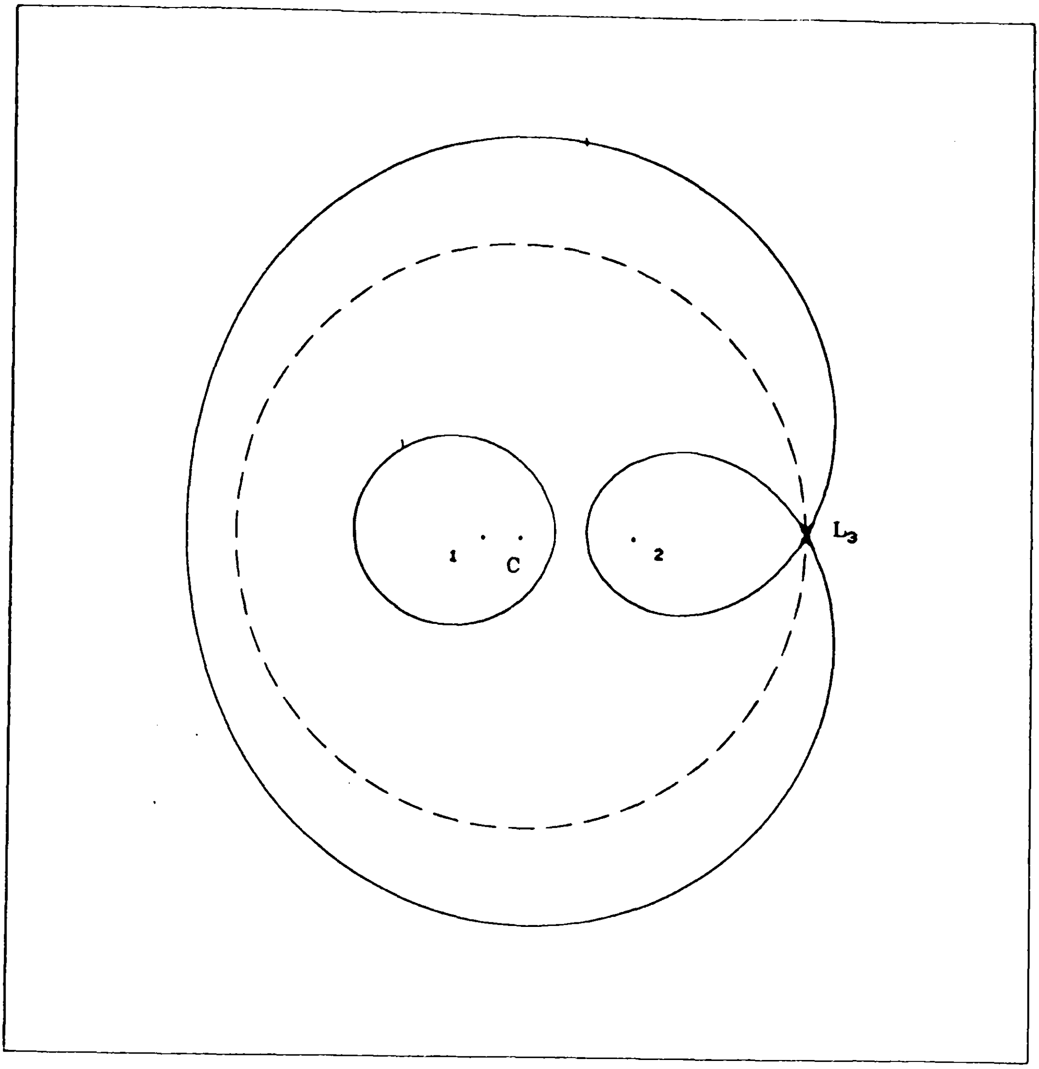


Figure 5.4 The contour curve and circle passing through L_3. The mass parameters are m_1=0.75, m_2=0.25 and m_3=1.

π, the contour passing through L_3 cannot cross over the circle; thus the outside contour curve is necessarily outside the circle everywhere.

On the other hand, as the contour extends from $\theta=0$ to $\theta=\pi$, it cannot fold back to close at a direction other than $\theta=0$, for this would produce at least a sixth critical point.

Thus the contour must stay outside the circle and close onto itself without any folding, as what is shown in Fig. 5.4. Therefore $\rho(\theta) \geq \rho(\theta=0) > 1$.

Because the contour passing through L_3 closes outside the circle, L_1 must lie inside this contour line (but nothing is known about its position relative the circle) and hence $Z(L_1) \leq Z(L_3)$. From this the statement quoted in step (1) follows immediately.

Similarly, one can show that the branch inside the circle always lie inside.

(5). Comments

One may feel that there is a need to prove the following as well: choose $\mathbf{R}_{13}$ as the reference line and its length the variable unit of distance, and then prove the result that $R_{12} < \rho$ (see Fig. 5.5). Although one can produce an independent proof, the result will be equivalent to what we have given above. Here we shall outline a proof using the notations of Fig. 5.5, thereby we do not have to do the technical calculations because of some simple relations between the two formulations of the question.

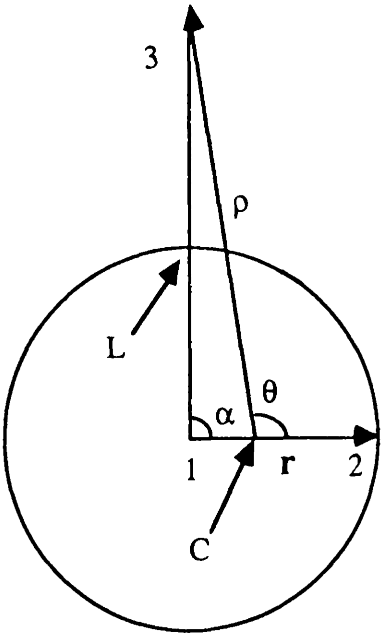


Fig. 5.5 The Jacobian vectors

Obviously, in this case the critical configuration is one of the two with the masses arranged in the order ($m_1m_2m_3$) and ($m_2m_1m_3$); the former is denoted as L in Fig. 5.5. From the result of step (1), it is evident that $R_{12} < \rho$ is true when m_2 is at L. Secondly, since the angle α is monotonic with θ, the property of Z with respect to α is similar to that with respect to θ. On a circle with its centre at m_1 and radius not greater than R_{1L}, the function Z has a primary maximum at α=0 and a secondary maximum at $\alpha=\pi$. Moreover there is no other critical point inside the circle passing through L except at $\alpha=\pi$. Because L is a nondegenerate saddle point, the contour curves bifurcate at L. Based on a similar argument, it follows that the critical oval around m_1 bounding the motion of m_2 lies completely inside the previous circle, thus $R_{12} < \rho$ is true in general. Other results follow similarly.

It is well-known that the critical points in the restricted problem, either circular or elliptic, are the limiting cases of the general 3-body problem with one of the masses tends to zero. It is also interesting to note that the property of the Hill curves of the coplanar circular restricted problem is similar to the Hill-type curves of the coplanar general problem, because of a similarity between the function Ω and Z, on whose level sets the curves are based. Where Ω is the effective potential of the restricted problem, that is,

$$\Omega = \rho^2 + \frac{2(1-\mu)}{\sqrt{\mu^2+\rho^2+2\mu\rho\cos\theta}} + \frac{2\mu}{\sqrt{(1-\mu)^2+\rho^2-2(1-\mu)\rho\cos\theta}}$$

using the notation of Fig. 5.1.

The analysis given in this section only depends on the function Z, so it also applies to the Hill-type stability found by Marchal & Bozis (1982) for the non-negative energy case. Thus the hierarchical stability condition HS-(C) is also guaranteed analytically, even if the total energy of the 3-body system is not negative.

Marchal et al (1984) studied the escape conditions within the Hill-type stability region by assuming that $r/\rho \le K$, where K is a constant in the region (0, $1+(m_2/m_1)$). From our result a closer estimation on the upper limit of r/ρ may be obtained, namely, $r/\rho \le k$, where $k=1/\rho(L_3) \in (0, 1)$.

Finally it is worth mentioning the work of Golubev (1968), in which the author obtained not only the Hill-type stability (cf. chapter 4) but also the statement quoted from Walker & Roy (1981). It is interesting to note that the method Golubev outlined is the same as the approach of this section, although the purposes are different.

However, the present author's approach is independent, because his purpose is to prove the relation between Hill-type stability and Roy's hierarchical stability. Many attempts have been made by the present author to modify the proof given by Walker & Roy (1981) before he realised that this apparently different question follow from the same arguments.

5.2 Results Based on Analysis of the Function C^2H

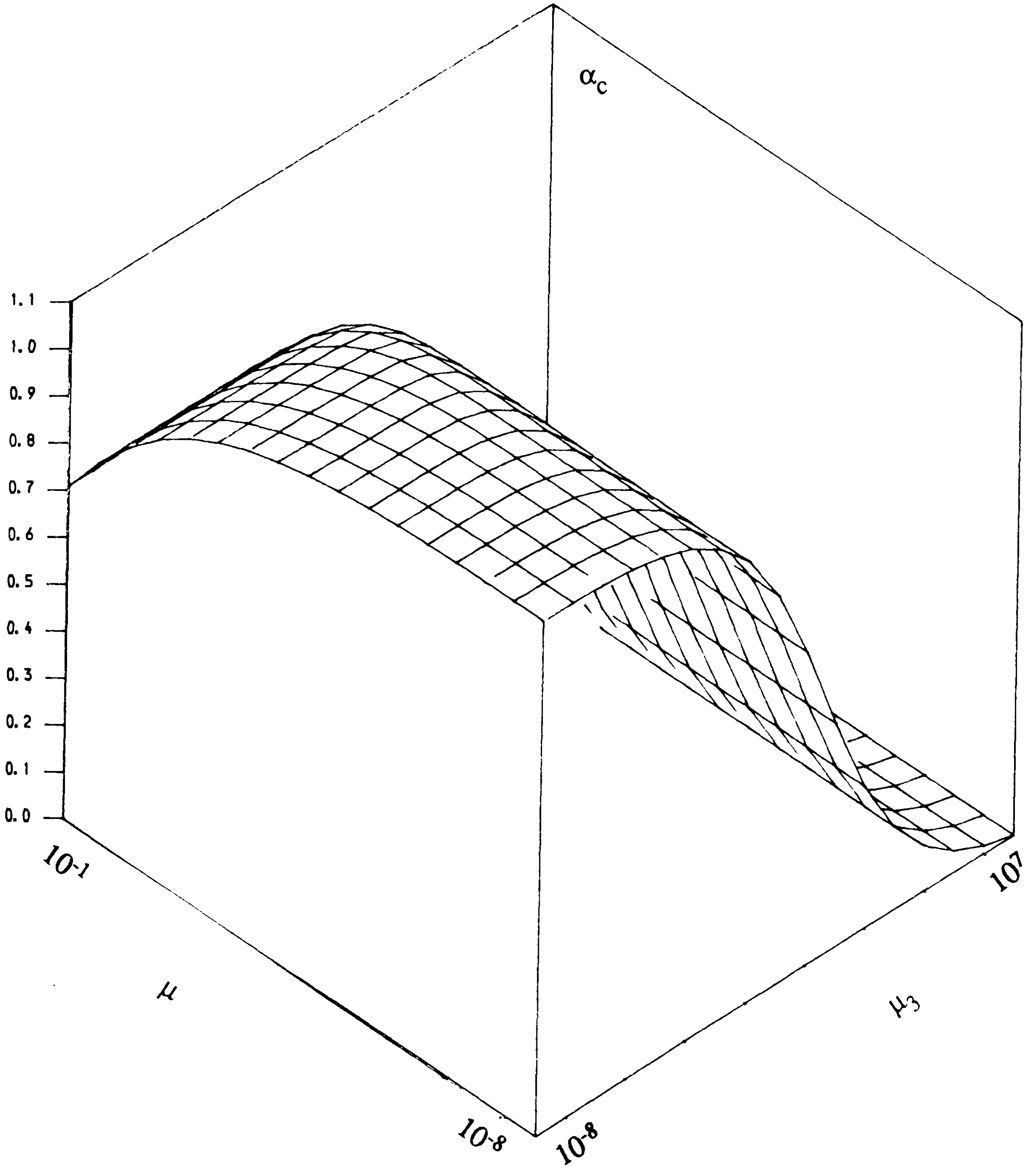
In this section we review some relevant analytical and numerical results obtained by earlier authors. Moreover, it will become clear that some widely held ideas may be proved or disproved based on straightforward but tedious analysis on the functions C^2H and IU^2; but the detailed proof will not be included.

We will use ρ_2 and ρ_3 to denote the Jacobian vectors describing the motion of the second mass around the first, and that of the third around the centre of mass of the first two masses respectively; they will be called the inner and outer orbits. Corresponding to this we use a_2, e_2, i_2, and f_2 to denote the semi-major axis, eccentricity, inclination and true anomaly of the inner orbit respectively, and a_3, e_3, i_3, and f_3 those of the outer orbit. The normalised masses μ and μ_3 will also be used.

If we use U_i, T_i, I_i and C_i (i=2, 3) to denote the potential energy, kinetic energy, moment of inertia and angular momentum of the orbit ρ_i respectively, then we have

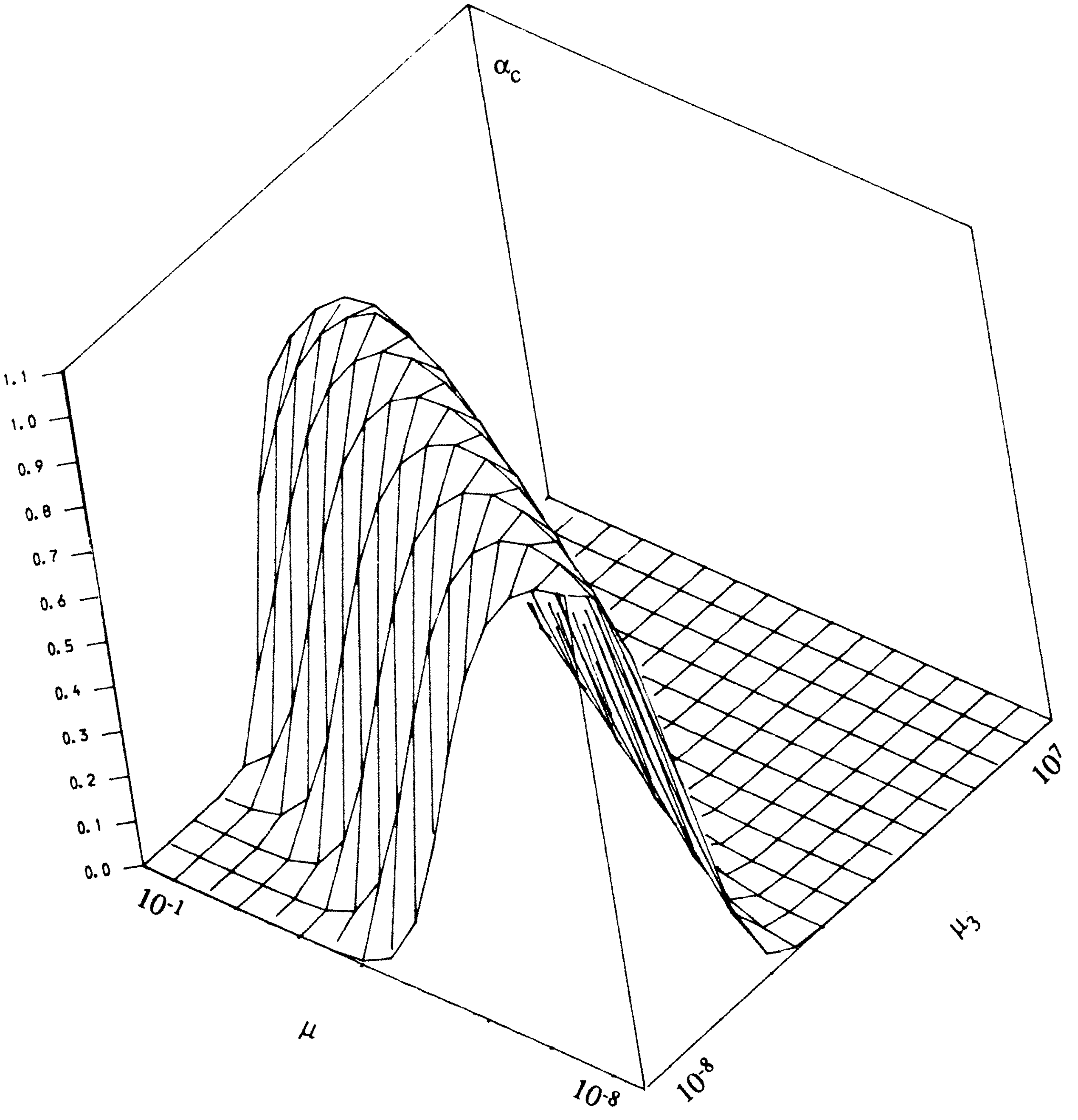
$$\begin{aligned} C^2 &= C_2^2 + C_3^2 + 2C_2C_3\cos i \\ &= \mu^2(1-\mu)^2 a_2(1-e_2^2) + \frac{\mu_3^2}{1+\mu_3} a_3(1-e_3^2) \\ &\quad + \frac{2\mu(1-\mu)\mu_3}{\sqrt{1+\mu_3}} \sqrt{a_2(1-e_2^2)a_3(1-e_3^2)}\cos i \end{aligned} \tag{5.2a}$$

$$\begin{aligned} H &= [(T_2+U_2)+(T_3+U_3)] + (U_{23}+U_{13}-U_3) = H_2 + H_3 + \delta H \\ &= -\frac{\mu(1-\mu)}{2a_2} - \frac{\mu_3}{2a_3} - \left[\mu\mu_3\left(\frac{1}{R_{23}} - \frac{1}{\rho_3}\right) + (1-\mu)\mu_3\left(\frac{1}{R_{13}} - \frac{1}{\rho_3}\right)\right] \end{aligned} \tag{5.2b}$$

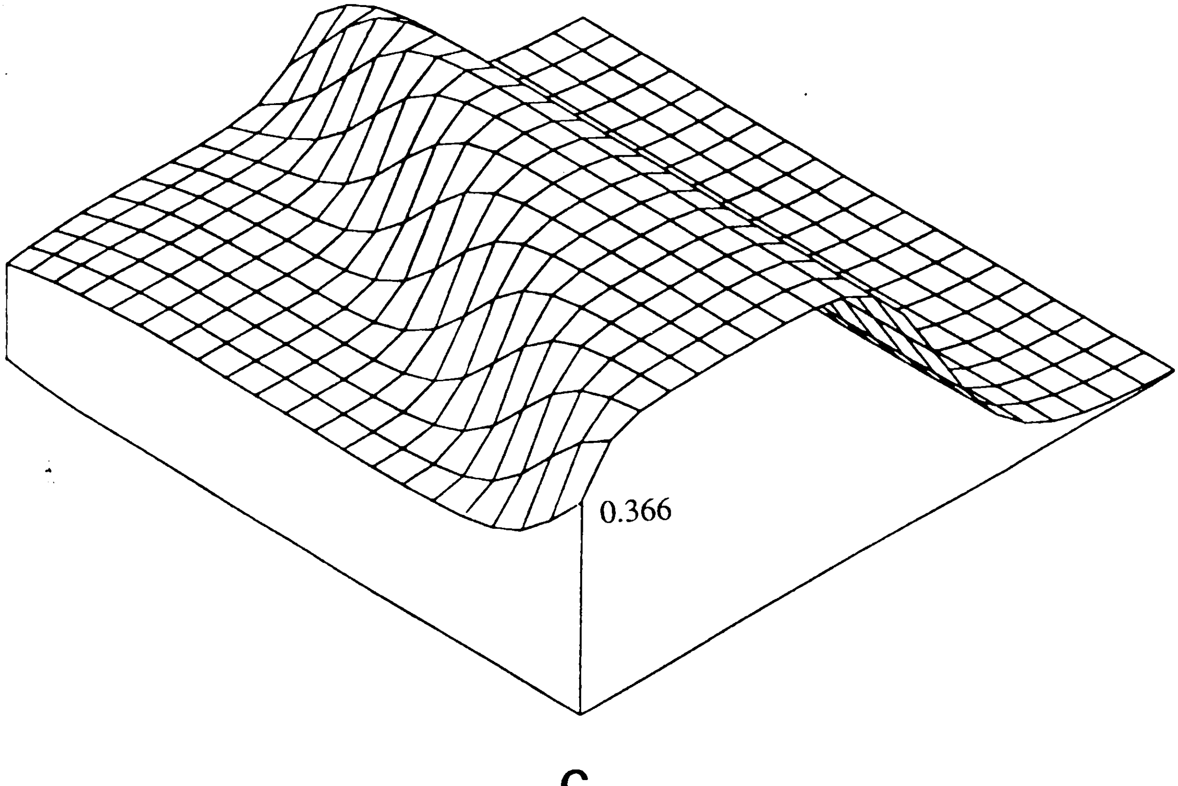


a

Figure 5.6 The critical stability surfaces of the 3-body problem in the O-$\alpha\mu\mu_3$ space with the origin at $(\alpha, \mu, \mu_3)=(0, 10^{-8}, 10^{-8})$. The O$\mu$ and Oμ_3 axes are logarithmic. (a). The surface is monotonic when $e_1=e_2=0$. (b). The surface is tunnel-shaped when $e_1=e_2=0.05$. (c). $e_1=e_2=0$ and $i=50^0$.



b



C

where i is the angle between C_2 and C_3.

Since the function C^2H is scale-free, the following ratios will be useful in simplifying equations (5.2a) and (5.2b), viz.

$$\alpha = \frac{a_2}{a_3} \quad , \quad \alpha_{23} = \frac{\rho_2}{\rho_3} \qquad (5.3)$$

(1). $(C^2H)_c$ and α_c stability criterion

It is shown in the last chapter that a Hill-type stability may be determined by a **C^2H stability criterion**, that is, by compare the actual value of C^2H and its critical value $(C^2H)_c$, where the critical value depends on the masses of the three bodies only. In addition, a more convenient version was introduced by Szebehely & Zare (1976), which is expressed in the ratio of the semi-major axis of the inner binary over that of the outer one, ie. $\alpha=a_2/a_3$. The critical value expressed in α may be solved from the critical value of C^2H, because the latter is a function of the masses and the orbital elements. Obviously this critical value, denoted as α_c, is in general a function of the masses and the other orbital elements. If the actual value of α of a real 3-body system is below α_c then the system is Hill-type stable; thus the stability criterion is sometimes called **α_c criterion** as well.

Walker et al (1980) introduced the critical stability surfaces, namely, the surfaces of α_c plotted against μ and μ_3 in the O-$\alpha\mu\mu_3$ space. They found that when the eccentricities vanish, the critical surfaces are monotonic in μ and μ_3 (see Fig. 5.6a).

(2). Tunnel effect of the eccentricities

It was found by Szebehely & Zare (1976) that the eccentricities and inclination are the most important parameters that influence the value of α_c. In their calculation, a 2-body approximation was used in the energy, namely, the δH term was neglected. Application to the triple star systems shows the validity of α_c in indicating actual stability.

In a later work using the exact expression, Valsecchi et al (1984) found that when eccentricities are introduced into the orbits, the critical stability surfaces suffer drastic collapse on the two side wing regions (see Fig. 5.6b); with e_2 being responsible for the collapse on the μ-side wing, while e_3 accounts for the collapse on the μ_3-side wing. This will be called the '**tunnel effect**' of eccentricities on the critical stability surfaces. Application of their calculation to real Planet-Satellite-Sun systems suggests that the analytical stability criterion should be taken as too restrictive.

Fig. 5.6c is an example showing the effect of the relative inclination of the two orbital plane, from which we see that in addition to an overall collapse, the effect of inclination is very like that of the inner eccentricity e_2.

(3). Upper bounds on α_c

Since the value of C only depends on the semi-major axes, eccentricities and the angle i, it is evident that the dependence of C^2H on θ, the angle between ρ_2 and ρ_3, is completely due to that of δH. A straightforward calculation shows that $\theta=0$ and π are two local minima of C^2H, and $\theta=\pm\theta_0$ (correspond to $R_{13}=R_{23}$) is a local maximum. Moreover, $\delta H(\theta=0) \leq \delta H(\theta=\pi) \leq 0$, hence the value of C^2H at $\theta=0$ is more negative than at $\theta=\pi$. This property is very similar to that of the function Z, and is useful in finding the orbital elements corresponding to the greatest and smallest value of α_c.

From the analysis of the last section we have already proved the result that the critical ratio of α_{23} is below $\rho_2/\rho_L <1$, where ρ_L is the distance of the critical configuration from the centre of mass of the first two masses. In fact based on analysis of the function C^2H, one can show that in the range $(0, \rho_2/\rho_L]$ there is one and only one solution, α_{23}, satisfying the equation $C^2H=(C^2H)_c$.

Let us denote the very simple but important position with orbital element values ($\theta=0$, $f_2=\pi$, $f_3=0$) by Ξ, namely, the inner masses are at their apocentres and the outer one at its pericentre. An important property of this position is that it is the unique position where δH, hence C^2H reaches its most negative value among all possible positions. Consequently, the critical value α_c has its greatest value at this position, based on which and the above bound on the critical ratio in α_{23}, we can obtain an upper bound for the value of α_c, viz.

$$\alpha_c \leq \alpha_c\Big|_{\Xi} \in (0, \frac{1-e_3}{1+e_2}\frac{\rho_2}{\rho_L}] \subset (0, \frac{1-e_3}{1+e_2}) \subset (0, 1). \qquad (5.4)$$

It is widely held that when a 3-body system satisfy the C^2H stability criterion, then both $\alpha_{23}\leq 1$ and $\alpha\leq 1$ are true. The above results show that they are in fact true; but there is no causal relation between them, one must establish them separately.

In their numerical experiments, Walker & Roy (1983) used a cross-over ratio instead of HS-(C) to check whether the hierarchy is still preserved, that is

$$(1-\mu)a_2(1+e_2) \leq a_3(1-e_3) \quad \Rightarrow \quad \alpha \leq \alpha'_x \equiv \frac{1-e_3}{1+e_2}\frac{1}{1-\mu} \qquad (5.5a)$$

However, when μ-> 0.5, this cross-over ratio can go beyond 1, which is very unlikely to correspond to any stable orbits. In the present approach we shall drop the factor $1/(1-\mu)$, thus defining the cross-over ratio as

$$\alpha_x \equiv \frac{1-e_3}{1+e_2} \leq 1. \qquad (5.5b)$$

Obviously, when a system is Hill-type stable, we have

$$C^2H \leq (C^2H)_c \quad \Rightarrow \quad \begin{cases} \alpha_{23} = \rho_2/\rho_3 < 1, & \alpha_c < 1 \\ \alpha_c < \alpha'_x, & \alpha_c < \alpha_x \leq 1 \end{cases} \qquad (5.6)$$

It is widely held that syzygy (aline or conjunction) position is a destructive position for the stability of a many-body dynamical system, particularly when the inner masses are at the apocentre and the outer mass at the pericentre. This position is exactly the position Ξ for a 3-body system, which gives the greatest critical stability ratio α_c. Thus such syzygy positions are the most favourable ones for Hill-type stability. However, it remains to investigate which is true.

It is widely held and was clearly stated by Harrington (1972) based on his observations on the results of numerical experiments that *the higher the values of eccentricities, the less stable the systems are.* A later work by Szebehely & Zare (1976) showed that the value of α_c always decreases with the eccentricities, hence confirming Harrington's observation by analytical criterion. However, we will see in later sections that there are exceptions to Harrington's observation; moreover, there are exceptions to the behaviour of α_c with eccentricities as well. By a careful analysis of equations (5.2a, b) at the position Ξ with i=0, one can construct a function K such that the solution to the equation $K=(C^2H)_c$ bounds the value of α_c from below, but equals α_c when the eccentricities vanish. One can show that the solution of the constructed equation increases with both eccentricities if their sizes are small. Thus the value of α_c at the position Ξ for the coplanar problem always increases with eccentricities at least in a small range.

Harrington (1972) also observed that for 3-body systems with given mass parameters, it seems that the 'best index of stability' is $a_3(1-e_3)/a_2$. The strongest upper bound on α_c defined by equation (5.4) may be written as $a_3(1-e_3)/a_2(1+e_2)=\rho_L/\rho_2$, where the right side is a scale-free function of the masses only. If one notices that in the case of three almost equal masses, the upper bound given by equation (5.4) is quite close to the actual value of α_c (cf. Fig. 5.6b), and that Harrington's experiments used such masses, then one sees that a possible index of stability may be of the form $a_3(1-e_3)/a_2(1+e_2)$. The less significant position of the factor $(1+e_2)$ in the denominator may explain why it was not noticed by Harrington.

These are the results obtained following the conventional role granted for the function C^2H; however, the numerical experiments of the following sections will suggest that such results can hardly be regarded as of general importance. In what follows we will apply the analysis of C^2H to a different use.

(4). Noticeable variation patterns of the a's and e's

It was found by Walker & Roy (1983) from their numerical integrations on direct CHT systems with initially circular orbits that *on commencing the numerical integration procedure the semi-major axis of the inner binary always decreases, whereas that of the outer binary always increases*. We will try to explain this simply using the conservation of C, H and their combination C^2H. The problem may be studied under the following two assumptions.

(i). If a 2-body approximation works well for the total energy H, then its conservation law requires that the two semi-major axes must change in opposite ways with time.

An independent use of C does not provide much useful information because all important orbital elements are involved. However, we can make use of the scale-free property of the quantity C^2H. The value of α_c is the solution to the equation $C^2H-(C^2H)_c=0$ (Fig. 5.6), which for a CHT system is mainly influenced by the eccentricities. The most significant effect is the decrease of α_c with eccentricities, especially when both eccentricities are changing in the same way. Obviously, this effect is also true for the solution to any equation $C^2H-(C^2H)_0=0$, with $(C^2H)_0 \leq (C^2H)_c$.

(ii). The second assumption is to assume that the above effect is also true
for certain $(C^2H)_0$ sufficiently close to but greater than $(C^2H)_c$, namely,
for systems not stable in the sense of Hill.

Combining the above two assumptions we obtain that most observable variations in the orbital elements are either of those shown in Table 5.1. It is worth noting that as long as the assumptions are satisfied, then this Table applies to long term trends, as well as short term changes.

Table 5.1

(a) :	$a_2\downarrow$	$e_2\uparrow$	$a_3\uparrow$	$e_3\uparrow$
(b) :	$a_2\uparrow$	$e_2\downarrow$	$a_3\downarrow$	$e_3\downarrow$

Here we must emphasise that Table 5.1 only includes most of the observable variations in the major orbital elements. They are only 'observable' ones because complicated small variations do exist; 'most of the observable' ones because not all observable variations are included.

The behaviours listed in Table 5.1 may be violated if either of the assumptions fails to be held by a system. Moreover, even if both were held, there are still observable changes when the two eccentricities are changing in opposite ways. This is especially true when one of the binaries has dominant masses.

The results found by Walker & Roy (1983) can be explained as follows. They found that (a) of the Table is always the case, because they made the observation only *immediately after commencing* the integration, when the only possibility for both eccentricities is to increase; and the eccentricities must change because of the perturbation. Moreover, the only possibility for the first noticeable change in α is to decrease (either gradually or suddenly), because the initial conditions they chose correspond to the greatest value of α.

Our numerical experiments on initially elliptic orbits in the next section will produce examples deviating from Table 5.1. For systems deeply inside the Hill-type stability region, ie. $\alpha<<\alpha_c$, usually the short term changes are not easily observed. But about 90% of the system's long term changes agree with Table 5.1. For systems inside or outside the Hill-type stability region, there are noticeable short term changes; among them about 80% agree well with those listed in Table 5.1. However, it remains to explain in detail why the two assumptions and Table 5.1 are satisfied by so many systems.

(5). The empirical stability ε parameters

Walker and Roy (1983) introduced the parameters ε_{23} and ε_{32} to characterise respectively the size of the disturbances of the inner orbit on the outer and the outer on the inner. They are defined by

$$\varepsilon_{23} = \mu(1-\mu)(\alpha_{23})^2 \quad , \quad \varepsilon_{32} = \mu_3(\alpha_{23})^3 \ . \qquad (5.7)$$

In order to compare our study with Walker & Roy's in the following sections, we need to calculate the parameters such as α_c and α_x in the $(\varepsilon_{23}\ \varepsilon_{32})$ space.

The critical stability surface may be first calculated in the O-$\alpha\mu\mu_3$ space by solving a set of algebraic equations, then transformed into the $(\varepsilon_{23}\ \varepsilon_{32})$ space. A property of the transformation was given in Walker et al (1980), namely, when the critical stability surface in the former space is mapped onto the critical stability surface in the latter space, no points can change from one side of the critical stability surface to the other side during the transformation. The proof was given for the circular case, but it is a general result because the Jacobian matrix of the transformation is non-singular except on part of the boundary enclosing the Hill-type stability regions. Another property is that all points below the critical stability surface in the former space are hierarchically stable, whereas in the transformed space those below the surface defined by

$$\alpha_{23} \leq 2\sqrt{\varepsilon_{23}} \quad , \quad \alpha \leq \alpha' \equiv 2\sqrt{\varepsilon_{23}}\,(1-e_3)\ /\ (1+e_2) \qquad (5.8)$$

are physically meaningless. However, this is simply a deformation of the transformation

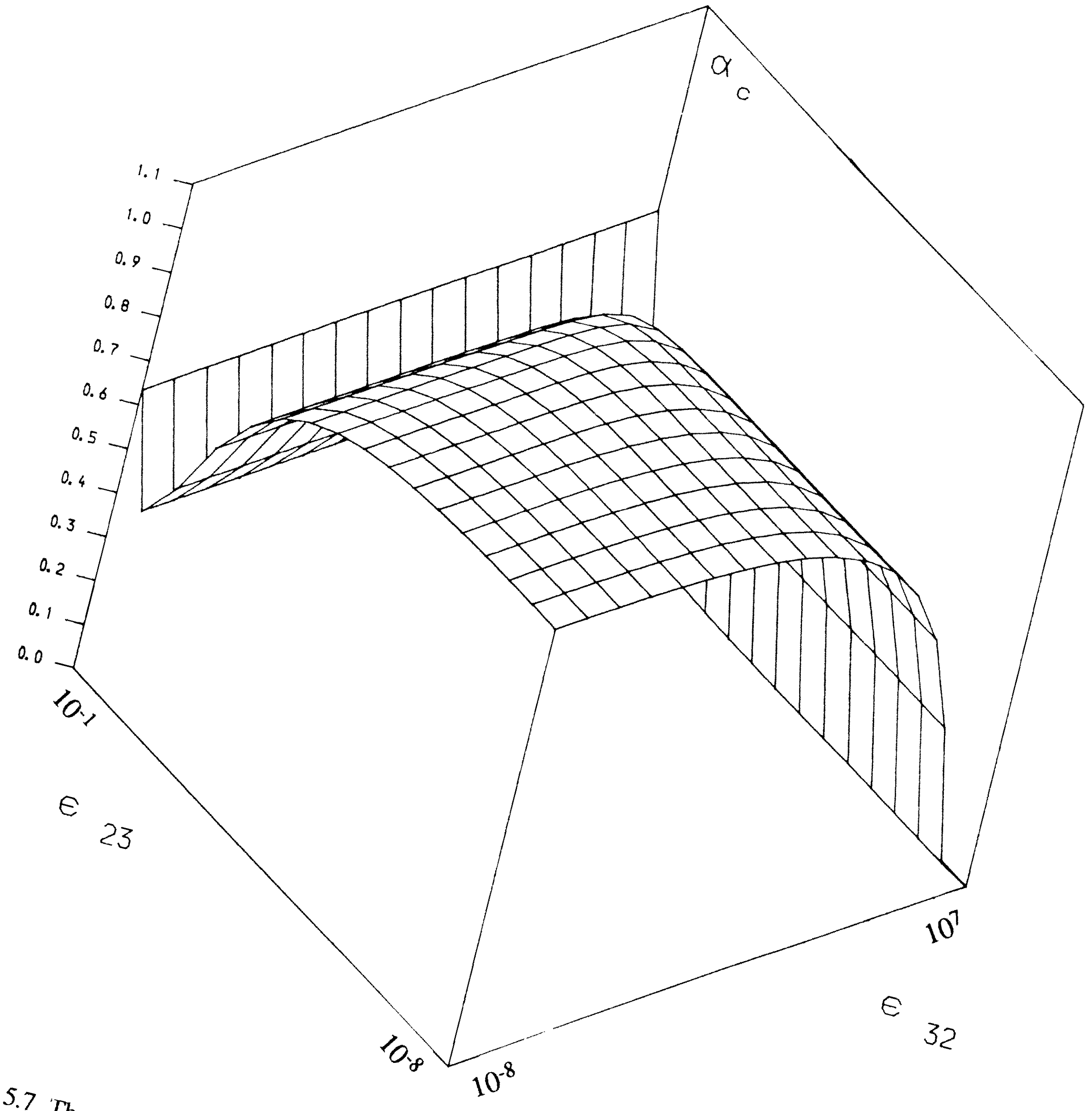
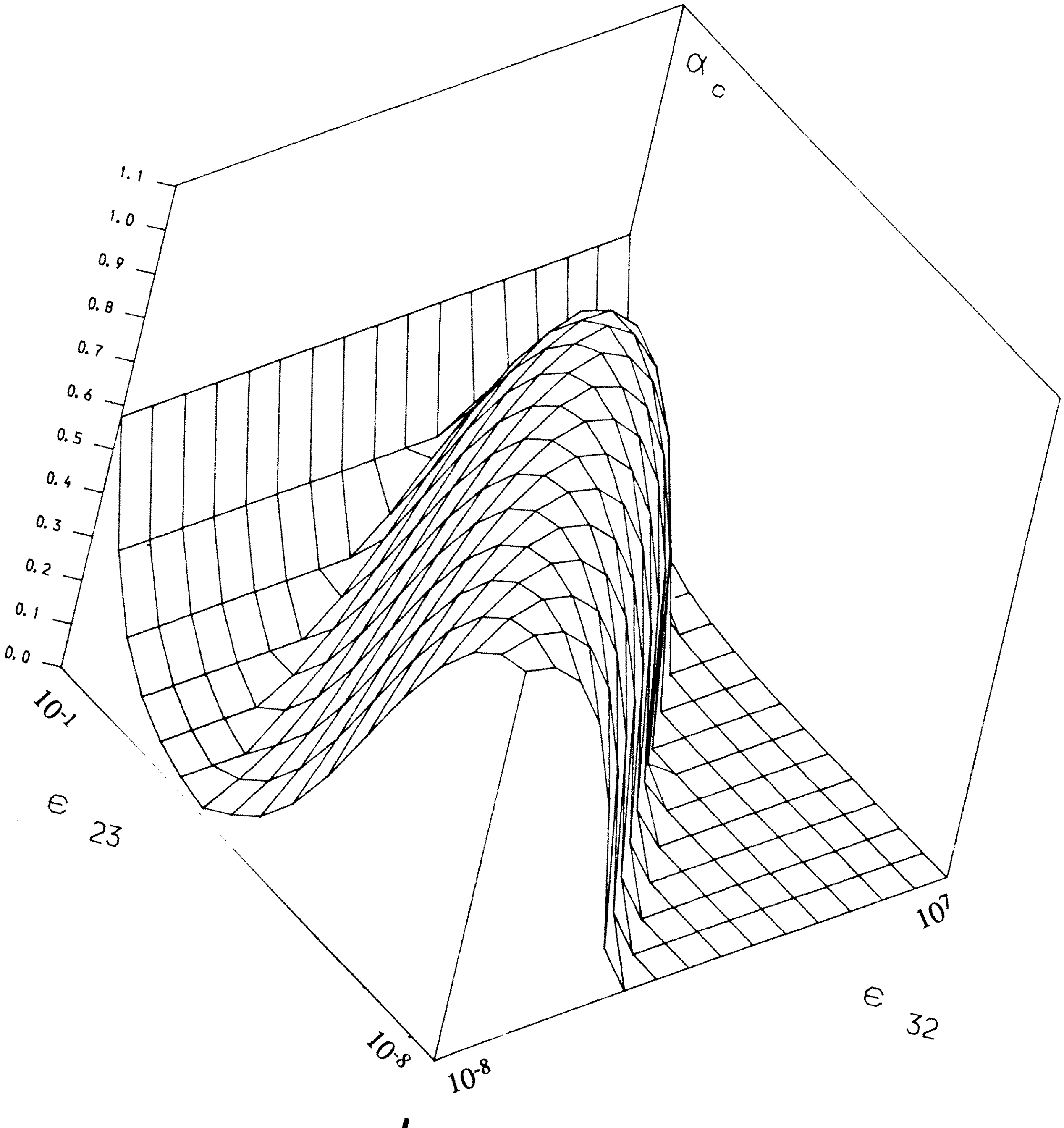


Figure 5.7 The critical stability surfaces in the O-$\alpha\varepsilon_{23}\varepsilon_{32}$ space with the origin at $(\alpha, \varepsilon_{23}, \varepsilon_{32})=(0, 10^{-8}, 10^{-8})$. The O$\varepsilon_{23}$ and Oε_{32} axes are logarithmic. (a). $e_1=e_2=0$. (b). $e_1=e_2=0.05$.



b

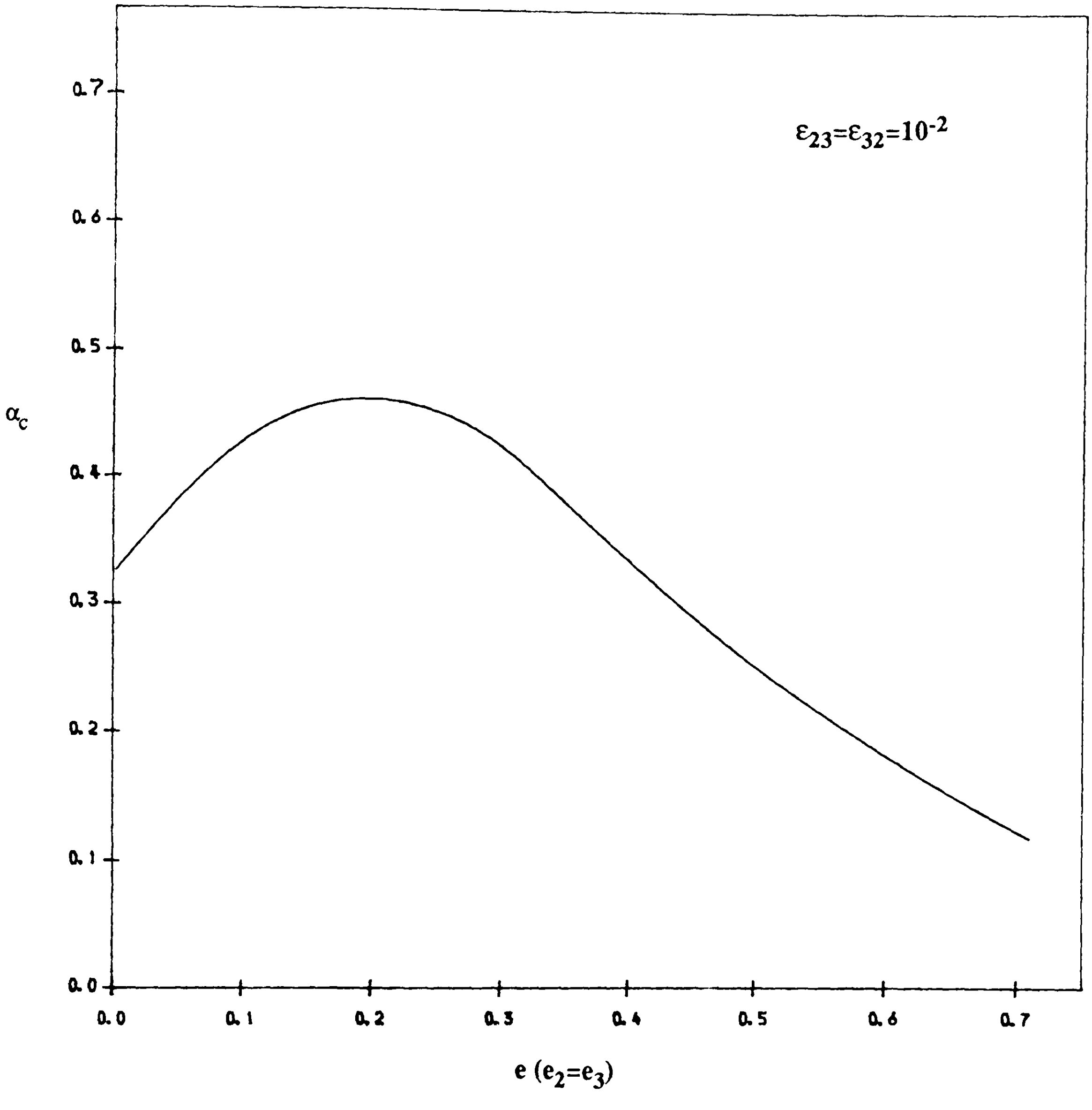


Figure 5.8 Illustrating the increase of α_c with eccentricities for a fixed (ε_{23}, ε_{32}) pair. This is due to a distortion produced by the transformation, equation (5.1).

(5.1), which transforms the region with $\mu > 0.5$ in O-$\alpha\mu\mu_3$ space into the region below the above surface. The shapes of the critical stability surfaces in the two spaces are similar (see Fig. 5.7), which is understandable if we look at their level sets.

As was shown analytically, the value of α_c can increase (slightly) as the eccentricities increase, and this phenomenon is more obvious in the ε-space (see Fig. 5.8). Although this is a distortion due to the transformation, it suggests that stability might increase as eccentricities increase. We will give some examples supporting this point from our numerical experiments. We will also give examples suggesting that aline configurations are not necessarily always the worst positions for stability, which has hitherto been taken for granted, but opposed by the analytical Hill-type stability criterion.

Based on equation (5.4) and Harrington's (1972) result we will use the following stability indicator which can be calculated directly from the results of Walker (1983), viz.

$$\alpha_I = \alpha_{co} \cdot (1 - e_3) / (1 + e_2) \qquad (5.9)$$

where α_{co} is the critical value of the ratio of the semi-major axes calculated for the circular system with the same mass parameters when the bodies are in conjunction. The essential feature of this stability indicator, α_I, is that the effect of eccentricities is simplified by neglecting the 'tunnel-effect', which will be shown in the next section to be not easily detectable by short term numerical integrations. However, the properties of this stability indicator are different in the ($\mu\mu_3$)-space and the ($\varepsilon_{23}\ \varepsilon_{32}$)-space due to the deformation of the transformation. For example, α_I is above α_c in the ($\mu\mu_3$)-space, but this is not necessarily true in the ($\varepsilon_{23}\ \varepsilon_{32}$)-space, nor is it necessarily close to it.

In the ($\varepsilon_{23}\ \varepsilon_{32}$)-space, the value of α'_x is more conveniently given in the ε's rather than μ's. This can be done by solving equations (5.7) for α,

$$\varepsilon_{23} = \mu(1-\mu)\alpha_{23}^2 \quad \Rightarrow \quad \alpha_{23}^2\,\mu^2 - \alpha_{23}^2\,\mu + \varepsilon_{23} \quad \Rightarrow$$

$$\Rightarrow \; 1-\mu = \frac{1}{2\,\alpha_{23}}\left(\alpha_{23} + \sqrt{\alpha_{23}^2 - 4\,\varepsilon_{23}}\right)$$

$$\Rightarrow \left(\alpha_{23} + \sqrt{\alpha_{23}^2 - 4\,\varepsilon_{23}}\right) = 2\,\alpha_{23}(1-\mu)\,;$$

$$\alpha \le \frac{1}{1-\mu}\cdot\frac{(1-e_3)}{(1+e_2)} \quad \Rightarrow \quad \alpha(1-\mu) \le \frac{(1-e_3)}{(1+e_2)} \quad \Rightarrow$$

$$\Rightarrow \qquad \alpha_{23}(1-\mu) \le \frac{(1-e_2)\;/\;(1+e_2\cos f_2)}{(1+e_3)\;/\;(1+e_3\cos f_3)} \equiv W\,.$$

$$\left(\alpha_{23} + \sqrt{\alpha_{23}^2 - 4\,\varepsilon_{23}}\right) \le 2W \quad \Rightarrow \quad \alpha_{23} \le W + \frac{\varepsilon_{23}}{W} \quad \Rightarrow$$

$$\Rightarrow \qquad \alpha \le \frac{(1-e_3)}{(1+e_2)}\cdot\left\{1 + \varepsilon_{23}\left(\frac{(1+e_3)\,/\,(1+e_3\cos f_3)}{(1-e_2)\,/\,(1+e_2\cos f_2)}\right)^2\right\} \equiv \alpha'_x\,.$$

From this we see that the properties of the two cross-over ratios defined by equations (5.5a) and (5.5b) are not changed due to the transformation. Thus we will choose to use that defined by the later equation, ie. α_c, whose expression is not changed either.

5.3 Numerical Experiments on 3-Body Systems I - A First Test of the Tunnel Effect

Valsecchi et al (1984) found the 'tunnel effect' of the eccentricities on the critical stability surfaces of the 3-body problem. Their application of the α_c stability criterion to the Planet-Satellite-Sun systems showed that the elliptic criterion should be taken as too restrictive, since the calculation gives α_c far below the actual α of these systems, and it is a well-known fact that the systems are quite stable. On the other hand it has been shown by Walker & Roy (1983) that the circular criterion is a very good indicator for a practical stability. Thus there is a need to investigate by numerical experiments whether the tunnel shaped stability surface does reflect the truth.

In this section we will make a very first investigation on the question. One has to carry out a systematic study to obtain any certain conclusion. The study of this section is

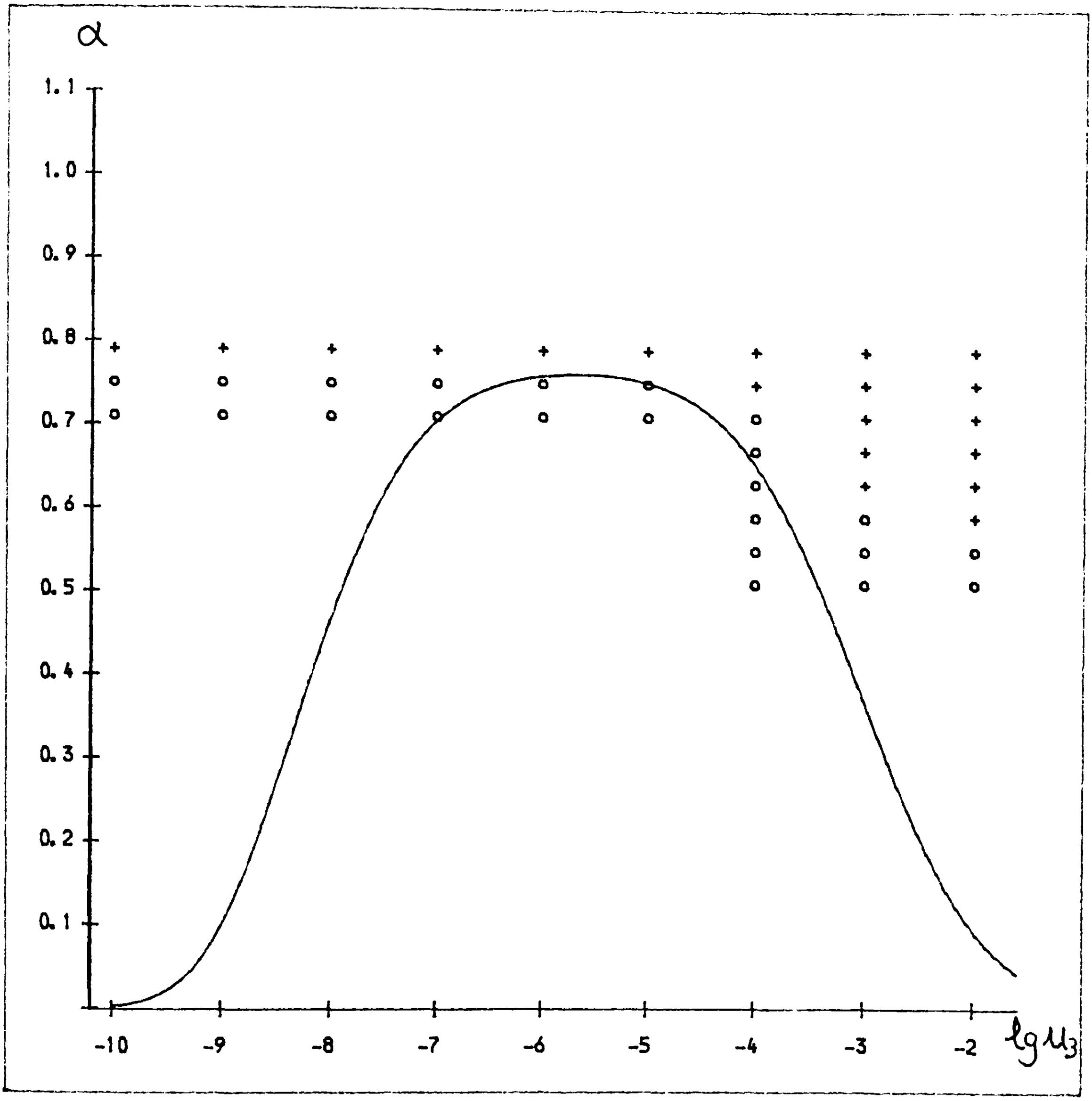


Figure 5.9 Numerical investigation (up to 600 synodical periods) on the tunnel-shaped surface of stability for a cross section close to the Sun-Neptune-Pluto subsystem. The mass parameter is μ=5.195×10^{-5}, and the initial eccentricities are e_2=0.008 and e_3=0.2.

limited to systems with parameters close to that of the Sun-Neptune-Pluto system. In Table 5.2 we give a class of the values of α_c by introducing more and more the real orbital elements of the Sun-Neptune-Pluto system into the calculation. It is seen that there is some uncertainty about the system's stability.

Table 5.2 Values of α_c for Sun-Neptune-Pluto system at Ξ

($\mu = 5.20\text{x}10^{-5}$ $\mu_3 = 5.00\text{x}10^{-9}$ $\alpha = 0.762$)

	$e_2 = 0$ $e_3 = 0$	$e_2 = 0.00858$ $e_3 = 0.250$
$i = 0^0$	0.915	0.314
$i = 17^0$	0.671	0.282

Shown in Fig. 5.9 are the numerical integration results for direct CHT systems with initial eccentricities close to those of the Sun-Neptune-Pluto system, namely, e_2=0.008, e_3=0.2. The integrated systems have the same parameter μ=5.195x10^{-5}, but different μ_3=10^{-10}, 10^{-9}, ..., 10^{-2}, and the initial α's take the values in a range covering 0.76. In the diagram, a '+' denotes a cross-over of orbits, a 'o' denotes a system which is stable up to 600 synodic periods, which is the limit of the numerical experiments.

It is seen that no tunnel effects appear on the diagram. This suggests that a monotonically decreasing stability surface (as obtained by the circular criterion) seems to be the clearer qualitative feature of the problem, while the tunnel-shaped elliptic criterion may be of little practical value. However, a longer-term investigation is required for a more decisive conclusion.

5.4 Numerical Experiments on 3-Body Systems II - Systematic Investigation of Elliptic Motions

In this section we will make a more complete and systematic numerical investigation of the behaviour of the initially elliptic CHT systems using the concept of the hierarchical stability conditions (cf. chapter 1). But first let us justify the stability conditions so that they can be implemented on the computer.

In the definition of hierarchical stability, instability by collision is not listed since the collision manifold is of measure zero (in contrast, little is known about the escape instability). However, we have given this word a different meaning in our numerical experiments. A collision is said to occur if one of the eccentricities goes up to so high a value that the accuracy of our numerical routine becomes unreliable. This instability may in fact be only a process preceding an escape. A second point is that both instability conditions HS-(A) and HS-(B) can occur in the Hill-type stability region, which implies a potential incompatibility between hierarchical stability (which by definition involves all three conditions) and Hill-type stability. This was not so serious in Walker et al's (1980, 1983) work, but can be fatal for the picture they have built, in which great consistency was displayed between these two notions of stability. Related to this is the crucial difficulty in how to decide whether a drastic change has occurred or not; this is a very subjective matter. Many such ambiguous changes have been observed during the numerical experiments not only outside the Hill-type stability region, but also inside it. The experimenter is faced with a choice; either he forgets about this ambiguity inside the Hill-type stability region in order to preserve the 'neat' picture, or he abandons the picture. The present author has chosen the latter because of the experiments like [0226] shown in Fig. 5.11, where assured instabilities (A) and (B) have been observed inside the Hill-type stability region. The sets of experiments, [1062] and [2062], were also helpful in choosing this decision. A detailed discussion on this matter is presented in this section.

Much analytical work has been carried out (see, for example, Marchal et al, 1984) concerning the condition of escape. However, it cannot be applied to the present study to determine, before the integration, which body will escape, since our numerical experiments are begun from mirror configurations. Escape might be judged to happen if one of the bodies is thrown far away from the centre of mass, but in fact we check the sign of the energy of the two-body subsystems. Subsequent recapture is possible, but since this is obviously not a stable situation it is not the major interest of this study. It may also be noted that instability (B) can on occasion be very severe such that a collision, or escape occurs, though it can also be less severe. The latter case is too difficult for the numerical routine to deal with, and it is up to the experimenter to judge from experience. Therefore the following formulation of instabilities was adopted so that numerical routines could be used to signal the following cases:

(a). escape - energy of any 2-body subsystem becomes nonnegative;

(b1). collision - eccentricity of any 2-body subsystem grows beyond 0.99;

(b2). close encounter - a close approach between two bodies occurs resulting in drastic changes in the semi-major axes and eccentricities of the orbits of the 2-body subsystems, but is not so severe as (a) and (b1);

(c). cross over - the pericentre distance of the outer orbit, ρ_3 , becomes less than the apocentre distance of the inner orbit, ρ_2.

Note that instability (c) is slightly different from the 'cross-over' version in Walker and Roy (1983). The advantage of condition (c) over HS-(C) is that the very unstable and less interesting cases with $\alpha > 1$ can be precluded beforehand. Another difference between our numerical experiments and Walker and Roy's (1983) is that in the present study we do not terminate computing when a less severe close encounter (b2) occurs. It should also be noted that (a) and (b1) must be preceded by one or several (b2), but (c) is not necessarily preceded by (b2). (b1), however, may signal the occurrence of (a). In this approach we present in Fig. 5.11 the complete results without noting instability (b2). Only after a detailed discussion based on careful observations of orbital stability, do we then schematically show in Table 5.2 the result with (b2) noted.

It will be seen that only one of the sixteen plots supports the neat empirical stability picture built up in the case where the numerical experiments begin from orbits initially circular. We will see that the introduction of initial eccentricities drastically complicates the behaviour of the systems. The motion turns out to be so irregular that in any particular experiment the uncertainty in any measure of the life-time reading is far bigger than one synodic period. This, together with the newly observed valley and plateau structures, makes it almost impossible to fit empirical stability curves to the data. In addition, many phenomena hitherto unencountered in this work were observed such that the 'close encounter' version is not easy to use. Instabilities in the Hill-type stability regions completely destroy the attractive picture of empirical stability regions outside the Hill-type stability regions. The empirical stability picture must therefore be modified drastically so that the empirical stability curves, if a curve-fitting procedure is possible, go straight into the Hill-type stability regions below α_c. Accordingly the 'tunnel effect' of eccentricities on the critical stability surface may be completely irrelevant.

The Numerical Method

We present here the result of several hundred numerical integration experiments on initially elliptic, corotational, coplanar 3-body systems. All the experiments were carried out on the ICL 3980 mainframe computer at Glasgow University, using the same numerical routine that Walker and Roy (1983) used. In this routine the mutual radius vectors are calculated by a tenth order Taylor series, where the derivatives are evaluated by recurrence relations. The programme incorporates an automatic step-length regulator which shortens or lengthens the integration length of the computer in order that the error caused by truncating the Taylor series after the tenth order is less then a given tolerance (10^{-12} in this approach).

As is known, for essentially any numerical method local truncation errors can be controlled; estimations for the accumulated truncation error after many steps are not often possible. In a chaotic system such as the N-body problem two trajectories with nearby initial conditions depart at an exponential rate, thus the integration error must grow in a manner more complicated than exponential divergence. For this reason, the orbits obtained on the computer may be very far apart from the real solution, but they do capture the reliable properties of the real orbits. The accuracy of the integration routine is also affected by round-off error of the computer. Though pure round-off error can be studied statistically, a rigorous analysis is impossible when modified by truncation error. With these facts in mind, we have to find another way to get some rough idea about error accumulation and for how long we can run the numerical integration. For example, we can run the programme for fictitious 3-body systems with ε's -> 0 or for the linearly stable equilateral motion, whose orbital elements should remain constant. Such an estimation gives the result of about 6000 synodic periods for an 0.1% relative error in the position. Programmes have been run up to 1000 synodic periods if no instability sets in before this time scale. Energy and angular momentum were used to check the integration error, though they are not very adequate for this role (an integral of motion is not sensitive to integration error even if the motion is irregular). The relative error of them on commencing and at the end of the integration is found always below 10^{-7}.

The initial conditions are chosen such that the masses form a mirror configuration on commencing the integrations. It is useful to be clear about the aspects of this choice. One consideration is based on the fact that the trajectories after that epoch are mirror images of their trajectories before that epoch (Roy & Ovenden, 1955). Therefore, by studying one direction of time we also gain knowledge of the other one, so that the time-scale is cut down. To further this point the initial conditions are actually chosen at what is believed to be the worst configurations (ie. Ξ): the body in the inner orbit at apocentre, with the outer mass at its pericentre, and all masses collinear. Secondly, commencing from a mirror configuration may give more chance of picking up stable periodic trajectories (which is very rare in irregular regions), because the occurrence of two mirror configurations guarantees periodicity of the motion. On the other hand, one advantage of this study is that we pursue a method of 'ensemble' study instead of a study of long time behaviour; but this choice of mirror configuration means we are only studying a very small subset of the complete 'ensemble'. Nevertheless we have confidence that this subset captures the general feature of the whole set.

Since the ambiguity of detecting instability is magnified due to the introduction of eccentricities, we will first present the result with only assured instabilities noted and then carefully compare the difference with Walker and Roy (1983). Only after this is it possible to make some comments on those ambiguous close encounters. The original plots of hierarchical stability lifetime of the orbits in synodic periods (denoted by N_S) versus the initial α values are presented in Fig. 5.11. The diagrams there are named by a set of number of the form [2062], being a shorthand for the values of $[e_2\ e_3\ \varepsilon_{23}\ \varepsilon_{32}]$, where the first number 2 means that e_2=0.2, the second number 0 means e_3=0.0, while the third number 6 stands for $\varepsilon_{23}=10^{-6}$, and the fourth number 2, $\varepsilon_{32}=10^{-2}$. In Fig. 5.12 some typical examples are shown of the variation with time of orbital elements observed during the integration; they will be referred to as a, b etc.

In what follows we will use some descriptive words for the size of eccentricities, viz., 'very small' for the region (0 , 0.1), 'small' (0.1 , 0.25), 'moderate' (0.25 , 0.55), 'high' (0.55 , 0.75), 'very high' (0.75 , 0.99) and 'collision' for values above 0.99. Of course such words can not be accurate, and there is an uncertainty of about 0.05 in the value of the dividing points. Nevertheless we find from the numerical experiments that this division is useful: a shift from one region to another is of significance for the stability of the systems. Based on this division, we will also use terms like 'stable mode', 'sub-stable modes', 'random stable mode' and 'comet-like orbit', which will be explained in the appropriate place and Fig. 5.12.

General Behaviour of Eccentricities and Semi-major Axes

It is obvious that for systems with very small values in both ε's (eg. $\leq 10^{-6}$), the variation of the orbital elements, a's and e's, is very small and smooth; and the variations are similar for different initial conditions. According to perturbation theory, such systems can be regarded as changing almost linearly with time. It is observed that, during a period of 1000 synodic period, the value of initially vanishing e grows up to the order 10^{-2} , while the change of the other e is below 10^{-3} and the a's stay almost constant.

Therefore if one of the initial orbits is circular, the other one having a small eccentricity, then during the time limit of integration the stability would depend mainly on how the value of the eccentricity grows in the initially circular orbit. In such a system

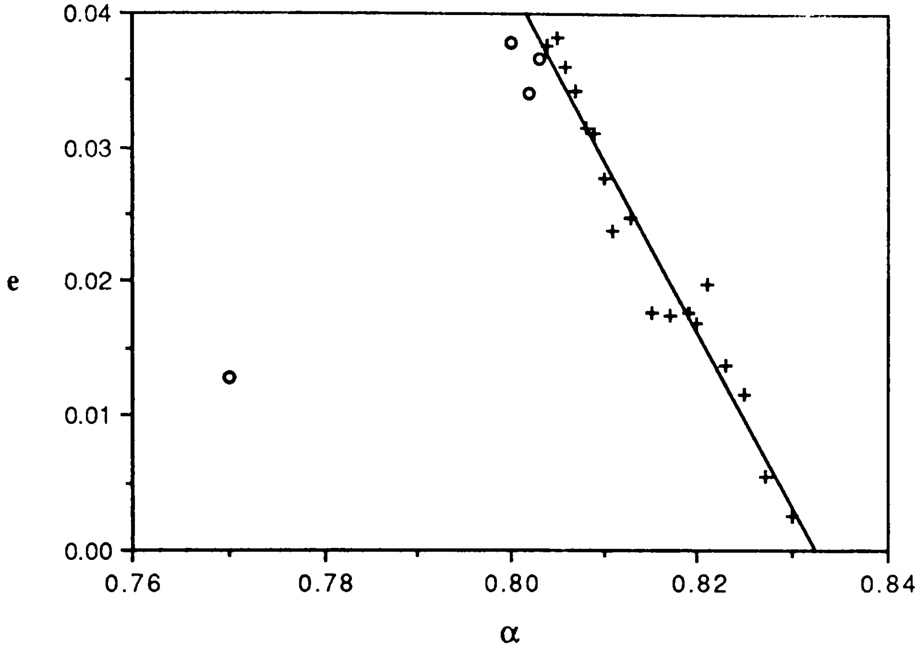


Fig. 5.10 The terminating value of e_3 vs. initial value of α for the set of experiments [2066]. A circle 'o' denote stability, and a cross '+' cross-over of orbits, as is the same in Fig. 5.11. A linear curve is fitted to the data, with the exception of the only stable system at left bottom corner of the diagram. Those stable systems indicated by the circles close to the line are very likely to suffer cross-over instabilities for slightly longer investigation, while the one in the corner is not.

the only instability is the cross-over of orbits; the accumulation of instability could not become large enough to lead to either collision or escape, while a close encounter is highly improbable. From equation (5.5b), it is seen that the terminating value of this specific eccentricity - that is the value of it when a cross-over is detected - is a linear function of the value of the initial ratio of semi-major axes ($\alpha \sim (1-e_i)$ to the first order in eccentricity). One example is shown in Fig. 5.10 for the set of experiments, [2066], of Fig. 5.11. In fact, this approximation is more adequate for initially circular outer orbits than for the inner ones. Recalling that the a's and e's vary almost linearly with time we see that for the diagrams denoted by [.....66], the validity of the above approximation is also reflected in the lifetime versus ratio of semi-major axes plot and a linear fitting is more reasonable than an exponential fitting. Moreover, it is seen that [1066] and [2066] show better linearity than [0166] and [0266].

A phenomenon worth noting concerns the systems with at least one $\varepsilon > 10^{-6}$. If both ε's are of the same order, it is understandable that the result given in section 5.2 concerning 'most observable' changes in a's and e's is a good approximation (cf. Fig. 5.12a). When one of the two perturbations dominates over the other (eg. one ε is 10^{-6}, the other 10^{-2}), some other possibilities are also expected. However, except in one of the cases with both $\varepsilon = 10^{-6}$, it is found that e_2 , a_3, e_3 almost always vary in the same manner, while a_2 varies in a direction opposite to them, if these changes are noticeable. This was analysed in the last section, the notable point being that this result captures the feature of almost all systems though the method to obtain it is by no means rigorous. It is also noted that this implies that energy and angular momentum must transfer in the same direction, either from inner orbit to outer or vice versa, while the parameter C^2H of the two subsystems must vary in the same way.

As a consequence of this observation, the behaviour of the four elements, a's and e's, can be understood by looking at one of them only. We choose one (sometimes two) of the eccentricities and consider its (or their) effect on stability.

Complexity of the Life-time vs. Ratio of Semi-major Axes Plot

The motion of an N-body system is chaotic. This complexity of phase space will be reflected in our life-time vs. ratio of major axes plots. A smooth curve fitting to samples each with only about thirty elements should therefore not be expected. The curves fitted in Walker and Roy (1983) must be regarded as statistical ones. Because of the considerable difference between initially circular and elliptical motion, a thirty-element sample, good enough in the former case to obtain reasonably fitted curves, is insufficient in the initially elliptic case; a much larger sample is required in the present study. Because of limits on computing time and expense, we test this only tentatively in the next section. No attempt is made to fit curves to the data in Fig. 5.11. Instead we will describe the result in this subsection and try to make some comments suggested by these experiments. The experimental sets are now considered in turn.

[1066] : This is one of the simplest sets of experiments. During the integration period the orbits of both 2-body subsystems suffer little change. Neither collision nor escape was detected up to the time limit of 1000 synodic periods (SP hereafter), nor is close encounter a useful concept for detecting instability. All instabilities are due to the slight increase in e_3 which eventually leads to a cross-over of orbit. It is observed that the value of e_3 for all orbits grows in almost the same way, and its terminating value is close to a linear function of the initial α's (cf. Fig. 5.10). The biggest value of the outer eccentricity is found for those stable orbits, which is about 5.0×10^{-2}. Should we try to fit a curve to this set of experiments, a linear rather than an exponential one is more suitable. A simple calculation by equation (5.5b) shows that for the orbit with initial $\alpha =$ 0.8 (which is below α_c) to cross over, a small value of 9×10^{-2} in the outer eccentricity will suffice. Therefore uncertainties exist for the stability of the systems which have survived up to 1000 SPs: the linear fitting may go straight into the Hill-type stability region, however, an exponential curve may be fitted if longer investigations are carried out.

[1062] : This is one of the most interesting sets of experiments, which is quite well approximated to by the circular restricted 3-body problem. The perturbation of the outer orbit on the inner one is strong, while the inner on the outer is weak. Instabilities are therefore due to the changes the inner orbit suffers. Again neither escape nor collision instability has been found up to the time limit of the integration; but close encounters do occur in many cases which are not yet strong enough to lead to a cross-over or other instabilities; they are thus not included in the plot. The outer orbit suffers little change, values of the outer eccentricity never grow above an order of 10^{-5}. Noticeable changes were observed in the inner orbits, a_2 and e_2 always changing in the opposite way: we choose to look at e_2 only.

One of the most apparent features of this plot is that there is an obvious valley in the range $\alpha \in (0.53, 0.55)$ and plateau in $\alpha \in (0.55, 0.59)$ so that no curve fitting would be successful. Let us denote 'plateau' by 'P' and 'valley' by 'V'. It is observed that the group of orbits denoted as PI are more stable than those in group P0, although inside each group the orbits on the right edges are less stable than the others. Up to 1000 synodic periods, the eccentricity of group PI changes almost periodically between 0.01 and 0.3 (termed a 'stable mode', see Fig. 5.12b, e and f); while that of group P0 behaves in this way for about one or two hundred synodic periods and then within one or a few synodic periods jumps to another mode such that the outer eccentricity varies fairly regularly within the moderate regions between 0.1 and 0.5 ('random stable mode', Fig. 5.12b). In a sense the eccentricity changes of the orbits in group P0 are similar to the unstable ones in the group V which suffer cross-over; but P0 survived because it has a smaller value of initial α. Since the orbit on the far left of the plot is already on the edge of the Hill-type stability region and still suffering early close encounters (jump of mode), we expect this to penetrate into the Hill-type stability region, thus close encounter is not consistent with the concept of Hill-type stability, nor with a study on condition HS-(C). Even if we note the close encounter instability, there is still no way of bringing down the stable plateau PI, hence a monotonic curve fitting would not be successful, nor can we fit a curve to points except group PI and explain PI by invoking commensurability as Walker and Roy (1983) did; because this is a wide 'band' instead of a sharp 'peak'. From what we will see in the following plots we conjecture that P0 may also have further complicated fine structures.

[1026] : In this case there is no α_c because of the distortion produced by the transformation from the μ-space to ε-space. Both escape and collision of the outer orbit have been found in addition to cross-over. Although a mixture of stable and unstable orbits is apparently displayed in the plot, those stable orbits, with the exception of the two on the far left, are believed to be unstable for a longer investigation: if close encounter is used, they can be brought down to reasonably shorter lifetimes. In all these experiments, it is always the outer orbit that suffers significant disturbance, while the inner one is only slightly disturbed - a situation for which the elliptic restricted 3-body problem is a good approximation. The size of a_3 and e_3 are always found to vary in the same trend. Again we choose the eccentricity of the outer orbit as the characteristic parameter to describe the variation of the outer orbit.

It is found that the first two orbits on the left of the plot are very stable, the size of the outer eccentricity being bounded below 0.15 during the integration period. In contrast to this, all other orbits, including those which survive up to 1000 SPs, show a similarity in the time variation of the outer eccentricity. It accumulates very slowly for the first few tens of SPs, then grows in a faster rate to a moderate value over another few tens of SPs before suddenly, due to a severe close encounter, the outer mass is thrown into a comet-like orbit with high or very high eccentricity and a bigger value of major axis. After this strong close encounter it may vibrate in an erratic manner between the very high region and moderate region, or it may stay in this comet-like orbit up to above 1000 SPs, or may even be thrown so far away that the energy of the outer 2-body subsystem becomes positive, which is regarded as an escape of the outer mass. As for collision, it is not easy to know whether an escape will occur or not if the integration is continued since the outer mass is already thrown into orbits of very high eccentricity and larger major axis. With this uncertainty in mind we simply call such orbits with the eccentricity above 0.99 but negative energy 'collision'. It is in fact quite arbitrary whether the orbit can survive or not after the close encounter.

For this set of experiments the procedure of Walker and Roy (1983) is possible since very stable systems have been observed. If close encounters are all treated as instabilities then a smooth exponential curve may be fitted to the data. However, since no analytical Hill-type stability stability region exists, the picture of an empirical hierarchical stability region outside the Hill-type stability region cannot be verified.

[1022] : In this case both orbits are strongly disturbed so that noticeable changes can be observed in all four orbital elements: they all obey the general rule given in Table 5.2. Neither collision nor escape have been observed. We choose the two eccentricities as characteristic parameters for describing the variation of the orbits since both change not only in the same direction but also with almost the same variation in size. Compared to other sets of experiments, this one shows more regularities with respect to the value of initial ratio of major axes, although the perturbations are stronger.

We summarise this point by starting from a bigger to smaller value of the ratio of major axes. For $\alpha > 0.65$, a close encounter occurs on commencing the integration such that both eccentricities jump into the high region and then vary irregularly in this region or enter the very high region; it is quite accidental whether the system can pass this close encounter and survive after this severe event without crossing over. The two 'stable' orbits are not really stable since they have entered the very high eccentricity region and survive accidentally.

For $0.55 < \alpha < 0.65$, a new mode suggesting stability begins to appear immediately after the integration begins. Before both e's begin to behave like the previous case, they stay below 0.3 for several SPs. The two 'stable' experiments in this region are again accidental. As the value of the ratio of major axes decreases, this new stability mode last for a few more SPs (Fig. 5.12d). This process continues until a value of about 0.475 in the ratio of major axes is reached, when suddenly the stability mode spreads up to above

1000 SPs. The variations of both eccentricities are now almost periodic, and the amplitude decreases to below 0.25. The transient region is very narrow as can be seen on the plot. From this point on, the smaller the value of the ratio of major axes, the more stable the system is: the amplitude of the e's' variation decreases with it.

If close encounters are noted, the study of Walker and Roy (1983) is applicable, with an empirical stability region found outside the Hill-type stability region.

[2066] : This is the simplest set of experiments. All features are like those of [1066] except that the behaviour is better. This is shown in Fig. 5.10 already. The accumulation of the outer eccentricity is found always to be below 4.0 x 10^{-2}.

[2062] : This interesting set of experiments displays a complicated structure. Essentially the features of this case are the same as [1062], but with more valleys and plateaus. Two valleys, VI and VII, have been detected, with VII on the edge of the Hill-type stability region. Although here P0 is completely within the Hill-type stability region, it is less stable than PI (in the sense that the orbital elements change more), while PII is the most stable group. Even within P0, stability decreases for a smaller ratio of major axes (observed from those systems in the plot). It is not clear whether P0 has got more fine structures or not, although as a general rule very stable systems should exist for very small ratios of the major axes. The accumulation in the outer e is below 10^{-5}. The representative parameter is the inner eccentricity, whose value is found to be below 0.3 for PII, below 0.6 for PI, while for P0 its value can go up to 0.7, and more irregularity being noticed. The other difference between P0, PI and PII is that there are two modes in the former groups, one below 0.3 and the other between 0.6 and 0.2, while in the latter group only the more stable mode exists up to 1000 SPs. The stable mode lasts longer in PI than in P0 after commencing the integration. Close encounter instability would contradict the concept of Hill-type stability and there is no sense in fitting curves no matter how close encounters are treated.

[2026] : In many ways this set of experiments has features which resemble [1026]: no critical ratio of major axes exists due to the distortion of the transformation. Escapes of the outer mass were detected, but no collision was found up to 1000 SPs, which is very likely to occur if the integration were continued for a longer time. Those 'stable' systems entangled with unstable ones are not actually stable since they spend most of the time with very high eccentricities for the outer orbits. Again a stable mode in the outer eccentricity with amplitude below 0.2 is found immediately after commencing the integration for almost all systems. This mode is followed by a very severe close encounter such that the outer mass is thrown to a more distant comet-like orbit with high or very high eccentricity. It is the extension of the stable mode over the relatively narrow region $0.485 < \alpha < 0.51$ that characterises the transition of orbits from instability to stability. Those systems with $\alpha < 0.48$ are all very stable and only almost periodic changes of the stable mode were observed in the outer eccentricity. The curve fitting technique of Walker and Roy (1983) is applicable here.

[2022] : Essentially the properties of this set of orbits are the same as [1022] with the exception that this set is less stable. The stable mode has got a bigger amplitude in both e's, and its duration is still much less than 1000 SPs even inside the Hill-type stability region. The most stable system of those run is the first one on the left, which is just inside the Hill-type stability region. However, the stable mode only lasts for a few hundred SPs, and then a close encounter follows which 'kicks' the system to a less stable mode: both e's vary in the high eccentricity region in an irregular way. Close encounter is in contradiction to the concept of Hill-type stability. Should we try to fit a curve using the procedure of Walker and Roy (1983) it is found to penetrate into the Hill-type stability region.

[0166] : This is essentially similar to [1066] and [2066], but with larger changes in e's. For those unstable orbits it is found that usually a cross-over instability occurs as the inner e grows to about 0.05, while the outer e decreases to about the same value. However, those stable cases may become unstable, since by the time 1000 SPs have elapsed, they usually have gained values of about 0.1 in the inner e, while the outer e has gone down to about 0.01. There therefore remain some uncertainties for this set of experiments. It is in this set of experiments that a violation of the general behaviour of eccentricities and semi-major axes (Table 5.1) is clearly observed.

[0162] : There is no critical ratio of major axes for this set of parameters. We have investigated those orbits with values of ratios of major axes down to about 0.5, where the orbits are found to be reasonably stable with random stable modes. The inner e is always bounded under 0.5. No strong close encounter occurs producing distinctly different modes around different values: changes are quite smooth and reversible. Even for unstable systems with bigger ratios of major axes, no severe close encounter has been found. Cross-over either occurs with the inner e below 0.5 or due to a moderate close encounter following the random stable mode, which brings the inner e up to the high region in about ten SPs.

[0126] : This set seems to be the best one for Walker and Roy's (1983) curve fitting procedure to be valid. Those systems inside the Hill-type stability region are all found to be very stable: the inner e is always below 10^{-4}, while the outer e varies almost periodically below 0.2. Those 'stable' systems out of the Hill-type stability region are quite different from the previous one, their behaviour show similarity to the escape systems: the system stays on the stable mode for a short period of time and then suddenly a very strong close encounter throws the outer mass to a comet-like orbit with very large new major axis and very high eccentricity, and stays until up to 1000 SPs without any more noticeable changes in a's and e's. It is seen on the plot that an exponential curve can be fitted to the data. However, following the same method as Walker and Roy (1983), no obvious empirical stability region exists out of the Hill-type stability region.

[0122] : This set of experiments, together with [0222], produces a new phenomenon in the behaviour of a's and e's. Again the two e's are chosen to describe the general behaviour of the orbits. The time variation of those systems within the Hill-type stability region consists of only a stable mode with low eccentricities, while those unstable ones outside usually consist of a set of 'sub-stable modes', with very small amplitudes, around various eccentricities, which modes are separated by some close encounters. Those 'stable' ones outside the Hill-type stability region are combinations of about 500 SPs stable mode followed by the above sub-stable modes (see Fig. 5.12c). The new phenomenon which has only been found in these two sets of experiments is an interesting variation of the e's (and the a's): some systems just outside the Hill-type stability region consist of a few hundred SPs stable mode plus a short unstable mode, followed by a new mode which is a combination of some short periodic changes with very small amplitude and a very long periodic change with reasonably large amplitude. This seems strange, hinting that an irregular orbit can eventually find its way to become regular. It is observed that most systems with initially highly eccentric orbits share this same feature.

[0266] : Essentially the same as [0166], but with apparently smaller variation in a's and e's: variation of outer e below 10^{-3}, accumulation of inner e of the order 10^{-2}. This does not imply that systems with higher values of e's are more stable than those with lower e's, because the ratio of major axes lies in different regions in the two plots. Because of the weakness of perturbations, it is not practical to use close encounters to determine stability. Therefore there is no reason to bring down the two stable points on the right hand side.

[0262] : No critical ratio of major axes exists for this set of parameters. It is essentially the same as [0162], but close encounters are slightly stronger such that collision of the inner orbit has been observed. In fact this is the only set of experiments with collision of inner orbit. It is observed that as the inner eccentricity goes up to a very high value the inner semi-major axis usually decreases by half. Variation of outer e is below 10^{-2}. The three stable systems on the right hand side are found to consist of only random stable mode bounded below 0.5, while the other stable ones are bounded by a lower value, about 0.3. This is presumably another example with complicated structures of valley and plateau.

[0226] : This seems to be the most exotic set of experiments which displays clearly the inconsistency of the concept of Hill-type stability and hierarchical stability used in Walker and Roy (1983). Not only collision and break-up of the outer orbit but also assured close encounter instabilities were found within the Hill-type stability region. This makes computing useless regarding an investigation of condition HS-(C) only; on the other hand, if we investigate all three hierarchical stability conditions and note instabilities regardless of the existence of the Hill-type stability region, then a curve may be fitted to the data but it would go into the Hill-type stability region below α_c. In any case there is absolutely no way to follow Walker and Roy (1983) to fit curves and investigate empirical stability regions outwith the Hill-type stability region.

This is in fact an intrinsic problem of the concept hierarchical stability itself, since Hill-type stability does not preclude collision nor escape of masses, it makes numerical integration studies on 'hierarchical' stability impossible even if we only check the stability of the hierarchy. Taking into account the fact that instability can cut short the time scale of numerical experiments, we chose to test the statistical conjecture by using this set of parameters. None of the 'stable' systems, except the only stable one on the far left, is stable, since they survive through on comet-like orbits.

[0222] : None of this set of experiments is really stable, because almost all of them have suffered a close encounter immediately after commencing the integrations, which increases the e's by about 0.3. After this a quasi-stable mode followed, which lasts from a few to a few tens of SPs, and then another close encounter kicks the system into the high eccentricity region. Exceptions have only been observed for $\alpha < 0.39$, where there is no noticeable close encounter on commencing the integration. It is also in this region that the new phenomenon noticed in [0122] was observed again, but with relatively longer multiple sub-stable modes. Even the system on the far left inside the Hill-type stability region suffers a strong close encounter which kicks eccentricities to about 0.5 and doubles the size of the outer major axis.

A comment is in order concerning the difference between e=0.1 and e=0.2. For systems with higher eccentricities, severe close encounters are quite common immediately after the integration has been begun, then it seems that, after this redistribution of energy and angular momentum between the two subsystems, the systems find more stable states (see Fig. 5.12g, h and i).

Comments on the Hierarchical Stability of Coplanar 3-body Systems

The above experiments clearly demonstrate the complexity of the general 3-body problem. In making these comments let us recall what our purpose is. The motivation of the present study can be traced back to estimating the life-time of our planetary system. Since there is no general analytical answer regarding its orbital stability, Walker and Roy (1983) tackled the problem by asking an apparently weaker question, based on hierarchical stability. The methodology of their study is to extrapolate the life-time of weakly perturbed systems (real systems usually are) by investigating fictitious systems with stronger perturbations so as to cut short the time-scale of the numerical integrations and circumvent the problem of reliability for long term behaviour, which is usually beyond the ability of our numerical routines. This was successful in Walker and Roy (1983) because of the smooth behaviour of the life-time against the parameter α. However, not too much should be read into such results since chaotic systems do not in general possess smooth properties. This point has been clearly shown in our investigation of initially elliptical systems. The reason why a completely different phenomenon has been observed in the two studies is that eccentricities make a great difference in considering the long term behaviour. Another reason is that Walker and Roy (1983) studied for a relatively shorter time scale and that since the initial e's were zero their values were not allowed to grow out of the very small e region, otherwise instability was noted. Therefore their study was limited to the very low e region not only at the beginning of the integration but also afterward because orbits were observed for relatively shorter times, in which situation chaos was not yet clearly manifested. This is also why commensurability was a plausible explanation for the existence of peaks in their graphs. Empirical stability regions were found generally to exist, and as the systems were begun with parameters nearer and nearer to the critical stability surfaces the time variation of a's and e's are found to become more periodic.

In contrast to this, the present study has been carried out not only for longer time scales but also for higher e's. The studies of Walker and Roy (1983) rely heavily on the C^2H criterion, which only guarantees condition HS-(C), neither HS-(A) nor HS-(B). The idea of an empirical stability region outside the Hill-type stability region was actually based on the observation that HS-(C) guarantee both HS-(A) and HS-(B) in their numerical experiments. However, the set of experiments [0226], among many others, contradicts this: instabilities (A) and (B) can occur when (C) is guaranteed. Because of this we came to regard moderate close encounters as instability. This point of view is confirmed by integrating some of the orbits backwards in time: orbits with substantial close encounters usually cannot be integrated back to the starting condition, unlike orbits without close encounters. The result is that in the full sense of Walker and Roy's definition of hierarchical stability, viz., condition HS-(A), HS-(B) and HS-(C) (see section 1.3), many systems in the Hill-type stability regions are not hierarchically stable, nor do the systems become more stable when α decreases. In fact only one of the sets of experiments, namely, [1022], supports the validity of the attractive picture of Walker and Roy (1983). All these are schematically shown in Table 5.4.

Can we hope to preserve the empirical stability picture if we only consider instability condition (C)? Certainly, within the Hill-type stability region condition HS-(C) holds; but what about outside it? The answer is still 'no', not only up to 1000 SPs, but in principle; because instabilities (A) and (B) makes it impossible for any numerical integration to be continued indefinitely. In fact, HS-(C) can be preserved even for systems with non-negative total energy (Marchal & Bozis, 1982). On the other hand, even if the above difficulty did not exist, nor was the very exotic [0226] plot observed, the simple picture of an empirical stability region outside the Hill-type stability region is not necessarily true. A good example is the plot [0126], where all systems inside the Hill-type stability region are quite stable, while the instabilities extend right up to the very edge of this analytical stability region. We see from the above comments that it is advantageous to keep the full definition of hierarchical stability and modify the former picture of empirical stability so that the empirical stability curves are allowed to go into the Hill-type stability regions.

The suggestion from [2062], among others, is that even if there are hierarchically stable systems (according to either HS-(C) only, or all three stability conditions) outside the critical stability surface, they do not follow the simple picture given by Walker and Roy (1983): they are found to be mixed with non-hierarchically stable systems. These are also the experiments which challenge the modified picture of empirical stability, nevertheless, the statistical interpretation given in the next section may justify this point.

At this point it is appropriate to make some other comments. Commensurability of mean motions does not seem to play a significant role even in non-coupled elliptical systems, since they will be very atypical. Therefore, Walker and Roy's (1983) explanation of the peaks by commensurability does not seem to apply to the results of the present study. Even if commensurability were important, it would only be for weakly coupled systems. There is therefore no simple explanation for our 'peaks' in Fig. 5.11, which are in fact not peaks if we consider close encounter. Only when close encounters are taken into account, could commensurability play a possible role. It is in fact observed that, with respect to the sets of data, there are no apparent sharp peaks like those found in Walker and Roy. What we have now are valleys and plateaus, with peaks (near α_c) taken to be degenerate valleys or plateaus. This cannot correspond to any isolate single value commensurability.

If we look at the sets [0062], [1062] and [2062], it is evident that the plateaus PI and

PII on the latter two are more stable than the systems with same initial α in the circular case in Walker and Roy (1983). Furthermore, between 0.53 and 0.55, P0 in [2062] is more stable than V in [1062]. These are examples which support the point that stability can increase with increasing the values of the eccentricities (see last section).

From Fig. 5.12, which is representative of the whole investigation, it seems that syzygy is not always the worst configuration, since changes in the elements do not seem larger near syzygy than far from it. In fact, the accumulation of instability builds up through a whole synodic period, if not at the last fatal one. It is also obvious from Fig. 5.12 that the quantization of reading lifetime in SP is not one in general, as was remarked by Walker; subjective reading can miss a great number of synodic periods.

A comparison between Walker and Roy (1983) and M^{c}donald (1986) shows that in general retrograde systems are more stable than prograde ones, which was also observed by Henon (1970) in the context of the restricted 3-body problem. It is observed in Fig. 5.12 that this is because retrograde motions can pass the worst configurations more quickly such that a close encounter of the same size will not last time enough for the retrograde orbits to build up instability, even though there are more close encounters during the same time interval.

An observation on escape given by Walker and Roy (1983) is also confirmed in our experiments. It is observed that, if the smallest mass is in the outer orbit, it is always this mass that is thrown far away which signals an escape. However, there is an uncertainty if the smallest mass is involved in the inner orbit, where the semi-major axis decreases as e grows to a very high value.

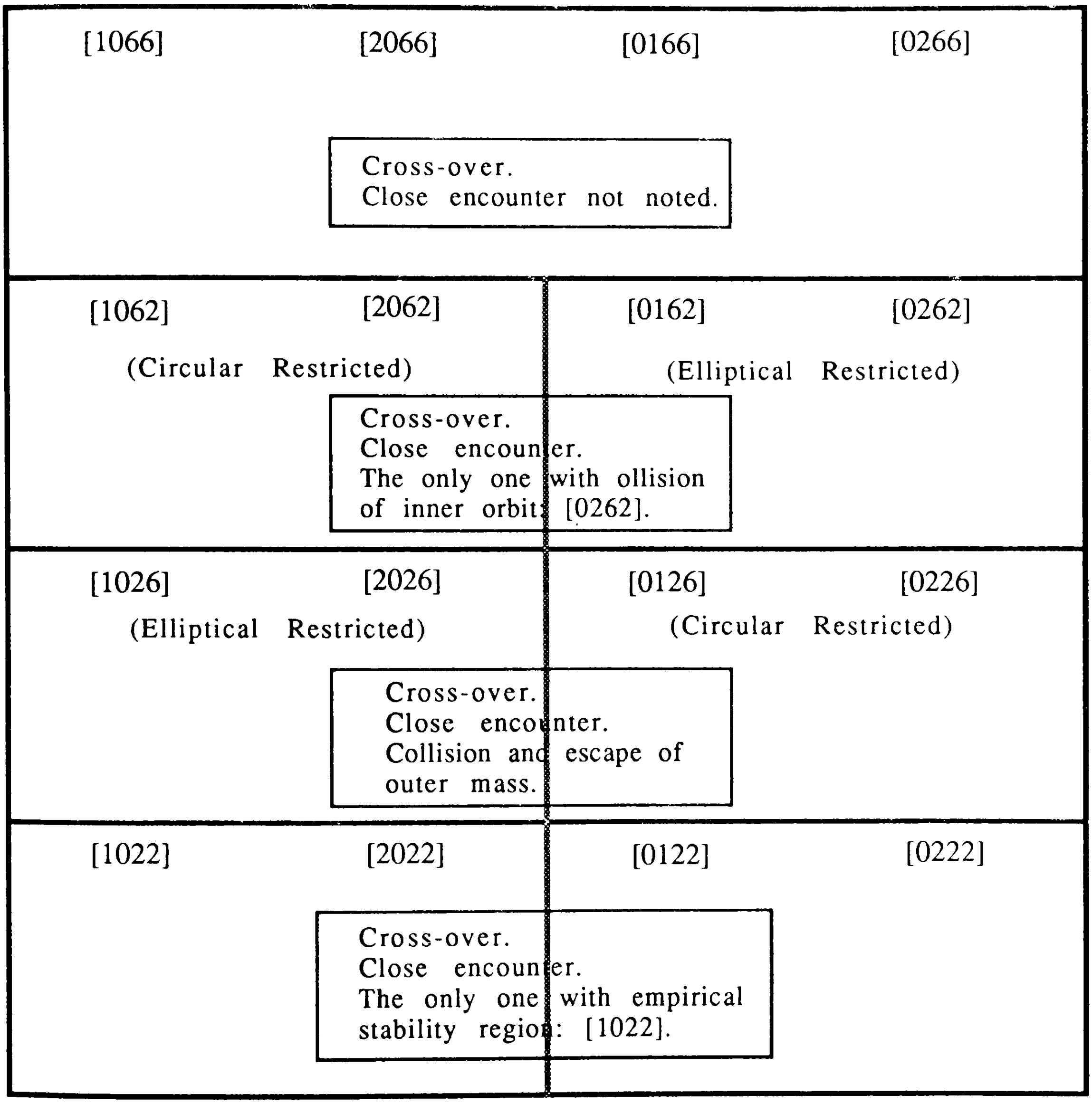
It is also notable that, from Table 5.4, there seems to be an apparent difference in that the 'circular restricted 3-body systems' show more valley and plateau structures than the elliptical ones. This may be explained as that the former case has more stable motions mixed with unstable ones in the presumably unstable regions. Note further that in the two circular cases, the systems with the smallest mass in the inner orbit show more structures than when the smallest mass is put in the outer orbit round both 'primaries'.

Finally, we point out that, though our experiments have been presented in the ε-parameter space which distorts the pictures, the general nature of the result should be the same in the μ-parameter space; because instabilities have been observed within the Hill-type stability regions, and stabilities outside such regions, which cannot be changed by the transformation.

Table 5.3 Value of the stability indicators. A '*' in the α_e column means that systems with initial α equal to or below the number are quite stable up to the time limit of the numerical investigation (1000 SPs); no stable system has been yet found for those sets of experiments without '*'.

e_2	e_3	ε_{23}	ε_{32}	α'	α_c	α_I	α_e	α_x
0.0	0.0	-6	-6	.002	.973	.973	0.974	1.000
		-6	-2	.002	.412	.412	0.516	
		-2	-6	.200	.520	.520	0.627	
		-2	-2	.200	.323	.323	0.475	
0.0	0.1	-6	-6	.002	.842	.875	< 0.860	0.900
		-6	-2	.002	▨	.371	< 0.500	
		-2	-6	.180	.546	.465	< 0.546*	
		-2	-2	.180	.369	.291	< 0.369*	
0.0	0.2	-6	-6	.002	.700	.778	< 0.770	0.800
		-6	-2	.002	▨	.330	< 0.380*	
		-2	-6	.160	.548	.416	< 0.440*	
		-2	-2	.160	.374	.259	< 0.370	
0.1	0.0	-6	-6	.002	.852	.884	< 0.865	0.909
		-6	-2	.002	.496	.375	< 0.500	
		-2	-6	.182	▨	.473	< 0.480*	
		-2	-2	.182	.376	.294	< 0.470*	
0.2	0.0	-6	-6	.002	.735	.810	< 0.800	0.833
		-6	-2	.002	.552	.343	< 0.540	
		-2	-6	.167	▨	.433	< 0.480*	
		-2	-2	.167	.407	.269	< 0.407	

Table 5.4 Schematic plot of the sets of experiments as close encounter instability is noted. Note the linear nature in the sets [...66], the valleys and plateaus in the sets [1062] and [2062], instabilities inside Hill-type regions in the sets [0226], [0222] and [2022]. [1022] is the only set showing empirical stability regions outside Hill-type stability region, while the two sets [0126] and [0122] have no empirical stability regions outside Hill-type region.



[1066] [2066] [0166] [0266]

Cross-over.
Close encounter not noted.

[1062] [2062] [0162] [0262]

(Circular Restricted) (Elliptical Restricted)

Cross-over.
Close encounter.
The only one with ollision of inner orbit: [0262].

[1026] [2026] [0126] [0226]

(Elliptical Restricted) (Circular Restricted)

Cross-over.
Close encounter.
Collision and escape of outer mass.

[1022] [2022] [0122] [0222]

Cross-over.
Close encounter.
The only one with empirical stability region: [1022].

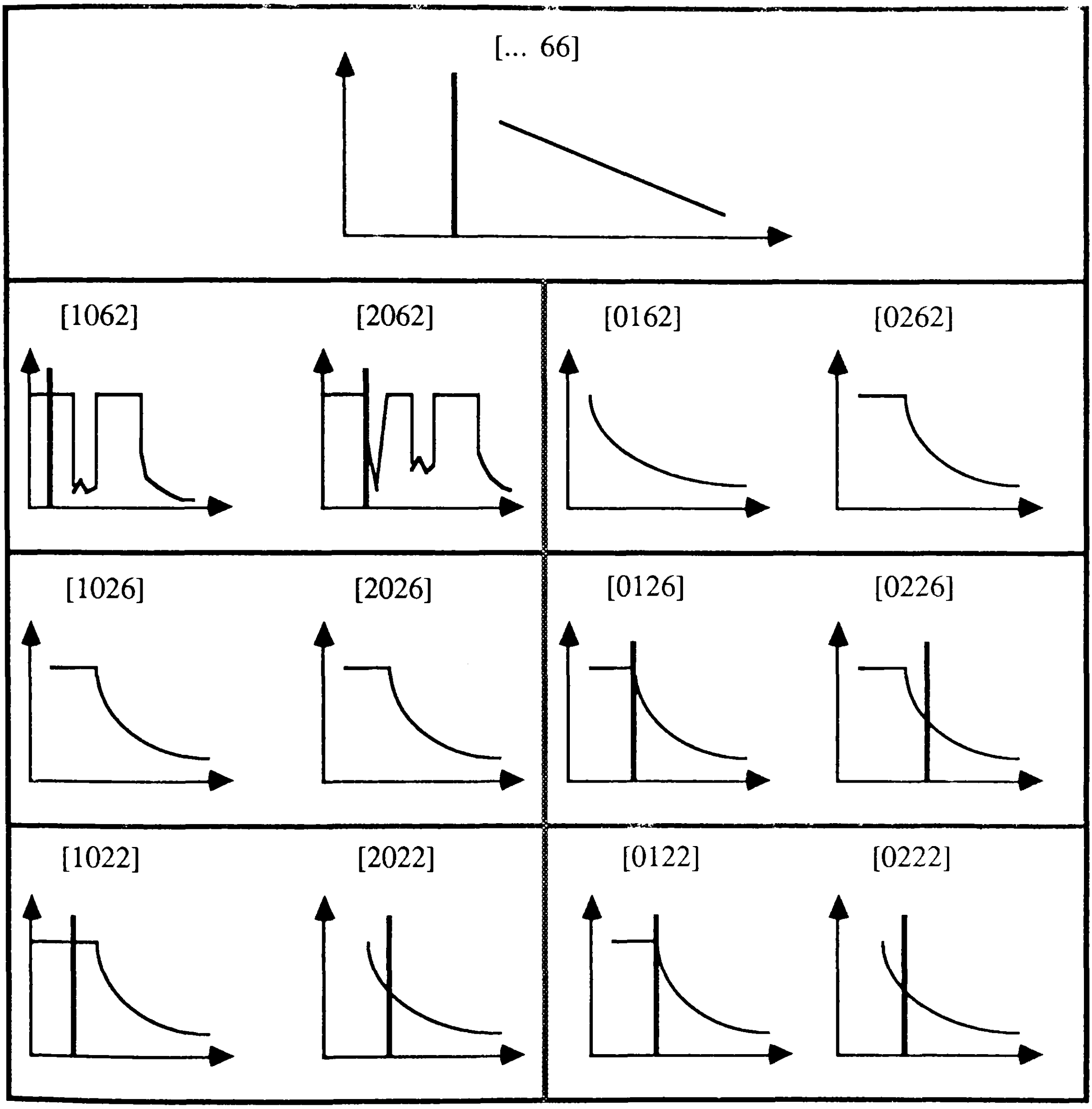

[... 66]
[1062]
[2062]
[0162]
[0262]
[1026]
[2026]
[0126]
[0226]
[1022]
[2022]
[0122]
[0222]

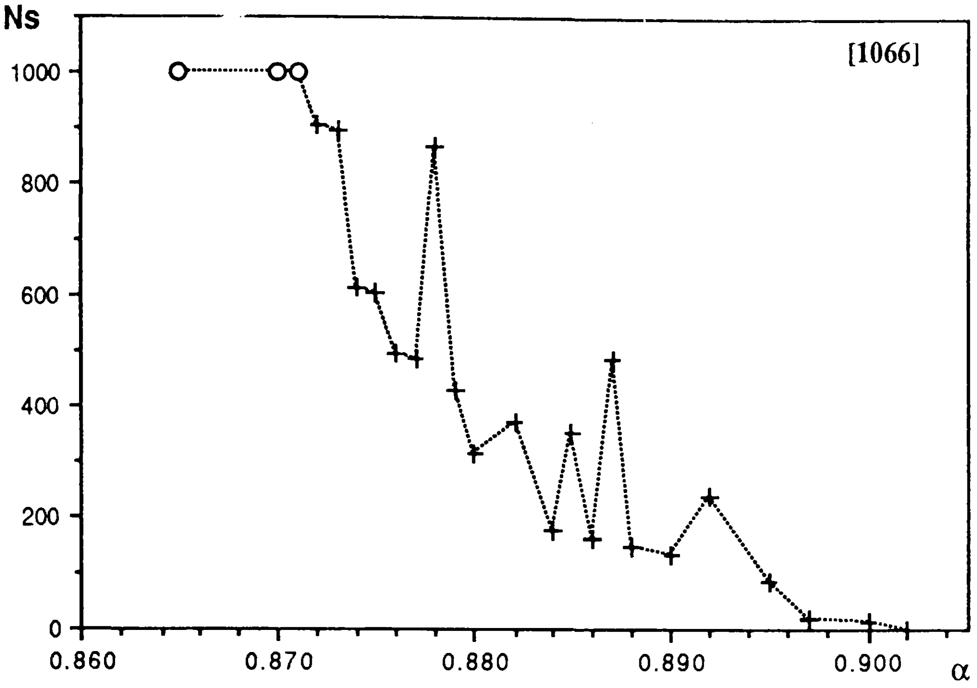
Fig. 5.11

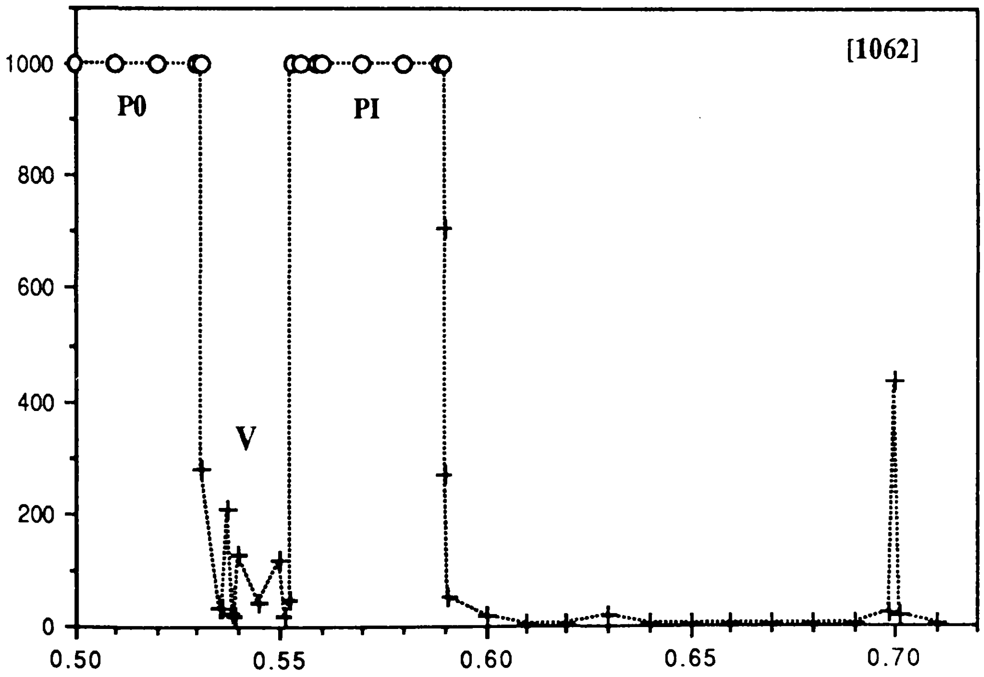
Presented in this set of diagrams is the result of several hundred numerical integration of fictitious systems. The close encounter instability is not noted here. The plots are the same as in Walker and Roy (1983), namely, the vertical axis is the life-time (Ns) of the systems in synodic period (SP), while the x-axis is the initial value of the ratio of semi-major axes (α). The symbols used in the plot are defined as

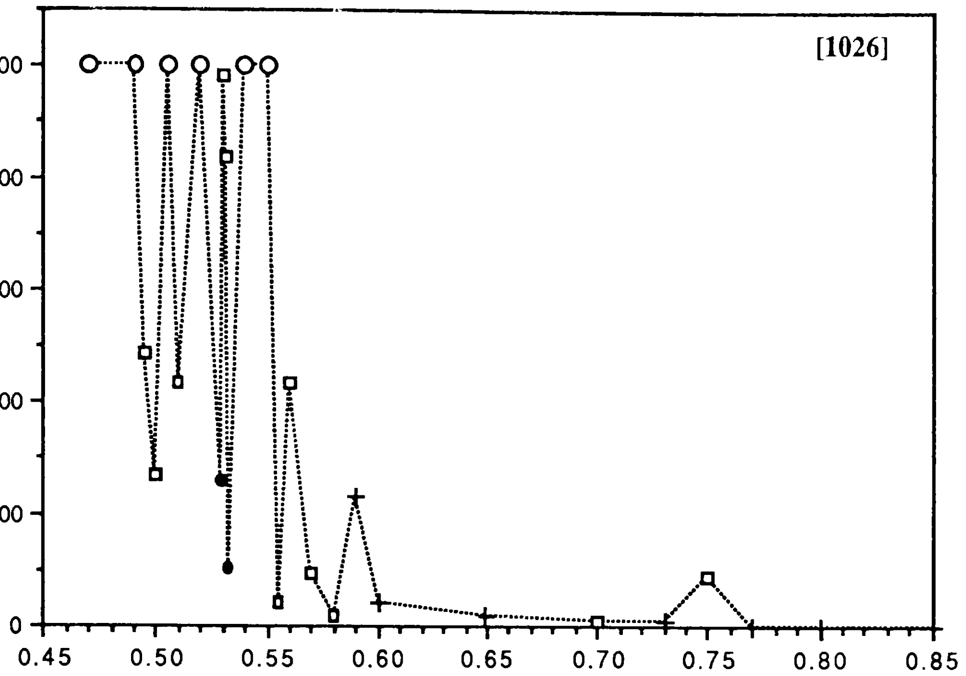
o .... stable up to 1000 SP
+ .... cross over of orbits
❑ .... energy of one orbit becomes non-negative
●.................. eccentricity of at least one orbit goes beyond 0.99.

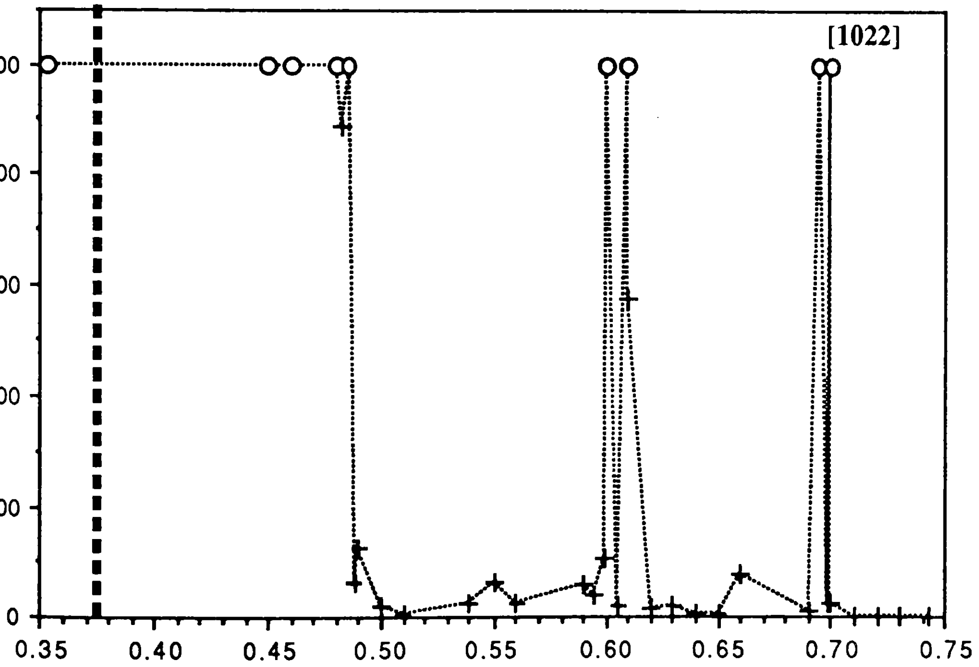
The points in the diagrams are joint by dotted lines to indicate the order of them. On some of the plots a dark broken vertical line is drawn to indicate the value of α_c.

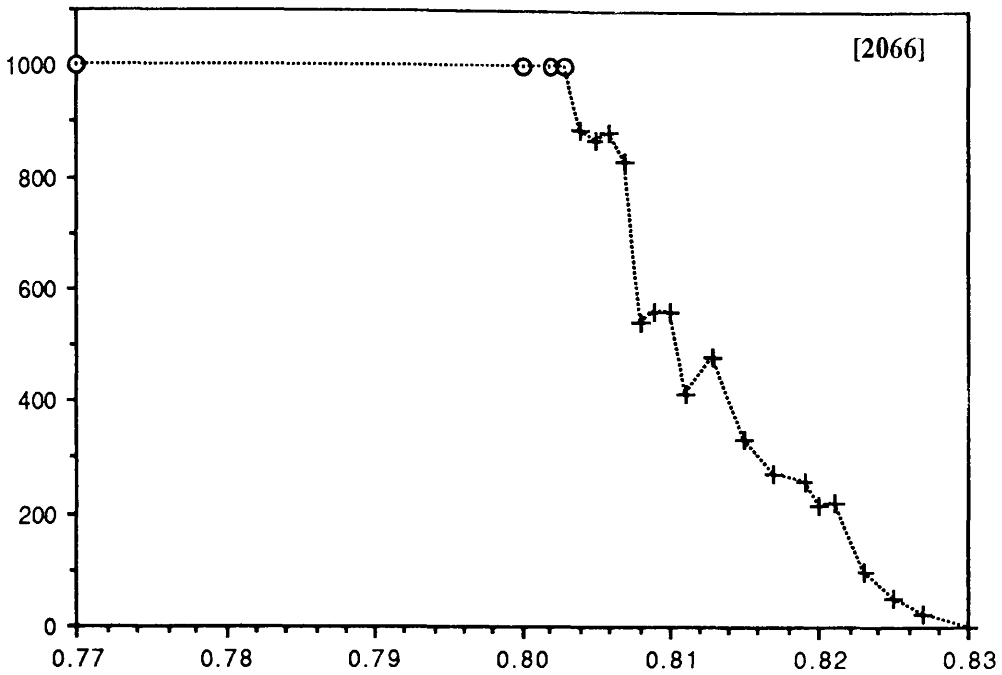
Also note that in these numerical experiments escapes (energy of one orbit becomes non-negative) are only observed for the outer orbit. Almost the same phenomenon is observed for collisions, with the only exception of the set [0262], where collisions are observed in the inner orbit.

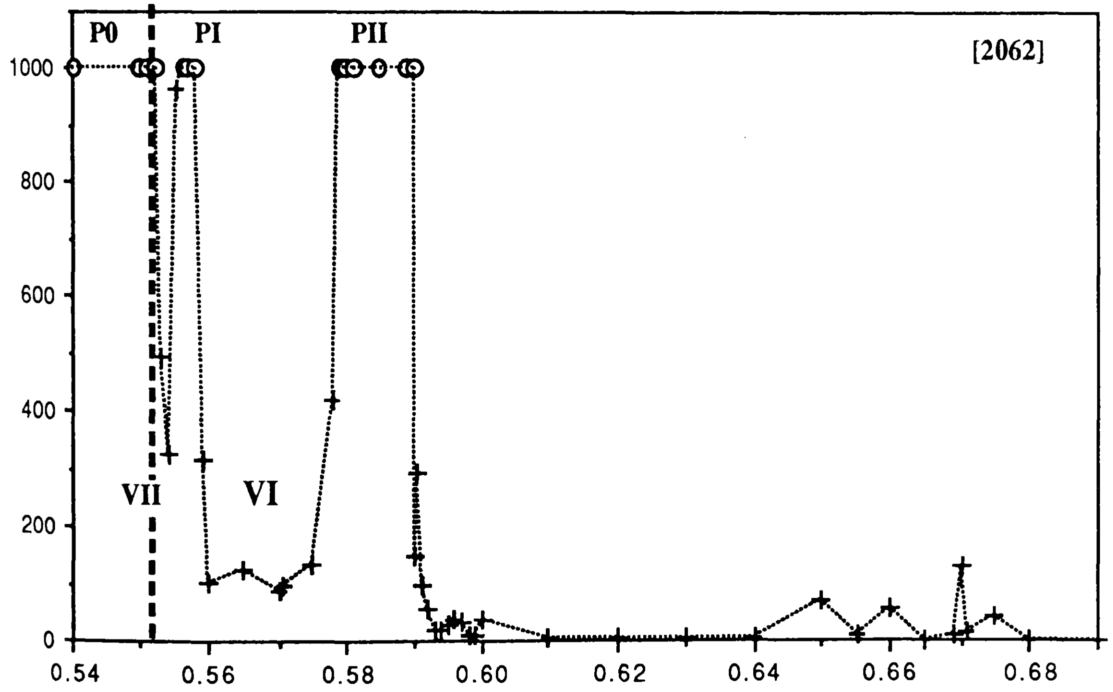

Ns
[1066]
1000
800
600
400
200
0
0.860
0.870
0.880
0.890
0.900
α

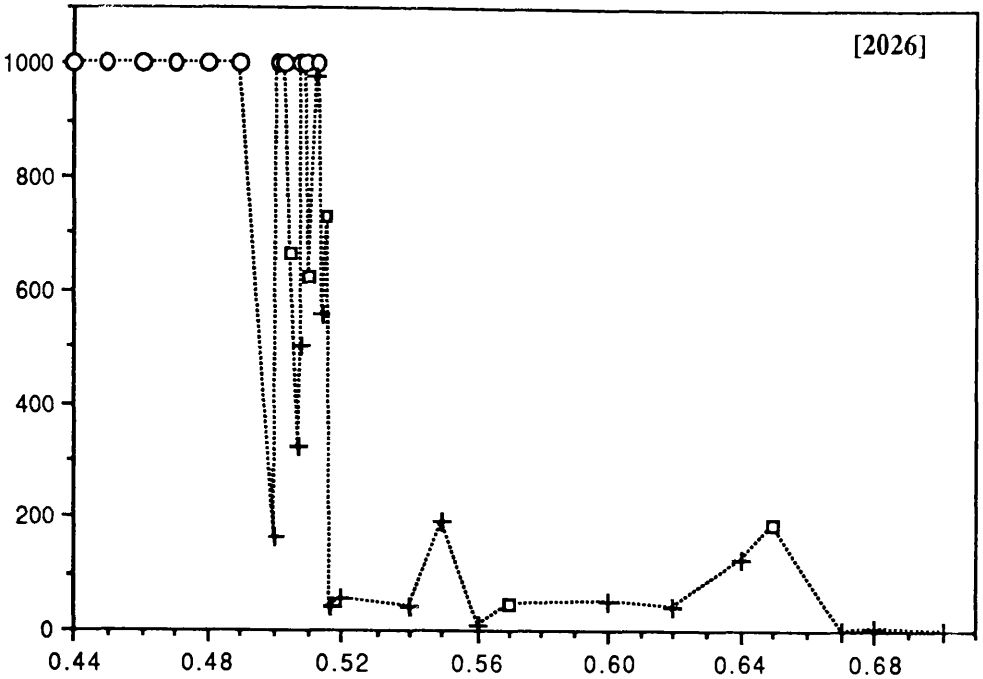

[1062]
P0
PI
V
1000
800
600
400
200
0
0.50
0.55
0.60
0.65
0.70

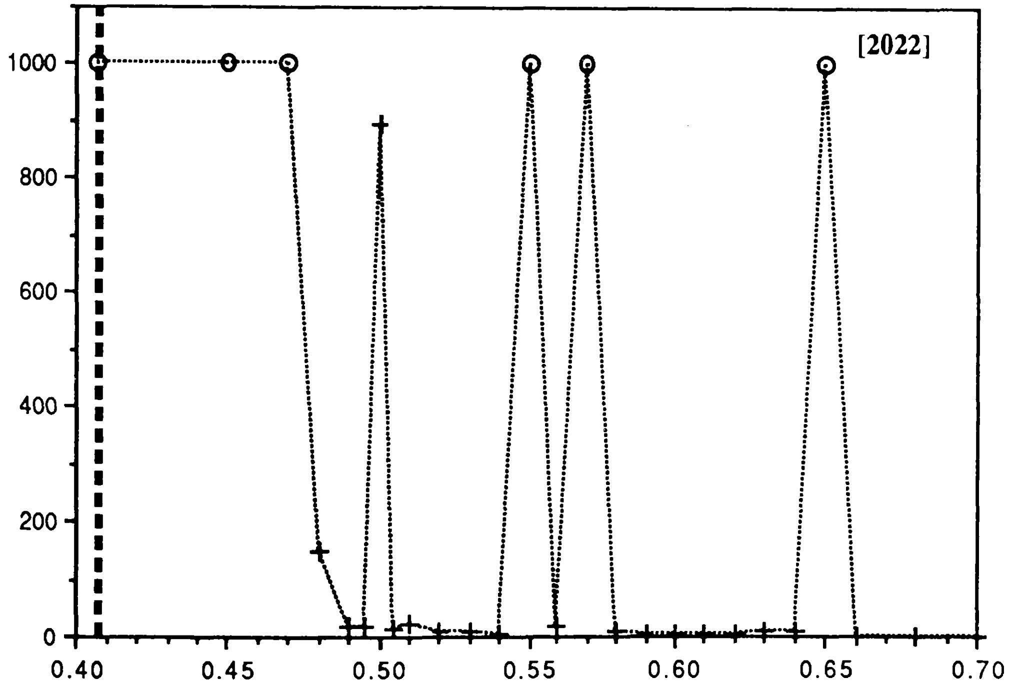

[1026]
1000
800
600
400
200
0
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85

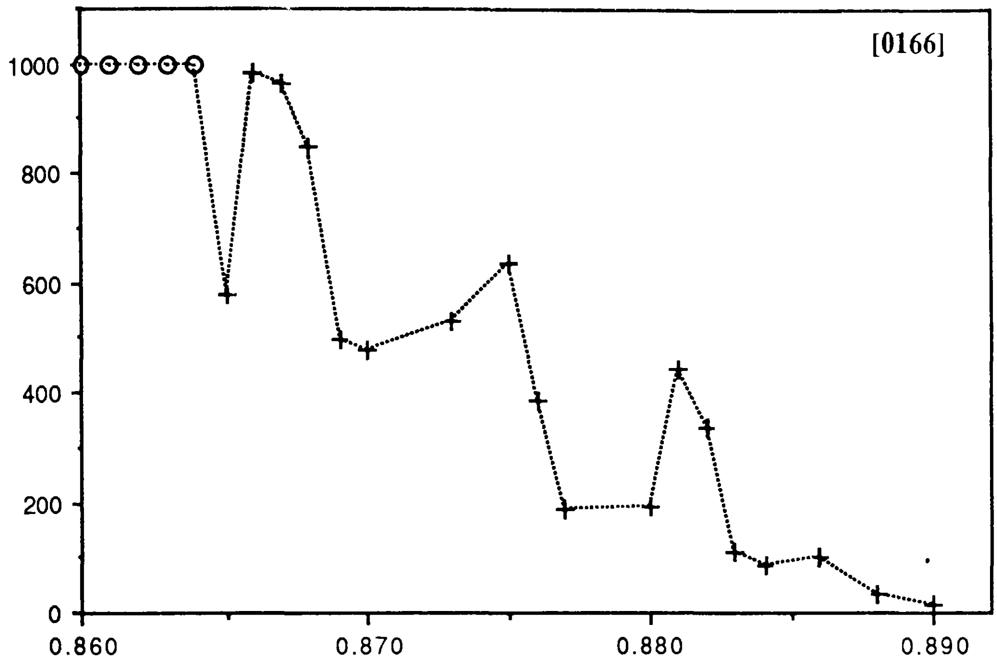

[1022]
1000
800
600
400
200
0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75

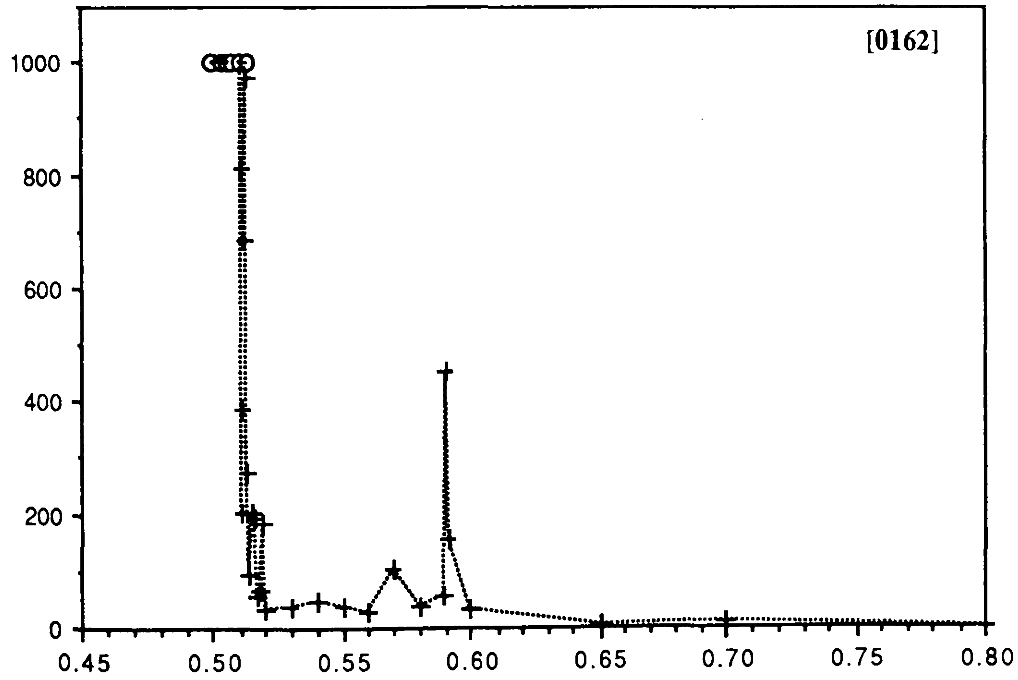

[2066]
1000
800
600
400
200
0
0.77
0.78
0.79
0.80
0.81
0.82
0.83

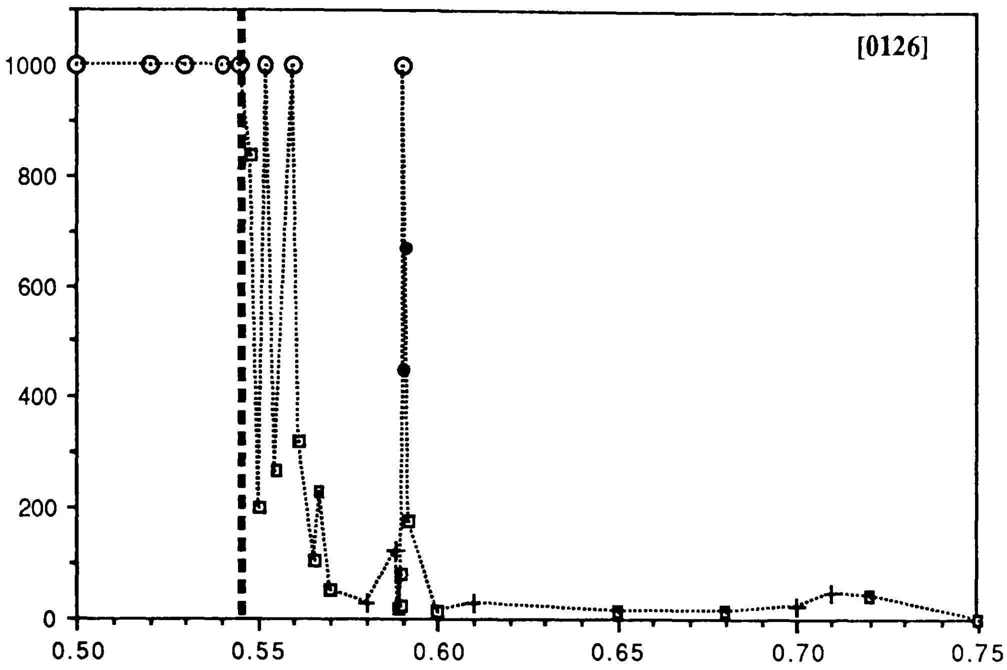

P0
PI
PII
[2062]
VII
VI
1000
800
600
400
200
0
0.54
0.56
0.58
0.60
0.62
0.64
0.66
0.68

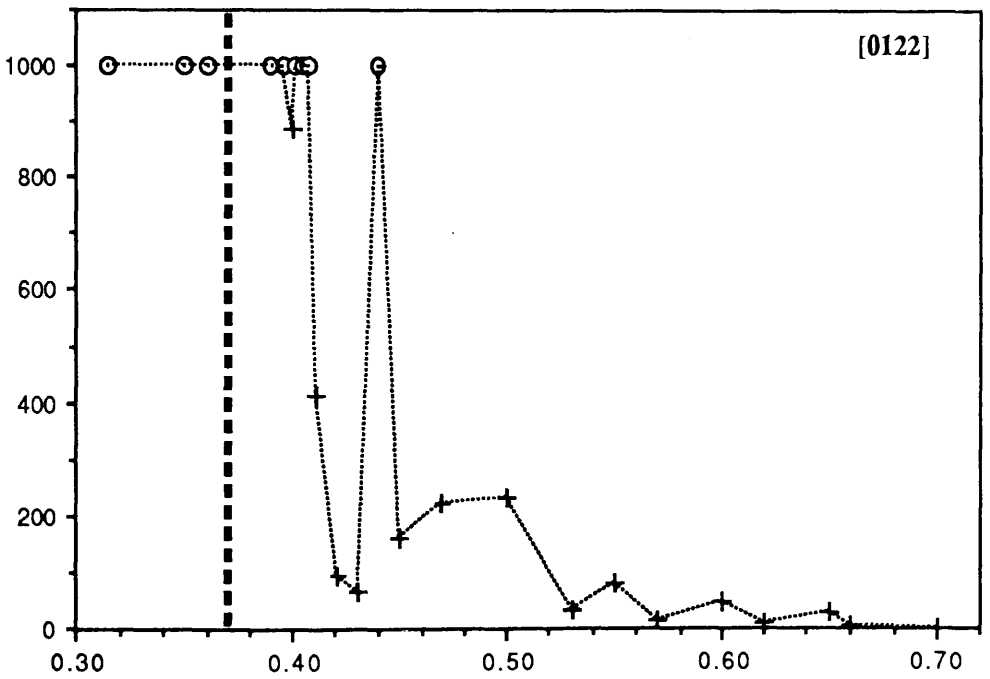

[2026]
1000
800
600
400
200
0
0.44
0.48
0.52
0.56
0.60
0.64
0.68

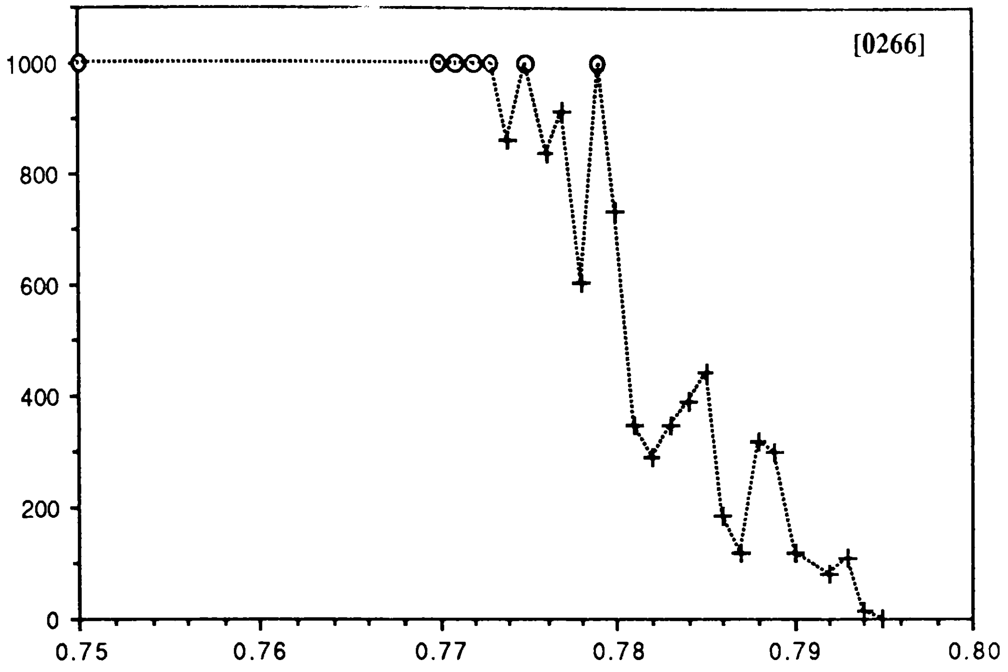

[2022]
1000
800
600
400
200
0
0.40
0.45
0.50
0.55
0.60
0.65
0.70

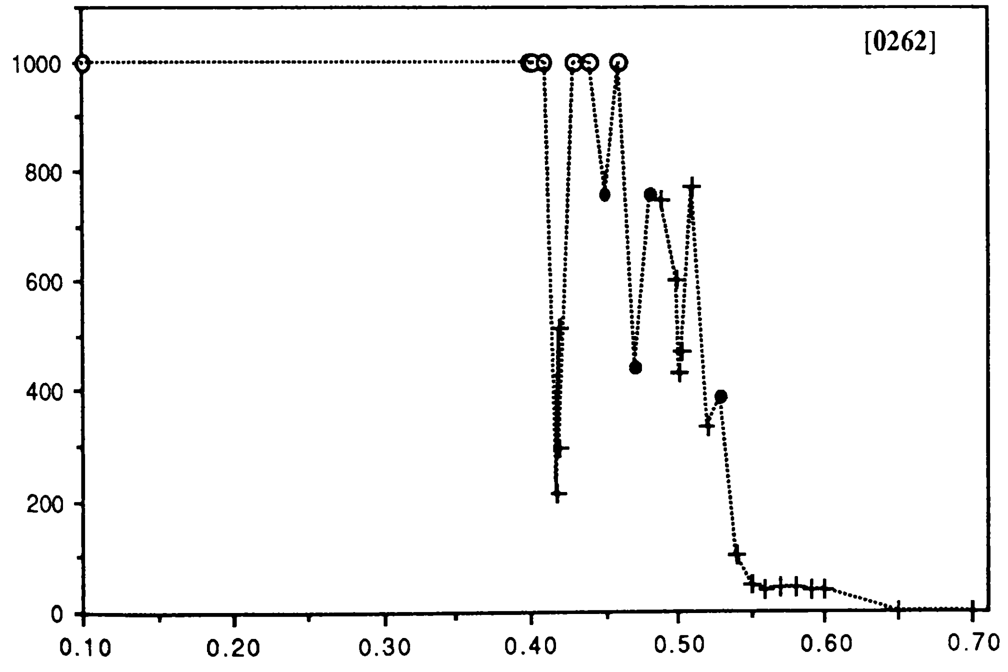

[0166]
1000
800
600
400
200
0
0.860
0.870
0.880
0.890

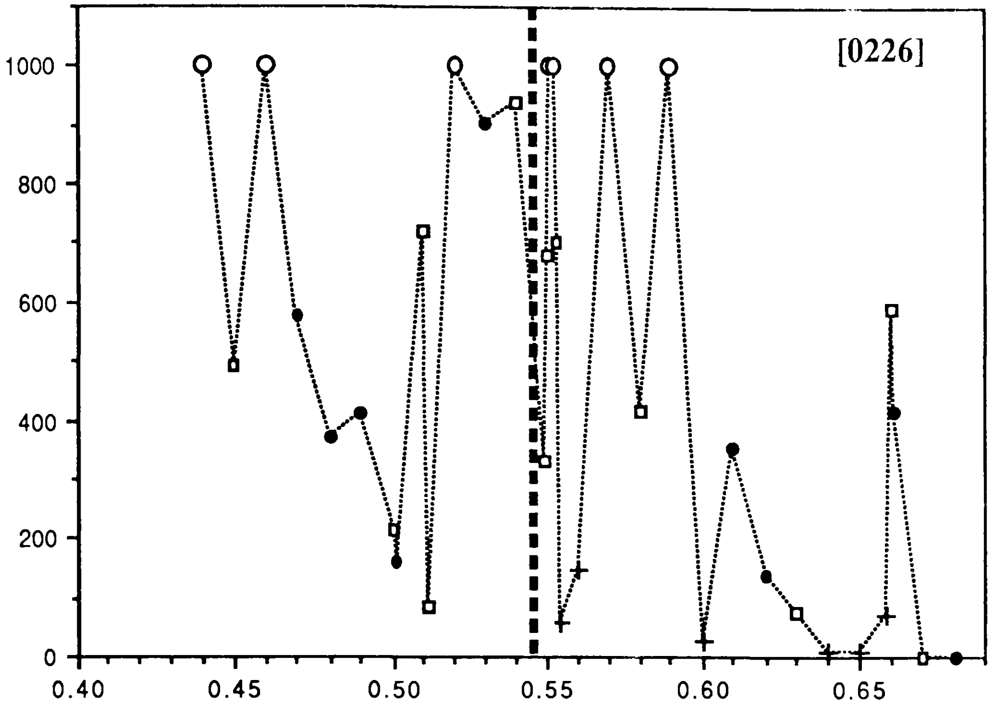

[0162]
1000
800
600
400
200
0
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80


[0126]
1000
800
600
400
200
0
0.50
0.55
0.60
0.65
0.70
0.75


[0122]
1000
800
600
400
200
0
0.30
0.40
0.50
0.60
0.70


[0266]
1000
800
600
400
200
0
0.75
0.76
0.77
0.78
0.79
0.80


[0262]
1000
800
600
400
200
0
0.10
0.20
0.30
0.40
0.50
0.60
0.70


[0226]
1000
800
600
400
200
0
0.40
0.45
0.50
0.55
0.60
0.65

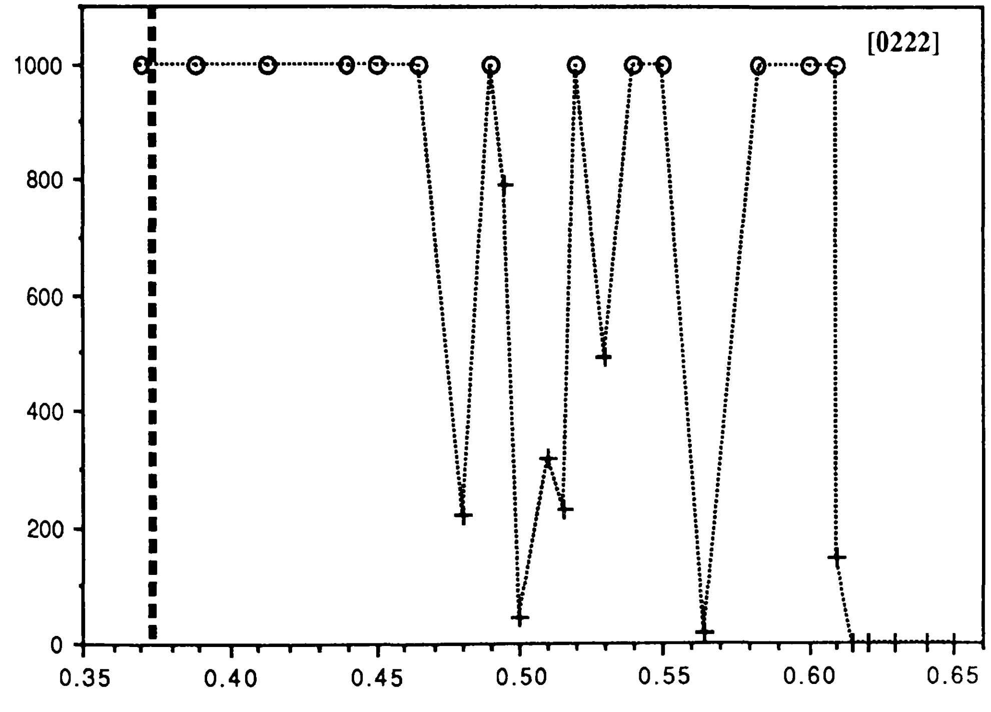

[0222]
1000
800
600
400
200
0
0.35
0.40
0.45
0.50
0.55
0.60
0.65

Fig. 5.12

Presented here are some typical examples of the variation of semi-major axes and eccentricities with respect to time (in fact we have used the number of integration steps 'NSTEP'). In the diagrams a star '*' indicates a conjunction of the three masses in the order of 1, 2 and 3.

(a). Typical correlated behaviour of the semi-major axes (in full curves) and eccentricities (in broken curves).

(b). A 'stable mode' followed by a 'random stable mode' in e_2.

(c). A 'stable mode' followed by some 'sub-stable modes' in e_3.

(d). An extension of the 'stable mode' preceding instability.

(e). A stable (almost) periodic motion.

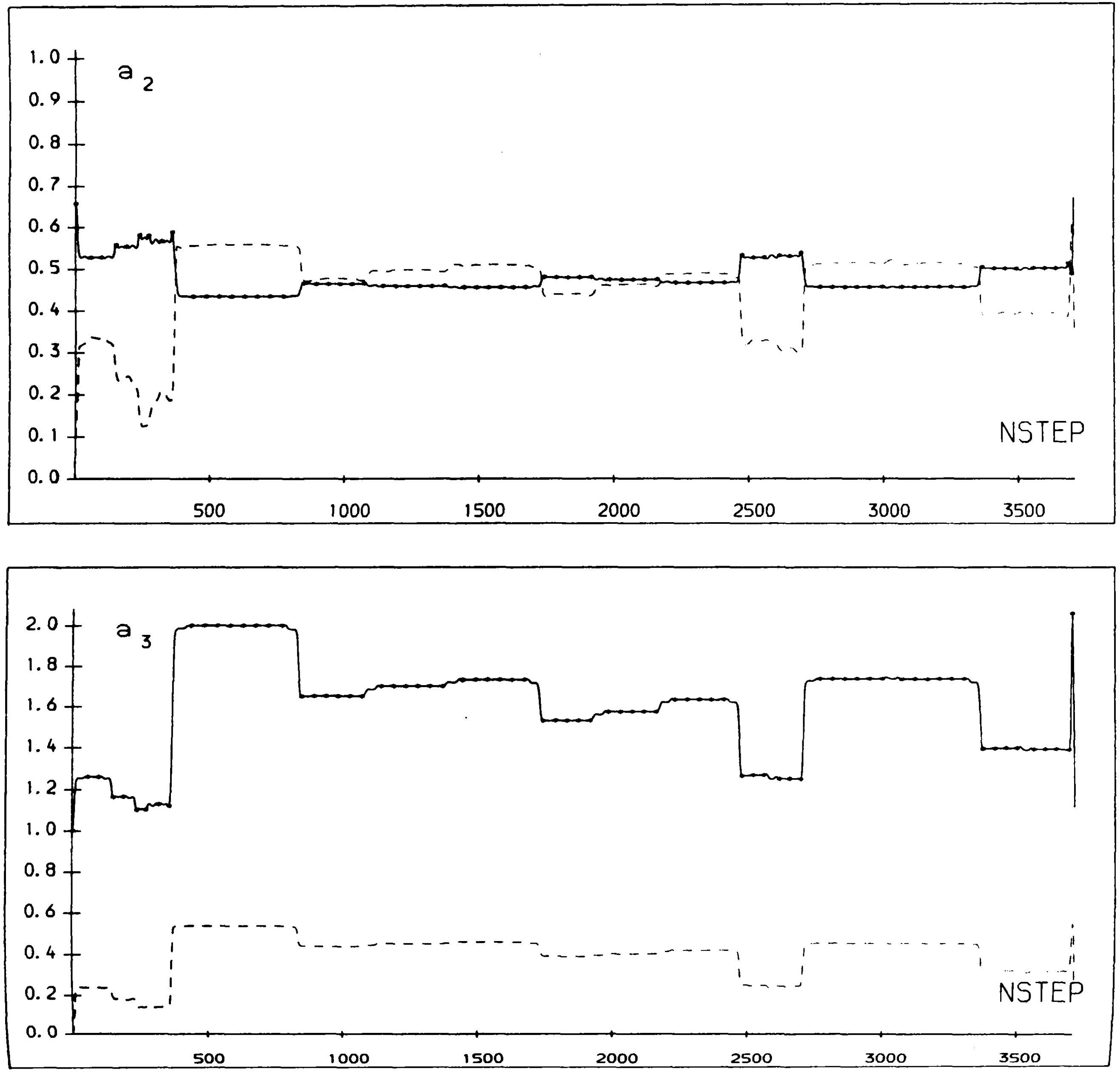
(f). Another stable (almost) periodic motion.

(g). An example of possible self-stabilisation: 'stable mode' + 'sub-stable modes' + 'long term periodic motion'.

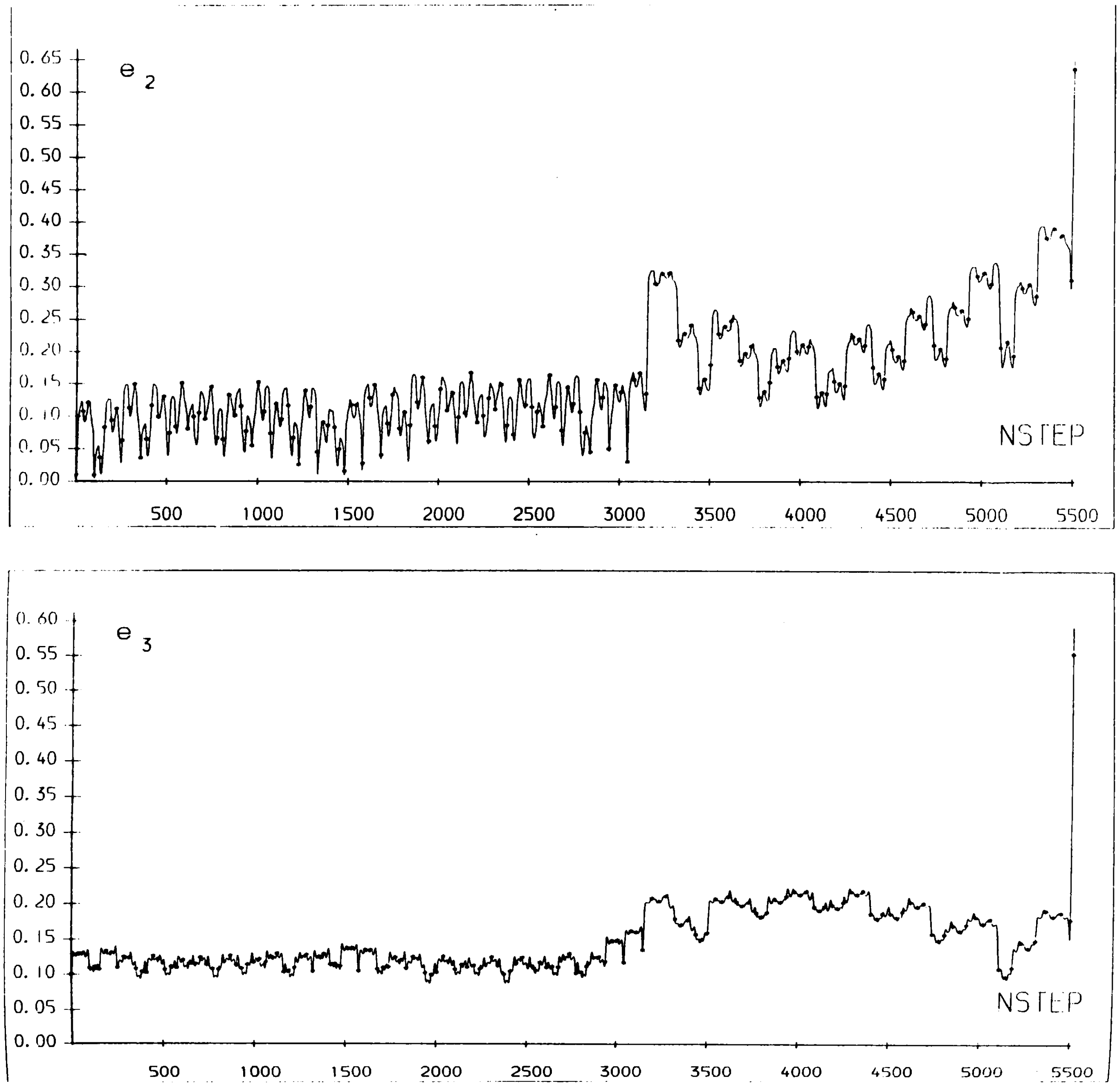
(h). A better example of possible self-stabilisation: 'strong close encounter' + 'irregular motion' + 'long term periodic motion'.

(i). An example of possible self-stabilisation: 'stable mode' + 'sub-stable modes' + 'long term periodic motion' + 'sub-stable modes'.

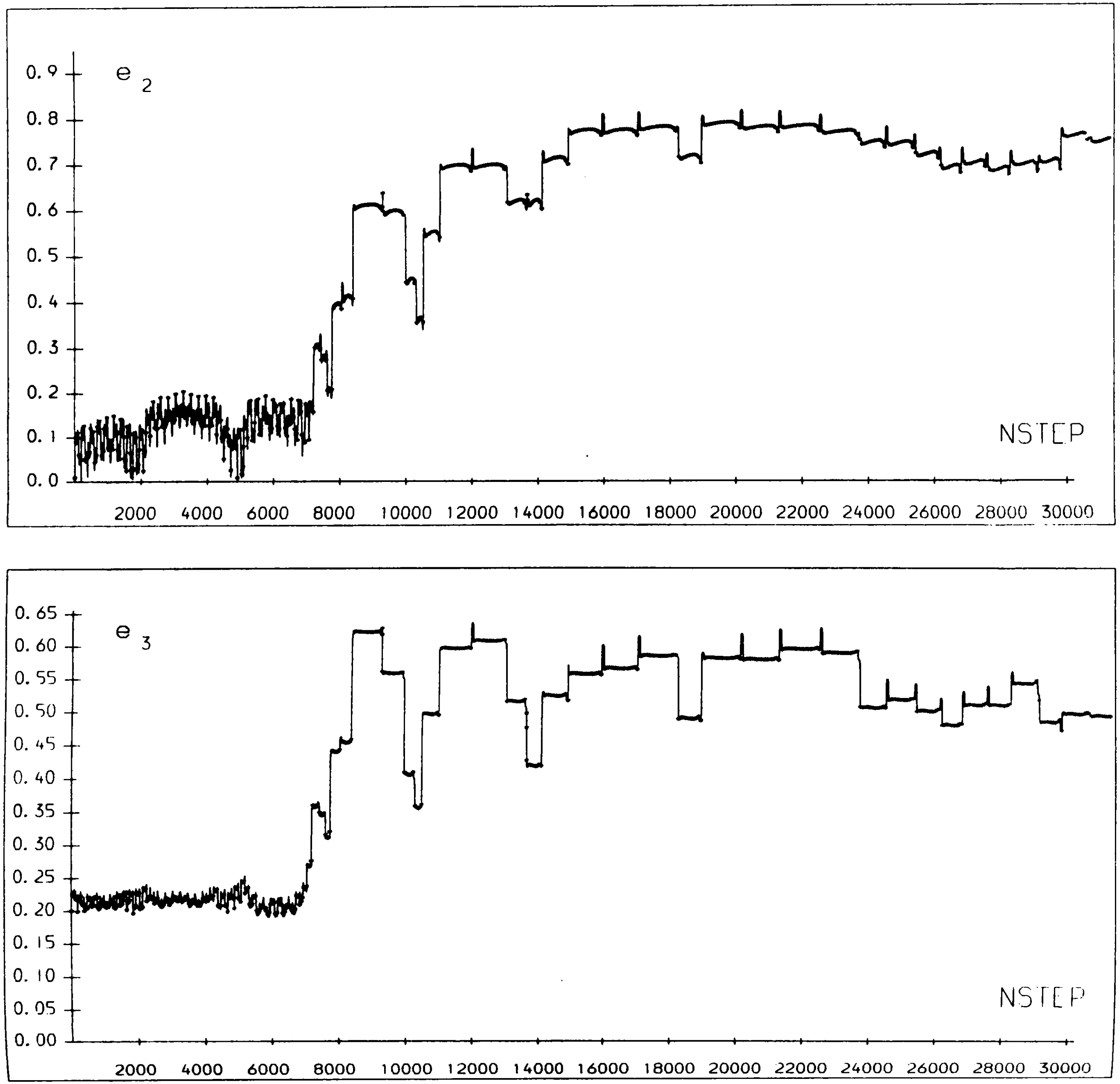
(j). A comparison of a prograde and corresponding retrograde motion. The diagrams only cover 38 SP in each case. The stable (almost) periodic motion of the retrograde system is maintained up to the time limit of the numerical integration, namely, 1000 SPs; while the direct system suffers close encounters after about 10 SPs and finally a cross-over instability occurs at 86 SPs.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
a 2
NSTEP
500
1000
1500
2000
2500
3000
3500
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
a 3
NSTEP
500
1000
1500
2000
2500
3000
3500

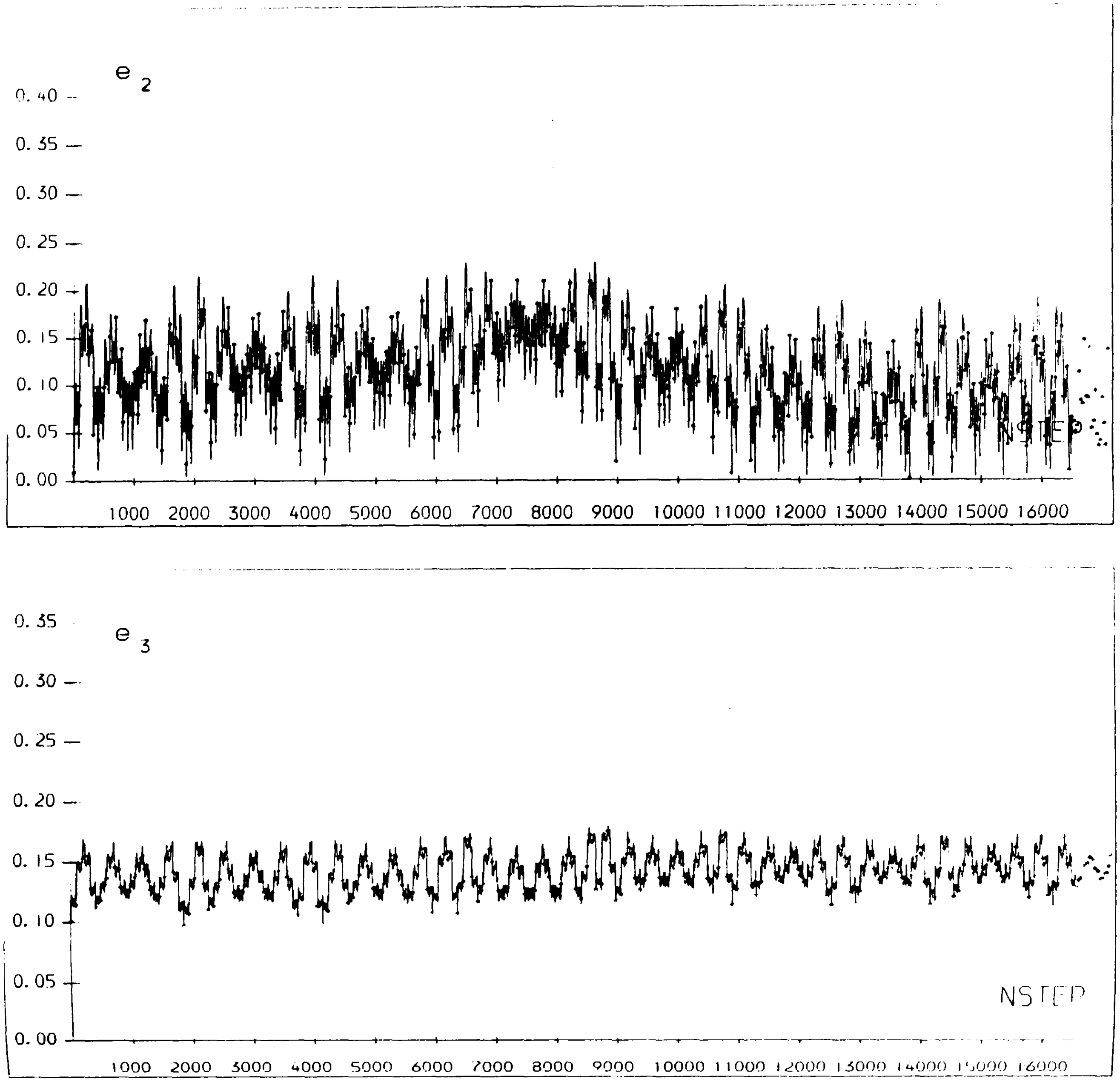
a


e 2
0. 00
0. 05
0. 10
0. 15
0. 20
0. 25
0. 30
0. 35
0. 40
0. 45
0. 50
0. 55
0. 60
0. 65
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
NSTEP
e 3
0. 00
0. 05
0. 10
0. 15
0. 20
0. 25
0. 30
0. 35
0. 40
0. 45
0. 50
0. 55
0. 60
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
NSTEP

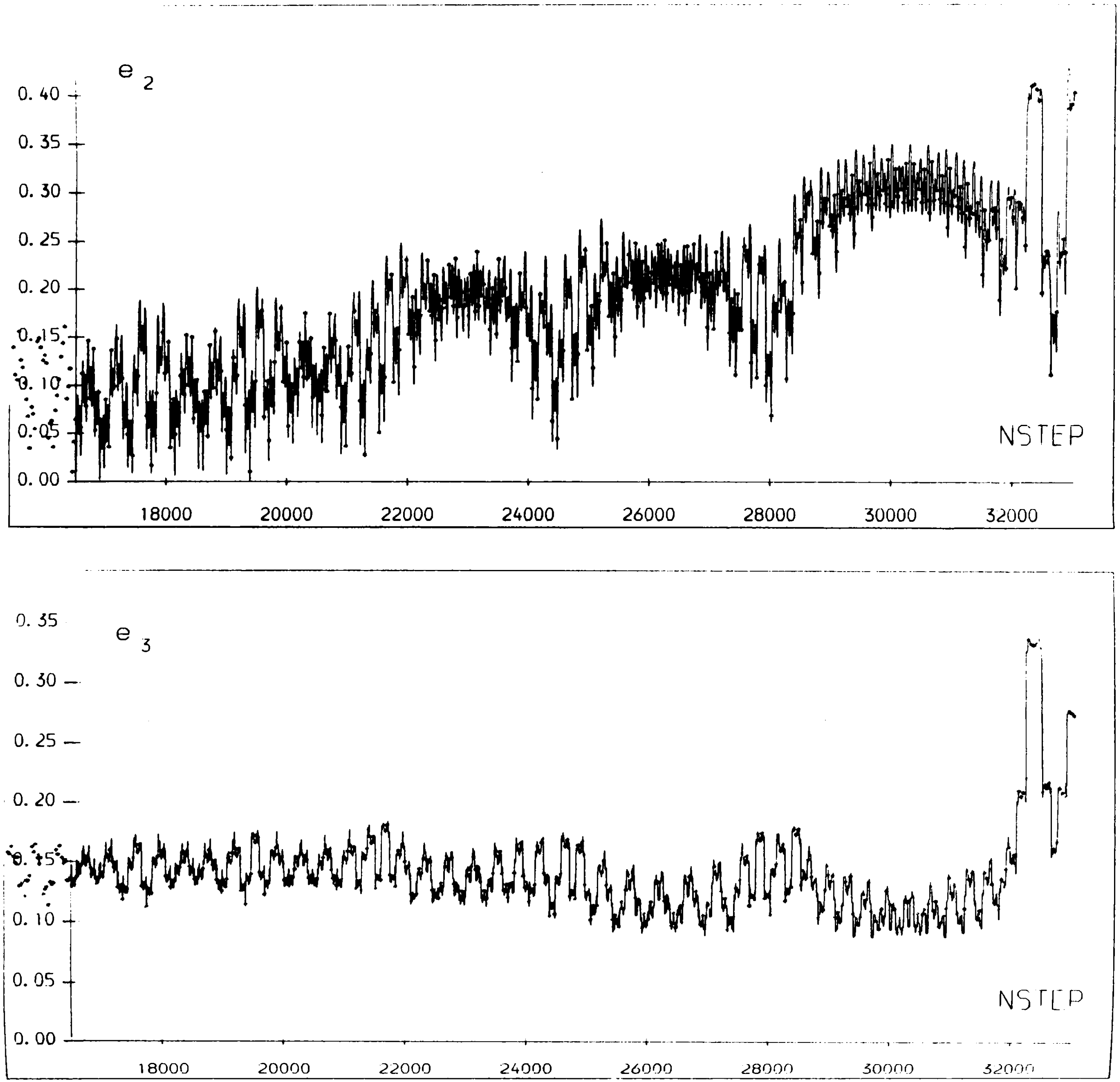
b


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
e 2
NSTEP
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
e 3
NSTEP
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000

C

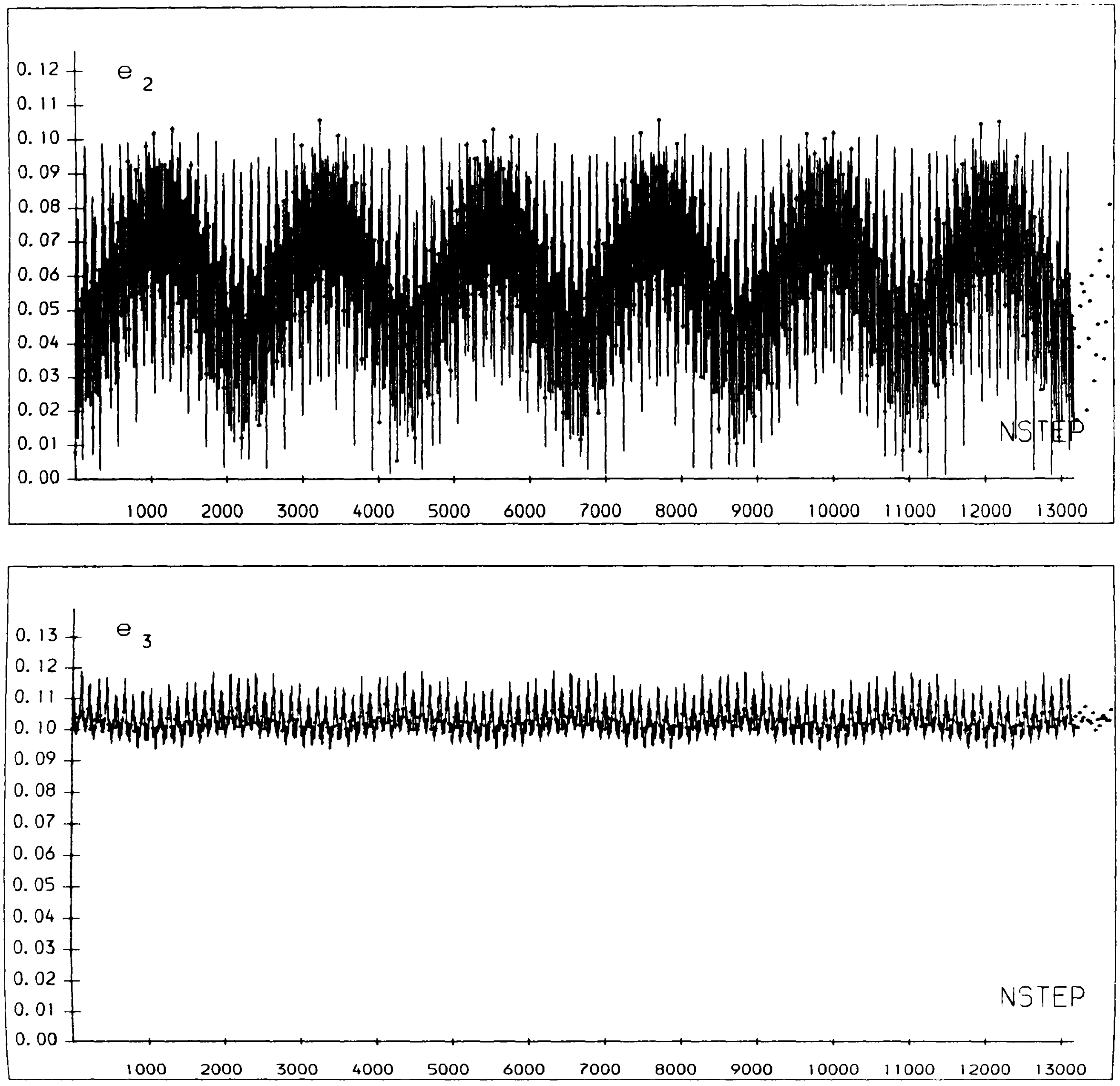

e 2
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
NSTEP
e 3
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
NSTEP

d

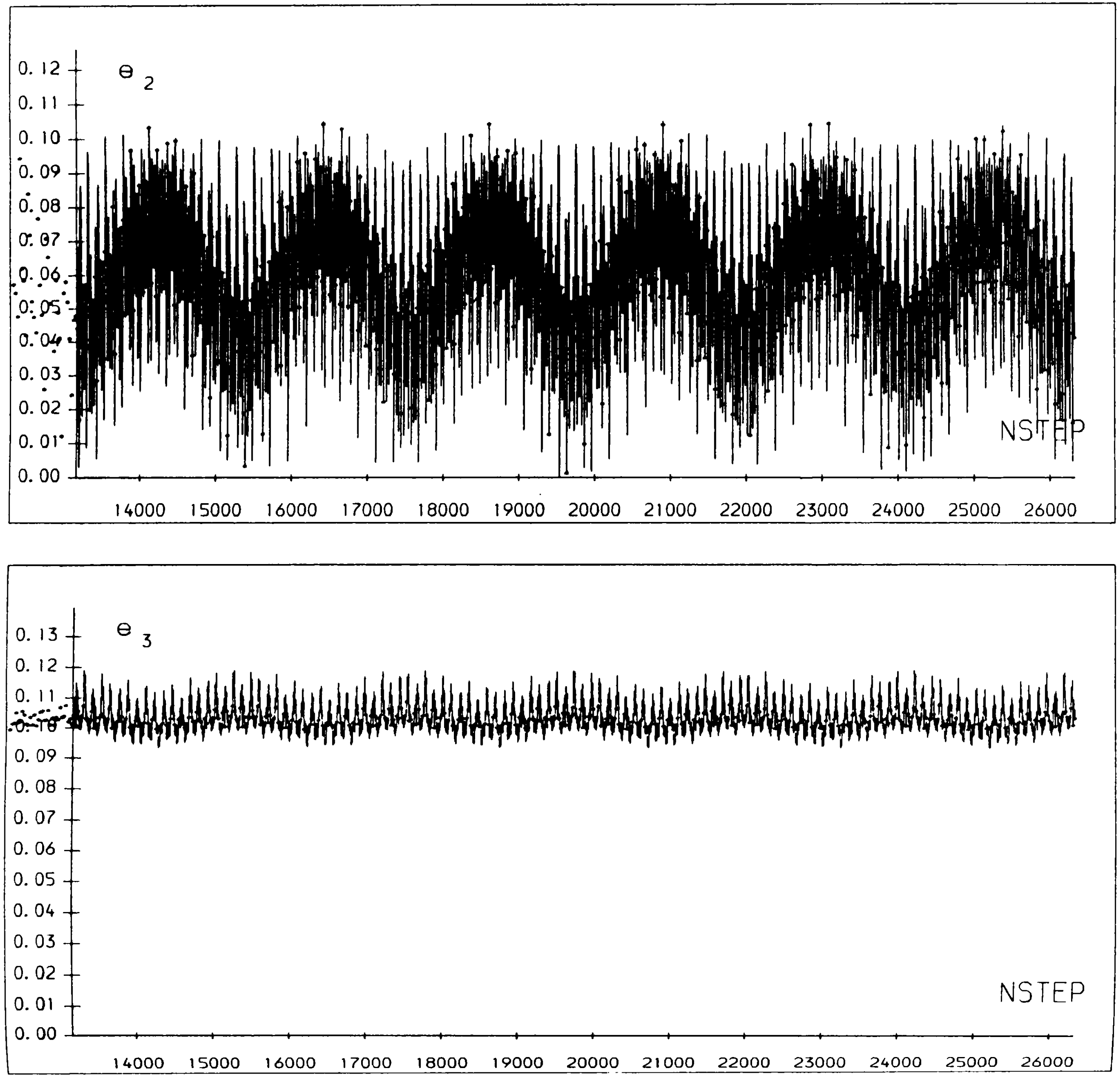

e 2
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
18000
20000
22000
24000
26000
28000
30000
32000
NSTEP
e 3
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
18000
20000
22000
24000
26000
28000
30000
32000
NSTEP

d

continued


e 2
0. 12
0. 11
0. 10
0. 09
0. 08
0. 07
0. 06
0. 05
0. 04
0. 03
0. 02
0. 01
0. 00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
NSTEP
e 3
0. 13
0. 12
0. 11
0. 10
0. 09
0. 08
0. 07
0. 06
0. 05
0. 04
0. 03
0. 02
0. 01
0. 00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
NSTEP

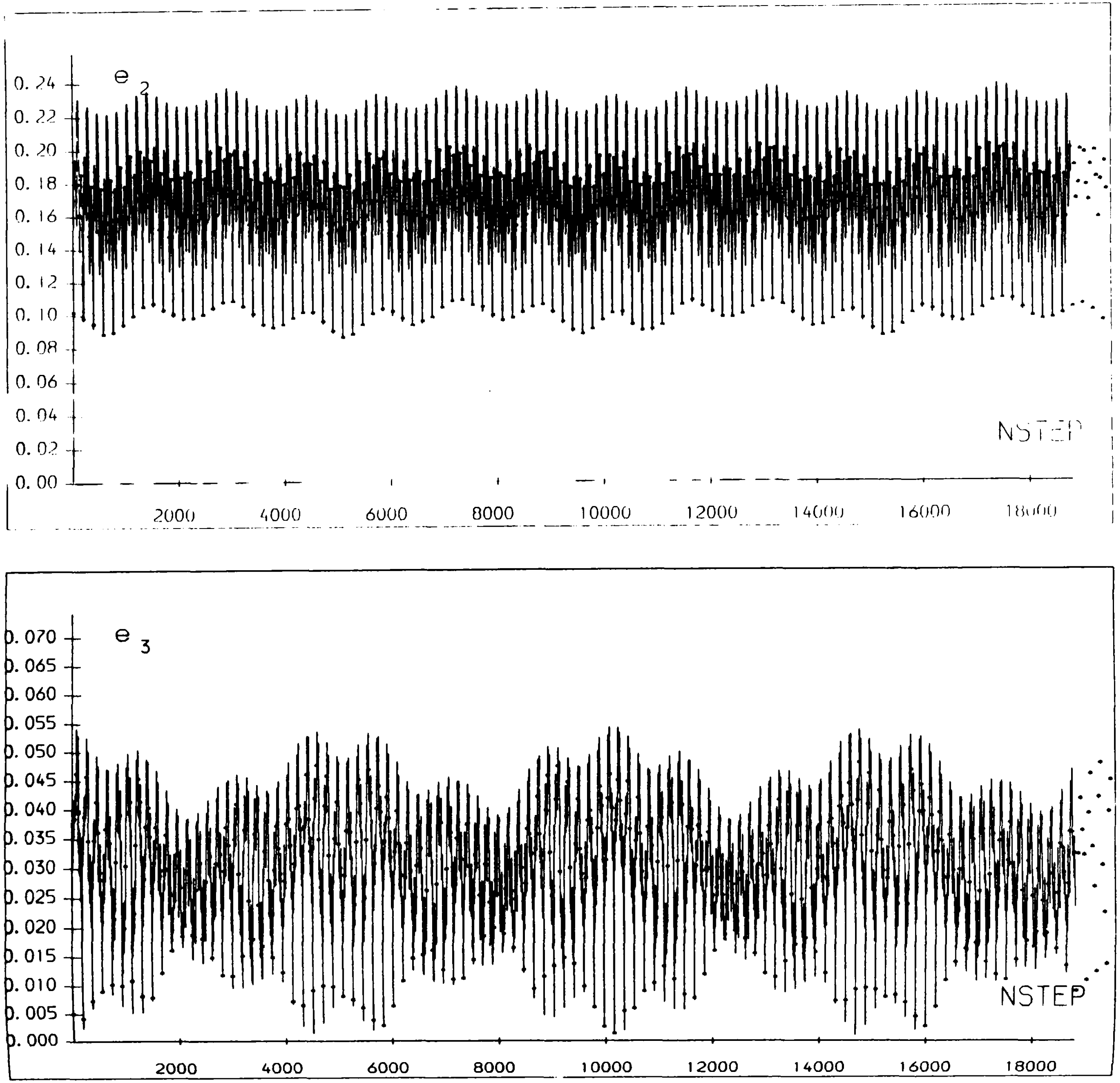
e



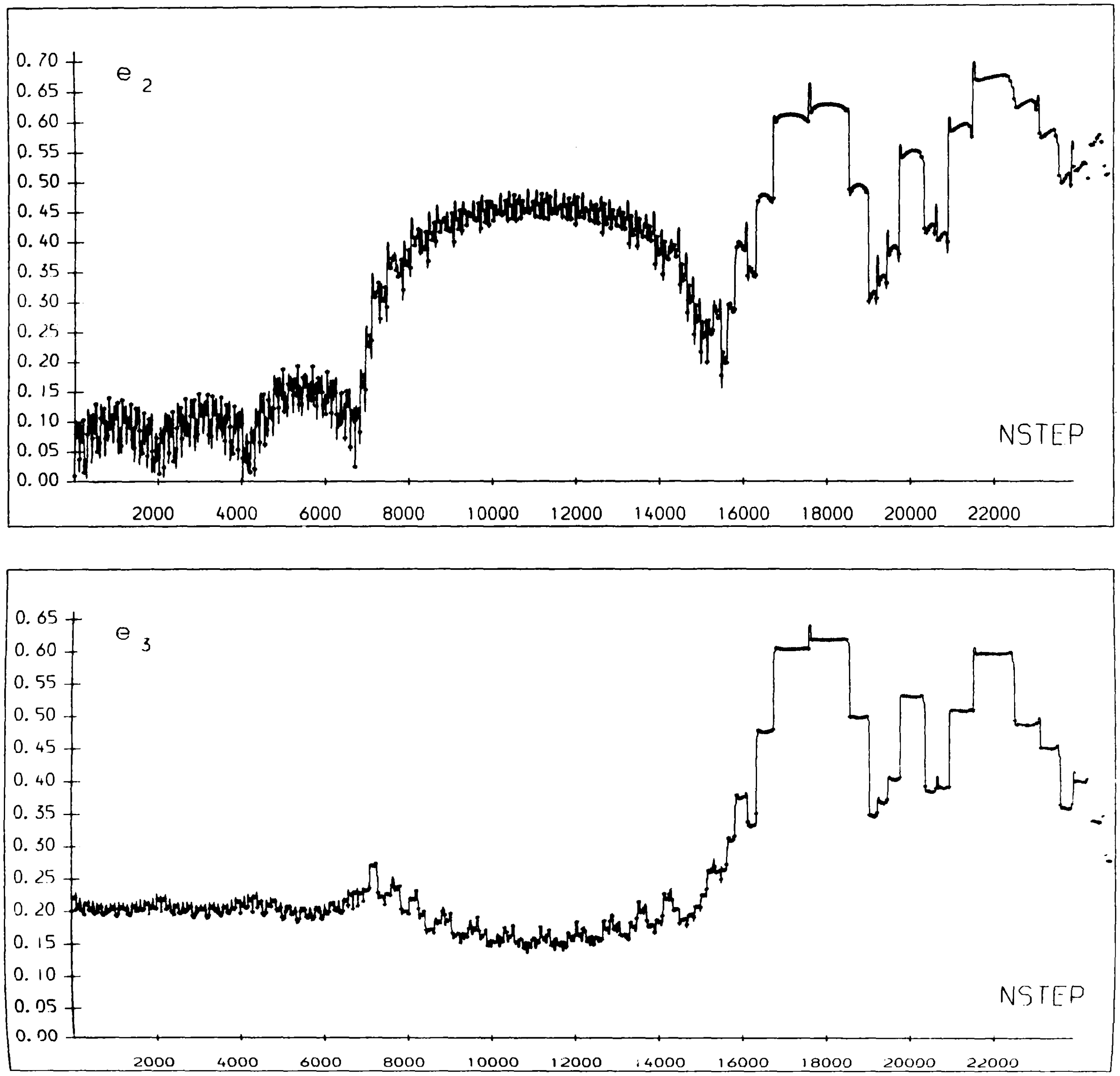
0. 12
0. 11
0. 10
0. 09
0. 08
0. 07
0. 06
0. 05
0. 04
0. 03
0. 02
0. 01
0. 00
e 2
NSTEP
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
0. 13
0. 12
0. 11
0. 10
0. 09
0. 08
0. 07
0. 06
0. 05
0. 04
0. 03
0. 02
0. 01
0. 00
e 3
NSTEP
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000

e

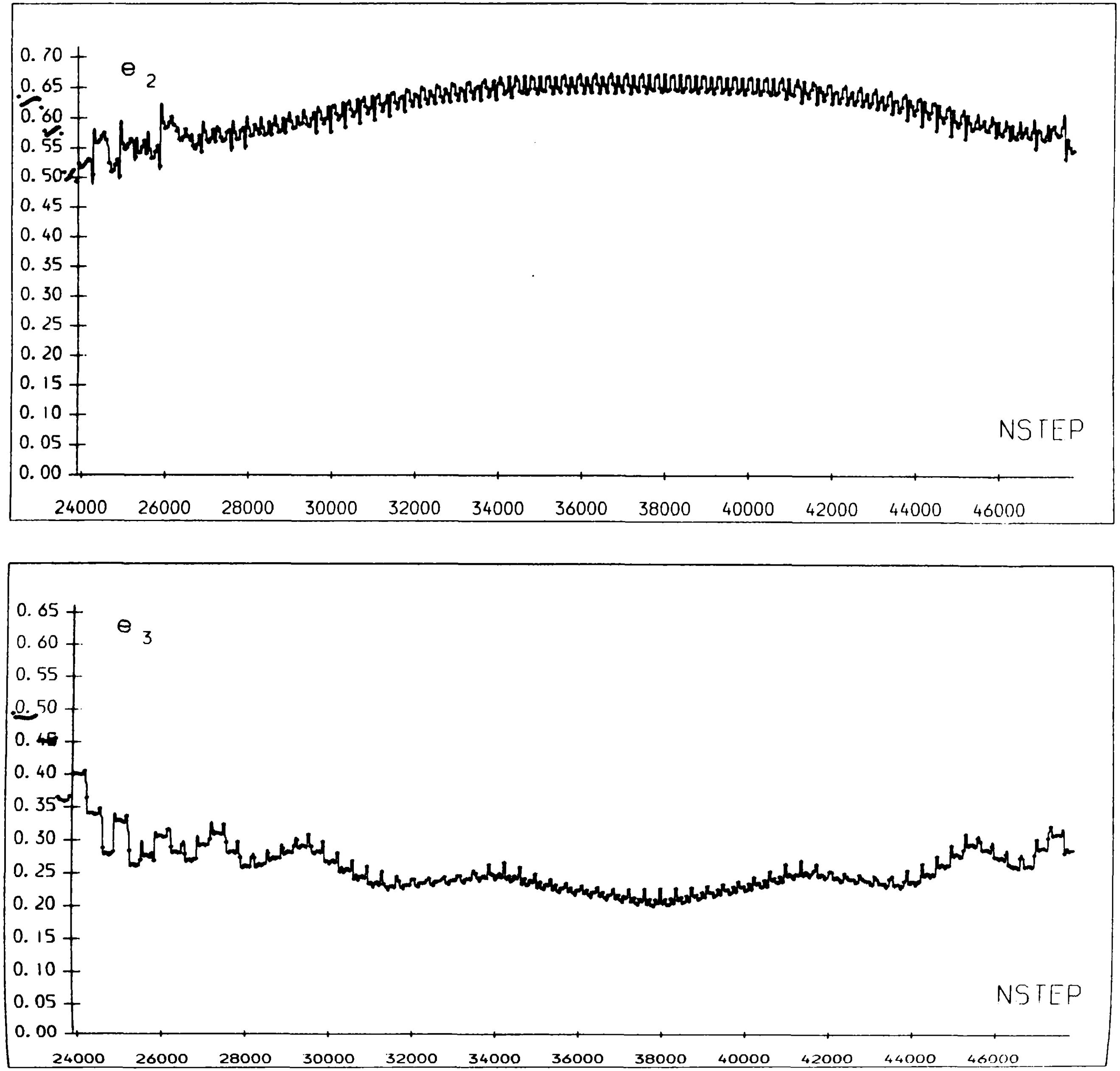
continued


e 2
0. 24
0. 22
0. 20
0. 18
0. 16
0. 14
0. 12
0. 10
0. 08
0. 06
0. 04
0. 02
0. 00
2000
4000
6000
8000
10000
12000
14000
16000
18000
NSTEP
e 3
0. 070
0. 065
0. 060
0. 055
0. 050
0. 045
0. 040
0. 035
0. 030
0. 025
0. 020
0. 015
0. 010
0. 005
0. 000
2000
4000
6000
8000
10000
12000
14000
16000
18000
NSTEP

f

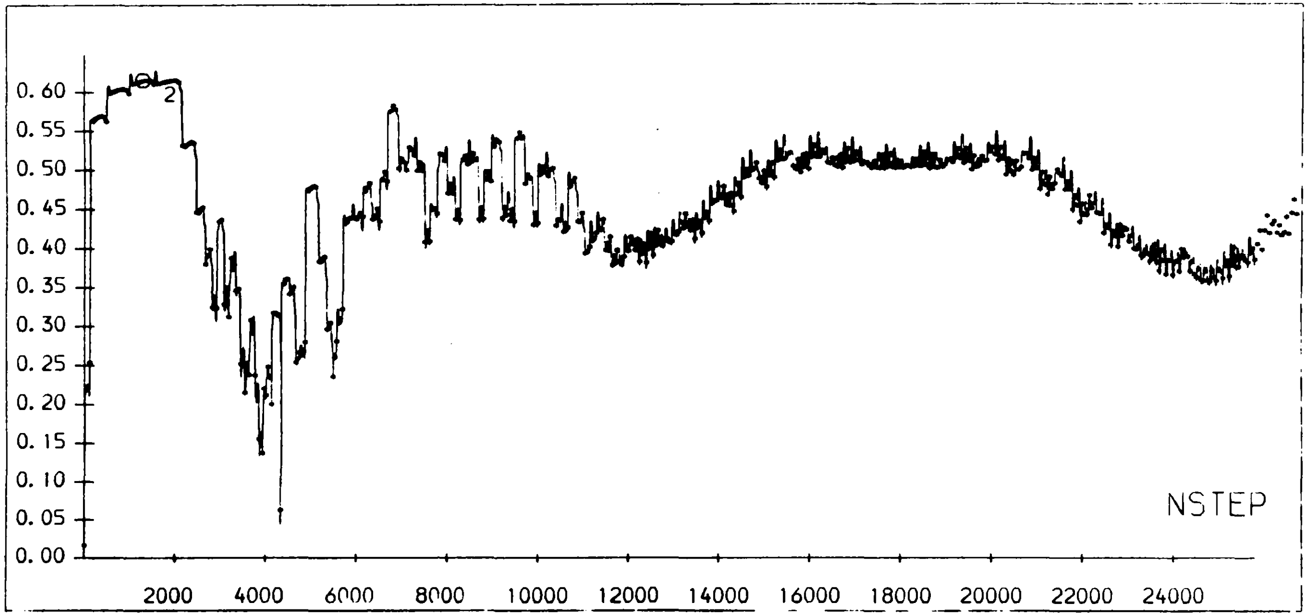

0. 70
0. 65
0. 60
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
e 2
NSTEP
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
0. 65
0. 60
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
e 3
NSTEP
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000

g

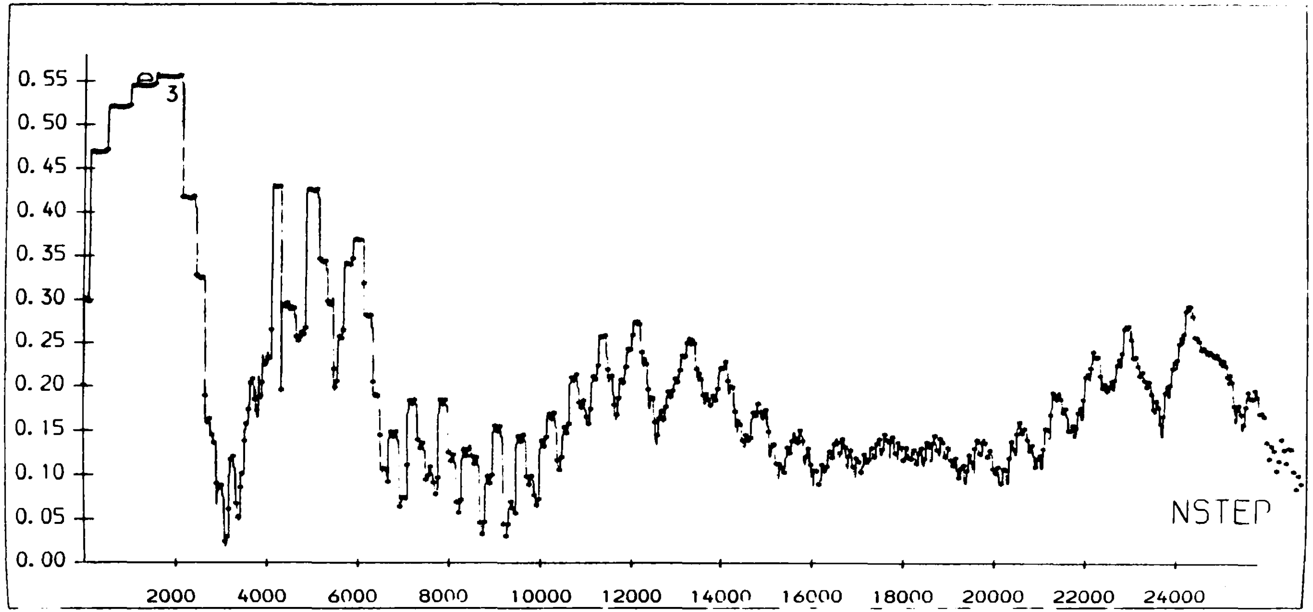

θ 2
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
24000
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
NSTEP
θ 3
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
24000
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
NSTEP

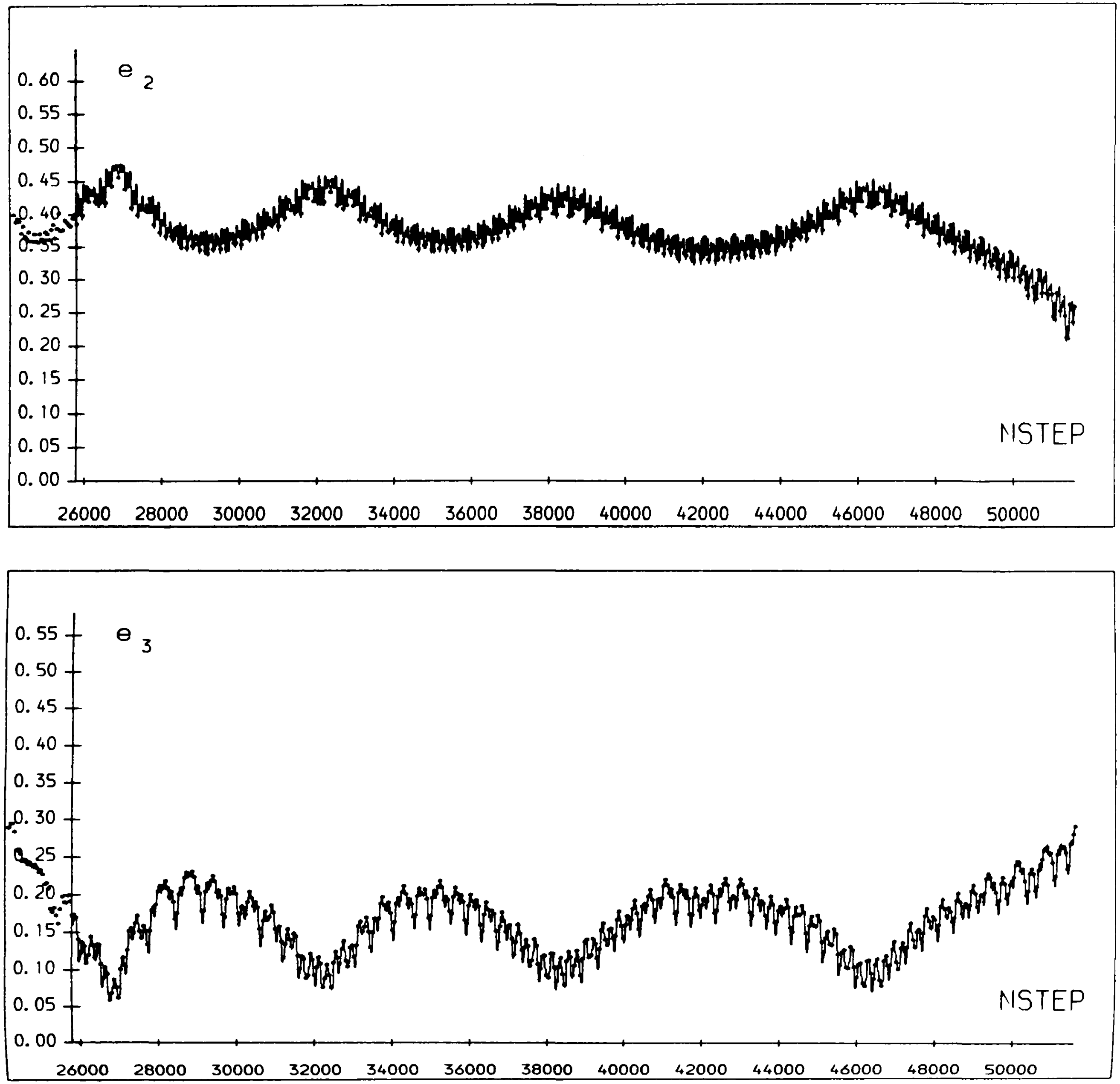
g

continued


0. 60
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
|e|
2
NSTEP

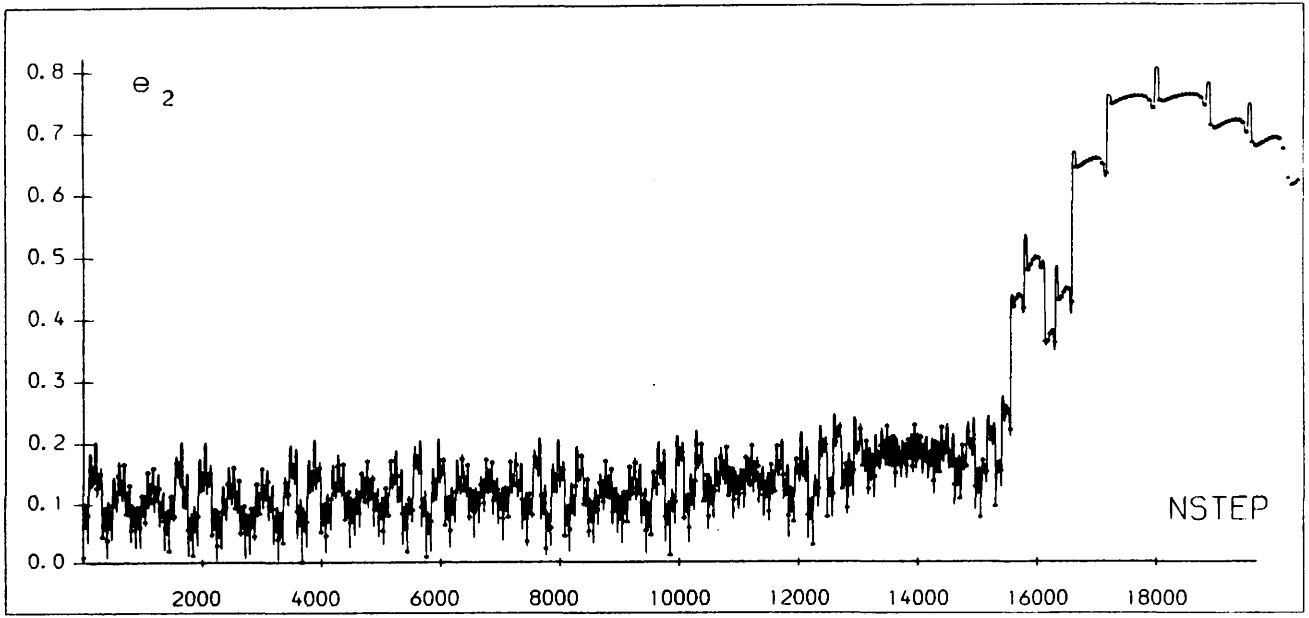
h

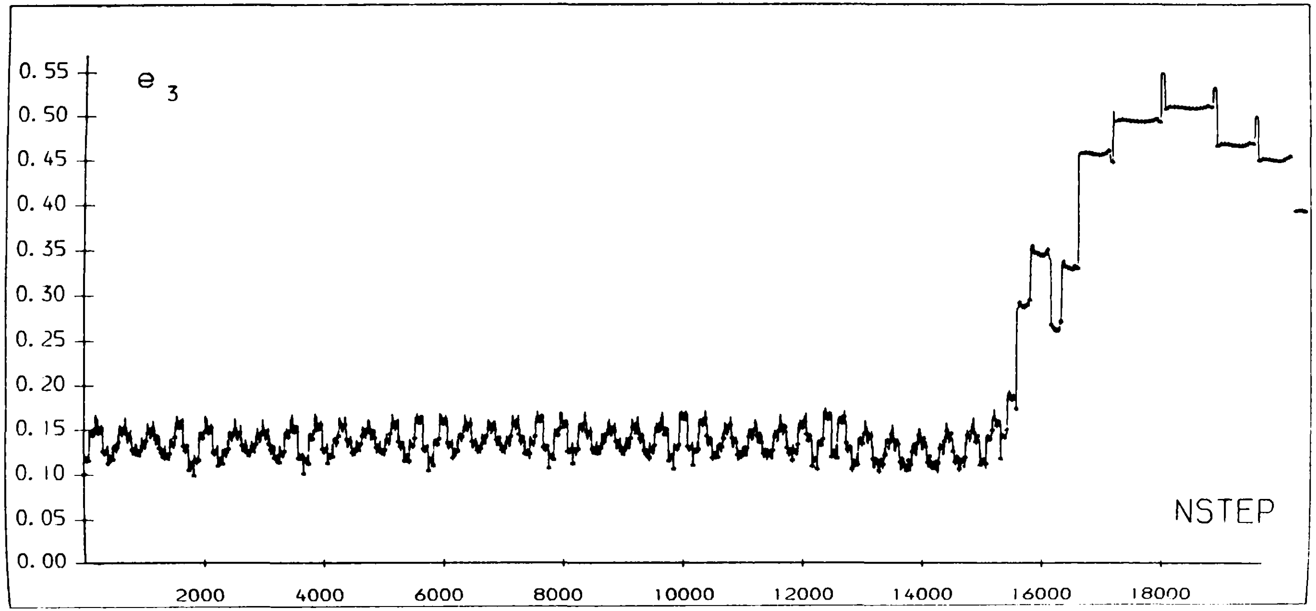

0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
e
3
NSTEP

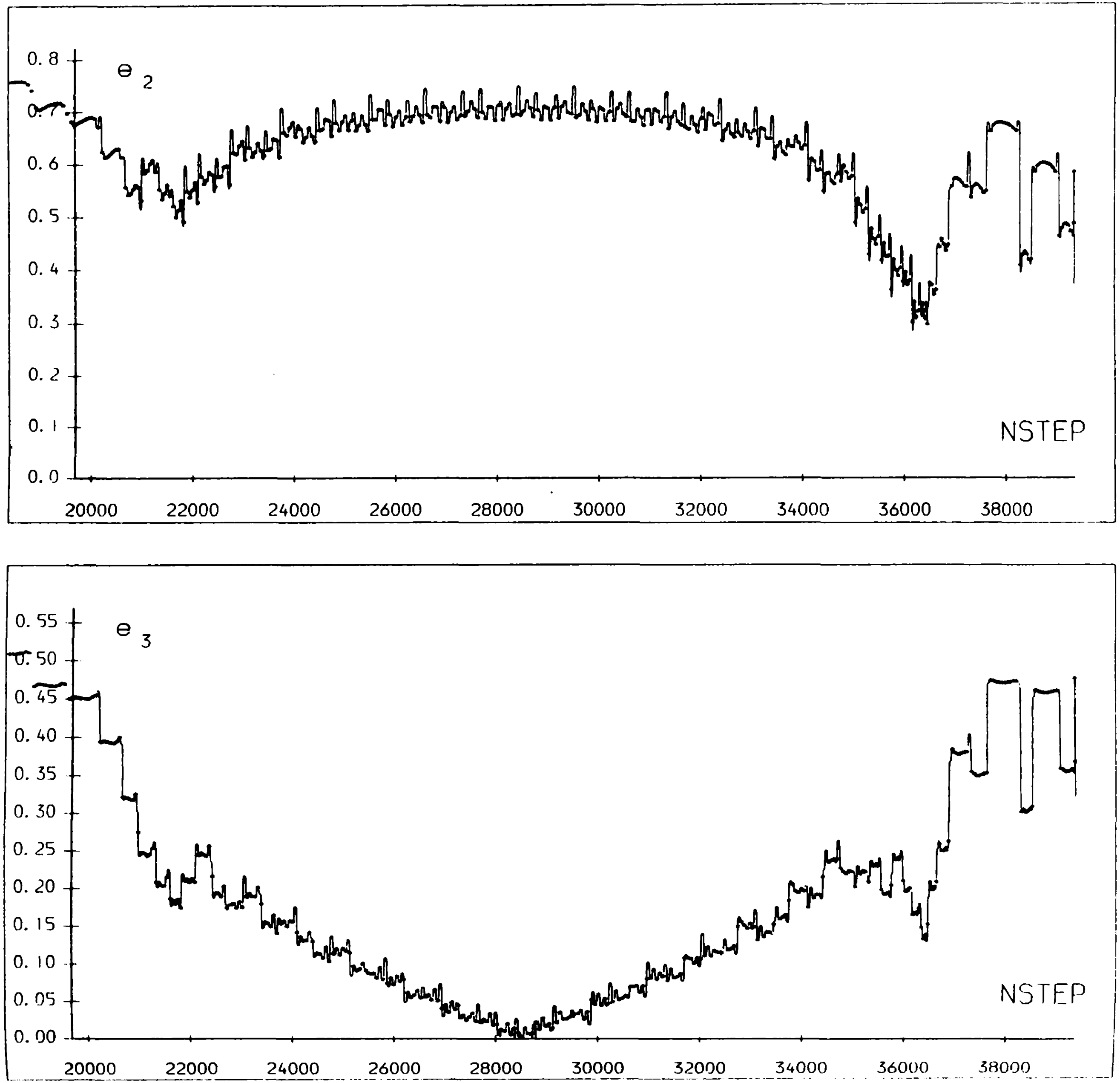

e 2
0. 60
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
50000
NSTEP
e 3
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
50000
NSTEP

h

continued

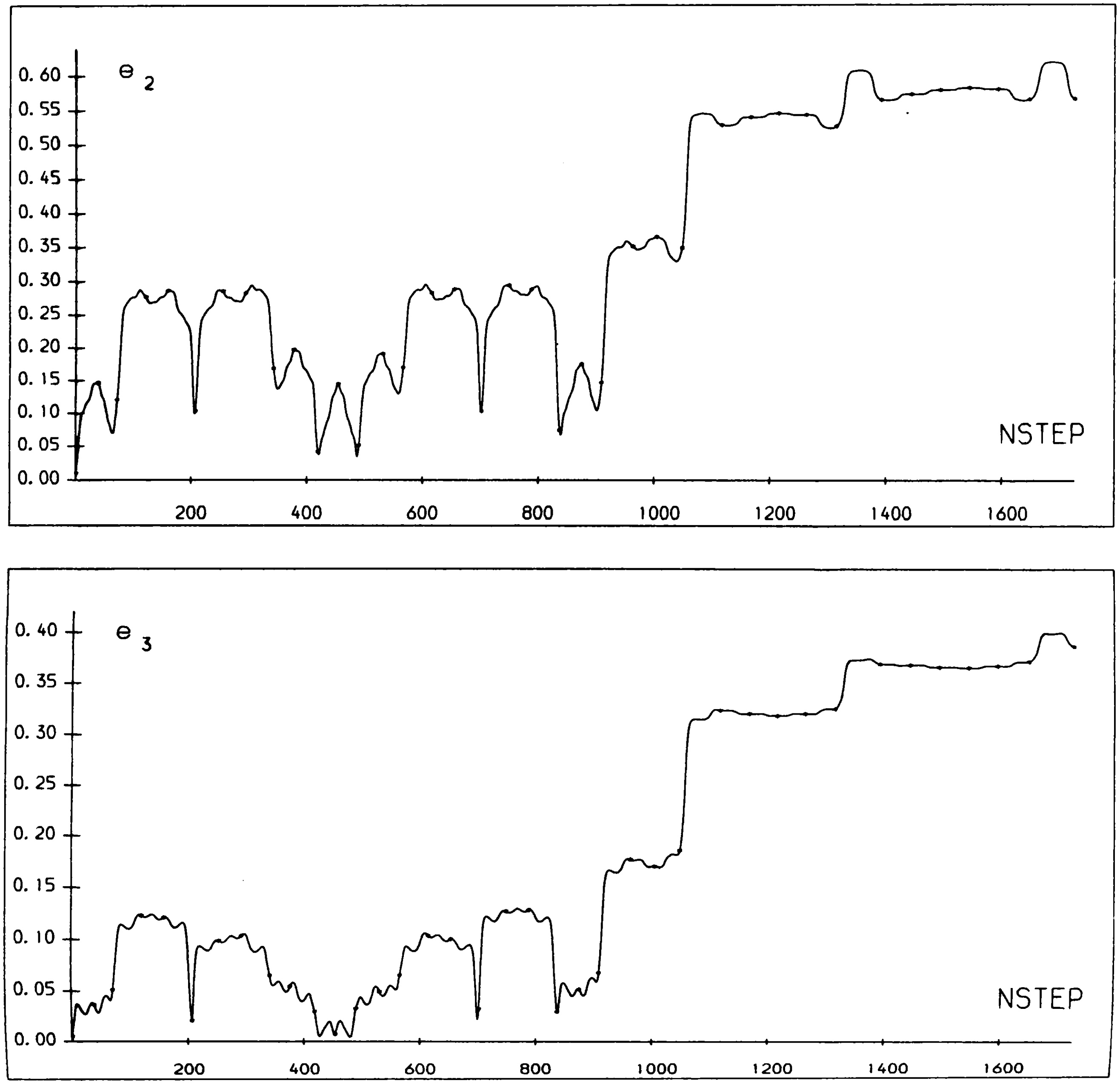

e 2
0. 8
0. 7
0. 6
0. 5
0. 4
0. 3
0. 2
0. 1
0. 0
2000
4000
6000
8000
10000
12000
14000
16000
18000
NSTEP


e 3
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
2000
4000
6000
8000
10000
12000
14000
16000
18000
NSTEP


e 2
0. 8
0. 7
0. 6
0. 5
0. 4
0. 3
0. 2
0. 1
0. 0
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
NSTEP
e 3
0. 55
0. 50
0. 45
0. 40
0. 35
0. 30
0. 25
0. 20
0. 15
0. 10
0. 05
0. 00
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
NSTEP

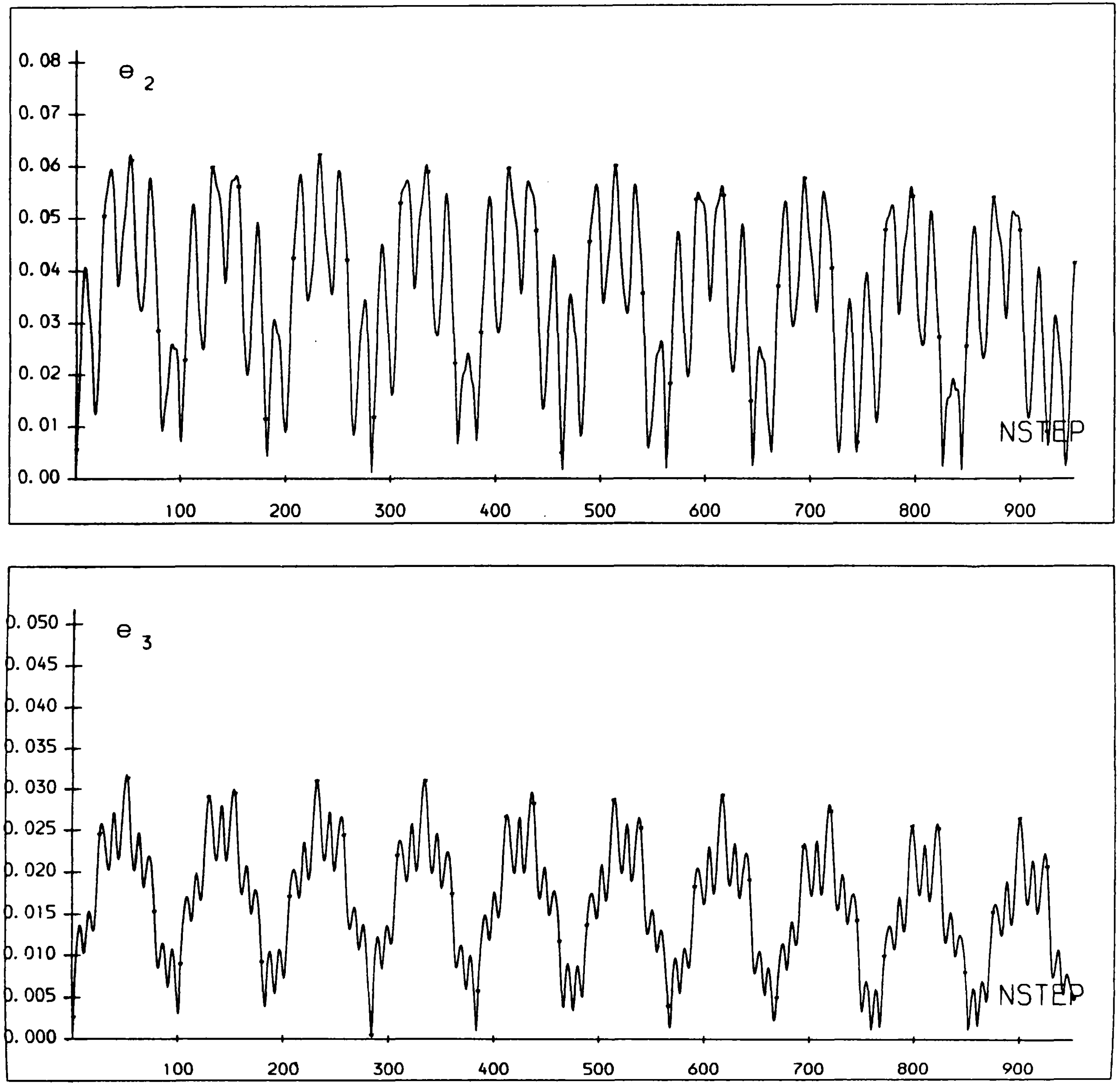
i

continued


e 2
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
200
400
600
800
1000
1200
1400
1600
NSTEP
e 3
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
200
400
600
800
1000
1200
1400
1600
NSTEP

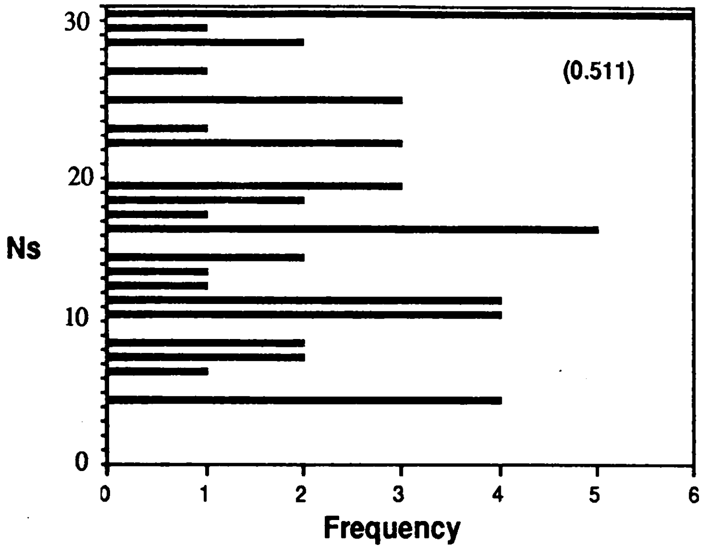
j

(direct)



j

(retrograde)



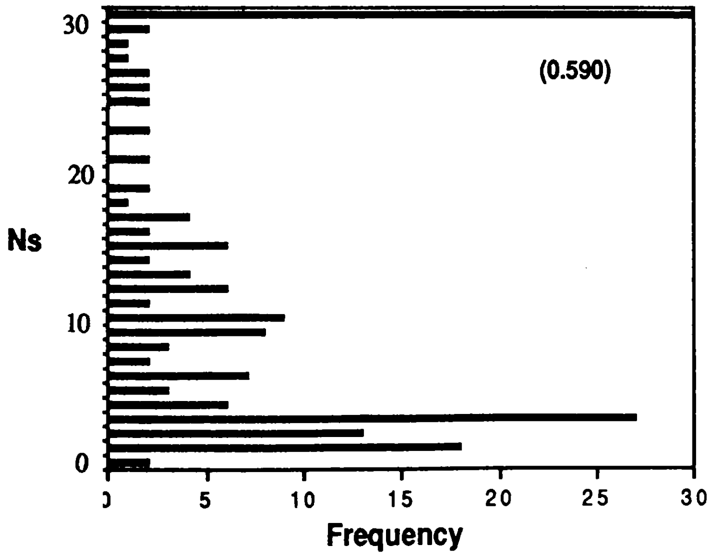
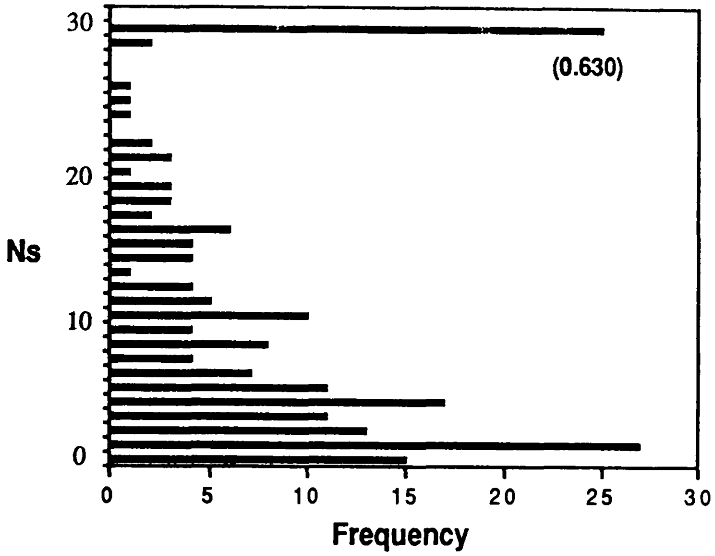


Figure 5.13 Distribution of systems with their lifetimes (Ns) for the same (ε_{23}, ε_{32}) parameters and eccentricities. The four coplanar 3-body systems are taken from the set [0226] of Figure 5.11. The number in the up right corner of each diagram indicates the value of the initial α. About 200 systems are investigated for each α, and the top 'bar' counts for the number of systems which last more than 30 SPs.



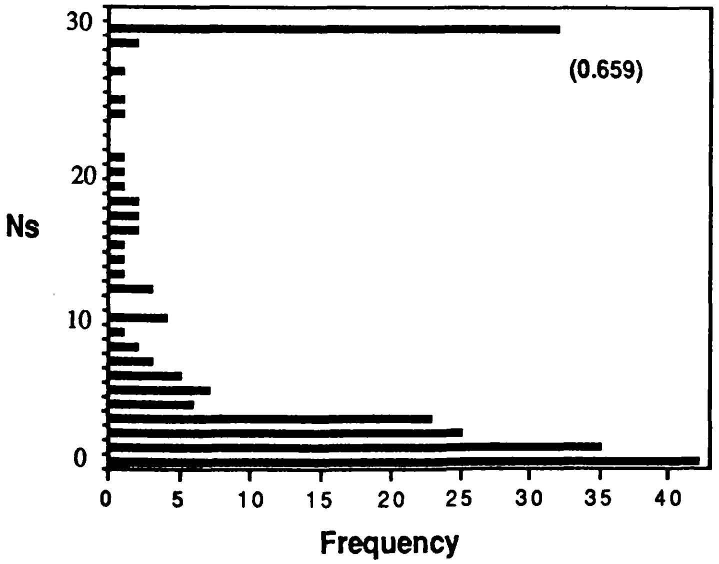


Fig. 5.13 (continued)

5.5 Numerical Experiments on 3-Body Systems III - A First Test of Roy's Statistical Stability Conjecture

Considering the practical value of the complete definition of hierarchical stability it still remains possible to fit smooth curves to the lifetime plot following Walker and Roy (1983), with a modification of allowing the curve to go into the Hill-type stability region given by the C^2H criterion. Furthermore, the fitted curves must be justified to carry a statistical interpretation: coplanar hierarchical 3-body systems with the same initial (e_2 e_3 ε_{23} ε_{32} α) parameters but different relative orientation of the two orbits and different initial positions on the orbits of the masses usually have correlated lifetime; the maximal possible lifetime for each initial α fits into a smooth curve.

In order to test or obtain this curve, however, one would need a much larger set of numerical integration experiments. In the present work, only about 200 systems are numerically integrated for each of the four initial α's taken out of the set [0226]. This set was chosen because of the instabilities below α_c, which make possible a test of systems inside the Hill-type stability region.

The results are presented in Fig. 5.13, where the range of life-time is chosen so as to cover the maximally possible life-time. It is already clear that the maximal possible lifetime shifts continuously toward larger values for smaller values of the initial α, and the distribution of the systems are more spread out over larger life-times. In fact, while most of the systems' life-time is below 30 SPs for initial α=0.659, 0.63, and 0.59, only about 40 of the 200 systems with initial α=0.511 have life-time below 30 SPs, although the maximal possible life-time is only about 15 SPs. We also note that for smaller value of initial α, the life-time is very unlikely to be zero. Thus we conjecture that, the smaller the value of initial α, the bigger the value of the shortest life-time.

We note from Fig. 5.11 that although the α=0.511 system is inside the Hill-type stability region, the cross-over instability was observed during this investigation. This is not a counter example for the result proved in sections 5.1 and 5.2 (see equation (5.4)); because the value of α_c is also a function of the relative orientation and the position of the masses on the orbits. The value of α_c indicated in Fig. 5.11 is in fact the greatest value for the given parameters specified there.

5.6 Summary

In this chapter we proved using a geometrical method that when a 3-body system (either planar or spatial) satisfies the Hill-type stability condition, then it preserves a geometrical hierarchy in the sense that not only $\rho_2 < \rho_3$ but also $\alpha_c < \alpha_x < 1$. An important statement concerning the size of IU^2 at the three collinear critical configurations was proved by Walker & Roy (1981) in a limited case. This is shown to be true in general as an immediate result of our proof.

In spite of the theoretical importance of the analytical Hill-type stability criterion and the applicability of the circular C^2H stability criterion (see eg. Szebehely and Zare, 1976; Walker and Roy, 1983), it is shown in the present work that the elliptical C^2H stability criterion is valueless in indicating stability of the prograde coplanar 3-body systems. The numerical experiment does not show obvious tunnel-shaped critical stability surfaces. Moreover, stable systems exist outside the analytical stability region and unstable systems exist inside it. New complicated valley and plateau structures are observed in the diagrams of Fig. 5.11, which are believed to be a reflection of the complexity of the phase space structure of a general nonlinear system.

As a future work we hope to carry out the numerical experiment using the μ-parameters, and investigate the 3-dimensional motion of the 3-body problem. It is also desirable to find some explanation for the phenomena observed here from chaotic dynamics.

CHAPTER 6

Symmetries and Conservation Laws in General Relativity

In the last two chapters we have discussed the restrictions on possible motion of classical gravitational N-body systems by the energy and angular momentum integrals. However, the result is only important for systems with not more than three bodies, and cannot be generalised to systems with more than three bodies. The purpose of this and the following chapters is to generalise the classical results into the framework of general relativity; in particular, bounded motion of isolated gravitational few-body systems in asymptotic flat spacetime (see Misner et al, 1973). Such an approach is completely novel. Although it is quite apparent that many of the general difficulties will be encountered in the course of this approach, we will obtain some important results.

To proceed we must modify the conventional way of achieving restrictions on allowed motion. Instead of discussing the constraints of conserved quantities (integrals) on the possible motion, as is usual and apparently obvious, we look for the relationship between forbidden motion and the underlying symmetries of the problem. In classical mechanics, it may look as though this change of view is trivial if not strange, for it is well-known that conservation laws are related to symmetries through Noether's theorem (Noether, 1918; see also Olver, 1986; Abraham & Marsden, 1978). In fact, Smale (1970) preferred to discuss the topology of mechanical systems in terms of 'symmetry' rather than 'integrals' although in this case the difference in approach was not important (see also Heisenberg, 1967).

However, it turns out to be necessary to make such a change of view in general relativity, because here there is in general no tensorial integral conservation laws (but see Dixon, 1979) since a spacetime metric does not usually admit Killing vectors. This problem caused some severe difficulties for several decades after the establishment of general relativity and is still attracting extensive interest today (see, for example, Ehlers (1979) for some review papers on the alternative formulation of conservation laws). The coordinate dependent formulation of conservation laws found by Landau & Liftshitz (1962) has gone some way to solving this problem in a restricted way but much work is still required. From the discussion on coordinate dependent conservation laws given in section 6.3, we see that some nontrivial uncertainties may exist concerning the relationship between conserved quantities such as energy and angular momentum and any symmetries which may exist. In contrast, it is quite natural and in fact necessary to put the system under investigation into an asymptotically flat spacetime and to assume some definite asymptotic symmetries. Asymptotic symmetry is always a definite concept and therefore its relationship to forbidden motion may be more fundamental, while the notion of conserved quantities may be regarded merely as of secondary importance but nevertheless useful. In fact, from the present discussion, it is found that it is not strange to relate these two concepts directly by ignoring the concept of integrals.

It is not surprising that forbidden motion may be found for some fictitious systems possessing symmetries (Killing vectors). However, an investigation of the relationship between forbidden motion and asymptotic symmetries in more general cases is not only important for the sake of determining ordered motion, but is also of interest by itself. The analysis of this chapter will provide the reason why we finally adopt the idea of asking for the relationship between forbidden motion and (asymptotic) symmetry; and in the next chapter we will show that the latter does impose restrictions on the motion of the system. We will discuss the constraints imposed by energy and angular momentum, but always keep in mind that it is the symmetry that is important.

The task of this chapter is to present a general discussion on symmetries and conservation laws in general relativity; and in particular, the system's energy and angular momentum. Most of the material is standard; but it will be formulated to favour the present approach. In the meantime, some questions are clarified by more profound answers. We have chosen to keep our notation in this and the following chapters as close as possible to Schutz (1980), namely, a bold face letter denotes tensor, barred letter a vector, tilded letter a 1-form.

In section 6.1 the relevant results from differential geometry are introduced. A fundamental difference between vector and 1-form on manifold is emphasised, which finds an application in later sections. In section 6.2 we discuss a relationship between Killing vectors and conserved quantities for geodesic motion on metric manifolds. A question concerning the importance of vectors and 1-forms in connection with conservation laws is answered. In section 6.3 Gauss' theorem is introduced using the language of manifolds, and applied to investigate the general relationship between (asymptotic symmetries) and conservation laws in general relativity. New formal conservation laws are constructed.

In section 6.4 we propose to study the integrability of relativistic systems and to discuss a possible way of determining relativistic chaos. The classical motion of one particle in an exterior field (cf. chapter 2) is formulated in the language of geodesic motion on Riemann manifolds; thus establishing a connection between the Poisson bracket and Lie bracket. In section 6.5 we make the first attempt to generalise the classical Sundman inequality approach into the study of bounded motion in relativity. This is applied to produce the simplest forbidden motion in general relativity, namely, that of the geodesic motion in the Schwarzschild geometry.

6.1 Elementary Differential Geometry

Although both Hamiltonian dynamics and relativity can be formulated using coordinate dependent languages, the coordinate-free geometric notion provides a great advantage in dealing with these problems. In approaches using geometric objects one often feels a lack of notation for what one wants to express; however, the freedom obtained by escaping from the constraints of a particular coordinate system often provide a more profound insight into the question.

In this section we will introduce the relevant concepts of topology, calculus on differential and Riemann manifolds. The main references are Bishop & Goldberg (1968), Misner et al (1973), Choquet-Bruhat et al (1977), and Schutz (1980).

Manifold, Vector, 1-Form and Tensor

The abstract concepts of modern mathematics are often purified versions of familiar concepts. To some extent, modern mathematics may be looked upon as an abstract building process that introduces more and more structures to the most basic concepts of set theory. A **topological space** may be interpreted as a set with a local **topology** structure, which defines a **neighbourhood** for each point of the set (see Janich, 1984). A topological space is still very abstract, since the topological structure, ie. neighbourhood, is not necessarily a continuous region. For example, a finite lattice forms a trivial topological space; whereas the normal distance can induce a topology on the linear space R^n. In order to define something more useful for the physics of the real world, more structures must be introduced one after another, the first one being continuity.

A **manifold** is a topological space with a local topology similar to that of the linear space R^n, whose local topology is defined in the usual way. Put in other words, a **chart** coordinate system $\{x^1, ..., x^n\}$ is defined for a neighbourhood of any point on a manifold; moreover, in the region where two chart coordinates overlap, there must be a 1-1 C^k transformation between the two chart coordinates. The second requirement implies that all the chart coordinates must have the same dimension, which is called the dimension of the manifold. Later we will see that a metric tensor may be introduced to form a Riemannian manifold.

Mappings between manifolds are also often encountered. A **homeomorphism** is a continuous 1-1 onto mapping between two continuous topological spaces such that the inverse mapping is also continuous. It is worth noting that the continuity of a 1-1 mapping does not guarantee the continuity of the inverse mapping. A **diffeomorphism** is a C^∞ 1-1 onto mapping between two continuous topological spaces such that the inverse mapping is also C^∞. In the same way one can define C^k differentiable mappings.

On manifold, one can talk about most physically useful concepts such as functions, curves and tangent vectors. A function on a manifold is a mapping from the manifold to a subset of R^1. A **curve** is defined as a differentiable mapping from an open set of R^1 into a manifold M. Thus one associates to each point on the curve with a number, say λ, in the open set of R^1; the curve is said to be parameterized by the parameter λ. Different parameterization of a single path will be considered as different curves.

It is well-known that in linear algebra, vectors and vector space are taken as the starting point, then 1-forms and tensors can be defined on the vector space as operators. The whole content of these can be defined on the manifold as well, but they will carry richer meaning. For example, in linear algebra vectors and 1-forms have equal position, they are different on a manifold. Although this linear structure can be introduced to a manifold either by starting from vectors or starting from 1-forms, it is conventional to follow the former. In developing this structure on a manifold, one always assumes the tensor algebra on linear space as preliminary.

In linear algebra, a **1-form** (denoted by a tilde '~' over a letter) is defined as a linear, real-valued operator on vectors. Moreover, we can define the addition of 1-forms and their multiplication by real numbers such that they form a linear space, called the **dual space** of the vector space. It can be proved that the dimension of the dual space is the same as that of the vector space; moreover, because of the linearity structure of both vectors and 1-forms, they have a symmetric dual position. The following notations will be used to denote the contraction of a 1-form with a vector, viz.

$$\tilde{\omega}(\bar{V}) = \bar{V}(\tilde{\omega}) = < \tilde{\omega}, \bar{V} > = < \bar{V}, \tilde{\omega} > \quad .$$

Thus in linear algebra, vectors and 1-forms are symmetric, that is, they are not distinguishable by their properties. However, we will see that this symmetry is broken on manifold.

In the linear space of 1-forms dual to that of the vectors, any n linearly independent 1-forms constitute a basis. However, once a basis has been chosen for the vectors, this induces a preferred basis for the 1-forms, called the **dual basis**. It is defined by

$$< \tilde{\omega}^i, \bar{e}_j > = \delta^i_j$$

for every i, j=1, ..., n.

A tensor (denoted by a bold face letter) is defined as a linear, real-valued operator on 1-forms and vectors. A tensor of type (N, N') takes as arguments N 1-forms and N' vectors, and it is conventional to put all the 1-forms before the vectors. Tensors of each type are also assigned a linear structure to form linear spaces. Tensors can be constructed from the **outer product** of tensors (denoted by ⊗), and the outer product of basis vectors and their dual basis 1-forms forms a basis of the tensor spaces. However, not all tensors can be formed by outer products. For example, Not all (2, 0) type tensors can be expressed as the outer product of two vectors (see Schutz, 1980, P59). We also note that a tensor is completely determined by its components on a basis.

A tangent vector on a manifold is a particular kind of linear operator, called derivation, which is not necessarily limited to a manifold. A **derivation** is a linear operator (or mapping) on an algebraic system (eg. linear space) which satisfies the Leibniz rule. An **antiderivation** is a linear operator on an algebraic system which satisfies the anti-Leibniz rule. For example, we can explicitly write out the derivation defined on the algebraic system of all tensors **T**, including scalar functions. Let the operator be denoted as $\mathbb{D}$, then being a derivation it must satisfy the axioms

$$\text{(a)}.\ \mathbb{D}\mathbf{T} = \text{a tensor of the same type as } \mathbf{T}\ ,$$

$$\text{(b)}.\ \mathbb{D} < \tilde{\omega}, \bar{V} > = < \mathbb{D}\,\tilde{\omega}, \bar{V} > + < \tilde{\omega}, \mathbb{D}\,\bar{V} > ,$$

$$\mathbb{D}\,c = 0\ ,\quad \mathbb{D} < \tilde{\omega}^i, \bar{e}_j > = \mathbb{D}\,\delta^i_j = 0,$$

$$\text{(c)}.\ \mathbb{D}(a\mathbf{A} + b\mathbf{B}) = a\,\mathbb{D}\mathbf{A} + b\,\mathbb{D}\mathbf{B} \qquad \text{(linear operator)},$$

$$\text{(d)}.\ \mathbb{D}(\mathbf{A} \otimes \mathbf{B}) = (\mathbb{D}\mathbf{A}) \otimes \mathbf{B} + \mathbf{A} \otimes (\mathbb{D}\mathbf{B}) \quad \text{(Leibniz rule)},$$

where a, b and c are constant numbers; and the conventional notations for vectors, 1-forms, tensors, basis vectors and its dual 1-form basis are used (eg. Schutz, 1980). A derivation can have its own linear structures to form a linear space. By using the first three axioms, it is straightforward to show that the Leibniz rule can be equivalently written as

$$\begin{aligned}\text{(d')}.\ \mathbb{D}[\mathbf{T}(\ldots\tilde{\omega}^i\ldots;\ldots\bar{e}_j\ldots)] = \ & (\mathbb{D}\mathbf{T})(\ldots\tilde{\omega}^i\ldots;\ldots\bar{e}_j\ldots) \\ & + \sum_i \mathbf{T}(\ldots\mathbb{D}\,\tilde{\omega}^i\ldots;\ldots\bar{e}_j\ldots) \\ & + \sum_j \mathbf{T}(\ldots\tilde{\omega}^i\ldots;\ldots\mathbb{D}\bar{e}_j\ldots)\ ,\end{aligned}$$

$$\begin{aligned}\text{(d'')}.\ \mathbb{D}[\mathbf{T}(\ldots\tilde{\omega}\ldots;\ldots\bar{V}\ldots)] = \ & (\mathbb{D}\mathbf{T})(\ldots\tilde{\omega}\ldots;\ldots\bar{V}\ldots) \\ & + \sum_i \mathbf{T}(\ldots\mathbb{D}\tilde{\omega}\ldots;\ldots\bar{V}\ldots) \\ & + \sum_i \mathbf{T}(\ldots\tilde{\omega}\ldots;\ldots\mathbb{D}\bar{V}\ldots)\ .\end{aligned}$$

Consider an n-dimensional manifold M, with a coordinate system $\{x^i\}$. One can show that an ordinary derivative at a point m on the manifold M along a λ-curve satisfies the above conditions, if the algebraic system operated on is the space of all analytic scalar functions. This is a very useful kind of derivation defined on a manifold; it is called the (tangent) **vector** at the point m∊M (denoted by a bar '-' over a letter). Such tangent vectors (with the usual linear structure) at a point m form a linear space, T_m, called the **tangent space** at m, whose dimension is equal to that of the manifold. One can also show that the tangent vectors along the coordinate lines $\{x^i\}$ form a natural basis of the tangent space at m; such a basis is a **coordinate basis**. The tangent vector of a λ-curve is usually denoted by d/dλ, and the coordinate basis by $\{\partial/\partial x^i\}$, or $\{\partial_i\}$, thus we have

$$\bar{V} = \frac{d}{d\lambda} = \frac{dx^i}{d\lambda}\frac{\partial}{\partial x^i} = V^i\frac{\partial}{\partial x^i}, \quad \bar{V}(f) = \frac{df}{d\lambda} = \frac{dx^i}{d\lambda}\frac{\partial f}{\partial x^i} \quad ,$$

where the summation convention is used.

A vector field refers to a rule for defining a vector at each point of M. **Linear independence** can be defined for both vectors at a point and vector fields; in the latter case, the independence is over scalar functions rather than constant numbers. A set of n linearly independent vectors (fields) forms a **basis** (fields), which need not be a coordinate basis. Given a C^1 vector field there is one and only one integral curve passing through each point, whose tangent vector is exactly the vector field.

A Lie bracket can be defined for two tangent vectors (not for vectors in linear algebra), which can be proved to produce a new vector,

$$\bar{U} \equiv \frac{d}{d\mu} \quad , \quad \bar{V} \equiv \frac{d}{d\lambda}$$

$$[\bar{U}, \bar{V}] = \bar{U}\,\bar{V} - \bar{V}\,\bar{U} = \frac{d}{d\mu}\frac{d}{d\lambda} - \frac{d}{d\lambda}\frac{d}{d\mu} \ .$$

One must note that although the Lie bracket of two vectors defines a vector, neither of the two terms is in general a vector. Therefore, the individual terms of the above equation are only defined on functions; but we will see that the bracket, as a vector, is also defined on 1-forms.

For vectors we have the following two very important results: (1). any nonsingular vector can be a basis field; (2). a set of n independent vectors forms a coordinate basis iff any two of them **commute**, ie. their bracket vanishes.

On manifolds, 1-forms can also be defined independent of tangent vectors. However, it is conventional to define a 1-form at a point m∊M based on the tangent vector space at

m. A **1-form field** is a rule which defines a 1-form at every point of the manifold. Then tensors at a point and tensor fields on the manifold can be developed in the same way. Again their linearity as either operator or linear space is defined over numbers at each point, ie. functions. It can be shown that the gradient of a function (with the usual addition and multiplication rule) is a 1-form on a manifold. However, not all 1-forms are gradients of a function. On a manifold the dual space is called the cotangent space at m, T^*_m. We have the following equations

$$\bar{V} \equiv \frac{d}{d\lambda} \quad , \quad \tilde{\omega} \equiv \tilde{d}f = f_{,i}\,\tilde{d}x^i$$

$$<\tilde{\omega},\bar{V}> \equiv <\tilde{d}f, \frac{d}{d\lambda}> \equiv \frac{df}{d\lambda} \quad , \quad <\tilde{d}x^i, \partial/\partial x^j> \equiv \frac{dx^i}{dx^j} = \delta^i_j \ .$$

From the last of the above equations we see that the 1-form gradients of the coordinates $\{x^i\}$ are in fact the basis dual to the coordinate basis vectors $\{\partial/\partial x^i\}$. We will see after the exterior derivative is defined that not every 1-form can be adapted to a set of coordinate basis 1-forms. The reason is that not all 1-forms are exact (nor closed), but coordinate basis 1-forms are necessarily exact. This is one of the examples showing the different properties of 1-forms and vectors on a manifold.

Properties of tensors and basis transformations are essentially the same as in linear algebra, hence notations will be specified in the context.

Lie Derivative, Exterior Derivative and Covariant Derivative

The Lie derivative is another example of a derivation on a manifold, which operates on tensors of any type. Although the Lie derivative can be defined in many standard ways (see the references mentioned at the beginning of this section), the following observation is more convenient. Since the manifold is defined as a space with coordinate charts covering it, the most natural and simplest derivatives of tensors on a manifold, as a generalisation of the everyday calculus, are partial derivatives of the components of a tensor with respect to the coordinates. The **Lie derivative** is just the coordinate-free, geometric version of this: consider the one of the coordinate vector fields as a coordinate-free vector field, and regard the partial derivatives of the components of a tensor in this coordinate system with respect to the chosen coordinate as a coordinate-free tensor of the same type. We thus have the Lie derivative of any tensor field with respect to any vector field

$$(\mathcal{L}_{\bar{A}}\mathbf{T})^{\cdots}_{\cdots} = \frac{\partial T^{\cdots}_{\cdots}}{\partial x^1} \qquad (\bar{A} = \partial_1) .$$

This interpretation is possible because any non-singular vector field can be a coordinate basis field. This point is very useful in rewriting component equations with partial derivatives with respect to coordinates into a tensorial form involving the Lie derivative and covariant derivative. Using the standard notations we have the following results

$$\mathcal{L}_{\bar{V}} f = \bar{V}(f) = \frac{df}{d\lambda} \quad , \quad \mathcal{L}_{\bar{V}} \bar{U} = [\bar{V}, \bar{U}] = - \mathcal{L}_{\bar{U}} \bar{V}$$

$$[\mathcal{L}_{\bar{V}}, \mathcal{L}_{\bar{U}}] = \mathcal{L}_{[\bar{V}, \bar{U}]} \quad , \quad \mathcal{L}_{\bar{V}} + \mathcal{L}_{\bar{U}} = \mathcal{L}_{\bar{V}+\bar{U}}$$

Moreover, a Lie derivative has all the properties of a derivation.

Now we are able to state a very important theorem about the submanifold, namely, Frobenius' theorem. An m-dimensional **submanifold** S of an n-dimensional manifold M is a set of points of M which are characterised in a coordinate system by $x^1 = ... = x^{n-m} = 0$. It is easy to prove that if two vectors are linear combinations (not necessarily with constant coefficients) of m vector fields, then their Lie bracket is a linear combination of the same m vector fields as well. **Frobenius' theorem** states its converse: if the Lie brackets of a set of m C^∞ vector fields with one another are linear combinations of the m vector fields, then the integral curves of the fields mesh to form a family of submanifolds.

On an n-dimensional manifold we can also define differential forms, integral calculus and exterior derivatives. A **p-form** ($p \geq 2$) is defined to be a completely antisymmetric (or, skew-symmetric) tensor of (0, p) type. Similarly, p-vectors can be defined. A 1-form is a (0, 1) tensor; a scalar function is a 0-form. The antisymmetric part of a (0, p) tensor is a p-form. In this work we will adopt the normalised antisymmetric (and symmetric) part, for example,

$$\tilde{\omega}_A(\bar{U}, \bar{V}) = \frac{1}{2!}[\tilde{\omega}(\bar{U}, \bar{V}) - \tilde{\omega}(\bar{V}, \bar{U})] , \quad (\tilde{\omega}_A)_A = \tilde{\omega}_A$$

$$\tilde{\omega}_S(\bar{U}, \bar{V}) = \frac{1}{2!}[\tilde{\omega}(\bar{U}, \bar{V}) + \tilde{\omega}(\bar{V}, \bar{U})] , \quad (\tilde{\omega}_S)_S = \tilde{\omega}_S \ .$$

Similarly, we use the following notations for antisymmetric (and symmetric) indices

$$\omega_{[ij]} = \frac{1}{2!}[\omega_{ij} - \omega_{ji}] = (\tilde{\omega}_A)_{ij} \quad , \quad \omega_{(ij)} = \frac{1}{2!}[\omega_{ij} + \omega_{ji}] = (\tilde{\omega}_S)_{ij} \ .$$

It is worth noting that the symmetry property of a tensor is coordinate independent. In other words, if the components of a tensor are symmetric (skew-symmetric) on one basis, then so are they on any basis.

It is well-known that any (0, 2) tensor (ie. matrix) can be decoupled as the sum of its symmetric part and antisymmetric part. However, this is not true for higher order tensors. One can also prove that all p-forms forms a linear subspace of the (0,p) tensor space, its dimension is

$$C^n_p = \frac{n!}{p!(n-p)!} \ .$$

Just as higher order tensors could be made from the outer product of lower order tensors, we can define a **wedge product** for constructing differential forms of higher degree, for example,

$$\tilde{p}\wedge\tilde{q} = \tilde{p}\otimes\tilde{q} - \tilde{q}\otimes\tilde{p} = (-1)^{pq}\tilde{q}\wedge\tilde{p} \ .$$

Any p-form can be decomposed on the wedge product of a set of basis 1-forms

$$\tilde{\alpha} = \frac{1}{p!}\alpha_{i\dots j}\tilde{\omega}^i\wedge\dots\wedge\tilde{\omega}^j = \frac{1}{p!}\alpha_{[i\dots j]}\tilde{\omega}^i\wedge\dots\wedge\tilde{\omega}^j \ .$$

Contraction of a vector with a p-form is defined as

$$\tilde{\alpha}(\bar{\xi}) \equiv \tilde{\alpha}(\bar{\xi},\dots) = \frac{1}{(p-1)!}\xi^i\alpha_{ij\dots k}\tilde{\omega}^j\wedge\dots\wedge\tilde{\omega}^k \ .$$

In studying differential and integral calculus on manifolds it is often convenient to introduce the completely antisymmetric **Levi-Civita symbols**

$$\varepsilon_{ij\dots k} = \varepsilon^{ij\dots k} = \begin{cases} +1 & \text{if}\{ij\dots K\}\text{ is an even permutation of}\{1, 2, \dots, n\} \\ -1 & \text{if}\{ij\dots K\}\text{ is an odd permutation of}\{1, 2, \dots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

One must note that they are not entries of any tensor. In fact they have the following meaning on any basis

$$\tilde{\varepsilon} \equiv \tilde{\omega}^1\wedge\tilde{\omega}^2\wedge\dots\wedge\tilde{\omega}^n = \frac{1}{n!}\varepsilon_{ij\dots k}\tilde{\omega}^i\wedge\tilde{\omega}^j\wedge\dots\wedge\tilde{\omega}^k \ .$$

Now we can define the integral of an n-form on n-dimensional manifolds. The dimension of the n-form space is exactly one, so in a coordinate system any n-form can be expressed as

$$\tilde{\omega} = \omega\tilde{\varepsilon} = \omega\tilde{d}x^1\wedge\tilde{d}x^2\wedge\dots\wedge\tilde{d}x^n$$

$$\tilde{d}x^1\wedge\tilde{d}x^2\wedge\dots\wedge\tilde{d}x^n(dx^1\partial_1, dx^2\partial_2, \dots, dx^n\partial_n) = dx^1dx^2\dots dx^n \ .$$

Thus a coordinate-free integral can be defined by

$$\int\tilde{\omega} \equiv \int\omega\tilde{\varepsilon} \equiv \int\omega dx^1dx^2\dots dx^n = \int\omega d^nx \ .$$

However, one must notice that both ω and the volume element d^nx are coordinate dependent. Moreover, no integral is defined for arbitrary tensors.

The exterior derivative is an operation inverse to integration, and is a very useful kind of antiderivation. The **exterior derivative** is defined to satisfy the following axioms

$\tilde{\alpha}$ is p-form, $\tilde{\beta}$ is q-form

(a). $\tilde{d}f$ = 1-form, $\tilde{d}\,\tilde{\alpha}$ = (p+1)-form

(b). $\tilde{d}(\tilde{\alpha}+\tilde{\beta})=\tilde{d}\,\tilde{\alpha}+\tilde{d}\,\tilde{\beta}$

(c). $\tilde{d}(\tilde{\alpha}\wedge\tilde{\beta})=(\tilde{d}\,\tilde{\alpha})\wedge\tilde{\beta}+(-1)^{p}\tilde{\alpha}\wedge\tilde{d}\,\tilde{\beta}$

(d). $\tilde{d}(\tilde{d}\,\tilde{\alpha})=0$.

From these one can show

$$\tilde{d}(f\,\tilde{d}g)=\tilde{d}f\wedge\tilde{d}g\ ;$$

$$\tilde{d}\,\tilde{\alpha}=\frac{1}{p!}\frac{\partial\alpha_{i\ldots j}}{\partial x^{k}}\tilde{d}x^{k}\wedge\tilde{d}x^{i}\wedge\ldots\wedge\tilde{d}x^{j}\quad(\text{with }\tilde{\alpha}\equiv\frac{1}{p!}\alpha_{i\ldots j}\tilde{d}x^{i}\wedge\ldots\wedge\tilde{d}x^{j})\ ;$$

$$\mathcal{L}_{\bar{V}}\tilde{\omega}=\tilde{d}[\tilde{\omega}(\bar{V})]+(\tilde{d}\,\tilde{\omega})(\bar{V})\ ,\quad\mathcal{L}_{\bar{V}}(\tilde{d}\,\tilde{\omega})=\tilde{d}(\mathcal{L}_{\bar{V}}\tilde{\omega})\ .$$

Moreover, on an n-dimensional manifold the Lie derivative of a volume n-form has some particular properties which are not shared by forms of different degree. For example

$$\left.\begin{array}{l}\mathcal{L}_{\bar{\xi}}\tilde{\omega}=\tilde{d}[\tilde{\omega}(\bar{\xi})]\\ \tilde{d}[\tilde{\omega}(f\bar{\xi})]=\tilde{d}[f\tilde{\omega}(\bar{\xi})]\end{array}\right\}\ \Rightarrow\ \mathcal{L}_{f\bar{\xi}}\tilde{\omega}=\mathcal{L}_{\bar{\xi}}(f\tilde{\omega})\,.$$

Based on the special properties of the volume n-forms a **divergence** can also be defined

$$\left.\begin{array}{l}\tilde{d}[\tilde{\varepsilon}(\bar{\xi})]=\xi^{i}{}_{,i}\tilde{\varepsilon}\\ (\mathrm{div}_{\tilde{\omega}}\bar{\xi})\,\tilde{\omega}\equiv\tilde{d}[\tilde{\omega}(\bar{\xi})]\end{array}\right\}\ \Rightarrow\left\{\begin{array}{l}\tilde{d}[\tilde{\omega}(\bar{\xi})]=(\omega\xi^{i})_{,i}\tilde{\varepsilon}=\frac{1}{\omega}(\omega\xi^{i})_{,i}\,\tilde{\omega}\\ \tilde{\omega}\text{-divergence of }\bar{\xi}:\mathrm{div}_{\tilde{\omega}}\bar{\xi}=\frac{1}{\omega}(\omega\xi^{i})_{,i}\end{array}\right.$$

A differential form is said to be **closed** if its exterior derivative vanishes. A p-form is **exact** if it is equal to the exterior derivative of a (p-1)-form. An exact form is always closed, but a closed form is exact only locally (this is called the Poincare lemma). The global question depends on the region being considered (see Schutz, 1980). (Nevertheless, on an 1-dimensional manifold a closed 1-form is always exact.)

Now let us take a look at another very useful derivation, the covariant derivative. Unlike the Lie derivative, the covariant derivative cannot be defined on the manifold we have been talking about so far. In addition to the differential structure we must introduce another structure onto manifolds, namely, the affine connection, which gives the manifold shape and curvature. By definition an **affine connection** (∇) is a rule for parallel transport of vectors along vector fields, viz.

$$\nabla_{\bar{U}}\bar{V}=o\quad\Leftrightarrow\quad\bar{V}\text{ is parallel-transported along }\bar{U}\ .$$

Using this parallel transport rule, we can easily define the covariant derivative of vector fields along a vector field (see Schutz, 1980), which is in principle not much different from the derivative of vectors in vector analysis.

We note that in the definitions of Lie and covariant derivatives, vector fields and 1-form fields are given another asymmetric position, since no similar derivative is defined along a 1-form field. This point is in fact a reflection of the previous fundamental difference of the two concepts, and shows that the analysis on manifold is really a generalisation of calculus (ie. differentiation is only defined with respect to a field that can always be related to coordinates).

The covariant derivative along a vector field ($\nabla_{d/d\lambda} \equiv \mathbb{D}$) satisfies the axioms of the derivation, in particular, the Leibniz rule, by which covariant derivatives can be defined for a tensor of any type. But the covariant derivative of a function is defined as

$$\nabla_{\bar{V}} f = \mathcal{L}_{\bar{V}} f = \bar{V}(f) \quad .$$

The covariant derivative does not only share with the Lie derivative their common properties belonging to the derivation, but has the following extra property,

$$\nabla_{f\bar{U}+g\bar{V}} \mathbf{T} = f\,\nabla_{\bar{U}} \mathbf{T} + g\nabla_{\bar{V}} \mathbf{T} \quad .$$

This property allows us to remove the vector field from the definition of covariant derivative, hence defining another tensor, the **gradient** of tensors. If **T** is an (N, N') tensor, then $\nabla\mathbf{T}$ is an (N, N'+1) tensor. This gradient operator (∇) is not a derivation, but one can show that it satisfies axioms (b), (c) and (d) of derivations, viz.

$$\nabla f = \tilde{d}f \,, \quad \nabla c = 0 \quad ,$$
$$\nabla[\tilde{\omega}(\bar{V})] = (\nabla\tilde{\omega})(\bar{V}| \quad) + (\nabla\bar{V})(\tilde{\omega}| \quad) \,,$$
$$\nabla(f\,\mathbf{A} + g\mathbf{B}) = f\,\nabla\mathbf{A} + g\nabla\mathbf{B} \,, \quad \nabla(\mathbf{A}\otimes\mathbf{B}) = (\nabla\mathbf{A})\otimes\mathbf{B} + \mathbf{A}\otimes(\nabla\mathbf{B}) \quad ,$$

where c is a constant number, f and g are arbitrary scalar functions. For the order of the arguments of a tensor, we have used the convention of putting the 1-forms before a semi-colon, vectors after it, and the vector along which derivatives is taken is put to the right side of a vertical bar.

However, properties (d') and (d") do not apply to this gradient operator, for by definition, tensors do not take higher order tensors as arguments. Although one might justify using the concept of contraction to develop (d') and (d") for the gradient, it would not be very useful. The value of these two properties is that they can develop component expressions for a geometric derivation, but for gradients this can always be obtained from the properties of the covariant derivative operator, $\nabla_{d/d\lambda}$.

Geodesic curves may be defined on manifolds with affine connections. A **geodesic** curve is a curve which parallel-transports its own tangent vector, viz.

$$\nabla_{\bar{V}}\bar{V} = 0 \;\; ; \;\; V^j V^i{}_{;j} = 0 \;\; ; \;\; \frac{d^2x^i}{d\lambda^2} + \Gamma^i_{jk}\frac{dx^j}{d\lambda}\frac{dx^k}{d\lambda} = 0.$$

The parameter of a geodesic curve is called affine parameter.

Since any tensor can be expressed as a linear combination of basis tensors, and these basis tensors are all derivable from the vector basis (not necessarily coordinate basis), the connection can be completely described by giving the gradients of the basis vectors. So we define

$$\nabla_{\bar{e}i} \equiv \nabla_i \,, \quad \nabla_i f \equiv \bar{e}_i(f) \equiv f_{,i} \,, \quad \nabla_i \bar{e}_j = \Gamma^k_{ji}\bar{e}_k \quad ,$$

where the order of the indices is important. The functions Γ's are called **Christoffel symbols**; they do not form the components of a tensor. Now the gradient of tensors can be written in component form

$$(\nabla\bar{V})^j_i \equiv V^j{}_{;i} = V^j{}_{,i} + \Gamma^j_{ki}V^k \,, \quad (\nabla\tilde{\omega})_{ij} \equiv \omega_{i;j} = \omega_{i,j} - \Gamma^k_{ij}\omega_k \,,$$

$$T^{i\ldots j}_{k\ldots l;t} \equiv (\nabla \mathbf{T})^{i\ldots j}_{k\ldots lt} = (\nabla \mathbf{T})(\tilde{\omega}^i, \ldots, \tilde{\omega}^j; \bar{e}_k, \ldots, \bar{e}_l \,|\, \bar{e}_t)$$

$$= (\nabla_t \mathbf{T})^{i\ldots j}_{k\ldots l} = (\nabla_t \mathbf{T})(\tilde{\omega}^i, \ldots, \tilde{\omega}^j; \bar{e}_k, \ldots, \bar{e}_l) \,,$$

$$(A^{\cdots}_{\cdots}B^{\cdots}_{\cdots})_{;t} \equiv [\nabla(\mathbf{A}\otimes\mathbf{B})]^{\cdots\,\cdots}_{\cdots\,\cdots;t} = [\nabla_{;t}(\mathbf{A}\otimes\mathbf{B})]^{\cdots\,\cdots}_{\cdots\,\cdots}$$

$$= A^{\cdots}_{\cdots;t}B^{\cdots}_{\cdots} + A^{\cdots}_{\cdots}B^{\cdots}_{\cdots;t} \;.$$

For an affine connection we can define a (1, 2) torsion tensor by

$$\mathbf{T}(\;\;;\bar{U},\bar{V}) = \nabla_{\bar{U}}\bar{V} - \nabla_{\bar{V}}\bar{U} - [\bar{U},\bar{V}] \;\; .$$

When the torsion tensor vanishes the connection is said to be symmetric, viz.

$$\nabla_{\bar{U}}\bar{V} - \nabla_{\bar{V}}\bar{U} = [\bar{U},\bar{V}] \;\; \Leftrightarrow \;\; \nabla \text{ is a symmetric connection}$$

$$\Leftrightarrow \Gamma^k_{ij} = \Gamma^k_{ji} \quad \text{in a coordinate basis.}$$

In the present approach only **symmetric connections** are considered.

When the connection is symmetric, then in any expression for the components of the

Lie derivative of a tensor, **all** commas can be replaced by semicolons (see Schutz, 1980, P208). In this statement the components are taken in a coordinate basis, because 'comma' and 'semicolon' can only be talked about in a coordinate system. One must also note that the traditionally termed covariant 'semicolon' relations are still coordinate dependent, although they are closer to tensorial relations than 'comma' ones.

One can prove that the operator (not necessarily symmetric connection)

$$\mathbf{R}(\bar{U}, \bar{V}) \equiv [\nabla_{\bar{U}}, \nabla_{\bar{V}}] - \nabla_{[\bar{U}, \bar{V}]}$$

is a (1, 1) tensor. Moreover, **R** can also be proved to be a (1, 3) tensor, called the **Riemann curvature tensor**. This is a very important tensor on Riemann manifolds and in general relativity. We use the standard convention for the order of the indices, namely,

$$\mathbf{R}(\tilde{\omega}; \bar{\xi} \mid \bar{U}, \bar{V}) = \mathbf{R}(\bar{U}, \bar{V})(\tilde{\omega}; \bar{\xi}) = \{[\nabla_{\bar{U}}, \nabla_{\bar{V}}]\,\bar{\xi} - \nabla_{[\bar{U}, \bar{V}]}\,\bar{\xi}\}(\tilde{\omega})$$

$$R^{i}_{\;jkl} \equiv \mathbf{R}(\tilde{\omega}^{i}; \bar{e}_{j} \mid \bar{e}_{k}, \bar{e}_{l}) = \{[\nabla_{k}, \nabla_{l}]\bar{e}_{j} - \nabla_{[k,l]}\,\bar{e}_{j}\}(\tilde{\omega}^{i}) \quad .$$

It is well known that the Euclidean space and Minkowski spacetime are flat manifolds. In the language here, a space is **flat** iff the Riemann tensor vanishes. A flat space has a global notion of parallelism: parallel transport of vectors becomes path independent. On a flat manifold, there exist coordinate systems in which all Christoffel symbols vanish everywhere; but it is possible to choose a coordinate system in which the Christoffel symbols do not vanish. However, general relativity uses a curved spacetime, on which there is no coordinate system in which the Christoffel symbols vanish. For the calculation of the Riemann tensor from the Christoffel symbols and other properties of this tensor, we refer to the standard references (eg. Schutz, 1980). Now let us turn to what is more important for general relativity, namely, a metric connection.

A **metric tensor** is a symmetric (0, 2) tensor, namely,

$$g(\bar{U}, \bar{V}) = g(\bar{V}, \bar{U}) = \langle \bar{U}, \bar{V} \rangle \quad .$$

From the matrix theory, we know that any linear space with a metric tensor has an orthonormal basis on which the metric tensor is diagonal with ±1 as entries. The trace of the canonical form is called the **signature** of the metric. Because the diagonal form of a continuous metric tensor field is the same everywhere on the manifold, the signature is a global constant. A positive definite metric is called a Riemannian metric. In particular, if it happens that on a basis $g_{ij}=\delta_{ij}$, then the metric is called an Euclidean metric, the basis a Cartesian basis. If the canonical form of an indefinite metric is a Lorentz metric, ie. (-1, 1, ..., 1), then the metric is called a Minkowski metric. The orthonormal basis of a Minkowski space is called a Lorentz basis.

As a tensor on manifolds, the metric tensor can induce many structures to the manifolds, which are stronger than any of those we have been discussing. For example, a metric tensor can induce a distance, and hence a topology; it can also induce a preferred volume form, a connection etc. Therefore on a manifold with various structures it is important that they must be compatible with one another. The strongest restrictions come from the compatibility with the metric tensor.

One can prove that the inverse matrix of a metric tensor (g) defines a symmetric (2, 0) tensor (g'). A metric tensor can map a 1-form to a vector, and vice versa, by

$$\tilde{V} = \mathbf{g}(\bar{V},\ ) = \mathbf{g}(\ ,\bar{V}), \quad \bar{V} = \mathbf{g}'(\tilde{V},\ ) = \mathbf{g}'(\ ,\tilde{V}) \ .$$

Similarly, the metric can map a tensor (including itself) by the so-called **index raising** and **lowering** to its associated tensor. It is conventional to write the metric in one of the following forms

$$ds^2 = g_{ij} dx^i dx^j, \quad \mathbf{g} = g_{ij} \tilde{d}x^i \otimes \tilde{d}x^j, \quad \mathbf{g}' = g^{ij} \frac{\partial}{\partial x^i} \otimes \frac{\partial}{\partial x^j} \ ,$$

where both g_{ij} and g^{ij} are symmetric matrices.

We have the following three main compatibility conditions between the metric, connection and volume n-form, namely,

$$\nabla \leftrightarrow \mathbf{g}: \ \begin{cases} \nabla[\mathbf{g}(\bar{A},\bar{B})] = 0 \ ; \quad \nabla \mathbf{g} = 0 \qquad (\Rightarrow \nabla \mathbf{g}' = 0) \\ \Gamma^i_{\ jk} = \frac{1}{2} g^{il}(g_{lj,k} + g_{lk,j} - g_{jk,l}) \quad \text{(in coordinate basis)} \end{cases}$$

$$\mathbf{g} \leftrightarrow \tilde{\omega}: \quad \tilde{\omega} = \sqrt{|g|}\ \tilde{\varepsilon} = \sqrt{|g|}\ \tilde{d}x^1 \wedge \tilde{d}x^2 \wedge \ldots \wedge \tilde{d}x^n \quad \text{(in coordinate basis)}$$

$$\tilde{\omega} \leftrightarrow \nabla: \ \begin{cases} \mathrm{div}_{\tilde{\omega}} \bar{\xi} = \frac{1}{\omega}(\omega \xi^i)_{,i} = \xi^i_{\ ;i} = (\nabla \bar{\xi})^i_{\ i} = (\nabla_i \bar{\xi})^i \\ \nabla \tilde{\omega} = 0 \ . \end{cases}$$

where g=det(g) (in general relativity | g | =-g). These compatibility conditions are a kind of equivalence relation, that is, if A is compatible with B, then so is B with A; if A is compatible with B, and B with C, then A is compatible with C.

The compatibility between the metric and volume n-form in fact defines a preferred volume n-form in an orthonormal basis, viz.

$$\mathbf{g} \leftrightarrow \tilde{\omega}: \quad \tilde{\omega} = \tilde{\omega}^1 \wedge \tilde{\omega}^2 \wedge \ldots \wedge \tilde{\omega}^n \quad \text{(in orthonormal basis)}$$

which in general relativity is the proper volume.

A manifold with structures satisfying the above compatibility conditions is what is important to general relativity, on which the Riemann tensor can be calculated by

$$R_{ijkl} = g_{im}R^m_{\ jkl} = \frac{1}{2}\,(g_{il,jk} - g_{ik,jl} + g_{jk,il} - g_{jl,ik})$$

in **normal** coordinates (ie. all coordinate curves are geodesic curves).

General relativity assumes a 4-dimensional spacetime manifold, which is locally flat and the metric tensor is locally equivalent to a Lorentz metric. But due to the existence of gravitation, the Riemann tensor is not zero and the spacetime is curved. The relation of matter distribution and geometry is governed by Einstein's field equation (the unit is such that c = G = 1)

$$\mathbf{G} + \Lambda\mathbf{g} = 8\pi\mathbf{T}\ , \quad G^{\alpha\beta} + \Lambda g^{\alpha\beta} = 8\pi\, T^{\alpha\beta}$$

where Λ is the cosmological constant, usually taken as zero. **T**, **g** and **G** are the stress-energy tensor, metric tensor and the Einstein tensor respectively. The Einstein tensor is defined by

$$\text{Einstein tensor: } G^{\alpha\beta} \equiv R^{\alpha\beta} - \frac{1}{2}Rg^{\alpha\beta}$$

$$\text{Ricci tensor: } R^{\alpha\beta} \equiv R^{\alpha\beta\mu}_{\ \ \ \ \mu}, \quad \text{Ricci scalar: } R \equiv R^{\alpha}_{\alpha}\ .$$

The detailed properties of these tensors are given in many textbooks (eg. Schutz, 1980).

However, it is worthwhile mentioning the differential conservation laws admitted by the above field equations. Due to the local flatness of the spacetime of general relativity, we have

$$G^{\alpha\beta}_{\ \ ;\beta} \equiv 0 \quad , \quad T^{\alpha\beta}_{\ \ ;\beta} \equiv 0\ .$$

6.2 Killing Vectors and Conservation Laws along Geodesics

In this section we study the relationship between symmetries that can be represented by Killing vectors and conservation laws for geodesic motion. Because of the complication in notation there is a limitation to the use of coordinate-free expressions; sometimes a component expression is shown to advantage especially when we are constructing complexes (tensors) from the contraction of several tensors. However, tensorial expressions are used as long as possible because of their conceptual clarity.

For example, we have the following useful relation on a metric manifold

$$\left.\begin{array}{l}\nabla_{\bar{A}}[\mathbf{g}(\bar{\xi},\bar{B})]=\mathbf{g}(\nabla_{\bar{A}}\bar{\xi},\bar{B})+\mathbf{g}(\bar{\xi},\nabla_{\bar{A}}\bar{B})\\ \nabla_{\bar{A}}\langle\tilde{\xi},\bar{B}\rangle=(\nabla_{\bar{A}}\tilde{\xi})(\bar{B})+(\tilde{\xi})(\nabla_{\bar{A}}\bar{B})\end{array}\right\}$$

$$\Rightarrow\ \mathbf{g}(\nabla_{\bar{A}}\bar{\xi},\ \ )\equiv\nabla_{\bar{A}}\tilde{\xi},\quad\text{where }\tilde{\xi}=\mathbf{g}(\bar{\xi},\ \ ).\qquad(6.1)$$

It is worth mentioning that there is not a similar relation for the Lie derivative, because the Lie derivative of g along an arbitrary vector field is not necessarily zero.

It is very useful to note two points here. The first point is a simple observation on the meaning of Lie derivative: the Lie derivative of a tensor is just the coordinate-free version of the partial derivatives of the components of tensor. This is useful in rewriting equations involving partial derivatives with respect to coordinates into a tensorial form involving the Lie derivative and covariant derivative. The second point is a property of the derivative of scalars. The Lie derivative and covariant derivative of a scalar with respect to a vector field are the same, and they are equal to the derivative of the scalar with respect to the parameter of the vector field, i.e. the scalar obtained by operating the vector on the scalar. This point is very useful as it enables us to convert between different kinds of derivatives and use their special properties to the best effect. In this section a particularly important scalar is the contraction of the metric tensor with vectors and/or 1-forms. The above property of this scalar is used to prove many relations.

Using the above techniques we can translate many relations, usually given in component form, into tensorial form. This is not just a trivial exercise; we will see that in this way many results can be simplified and interpreted in a proper way. For example, using equation (6.1) and the above two properties we obtain the following results for metric spaces (see Appendix D for a geometrical proof)

$$(\mathcal{L}_{\bar{\xi}}\mathbf{g})(\bar{A},\bar{B})=(\nabla_{\bar{A}}\tilde{\xi})(\bar{B})+(\nabla_{\bar{B}}\tilde{\xi})(\bar{A})=2(\nabla\tilde{\xi})_s(\bar{B}\,|\,\bar{A})\qquad(6.2a)$$

$$\begin{aligned}(\mathcal{L}_{\bar{\xi}}\mathbf{g})(\bar{A},\bar{B})&=(\nabla_{\bar{A}}\bar{\xi})(\tilde{B})+(\nabla_{\bar{B}}\bar{\xi})(\tilde{A})\\&=(\nabla\bar{\xi})(\tilde{B}\,|\,\bar{A})+(\nabla\bar{\xi})(\tilde{A}\,|\,\bar{B})\end{aligned}\qquad(6.2b)$$

$$\mathcal{L}_{\bar{P}}[\mathbf{g}(\bar{P},\bar{\xi})]=\nabla_{\bar{P}}[\mathbf{g}(\bar{P},\bar{\xi})]=\frac{1}{2}(\mathcal{L}_{\bar{\xi}}\,\mathbf{g})(\bar{P},\bar{P})+\mathbf{g}(\nabla_{\bar{P}}\bar{P},\bar{\xi})\qquad(6.3a)$$

$$\mathcal{L}_{\bar{P}}[\tilde{P}(\bar{\xi})]=\mathcal{L}_{\bar{P}}[\bar{P}(\tilde{\xi})]=\frac{1}{2}(\mathcal{L}_{\bar{\xi}}\,\mathbf{g})(\bar{P},\bar{P})+\mathbf{g}(\nabla_{\bar{P}}\bar{P},\bar{\xi})\qquad(6.3b)$$

where

$$\tilde{P} \equiv g(\bar{P},\ ), \quad \tilde{\xi} \equiv g(\bar{\xi},\ )$$

$$\mathcal{L}_{\bar{P}}[g(\bar{P},\bar{\xi})] \equiv \bar{P}[g(\bar{P},\bar{\xi})] \equiv \nabla_{\bar{P}}[g(\bar{P},\bar{\xi})] \ .$$

Equations (6.3a) and (6.3b) are very general relations; they may be simplified when the derivatives are taken along a geodesic curve, viz.

If $\nabla_{\bar{P}}\bar{P} = 0$, then

$$\mathcal{L}_{\bar{P}}[g(\bar{P},\bar{\xi})] = \nabla_{\bar{P}}[g(\bar{P},\bar{\xi})] = \frac{1}{2}(\mathcal{L}_{\bar{\xi}}\, g)(\bar{P},\bar{P}) \tag{6.4a}$$

$$\mathcal{L}_{\bar{P}}[\tilde{P}(\bar{\xi})] = \mathcal{L}_{\bar{P}}[\ \bar{P}(\tilde{\xi})] = \frac{1}{2}(\mathcal{L}_{\bar{\xi}}\, g)(\bar{P},\bar{P}) \ . \tag{6.4b}$$

Killing Vectors and Conservation Laws along Geodesics

It is very important to define Killing vector fields on metric manifolds, since a Killing field defines an isometry of the metric, thus a symmetry of the manifold. Combining this with Noether's theorem, we see that conservation laws can be related with Killing fields.

From the previous equations (6.1 6.2 6.3 6.4) it follows immediately that the Killi

$$\mathcal{L}_{\bar{\xi}}\, g \equiv 0 \Leftrightarrow \nabla\xi \text{ is a 2-form}$$

$$\Leftrightarrow \ \xi_{i;j} = \xi_{[i;j]} \ \Leftrightarrow \ \xi_{(i;j)} = 0 \ \Leftrightarrow \ \xi_{i;j} + \xi_{j\ ;i} = 0$$

$$\Leftrightarrow \Leftrightarrow \ \mathcal{L}_{\bar{\xi}}[g(\bar{A},\bar{B})] = 0 \ \text{ for } \{\text{all } \bar{A},\bar{B}: \mathcal{L}_{\bar{\xi}}\bar{A} = \mathcal{L}_{\bar{\xi}}\bar{B} = 0\}$$

$$\Leftrightarrow \Leftrightarrow \ g_{ij\,,\,k} = 0 \ \ (\text{but } g^{ij}{}_{,\,k} \neq 0) \text{ in a coordinate system with } \bar{K} = \partial_k \ .$$

$$\Leftrightarrow \ g_{ij\,,\,k} = 0 \ \ (\text{but } g^{ij}{}_{,\,k} \neq 0) \text{ in a coordinate system with } \bar{K} = \partial_k \ .$$

Geodesic motion in a prescribed geometry is a very important subject in general relativity. There are quantities conserved along geodesics if the metric admits some Killing vector fields. To see this let us consider a geodesic curve with parameter τ, defined by

$$V \equiv d/d\tau\ , \quad V^{\beta}{}_{;\alpha}V^{\alpha} = 0\ ; \quad \bar{P} = m\bar{V}\ , \quad P^{\beta}{}_{;\alpha}P^{\alpha} = P_{\beta;\alpha}P^{\alpha} = 0$$

where m is a positive constant number. From the above geodesic equation, one can show

$$\frac{dP_{\beta}}{d\tau} = \frac{1}{2m}\, g_{\mu\alpha,\beta}\, P^{\alpha}P^{\mu} \tag{6.5a}$$

Thus P_{β} is a conserved quantity along the geodesics if the β-coordinate curve is a Killing vector. The proof of equation (6.5a) is usually given in a coordinate system, viz.

$$P_{\beta;\alpha}P^{\alpha} = P_{\beta,\alpha}P^{\alpha} - \Gamma^{\gamma}_{\beta\alpha}P^{\alpha}P_{\gamma} = 0\ ,\quad P_{\beta,\alpha}P^{\alpha} = \bar{P}(P_{\beta}) = m\frac{dP_{\beta}}{d\tau}$$

$$\begin{aligned} m\frac{dP_{\beta}}{d\tau} &= \Gamma^{\gamma}_{\beta\alpha}P^{\alpha}P_{\gamma} = \frac{1}{2}\,g^{\gamma\mu}\,(g_{\mu\beta,\alpha} + g_{\mu\alpha,\beta} - g_{\alpha\beta,\mu})\,P^{\alpha}P_{\gamma} \\ &= \frac{1}{2}\,(g_{\mu\beta,\alpha} + g_{\mu\alpha,\beta} - g_{\alpha\beta,\mu})\,P^{\alpha}P^{\mu} \\ &= \frac{1}{2}\,g_{\mu\alpha,\beta}P^{\alpha}P^{\mu}\ . \end{aligned}$$

However, here we favour a completely geometric equation. Such a coordinate-free equation can be constructed from equation (6.5a) by using the remarks on the properties of Lie and covariant derivatives which is made at the beginning of the section. The key idea is to regard the β-coordinate curve as that of a coordinate-free vector, so equation (6.5a) may be directly rewritten as

$$m\frac{d}{d\tau}[\,g(\bar{P},\bar{K})\,] = \bar{P}[\,g(\bar{P},\bar{K})\,] = \mathcal{L}_{\bar{P}}[\,g(\bar{P},\bar{K})\,] = \frac{1}{2}\,(\mathcal{L}_{\bar{K}}\,g\,)(\bar{P},\bar{P}) \qquad (6.5b)$$

Therefore if the K-field is a Killing vector field, then one obtains a conserved quantity. In fact equation (6.5b) is exactly equation (6.4a), which has already been proved in a geometrical way. Now let us discuss a simple question to show the advantage of geometric equations over component ones.

In general relativity, a timelike Killing vector is of particular importance; because a space-time geometry is called **stationary** if the metric admits such a field. In this case one can always choose a coordinate system such that the metric components are independent of the t-coordinate. It is worthwhile mentioning that such a choice is not unique. A special case of a stationary space-time is one for which the timelike Killing vector is normal to a family of spacelike hypersurfaces; such a spacetime is called **static**. In a static spacetime there exists a coordinate system, which is adapted to the timelike Killing vector field, in which both of the conditions, $\partial g_{\mu\nu}/\partial x^0=0$ and $g_{0k}=0$, are satisfied.

Vector or 1-Form?

The following question arises if we only consider the component relation, (6.5a). The conserved quantities are determined by the 'lowered' components of the momentum vector, rather than 'raised' components. Since the component proof shows that this particular preference is related to the skew-symmetric feature of the Christoffel symbols as functions of the metric components and their first derivatives, one may conclude that the conservation of 'lowered' rather than 'raised' components is due to the compatibility between connection and metric.

This is partly correct. With the above interpretation, one could come to the wrong conclusion that 1-forms are more important than (their associated) vectors in relation to isometry and conservation laws. This confusion is easily clarified by looking at the coordinate-free equation, (6.5b), and the answer turns out to be more profound. Note that equation (6.5b) is true no matter whether the K-field is a Killing vector or not. The above question of how to interpret the importance of vectors and 1-forms according to equation (6.5a) is only related to the right side quantity of equation (6.5b), namely the interpretation of

$$\mathbf{g}(\bar{P},\bar{K}) = \langle \tilde{P}, \bar{K} \rangle = \tilde{P}(\bar{K}) = \langle \bar{P}, \tilde{K} \rangle = \bar{P}(\tilde{K}) \quad . \tag{6.6}$$

It is seen that this quantity is symmetric in the P-field and K-field. Whether this term is interpreted as a 'lower' or 'raised' component depends on whether we can find a coordinate system such that the Killing vector is a coordinate basis vector field, or that its associated 1-form is a coordinate basis 1-form (one of the complete set of dual bases of the coordinate vector fields). It may happen that both are possible in a specific problem. However, the general difference between vectors and 1-forms on a manifold, as was mentioned before, is manifest here.

On a manifold, any non-singular vector field can be a coordinate basis, while only an exact 1-form can be a coordinate 1-form. Therefore not all 1-forms can be chosen as coordinate forms.

If we look at the local property of any smooth manifold, then although a closed p-form is exact as a result of the Poincare lemma, not all p-forms are closed; whereas a general class of vectors satisfies the above coordinate condition.

If we look at the global question, then it may happen that no vector field satisfies the coordinate condition because of the nontrivial structure group of the manifold. For example, on a 2n-dimensional sphere or a Mobius band there is no continuous, singularity-free vector field. So it is not always possible to choose a vector field as a global coordinate basis field. Nonetheless, in this case, a closed p-form is not necessarily exact (see Schutz, 1980) even if the manifold considered has a trivial structure group (eg. a region of R^n which is not simply connected); thus providing even less choice of coordinate 1-forms.

In the particular question considered here, the associated 1-form of a Killing vector field is not necessarily exact, so it is not always possible to interpret the above term as a 'raised' component.

Thus the answer to the question of conserved quantities relates to the very fundamental property of vectors and 1-forms; it is an outcome of the differential structure of manifolds, but not that of the metric connection. However, this difference only occurs on a manifold; linear vector spaces are not rich enough to produce this difference.

6.3 Gauss' Theorem and Integral Conservation Laws

In last section we discussed conserved quantities along geodesics. In this section we will discuss the more general form of conservation laws admitted by general relativity, which are usually given as differential equations, namely, $T^{\mu\nu}{}_{;\nu}=0$. It is important to obtain their integral counterparts. To do this Stokes' theorem and Gauss' theorem must be invoked. Thus in this section we will discuss the theorems in the language of differential manifold, and then apply them to the study of conservation laws. In doing this, particular attention is paid to those important in general relativity.

Gauss' Theorem and Stokes' theorem

Let ∂U be a smooth orientable boundary of an n-dimensional region U on a manifold, then **Stokes' theorem** may be written

$$\tilde{\alpha} \equiv (n\text{-}1)\text{-form} \quad , \quad \tilde{\omega} \equiv n\text{-form}$$

$$\int_U \tilde{d}\tilde{\alpha} = \oint_{\partial U} \tilde{\alpha} \quad , \quad \int_U \tilde{d}[\tilde{\omega}(\bar{\xi})] = \int_U (\mathrm{div}_{\omega}\bar{\xi})\,\tilde{\omega} = \oint_{\partial U} \tilde{\omega}(\bar{\xi}) \ .$$

Gauss' theorem (or Green's) may be obtained from the second expression of Stokes' theorem by decoupling the volume n-form into the wedge product of an (n-1)-form and a 1-form normal to ∂U, viz.

$$\tilde{\eta} \equiv 1\text{-form normal to } \partial U, \quad \tilde{\beta} \equiv (n\text{-}1)\text{-form}$$

$$\tilde{\omega} = \tilde{\eta} \wedge \tilde{\beta}, \quad \tilde{\omega}(\bar{\xi}) = \tilde{\eta}(\bar{\xi})\,\tilde{\beta}\Big|_{\partial U} \quad \Rightarrow \quad \int_U (\mathrm{div}_{\omega}\bar{\xi})\,\tilde{\omega} = \oint_{\partial U} \tilde{\eta}(\bar{\xi})\,\tilde{\beta}\Big|_{\partial U} \ .$$

A complete proof of these two theorems may be found in Schutz (1980).

In Gauss' theorem, we have used the concept of an 1-form normal to a hypersurface, by which we mean that the associated vector of the 1-form is normal to the hypersurface. For example the 1-form

$$\tilde{d}f = \partial f/\partial x^i\, \tilde{d}x^i$$

is normal to the hypersurface f(x)=const.

Gauss' theorem is often written in a special coordinate system $\{x^i\}$ (i=1, 2, ..., n) of

U such that $\{x^i\}$ (i=2, ..., n) mesh to form the coordinates of ∂U, thus

$\tilde{d}x^1$ is normal to ∂U.

In this coordinate system, the decomposition of the volume n-form in the previous expression of Gauss' theorem may be carried out explicitly

$\tilde{\eta}$ is determined up to a function F, $\tilde{\eta} = F\tilde{d}x^1$, where $F = \tilde{\eta}(\partial_1)$;

$$\text{and } \tilde{\beta} = \frac{\tilde{\omega}(\partial_1, \dots, \partial_n)}{\tilde{\eta}(\partial_1)} \tilde{d}x^2 \wedge \dots \wedge \tilde{d}x^n .$$

If we choose $\tilde{\eta}(\partial_1) = 1$, and $\tilde{\varepsilon}^* \equiv \tilde{d}x^2 \wedge \dots \wedge \tilde{d}x^n$, then

$$\int_U (\mathrm{div}_{\tilde{\omega}}\, \bar{\xi})\, \omega\tilde{\varepsilon} = \oint_{\partial U} \tilde{\eta}(\bar{\xi}) \frac{\tilde{\omega}(\partial_1, \dots, \partial_n)}{\tilde{\eta}(\partial_1)}\, \tilde{\varepsilon}^* = \oint_{\partial U} \tilde{\eta}(\bar{\xi})\, \omega\tilde{\varepsilon}^* ,$$

$$\int_U \xi^i_{\ ;i}\, \omega\tilde{\varepsilon} = \int_U (\omega\,\xi^i)_{,i}\, \tilde{\varepsilon} = \oint_{\partial U} \tilde{\eta}(\bar{\xi})\, \omega\,\tilde{\varepsilon}^* \equiv \oint_{\partial U} \tilde{\eta}(\bar{\xi})\, d\Sigma \equiv \oint_{\partial U} \bar{\xi}(\tilde{d}\Sigma) .$$

As we see, on an n-dimensional manifold an integral is only defined for an n-form, and there is no integral defined for a general (N, N') tensor. To generalise the integral of a tensor met in calculus onto a manifold, one has to define how the tensors at every point are to be transported to a single point. One sees that such an integral would depend on the path of the transportation and the point to which the tensors are transported.

It is well-known that Stokes' and Gauss' theorems are very important in transferring differential conservation laws to integral ones. On a manifold, however, if one pursues a tensorial expression, then in general there is no integral conservation laws corresponding to differential ones involving 'semicolon' derivatives, except for those of a vector. Nevertheless, we can still work for it in a coordinate system by constructing a vector for a tensor, namely, contracting the tensor with some basis vectors and 1-forms.

$$\mathbf{T} \equiv (N, N') \text{ tensor} , \quad V^i \equiv T^{(\dots)i(\dots)}_{(\dots\dots\dots)} ; \quad T^{\dots i \dots}_{\dots\dots ;i} = 0 \Rightarrow V^i_{\ ;i} = 0 ,$$

where (...) indicates that the indices are not to be regarded as tensor indices as usual, but as contraction of the tensor with the indicated basis vectors and 1-forms. Then we can apply Gauss' theorem to the vector components $\{V^i\}$ constructed in the working coordinates,

$$\int_U V^i_{\ ;i}\,\omega\tilde{\varepsilon} = \int_U (\omega V^i)_{,i}\,\tilde{\varepsilon} = \int_U (\omega T^{\dots i\dots}_{\dots\dots\dots})_{,i}\,\tilde{\varepsilon}$$

$$= \int_U \frac{1}{\omega}(\omega T^{\dots i\dots}_{\dots\dots\dots})_{,i}\,\omega\tilde{\varepsilon} = \int_U (T^{\dots i\dots}_{\dots\dots\dots} + \Gamma^i_{ki} T^{\dots k\dots}_{\dots\dots\dots})\,\omega\tilde{\varepsilon}$$

$$= \oint_{\partial U} T^{\dots i\dots}_{\dots\dots\dots}\eta_i\,\omega\tilde{\varepsilon}^{*} = \oint_{\partial U} T^{\dots i\dots}_{\dots\dots\dots}\eta_i\,d\Sigma = \oint_{\partial U} T^{\dots i\dots}_{\dots\dots\dots}\,d\Sigma_i \quad .$$

Note that this set of equations are coordinate-dependent; they are true in any coordinate system, but their values differ in different coordinate system. Moreover, all these equations the integrals must be estimated in the same coordinate system.

We also point out that the above 'Gauss' theorem' for tensors is not in agreement with standard textbooks in form because of the existence of the ω-factor and the Γ-term. However, this does not bring ours into contradiction with theirs. The ω-factor and Γ-term have been retained in order that the theorem for the constructed vector is a coordinate-free equation. But since this is impossible for tensors, the above 'Gauss' theorem' for tensors is in fact coordinate dependent. Therefore, there is no need to keep the ω-factor. Starting with the volume n-form (ω=1)

$$\tilde{\varepsilon} = \tilde{d}x^1 \wedge \dots \wedge \tilde{d}x^n$$

one can easily obtain the standard coordinate dependent Gauss' theorem for tensors without the ω-factor, nor the Γ-term.

Asymptotic Symmetries and Integral Conservation Laws

As can be seen from the previous discussion, if the space time possesses no symmetry, then usually there is no tensorial integral conservation laws corresponding to the differential one $T^{\mu\nu}{}_{;\nu}=0$. However, there are always (scalar) integral conservation laws for $A^{\mu}{}_{;\mu}=0$. So whenever a Killing vector field is admitted by the metric, a (scalar) integral conservation law can always be constructed in the following way (see Hawking & Ellis, 1973)

If $\bar{K}$ is a Killing-vector, $\tilde{K} \equiv g(\bar{K},\ )$, construct $\bar{P} \equiv T(\tilde{K},\ )$, then

$$P^{\alpha}_{\ ;\alpha} = (T^{\mu\alpha}K_{\mu})_{;\alpha} = T^{\mu\alpha}{}_{;\alpha}K_{\mu} + T^{\mu\alpha}K_{\mu;\alpha}$$

$$P^{\alpha}_{\ ;\alpha} = 0 \quad \Leftarrow \quad T^{\mu\alpha}{}_{;\alpha} = 0\ ,\ \ T^{[\mu\alpha]} = 0\ ,\ \ K_{(\mu;\alpha)} = 0$$

$$P = \int P^0\sqrt{-g}\,d^3x = \int T^{0\mu}K_{\mu}\sqrt{-g}\,d^3x \ = \text{const} \qquad (6.7a)$$

If we choose a coordinate system such that

$$\bar{K} = \partial_\alpha, \text{ then } K_\mu = g_{\mu\alpha}$$

$$P = \int T^0_\alpha \sqrt{-g}\, d^3x = \text{const} \quad . \qquad (6.7b)$$

This formulation of conservation laws imposes too much restriction on the space-time, hence it loses its generality. However, following Landau & Lifshitz (1962), a more general formulation of integral conservation laws can be obtained, (although we should note that while the conservation laws are not tensorial in this formulation, they only depend on the asymptotic symmetries of the spacetime). The main feature of this formulation is to rewrite the Einstein field equations such that the 'semi-colon' in the differential conservation laws is replaced by a 'comma'. The standard Landau-Lifshitz formulation is usually expressed in any asymptotic Minkowski (Lorentz, inertial, universal rest) coordinate system (Misner et al, 1973). Here we give a more general formulation in any coordinate system so long as the time-coordinate is time-like everywhere and the spacetime is asymptotically flat. Although this is only a formal generalisation, it raises some questions on the importance of these coordinate dependent integral conservation laws.

Note that the crucial step of constructing the above integral conservation laws is to construct a vector from **T** and the associated 1-form of a Killing vector. The difficulty lies in the occurrence of 'semi-colon' instead of 'comma'. Let us observe the above procedure the other way round. We see that if the semi-colon is replaced by a comma, then the only obstacles are that we need a 'comma' formed conservation law of some symmetric complex $\mathcal{T}^{\mu\alpha}{}_{,\alpha}=0$ and a **Killing pseudo-vector** field $\mathcal{K}^\alpha$. If $\mathcal{K}_\mu$ is the associated pseudo-form of a Killing pseudo-vector, defined by $\mathcal{K}_{(\mu,\alpha)}=0$, then we can construct

$$\mathcal{P}^\alpha \equiv \mathcal{T}^{\mu\alpha}\mathcal{K}_\mu \ ,$$

$$\mathcal{P}^\alpha{}_{,\alpha} = (\mathcal{T}^{\mu\alpha}\mathcal{K}_\mu)_{,\alpha} = \mathcal{T}^{\mu\alpha}{}_{,\alpha}\mathcal{K}_\mu + \mathcal{T}^{\mu\alpha}\,\mathcal{K}_{\mu,\alpha}$$

$$\mathcal{P}^\alpha{}_{,\alpha} = 0 \quad \Leftarrow \quad \mathcal{T}^{\mu\alpha}{}_{,\alpha} = 0\ , \quad \mathcal{T}^{[\mu\alpha]} = 0\ , \quad \mathcal{K}_{(\mu,\alpha)} = 0$$

$$\mathcal{P} = \int \mathcal{P}^0\, d^3x = \int \mathcal{T}^{0\mu}\,\mathcal{K}_\mu d^3x \ = \text{const} \qquad (6.8)$$

where the factor $(-g)^{1/2}$ has been dropped, since this is not essential when the integrals do not define tensorial quantities (scalar here). In contrast to equation (6.7b), here we cannot always choose a coordinate system such that the pseudo-vector is a coordinate basis field,

$\{\mathcal{K}^{\mu}\}=\partial_{\alpha}$; because the pseudo-vector is determined by some coordinate conditions in a coordinate system.

Now let us observe the meaning of the above two required equations. We know that in a flat spacetime there is a global Minkowski (Cartesian) coordinate system in which the Christoffel symbols vanish, thus 'semi-colon' is equivalent to 'comma'. Therefore in such Minkowski coordinate systems the above equations define real Killing vectors; in asymptotic Minkowski coordinate systems they define pseudo-vector fields which are Killing vectors in an asymptotic flat region. But it must be noted that these equations only define pseudo-vectors, because, in contrast to the 'semi-colon' equations, the above equations always admit the following solution (see Carmeli, 1982)

$$\mathcal{K}_{\mu,\alpha} + \mathcal{K}_{\alpha,\mu} = 0 \quad \Leftrightarrow \quad \mathcal{K}_{\mu} = A\varepsilon_{[\mu\alpha]}x^{\alpha} + B_{\mu} \qquad (6.9)$$

where A and the B's are constants. Thus in any coordinate system ten such independent (Killing) pseudo-forms can be found; however, we do not know in general how many independent (Killing) pseudo-vectors these forms correspond to. In fact the pseudo-vectors are not required for the construction of conserved quantities. If the 'comma' conservation laws can be formulated, ten independent conserved quantities can be obtained in any coordinate system. We also notice that even in the global Minkowski coordinate system of special relativity, polar or spherical coordinates can be used for the spatial part, and there are certainly integral conservation laws in these coordinates which are different from those defined in Cartesian coordinates. The meaning of these unfamiliar conservation laws and their relations with the conventional one needs to be investigated further.

Let us now consider the possibility of 'comma' conservation laws. Define a complex and an effective energy-momentum tensor following Landau-Lifshitz (see Misner et al, chapters 19 and 20) viz.

$$H^{\mu\alpha\nu\beta} \equiv (-g)(g^{\mu\nu}g^{\alpha\beta} - g^{\alpha\nu}g^{\mu\beta}) \ ;$$

$$H^{\mu\alpha\nu\beta} = H^{\nu\beta\mu\alpha} = H^{[\mu\alpha][\nu\beta]} \ , \quad H^{\mu[\alpha\nu\beta]} = 0 \ ;$$

$$H^{\mu\alpha\nu\beta}{}_{,\alpha\beta} \equiv 16\pi \, T^{\mu\nu}_{\mathrm{eff}} \equiv 16\pi \, (-g)(T^{\mu\nu} + t^{\mu\nu}) \ ,$$

$$0 \equiv H^{\mu\alpha\nu\beta}{}_{,\alpha\beta\nu} = T^{\mu\nu}_{\mathrm{eff}\;,\nu}$$

where the stress-energy pseudotensor $t^{\mu\nu}$ is defined by

$$16\pi \, (-g)\, t^{\mu\nu} \equiv H^{\mu\alpha\nu\beta}{}_{,\alpha\beta} - 16\pi(-g)\, T^{\mu\nu} = H^{\mu\alpha\nu\beta}{}_{,\alpha\beta} - 2 \, (-g)\, G^{\mu\nu} \ .$$

This formal generalisation is based on the observation that the differential conservation laws of the effective T is a sole result of the symmetry and skew-symmetry characters of the constructed complex H. Because the g's and G's are components of tensors, the above formulation is valid in any coordinate system: the required symmetry properties of the complex H are held, the t's are satisfactory functions of metric components and hence of the gravitation feature. In addition, the factor (-g) may be neglected without changing the essential result.

Therefore in any coordinate system there are ten independent $\mathcal{K}_\mu$ satisfying the the set of equations $\mathcal{K}_{(\mu,v)}=0$, and a 'comma' conservation law of the symmetric effective energy-momentum. From these ten independent conserved quantities are obtained in any coordinate system,

$$P^\mu = \int T^{\mu 0}_{eff}\, d^3x \quad ; \quad J^{\mu\nu} = \int (x^\mu T^{\nu 0}_{eff} - x^\nu T^{\mu 0}_{eff})\, d^3x . \qquad (6.10)$$

When we are working in asymptotic Minkowski coordinate systems the above Killing pseudo-vectors become the ten independent asymptotic Killing vectors, and the conservation laws become the conventional Landau-Lifshitz conservation laws.

Some comments on asymptotic features of spacetime (symmetries, flatness) are in order. It is pointed out in Misner et al (1973) that asymptotic flatness is required for any meaningful concept like mass and (angular) momentum. Although they stress that this is because of the measurement required, a careful observation of the above procedure shows that asymptotic flatness is not necessary for a formal definition of such quantities. In the defining procedures, asymptotic flatness is important only because this is a sufficient condition for the convergence of the above formal integrals.

Another point arises with the choice of stating the relationship between forbidden motion and (asymptotic) symmetries, although it is conventional to relate to conservation laws. It has been noted for a long time that there is a general correspondence between conservation laws and symmetries (Noether, 1918), and that the latter is more convenient to study. For example, in classical mechanics, it is much easier to obtain the maximal number of independent first integrals by using symmetries rather than conservation laws (Weinberg, 1972). It is not apparent that the (asymptotic) symmetries of the problem determine forbidden motion and hence ordered motion. But this relation may be more fundamental. In general relativity, we see that the asymptotic symmetries are definite, while the conserved quantities are not. Therefore, some of the above conserved quantities in a bad coordinate system may even, in principle, have no value at all. For definiteness we must, therefore, ask for the relationship between forbidden motion and (asymptotic) symmetries.

6.4 Integrable Motion and Relativistic Chaos

In the last section we have discussed the relations between apparent (asymptotic) symmetries of space-time and conservation laws in general relativity, and we know that sufficient conserved quantities can lead to integrability of a classical system, and relativistic system as well (eg. the classical 2-body problem and geodesic motion in Schwarzschild geometry). In this section we give a Hamiltonian formulation for the geodesic motion and show its equivalence to the geometrical formulation given in section 6.2. Because of this we can look at the integrability question using different mathematical languages. The idea of integrals related to no apparent symmetries and a possible way of producing chaos based on truncation of relativistic problems are introduced. This investigation has some importance in understanding the integrability conditions and the general approximation method often used in treating relativistic problems.

A geodesic motion on metric manifolds can also be formulated in a Hamiltonian form by introducing a super-Hamiltonian (Misner et al, 1973, P645; Chandrasekhar, 1983) in the (x^{μ}, P_{μ}) phase space, viz.

$$\mathcal{H} = \frac{1}{2m} g(\bar{P},\bar{P}) = \frac{1}{2m} g^{\mu\alpha}(x) P_{\mu} P_{\alpha} , \quad \bar{P} \equiv m\bar{V} \equiv m\frac{d}{d\tau}$$

$$\begin{cases} \dfrac{dx^{\mu}}{d\tau} = \dfrac{\partial \mathcal{H}}{\partial P_{\mu}} \\ \dfrac{dP_{\mu}}{d\tau} = -\dfrac{\partial \mathcal{H}}{\partial x^{\mu}} \end{cases} \Leftrightarrow \begin{cases} V^{\mu} = \dfrac{1}{m} g^{\mu\alpha} P_{\alpha} \\ \dfrac{dP_{\mu}}{d\tau} = -\dfrac{1}{2m} g^{\alpha\beta}{}_{,\mu} P_{\alpha} P_{\beta} \; (\Leftrightarrow \nabla_{\bar{P}} \bar{P} = 0) \end{cases}$$

To see that this set of equations of motion is equivalent to equations (6.4a), (6.5a) and the standard geodesic equation, we only need to show the following relation

$$g^{\alpha\beta}{}_{,\mu} P_{\alpha} P_{\beta} = - g_{\alpha\beta,\mu} P^{\alpha} P^{\beta},$$

which is just a particular case of the more general relation

$$g^{\alpha\beta}{}_{,\mu} A_{\alpha} B_{\beta} = - g_{\alpha\beta,\mu} A^{\alpha} B^{\beta} \quad \text{or} \quad (\mathcal{L}_{\bar{\xi}} g)(\bar{A}, \bar{B}) = -(\mathcal{L}_{\bar{\xi}} g')(\tilde{A}, \tilde{B}) \quad ,$$

where the three fields, **A**, **B** and ξ, are arbitrary. The tensorial equation follows immediately if one applies the Leibniz rule to the following four identical scalars

$$\mathcal{L}_{\bar{\xi}} < \tilde{A}, \bar{B} > \quad , \quad \mathcal{L}_{\bar{\xi}} < \bar{A}, \tilde{B} > \quad ; \quad \mathcal{L}_{\bar{\xi}} [g(\bar{A}, \bar{B})] \quad , \quad \mathcal{L}_{\bar{\xi}} [g'(\tilde{A}, \tilde{B})] \quad .$$

As a result of this Hamiltonian formulation, the conserved quantities along geodesics in a metric space (**g** not necessarily positive definite) can be regarded as integrals of a canonical Hamiltonian system. On the other hand, a large class of classical dynamical problems can also be formulated as geodesics of a metric space (cf. chapter 2). In the language of differential geometry, the function **g(P, K)** is an integral of motion, if **K** is a Killing vector of the metric tensor. If this **K** vector field is adapted to a coordinate system, say as the x^{μ}-coordinate, then this coordinate does not occur in the Hamiltonian $\mathcal{H}$ and the integral becomes P_{μ}, the conjugate momentum of x^{μ}.

When there is an additional potential field acting on the particle, then its motion is no longer geodesic. The equation of motion becomes equation (6.3) and the potential energy must be added to the above Hamiltonian. In this case, **K** being a Killing field does not guarantee **g(P, K)** being an integral.

Let us now show a relationship between the Poisson bracket (of two scalars) and the Lie bracket (of two vectors), which was used in chapter 2 to view Liouvillie's integrability conditions in this language. Using the same notations as above, let us consider the problem in the (x^{μ}, P_{μ}) phase space, then a direct calculation shows

$$\left\{ g(\bar{P},\bar{A}), g(\bar{P},\bar{B}) \right\} = -g(\bar{P},[\bar{A},\bar{B}]) .$$

where **A** and **B** are arbitrary vector fields. This result does not require the metric tensor to be positive definite. However, if the metric tensor is definite, then the vanishing of the Poisson bracket and Lie bracket are equivalent. Therefore, two integrals being in involution means that their Killing vectors commute.

As seen from its application in chapter 2, although the present results are not of general applicability yet (to deal with more complicated problems, the more advanced notions of symplectic manifold and Poisson manifold have to be invoked), they provide much clearer view to many questions in classical dynamics. Moreover, a use of the Lie algebra of the vectors (or the scalars) or a direct application of Frobenius' theorem on the vectors provides the conditions under which one can construct n integrals in involution from n integrals which are not.

So far we have only discussed conserved quantities corresponding to apparent symmetries. However, not all conserved quantities may be related to obvious symmetries, nor Killing vectors. A well-known classical example is the extra integral that leads to the integrability of the Toda lattice (see chapter 2); while the integrability of the geodesic motion in Kerr geometry is a very good example in general relativity.

It was first shown by Carter (1968) using the technique of separating the Hamilton-Jacobi equation that there is an additional independent integral in the second example. Later this was also successfully established using the technique of Killing tensors (Walker & Penrose, 1970) and Newman-Penrose tetrad formalism (see

Chandrasekhar, 1983). However, the simplest way of constructing this integral turns out to be the most elementary technique familiar in solving the classical 2-body problem. To do this, one may write the equation of motion of $dP_\theta/d\tau$ in the standard Boyer-Lindquist coordinates using the above Hamiltonian formulation, and then multiply both sides of the equation by P_θ. A simple calculation results in the additional integral involving $(P_\theta)^2$. Because of the existence of such a 'hidden' symmetry, geodesic motion in the Kerr geometry becomes completely integrable.

In addition to the linear integrals corresponding to Killing vectors, there may exit more independent higher order integrals related with a class of Killing tensors (see Walker & Penrose, 1970; Woodhouse, 1975; Dolan et al, 1989). From the discussion of section 6.2, a Killing vector defines an integral linear in the 4-momentum (cf. equation 6.2a)

$$(\nabla\tilde{K})_s \equiv 0 \qquad \Rightarrow \qquad \mathcal{L}_{\bar{P}}[\tilde{K}(\bar{P})] = \nabla_{\bar{P}}[\tilde{K}(\bar{P})] = 0 \quad .$$

Similarly, a Killing tensor **F** defines a higher order integral (cf. equation 6.3b)

$$(\nabla\mathbf{F})_s \equiv 0 \qquad \Rightarrow \qquad \mathcal{L}_{\bar{P}}[\mathbf{F}(\bar{P}, \ldots, \bar{P})] = \nabla_{\bar{P}}[\mathbf{F}(\bar{P}, \ldots, \bar{P})] = 0 \quad .$$

It is easy to verify that the metric tensor is a Killing tensor. Moreover, we have

$$\left\{\tilde{A}(\bar{P}),\, C(\bar{P}, \ldots, \bar{P}) \right\} = -(\mathcal{L}_{\bar{A}}\, C)(\bar{P}, \ldots, \bar{P}) \quad ,$$

where **A** and **C** are an arbitrary vector and tensor respectively. Although we have not been able to establish a similar relation for the Poisson bracket of two integrals both corresponding to Killing tensors, the above relations suffice to show that the four independent integrals of the Kerr geodesic motion are in involution.

It is worth noting that in Schwarzschild geometry the hidden symmetry still exists, however, it degenerates into a linear function of the Killing vectors.

The Newman-Penrose formalism using a null **tetrad** bases is also a very important way of obtaining integrals of geodesic motions. A tetrad formalism uses linearly independent vector fields, which do not necessarily commute with one another, as a basis; this can favour the inherent symmetries of the space-time. The value of choosing null vectors, instead of the conventional orthonormal vectors, as basis fields lies in the fact that the essential element of a space-time is its light-cone structure. For the application of this formalism to obtain the additional integral of geodesic motion in the Kerr geometry see Chandrasekhar (1983, P343).

Because of integrability, there is no chaotic geodesic motion in either Schwarzschild or Kerr geometry. However, it is interesting to study the approximation of these problems. Although this would not lead to chaos in the Schwarzschild spacetime, this may shed some light on the relation between complete relativistic problems and their approximations. This is important because the approximation method is usually unavoidable in studying relativistic questions, and a truncation of an infinite series may lead to a system which is topologically different (an expansion is usually made with respect to some *coordinates*, hence only Killing *vectors* are respected). Because the extra integral of geodesic motion in the Kerr spacetime are related to an independent irreducible Killing tensor, which cannot be expressed as a coordinate, a truncation in this case is very likely to produce nonintegrability and thus chaos, as is similar to the Toda lattice problem.

Another possible example which may have completely relativistic chaos is the geodesic motion of an uncharged mass in the gravitational field of two fixed black holes. In this problem the gravitational attraction of the two black holes is balanced by the electric repelling force. This is an exact solution to the Einstein-Maxwell equations, called the Majumdar-Papapetrou solution (Chandrasekhar, 1983, P591). It was shown by Chandrasekhar (1989) that in his coordinates the Hamilton-Jacobi equation is not separable for the meridian geodesics. He pointed out that the question is very unlikely to be separable in any coordinates. Later Contopoulos (1990) found by numerical integration that the trapped (stable) and escape (unstable) geodesic solutions depend sensitively on initial conditions. In fact the initial conditions of the two cases are mixed like a Cantor set (see chapter 2). This is a signal of chaos.

Moreover, the 2-body problem in general relativity has not been solved; and it is very unlikely to be integrable. It seems that the integrability of the post-Newtonian 2-body problem has not been studied either. One may infer from the situation of the classical N-body problem that the post-Newtonian N-body problems ($N \geq 3$) are very unlikely to be integrable. Thus all these systems are good candidates for the production of relativistic chaos which have not yet been investigated extensively.

Finally it will be useful to put together the simple problems, say the classical 2-centre problem, the geodesic motion in Schwarzschild and Kerr geometries, and investigate their integrability conditions in a unified way. It remains to study whether the first order post-Newtonian 2-body and more body problems and the geodesic motion in the field of two fixed black holes are integrable or not. Chaos will definitely occur if they are not integrable, and this is a very interesting question.

6.5 Bounded Geodesic Motion in Schwarzschild Geometry

The purpose of this and the following chapter is to establish relations between bounded motion and symmetries of spacetime, in this section we will study the simplest case of relativistic problems, namely, geodesic motion in Schwarzschild geometry. It is well-known that this problem is completely integrable, and simple bounded motion exists (Schutz, 1980; Chandrasekhar, 1983). However, we will generalise the classical inequality method discussed in chapter 4 and apply it to this problem. In this way we can make the study of bounded motion in this relativistic example parallel to the classical study. This is the first example showing the possibility of studying bounded motions in relativity using the inequality method.

The Inequality

For the study of this section, we need to generalise Sundman's inequality so that it will not only be valid in a 3-dimensional Euclidian space but also in higher dimensional linear spaces. It turns out that its validity is independent of the definition of an inner product; this is a very useful point in studying relativistic problems using classical methods. The inequality is

$$2\sum_{i,k}(mA^{[i}B^{k]})^2 \equiv \frac{1}{2}\sum_{i,k}(mA^iB^k-mA^kB^i)^2$$
$$\leq\left[\sum_i m(A^i)^2\right]\left[\sum_i m(B^i)^2\right] \qquad (6.11)$$

where $\{A^i\}$ and $\{B^i\}$ are two arbitrary vectors and m is a constant. In introducing these vectors we have chosen to work with the antisymmetric part of the geometric object chosen since this is closely related to vorticity and rotation. We will show in the examples to follow that these terms can indeed represent some measure of the angular momentum in a general relativistic system.

Standard Bounded Geodesic Motion in Schwarzschild Coordinates

Schwarzschild spacetime has a preferred coordinate system, in which the components of the metric tensor take the simplest form

$$ds^2=-\left(1-\frac{2M}{r}\right)dt^2+\left(1-\frac{2M}{r}\right)^{-1}dr^2+r^2(d\theta^2+\sin^2\theta d\varphi^2) \qquad (6.12a)$$

$$\mathbf{g}=-\left(1-\frac{2M}{r}\right)\tilde{d}t\otimes\tilde{d}t+\left(1-\frac{2M}{r}\right)^{-1}\tilde{d}r\otimes\tilde{d}r$$
$$+r^2(\tilde{d}\theta\otimes\tilde{d}\theta+\sin^2\theta\tilde{d}\varphi\otimes\tilde{d}\varphi) \qquad (6.12b)$$

$$g' = -\left(1-\frac{2M}{r}\right)^{-1}\frac{\partial}{\partial t}\otimes\frac{\partial}{\partial t}+\left(1-\frac{2M}{r}\right)\frac{\partial}{\partial r}\otimes\frac{\partial}{\partial r}$$
$$+\frac{1}{r^2}\left(\frac{\partial}{\partial\theta}\otimes\frac{\partial}{\partial\theta}+\frac{1}{\sin^2\theta}\frac{\partial}{\partial\varphi}\otimes\frac{\partial}{\partial\varphi}\right) \qquad (6.12c)$$

where M is the mass, t is the coordinate time, and (r, θ, φ) are the spherical coordinates. The following coordinates are often used as well

$$ds^2 = -\left(\frac{1-\frac{M}{2r^*}}{1+\frac{M}{2r^*}}\right)^2 dt^2+\left[1+\frac{M}{2r^*}\right]^4\left(dr^{*2}+r^{*2}d\theta^2+r^{*2}\sin^2\theta d\varphi^2\right) \qquad (6.13a)$$

$$ds^2 = -\left(\frac{1-\frac{M}{2r^*}}{1+\frac{M}{2r^*}}\right)^2 dt^2+\left[1+\frac{M}{2r^*}\right]^4\left(dx^2+dy^2+dz^2\right) \qquad (6.13b)$$

$$ds^2 = -\left(\frac{1-\frac{M}{2r^*}}{1+\frac{M}{2r^*}}\right)^2 dt^2+\left[1+\frac{M}{2r^*}\right]^4\left(d\rho^2+\rho^2d\varphi^2+dz^2\right) \qquad (6.13c)$$

where the spatial coordinates used in equation (6.13b) are called **isotropic** coordinates. The transformation between the coordinates (6.12a) and (6.13a) is

$$\left\{\begin{array}{l} (r \geq 2M) \\ t = t \\ r = r^*\left(1+\frac{M}{2r^*}\right)^2 \\ \theta = \theta \\ \varphi = \varphi \end{array}\right. \Rightarrow \left\{\begin{array}{l} 1-\frac{2M}{r} = \left(1-\frac{M}{2r^*}\right)^2 / \left(1+\frac{M}{2r^*}\right)^2 \\ \Lambda^r_{r^*} = \frac{\partial r}{\partial r^*} = \left(1-\frac{M}{2r^*}\right)\left(1+\frac{M}{2r^*}\right) \\ g_{r^*r^*} = \Lambda^r_{r^*}\Lambda^r_{r^*}g_{rr} = \left(1+\frac{M}{2r^*}\right)^4 \end{array}\right.$$

and the transformations between (6.13a), (6.13b) and (6.13c) are

$$\left\{\begin{array}{l} (6.13b) \leftrightarrow (6.13a) \\ t = t \\ x = r^*\sin\theta\cos\varphi \\ y = r^*\sin\theta\sin\varphi \\ z = r^*\cos\theta \\ (r^{*2} = x^2+y^2+z^2) \end{array}\right. \quad \left\{\begin{array}{l} (6.13c) \leftrightarrow (6.13a) \\ t = t \\ \rho = r^*\sin\theta \\ z = r^*\cos\theta \\ \varphi = \varphi \\ (r^{*2} = \rho^2+z^2) \end{array}\right. \quad \left\{\begin{array}{l} (6.13b) \leftrightarrow (6.13c) \\ t = t \\ x = \rho\cos\varphi \\ y = \rho\sin\varphi \\ z = z \\ (r^{*2} = \rho^2+z^2). \end{array}\right.$$

For the geodesic motion in the above spacetime, it is a standard result that the conservation of energy and angular momentum impose restrictions on the possible motions. There exist two standard approaches, one is geometric (eg. Schutz, 1980), the other uses a Hamiltonian formulation. However, the basic equations are the same, namely,

$\bar{P} \equiv m\bar{V} \equiv m\frac{d}{d\tau}$ is the 4-momentum, $\bar{K}$ is a Killing vector

$$g(\bar{P},\bar{P}) = -m^2 \qquad \Leftrightarrow \qquad g_{\mu\nu}P^{\mu}P^{\nu} = g^{\mu\nu}P_{\mu}P_{\nu} = -m^2 \qquad (6.14a)$$

$$\mathcal{L}_{\bar{P}}\{g(\bar{P},\bar{K})\} = \frac{1}{2}(\mathcal{L}_{\bar{K}}g)(\bar{P},\bar{P}) \qquad \Leftrightarrow \qquad m\frac{dP_{\alpha}}{d\tau} = \frac{1}{2}g_{\mu\nu,\alpha}P^{\mu}P^{\nu}. \qquad (6.14b)$$

However, a preferred coordinate system is needed to deal with these equations: one must put equations (6.14a, b) in a coordinate system which contains as many Killing vectors as possible as the basis vector fields, namely the Schwarzschild coordinates. It is also straightforward to verify that in the Newtonian limit, these equations reduce to the classical conservation laws and Sundman's inequality, equation (4.4b).

It follows from equation (6.14b) that the t-coordinate and φ-coordinate basis vector fields are two independent Killing vectors. Therefore equation (6.14b) shows that the geodesic motion possesses two conserved quantities $P_t = -H$ and $P\varphi = C$ (these quantities, energy H and angular momentum C, are defined in a coordinate system in which if θ=90^o and $d\theta/dt=0$ at one moment then θ is a constant for all time. see Papapetrou, 1974). Then the component form of equation (6.14a) gives the forbidden motion results, viz.

$$-\left(1-\frac{2M}{r}\right)^{-1}H^2 + \left(1-\frac{2M}{r}\right)P_r^2 + \frac{1}{r^2}(P_\theta^2 + C^2) = -m^2$$

$$\Rightarrow \quad C^2 \le r^2\left\{\left(1-\frac{2M}{r}\right)^{-1}H^2 - m^2\right\}. \qquad (6.14c)$$

This is the normal approach, but it may be observed that, in the same spacetime, the tensorial equations (6.14a, b) must determine the same forbidden region, independent of the coordinates used. We will now adopt an isotropic coordinate system, equation (6.13b), and reproduce the same results.

Bounded Geodesic Motion in Isotropic Coordinates

Before a successful method is found to put this study into a completely coordinate-free form, let us reproduce the results by using the inequality method reviewed in chapter 4 and generalised in Appendix B. Although this approach is still not tensorial - we prefer to work in a coordinate system with 'Cartesian' (orthonormal) spatial coordinates - it is more generally applicable because almost all standard results in general relativity are given in an asymptotic Minkowski coordinate system. The concepts defined by equations (6.14a, b) should remain unchanged but it is not immediately obvious, in this coordinate system, how to carry out the appropriate calculation. Use of equation (6.11) makes it apparent: equation (6.14b) supplies the conserved quantities upon which we write Sundman's inequality, while the normalisation equation (6.14a) plays the role of replacing the 'kinetic energy' by conserved 'total energy' and 'potential energy'.

Let us observe that working in coordinates (6.13a) does not make much difference from working in (6.12a): Killing vectors and conserved quantities are not changed under this coordinate transformation. However, in the coordinate system (6.13b), only the t-coordinate basis is a Killing vector, thus only the energy is explicitly conserved. Other Killing vectors, and correspondingly conserved quantities, must in general be studied by solving the Killing equation. However, since the Killing vectors can also be obtained by a coordinate transformation method, the difficulty here does not lie in finding all the Killing vectors. The problem is how to study the forbidden motion defined by equations (6.14a, b) in coordinate system (6.13b); that is, how to obtain equation (6.14c) from equations (6.14a, b) if the Killing vectors are known. This is by no means easy unless the generalised inequality (6.11) is used. In this way the study of forbidden motion in general relativity can be made parallel to that in the classical study.

To obtain the required Killing vectors we observe that the φ-coordinate basis vector field is a Killing vector. In the coordinate system of (6.13b) this vector field is found to be

$$\frac{d}{d\varphi} = \frac{\partial}{\partial\varphi} = \frac{\partial x^{\mu}}{\partial\varphi}\frac{\partial}{\partial x^{\mu}} = \frac{\partial x}{\partial\varphi}\frac{\partial}{\partial x} + \frac{\partial y}{\partial\varphi}\frac{\partial}{\partial y} = - y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$$

If the symmetry in the coordinates x, y, and z is noted, then by constructing coordinates like (6.13a) it is seen that the spacetime possesses the following three independent Killing vectors in addition to the t-coordinate basis field:

$$\bar{K} = \pm\,(x^{i}\partial_{k} - x^{k}\partial_{i})$$

Accordingly, the three corresponding conserved quantities (angular momenta) are

$$g(\bar{K},\bar{P}) = K^{\mu}P_{\mu} = K^{k}P_{k} + K^{i}P_{i} = x^{i}P_{k} - x^{k}P_{i} = \left(1 + \frac{M}{2r^{*}}\right)^{4}(x^{i}P^{k} - x^{k}P^{i}) \quad .$$

The generalised Sundman's inequality on these quantities reads

$$C^{2} \equiv \frac{1}{2}\sum_{j,k}(x^{j}P_{k} - x^{k}P_{j})^{2}$$

$$= 2\sum_{j,k}(x^{[j}P_{k]})^{2} \;\leq\; \left(\sum_{k}(x^{k})^{2}\right)\left(\sum_{k}(P_{k})^{2}\right) \qquad (6.15a)$$

where we have denoted the left side of equation (6.11) by C to provide a formal association with the angular momentum in the classical Sundman inequality. The first factor on the right side of (6.15a) is a function of the coordinates already. The normalisation equation (6.14a) is now used to decouple the 'kinetic' and 'potential' energy, hence the second factor,

$$-m^2 = g^{\mu\nu} P_\mu P_\nu = \frac{1}{g_{tt}} H^2 + \left(1 + \frac{M}{2r^*}\right)^{-4} \sum_k (P_k)^2$$

$$\Rightarrow \quad \sum_k (P_k)^2 = \left(1 + \frac{M}{2r^*}\right)^{4} \left(- \frac{1}{g_{tt}} H^2 - m^2\right) .$$

Finally we obtain Sundman's inequality in general relativity capable of studying forbidden motion, viz.

$$C^2 \le \left(1 + \frac{M}{2r^*}\right)^{4} (r^*)^2 \left(- \frac{1}{g_{tt}} H^2 - m^2\right)$$

$$= r^2 \left\{ \left(1 - \frac{2M}{r}\right)^{-1} H^2 - m^2 \right\} \qquad (6.15b)$$

which is exactly the same equation as (6.14c). Thus in an isotropic coordinate system we have obtained a result equivalent to that in the coordinate system (6.12a).

This study is formally similar to the classical approach: equation (6.14b) supplies the definition for the components of the angular momentum, while equation (6.14a) provides a decoupling between potential and kinetic energy. However, we will see in the next chapter that in general relativity a complete decoupling is in general impossible.

A further generalisation of the above approach into arbitrary coordinates is still desired. We also hope to apply our method to the study of bounded geodesic motion in an axially symmetric spacetime (Schutz, 1980; Chandrasekhar, 1983).

6.6 Summary

In this chapter we have discussed the general problem of formulating conservation laws on a manifold, in particular, the spacetime geometry of general relativity. It is shown that some of the conservation laws can be regarded as a result of obvious symmetries and asymptotic symmetries of the spacetime manifold. Such relations are established through the Killing vector fields or Gauss' theorem.

In the course of this study, we have clarified several points. Firstly, we asked the question: which is more important to conservation laws, a vector or 1-form? It is found that the answer to this seemingly artificial question is surprisingly fundamental; the answer has nothing to do with the metric connection of the manifold, but is a result of the most basic properties of vectors and 1-forms, and the differential structure of the manifold. The differential structure of a manifold introduces the differences of the two concepts which do not exist on linear vector space.

Secondly, we found that the coordinate dependent Gauss' theorem for a tensor can have expressions different from that used in standard textbooks. Since the equations are not tensorial in any case, the extra term and factor may be switched on/off in the equations, thus providing more flexibility. It is also useful to note a technical relation used in our formulation of Gauss' theorem, that is,

$$\frac{1}{\sqrt{-g}}\left(\sqrt{-g}\, T^{\cdots\mu\cdots}_{\cdots\cdots}\right)_{,\mu} = T^{\cdots\mu\cdots}_{\cdots\cdots,\mu} + \Gamma^{\mu}_{\beta\mu} T^{\cdots\beta\cdots}_{\cdots\cdots} \neq T^{\cdots\mu\cdots}_{\cdots\cdots;\mu} \qquad (6.16)$$

is always true for any tensor **T** in general relativity. This is solely a result of the compatibility conditions between the metric, volume form and connection. A direct proof of this equation would be very lengthy. The integral conservation laws related to Killing vectors may also be obtained using this equation (see Landau & Lifshitz, 1962, P341).

Thirdly, we have found another class of coordinate dependent conservation laws, as a generalisation of the Landau-Lifshitz formulation. It is a standard result that the original Landau-Lifshitz formulation is related to asymptotic symmetries of the spacetime. However, it is not clear how our generalised conservation laws are related to symmetries. Our results apply to any coordinates, and in general lead to different conserved quantities in different coordinates. This result raises some questions concerning the relationship between symmetries and conservation laws. Because of this, in the discussion of the next chapter we will ask for the relationship between bounded motions and symmetries instead of conservation laws.

In this chapter we also made the first generalisation of the classical Sundman's inequality method to investigate restrictions of integrals on possible motion of relativistic systems. The standard bounded geodesic motion in Schwarzschild spacetime is reproduced using the generalised inequality method. In the next chapter will make a further generalisation and apply it to investigate more complicated problems.

CHAPTER 7

Symmetries of Spacetime, Conservation Laws and Forbidden Motion

The last chapter presented a general discussion on the symmetry of spacetime and conservation law in general relativity using the modern language of differential manifolds. At the end of the chapter we demonstrated a successful use of the generalised Sundman's inequality method in general relativity to establish constraints on possible motion imposed by symmetries and conservation laws. The standard bounded geodesic motion in Schwarzschild spacetime was reproduced using our inequality method. It is the purpose of this chapter to further generalise this method and apply it to more complicated relativistic systems. In spite of the simplicity of the method, we are able to present some new results.

We propose to study the restrictions on the possible motion of isolated few-body systems in asymptotic flat spacetime imposed by symmetries of spacetime. As in the classical systems studied in chapters 4 and 5, such restrictions can lead to some interesting bounded motion and hierarchical orbital motion. Such an approach is important because it is now realised, owing to the progress made in understanding deterministic chaos in nonlinear systems, that in most cases the general equations of motion are not soluble in closed-form in Newtonian mechanics, nor in relativity. Therefore if an approximation method is not adopted, all one can say about the behaviour is based on conservation laws and symmetries of the problem under investigation.

Our effort in the present chapter will be focused on the restrictions arising from conserved quantities, especially in those situations where the configuration space is divided to manifest some ordered hierarchical geometrical structure. However, it is impossible to generalise this study further into systems with more than three bodies, because the number of conserved quantities is limited. (In general a nonlinear dynamical system defined on a finite dimensional space possesses only a limited number of symmetries, or integrals.) Therefore only a very limited group of systems are solvable in closed form, while a slightly larger group of systems possess ordered motion due to the presence of integrals. For this reason we shall restrict our attention to the 'few-body' problem and consider a generalisation of ordered motion of gravitational few-body systems in the context of general relativity.

We ask the following question: Can the attractive results of bounded motion be taken over into the framework of general relativity? Put differently, do the conservation laws in general relativity impose restrictions on the motion of the participating bodies such that under some conditions the conservation laws determine some connected forbidden world tubes and the motion of the bodies are restricted to move within some separated possible world tubes? We shall confine ourselves to the case where a space + time split is possible and assume that the typical size of the bodies be small compared with the typical distance between them. Because of the many difficulties in general relativity, a negative first response to this consideration would not be unexpected. For example the existence of gravitational radiation means that energy and angular momentum will be carried away from the system. Nor is it yet clear whether a point-mass (or alternatively, centre of mass) idealisation is possible in general relativity (this is a very important consideration if useful forbidden motion is to be expected). Thus, no easy generalisation of the classical results can be made immediately. The situation, however, is not completely hopeless. Although this problem is far from being completely solved, much more useful results than expected are obtained. It will become clear, by attempting to relax the validity condition of the classical analysis, that gravitational radiation is not a serious difficulty (radiation is by no means unique to general relativity), but that the main difficulty lies in the curved spacetime and the highly nonlinear nature of the Einstein field equations.

As was shown in the previous chapter, the coordinate dependent Landau-Lifshitz (1962) formulation of conservation laws can be generalised to an arbitrary coordinate system. However, since symmetries and asymptotic symmetries of a spacetime are always a definite concept, our generalisation posed some doubts about the uniqueness of conservation laws. Therefore in order to avoid this ambiguity, we shall proceed by modifying the conventional way of achieving forbidden motion. Instead of discussing the constraints of conserved quantities (integrals) on the possible motion, as is usual and apparently obvious, we look for the relation between forbidden motion and the underlying (asymptotic) symmetries of the problem. We will still talk about the constraints imposed by energy and angular momentum, but always keep in mind that it is the symmetry that is important.

From the forbidden motion studies in classical mechanics we select the method which best suits our purpose. It must not only be as simple as possible, but also rely as little as possible on the specific problem and Newtonian concepts. Ideally the best ones should be those which depend only on the (asymptotic) symmetry of the problem. Although a coordinate-free method is desired, at present a method compatible with an asymptotic Minkowski coordinate system must be employed since it is almost always most convenient to work in such coordinates, and as most existing standard results have been formulated in them.

We now consider how forbidden motion is studied in classical mechanics. There are essentially two kinds of forbidden motion analyses, namely those which are determined by a constraint of the energy only, and those which are determined jointly by both the energy and angular momentum. Although the study of constraint by energy alone is of some interest, we will concentrate on those motions restricted by both energy and angular momentum (in general relativity the normalisation condition, equation (6.14a) must also be added). The use of the angular momentum improves the forbidden motion in three senses (all of them may occur in a single problem). Firstly, the regions forbidden by energy constraints alone are enlarged by making use of the angular momentum. Examples include the motion of a point mass in the potential field of fixed bodies possessing rotational symmetry, where the mass of the bodies can be distributed in a uniform sphere, spheroid, ring or disc. Secondly, when the energy is not enough to provide bounds on the motion, the introduction of angular momentum strengthens the restriction so that the motion may become bounded. A good example is the general 3-body problem. Although in this case only one integral is effectively used and a weaker inequality is applied the restrictions are obviously stronger than those obtained from the energy constraint, $H \geq U$, alone. Put another way, the combined first integral C^2H imposes a restriction which is stricter than that imposed by the energy. Thirdly, angular momentum also keeps the system away from simultaneous collision or, in the case of only one moving point mass, from approaching the axis of the rotational symmetry. An example is the aligned-many-centre problem. For detailed results of forbidden motion of the 2-centre problem see Born (1927). Two methods are of concern; the effective potential (canonical transformation) method and Sundman's inequality method.

The effective potential method is based on the fact that the Hamiltonian of a system is positive definite in the generalised momenta. By choosing generalised coordinates which contain as many ignorable coordinates as possible, we see that the corresponding conserved quantities will impose restrictions on the possible values of the remaining coordinates, hence there will exist possible and forbidden motion. If the dimension of the system is low then connected forbidden regions can be formed, thus defining some sort of order in the system. However, this method has some disadvantages although in principle it might be of general use. For example, as a general rule all conserved quantities may contribute to set up restrictions, however, usually no single coordinate system can be simultaneously adapted to all independent conserved quantities (because they are not necessarily in involution), and hence not all independent integrals can be made explicitly corresponding to ignorable coordinates at the same time (in either

Newtonian mechanics or general relativity). Even if the above procedure is possible the amount of algebra would grow enormously with the dimension of the system. However, from the experience in Newtonian mechanics, a complete study of all restrictions is not necessary for the purpose of determining hierarchy. The crucial conservation laws relevant to the problem are angular momentum and energy. Because of this we will concentrate on the second method (Sundman inequality) which is much simpler and more straightforward. This method contains less information but enough to define hierarchy. The Sundman's inequality is not only a very general relation but is more flexible. It will be generalised to a mathematical relation regardless of its physical content.

The quantities involved in this approach are *angular momentum*, moment of inertia, kinetic energy and potential energy. In Newtonian mechanics, all these quantities are well defined. The inequality is valid for both the classical N-body system and continuous systems, as well as parts of the system. It is also valid for a system of charged particles which radiate. The inequality is also relevant when referred to the centre of mass. In general relativity, usually there are no coordinate-free conserved quantities, hence the norm of angular momentum is not well defined. However, we can try to construct, from the ***coordinate-dependent*** conserved quantities, a generalised Sundman's inequality. This allows us to study forbidden motion and hierarchy despite the fact that the inequalities chosen may not be the only possible ones. The undefined quantities, moment of inertia and kinetic energy, are not essential to the investigation so long as the former is a function of coordinates only, while the latter can be replaced by 'energy' and a function of coordinates.

In section 7.1 we generalise the classical Sundman's inequality to become the central mathematical tool for the present study. Appendix B collects the important inequalities and their proofs, and is closely relevant to this section. Section 7.2 is a discussion on the general problem, where we hope to establish the existence of forbidden motion based on asymptotic symmetry. In section 7.3 we apply the method to the first order post-Newtonian approximation of general relativity and obtain inequalities which determine useful bounded motion for the 2- and 3-body problems.

7.1 Generalised Sundman's Inequality

In the previous chapter we have already seen the first generalisation of Sundman's inequality, which extends the inequality with one body involved from 3-dimensional Euclidian space to higher dimensional linear spaces (the inner product is irrelevant for its validity), viz.

$$2\sum_{i,k}(mA^{[i}B^{k]})^2 = \frac{1}{2}\sum_{i,k}(mA^iB^k - mA^kB^i)^2$$
$$\le \left[\sum_i m(A^i)^2\right]\left[\sum_i m(B^i)^2\right] \qquad (7.1)$$

where, $\{A^i\}$and $\{B^i\}$are two arbitrary vectors and m a constant.

A further generalisation of the Sundman's inequality is from that involving one point to that involving many points on the manifold, viz.

$$2\sum_{i,k}\left(\sum_b m_b A_b^{[i}B_b^{k]}\right)^2 = \frac{1}{2}\sum_{i,k}\left[\sum_b m_b A_b^i B_b^k - m_b A_b^k B_b^i\right]^2$$
$$\le \left[\sum_{i,b} m_b (A_b^i)^2\right]\left[\sum_{i,b} m_b (B_b^i)^2\right] \qquad (7.2)$$

where at each point b, there are two arbitrary vectors A_b and B_b, and a non-negative number m_b. In Appendix B, equation (7.2) is proved with $m_b=1$, the proof for arbitrary nonnegative m_b is straightforward. In fact the proof given is more general than the above interpretation. Also we note that all these equations can be applied to continuous systems if the sum over points is replaced by integral over a positive measure element.

We observe that although the cross product operation only belongs to 3-dimensional space and is accidental to this dimension, a formal generalisation of the Sundman's inequality involving the 'norm of the (wedge) cross product' to higher dimensional space is possible. Let us now take a look at some relaxation of the standard Sundman's inequality from the point of view of equations (7.1) and (7.2). Firstly, the total angular momentum and the total energy of the system need not be conserved for its validity. For example, it is not only valid for a subsystem but also for a system of charged particles where electromagnetic radiation carries angular momentum and energy away from the system. This is why we believe that gravitational waves would not be a problem for generalising these equations. However, a good understanding of conservation laws is relevant because we must apply our tool to some physical quantities for meaningful results. Secondly, if a centre of mass can be defined for each body involved, then an inequality can be constructed for the centre of mass. A separate study of how these centre of mass quantities change and interact with the interior motion of the bodies is needed, but this is not the main interest of the present paper, because we can always reasonably assume that this interaction is weak.

The second problem raises a difficulty in relativity, because a useful centre of mass is not easily defined (see Carmeli, 1982). The difficulty of this study is therefore related to the broader difficulties in general relativity: conservation laws and centre of mass.

Any progress in this general approach would be useful to our study. On the other hand, any definition for angular momentum must reflect the underlying rotational symmetry (thus involving the wedge product in some way); the establishment of some new formal relations involving angular momentum may also contribute to the approach on the above general difficulties in relativity.

Considering the difficulty of decoupling the problem of motion into external and internal in general relativity (see the review paper by Damour, 1987) together with other difficulties, we will follow one of two formulations.

(a). The whole continuous system

a1. Formally define energy, angular momentum etc., which are conserved when related to Killing vectors, but which nonetheless capture the conceptual meaning even if not conserved.

a2. Apply the Sundman inequality to these quantities of the whole system.

a3. Decouple the energy into 'effective' potential + kinetic energy. ('effective' will be defined more rigorously later)

(b). 'Mass-centre' part

b1. Construct a mass-centre and decouple the motion into orbital + interior + coupled terms.

b2. Apply Sundman's inequality to the quantities of the mass-centre.

b3. Split energy into 'effective' potential + kinetic.

Both of these formulations can be successfully carried out in Newtonian mechanics. The second procedure in either formulation will be the main approach adopted in the present study. This can always be done without much difficulty. The full potential of this method has only been touched upon in this chapter, however, the inequalities constructed and proved in Appendix B are already strong enough to show that asymptotic symmetries do impose restrictions on the motion of a continuous system.

In the full relativistic theory, an immediate application of the generalised Sundman's inequalities for the centre of mass is not possible since mass-centre is not well-defined. Therefore, we cannot obtain any hierarchical orbital motion in this situation. However, we can investigate the general question of the restriction of (asymptotic) symmetry on motion, following the first formulation above. To do this we choose the Landau-Lifshitz formulation of conservation laws because of its apparent simplicity and mathematical tractability. Thus the first two stages, of defining conserved quantities and applying the generalised inequality method, can be solved formally, but the decoupling of the total energy is a much more difficult problem. However, since these three procedures are not independent, a good approach to the first two stages may facilitate a solution to the problem encountered in the third.

The formulation of the mass-centre and conservation laws by Dixon (1979) (see also DeWitt & Brehme, 1960; Bailey & Israel, 1980) might be a better alternative to those considered here. These authors used a two-point tensor to express conservation laws and obtain a unique mass-centre, and thus their approach is also compatible with the second formulation above. One other advantage of theirs is that the conserved quantities are tensorial. These tensorial quantities are not conserved if the spacetime does not admit symmetries, as is usually the case in general relativity. However, one can still arrive at useful results if these quantities are changing slowly and the rate of change can be estimated since the inequalities are maintained even if the quantities involved are not conserved. What is important is that the object considered must capture the meaning of such quantities as angular momentum. Dixon's formulation has not been carried out in the present study and is left to a future work.

Both formulations can be carried out in the post-Newtonian limit, because a conserved mass and mass-centre can be defined to this order. It will be shown in section 7.3 that the classical hierarchical orbital motion is still held to this order.

Though the above two equations are valid in any coordinate system, they are relations between different quantities in different coordinates. In order to be related to physically meaningful conservation laws, a preferred coordinate system is required, eg., asymptotically Minkowski coordinates. Thus a further generalisation into tensorial form is required for future research.

7.2 General Discussion of the Full Relativistic N-Body Problem

In this section we are concerned with a general analysis of the problem in general relativity. Nevertheless, special relativity is also discussed because of the formal similarity between the integral conservation laws in both the special and general theories. If the Landau-Lifshitz formulation is followed and a pseudo energy-momentum tensor is defined for gravitation (the factor -g is neglected here, cf. section 6.3) then we have the following conserved 4-momentum and angular momentum in an asymptotic Minkowski coordinate system,

$$P^{\mu} = \int T^{\mu 0}_{\text{eff}} d^3x, \quad J^{\mu\nu} = 2 \int x^{[\mu} T^{\nu]0}_{\text{eff}} d^3x\,, \quad (T^{\mu\nu}_{\text{eff}} \equiv T^{\mu\nu} + t^{\mu\nu}) \qquad (7.3)$$

where $t^{\mu\nu}$ is a function of metric, which in turn is formally only a function of coordinates. This quantity vanishes in a global Minkowski spacetime. Now let us apply equations (7.1), (7.2) and those given in Appendix B to the quantities defined by equation (7.3).

A Formal Study in General Relativity and Special Relativity

If we assume that the spacetime is such that there is no coordinate singularity, then we can apply the generalised Sundman's inequality, equation (7.2) and (B8), to the conserved components of the angular momentum. Although we can sum over all the components, we believe that a physically useful construction is to sum over spatial parts only. (This can be done formally by projecting any variations in the components of angular momentum onto a spatial hypersurface orthogonal to the 4-velocity since we are primarily interested in how the spatial components behave). Hence, for a system of widely separated bodies, we have

$$C^2 \equiv \sum_{j,k} (J^{jk})^2 = 2 \sum_{j,k} \left[\int x^{[j} T_{\text{eff}}^{k]0} \, d^3x \right]^2$$

$$\leq \left(\int r^2 d^3x \right) \cdot \left[\int \sum_j (T_{\text{eff}}^{j0})^2 \, d^3x \right] \qquad (7.4)$$

$$C^2 \equiv \sum_{j,k} (J^{jk})^2 = 2 \sum_{j,k} \left[\int x^{[j} T_{\text{eff}}^{k]0} \, d^3x \right]^2$$

$$= 2 \sum_{j,k} \left(\int x^{[j} T^{k]0} d^3x + \int x^{[j} t^{k]0} d^3x \right)^2$$

$$\leq \left(2 \int r^2 \, d^3x \right) \cdot \left[\int \sum_j (T^{j0})^2 \, d^3x + \int \sum_j (t^{j0})^2 \, d^3x \right] \qquad (7.5)$$

where $r^2 \equiv \sum_j (x^j)^2$.

(Note that we have again defined C formally here).

These are two examples of the inequalities. Now let us observe from equation (7.5), say, that the LHS and the first factor on the RHS are in a satisfactory form for the definition of forbidden motion. The problem is then how to use the conservation of the 4-momentum and normalisation condition to replace the first term of the second factor by quantities involving only conserved quantities and functions of coordinates. This was successfully carried out in the last section for geodesic motion. We cannot do it in the same way here, because the 'energy equation' of P^0 in equation (7.3) is not useful to us since it does not involve the $(T^{j0})^2$ term. This is a common difficulty in general relativity. However, by utilising the flexibility of our method, the unwelcome term being positive definite may be effectively removed from the analysis. It must be kept in mind that in doing this the inequality is weakened and, in fact, may not be physically reasonable, since it is possible that a physically significant term could be neglected in this way. However, some form of forbidden motion may still be obtained.

A first approach is to use the flexibility of one of the many possible generalisations of Sundman's inequality, equation (B8) of Appendix B, so that we can make the first term of the second factor in equation (7.5) very small, viz.

$$\begin{aligned} C^2 &= 2 \sum_{j,k} \left[\int (x^{[j} T^{k]0} + x^{[j} t^{k]0}) d^3x \right]^2 \\ &\leq \left(\int (A+1) r^2 d^3x \right) \cdot \left(\int A^{-1} \sum_j (T^{j0})^2 d^3x + \int \sum_j (t^{j0})^2 d^3x \right) \\ &\approx \left(\int r^2 d^3x \right) \cdot \left(\int \sum_j (T^{j0})^2 d^3x + A \int \sum_j (t^{j0})^2 d^3x \right) \\ &\approx A \left(\int r^2 d^3x \right) \cdot \left(\int \sum_j (t^{j0})^2 d^3x \right) \end{aligned} \qquad (7.6)$$

where A is a positive number or function. There always exists a large enough A to make the last two approximations possible. Therefore, the RHS is a function of coordinates only. This must carry some information about the restriction on coordinates by conservation laws although only the angular momentum is used in this inequality. However, it gives useful forbidden motion only if the number of coordinates involved is small. From experience gained in the classical study, this means only if a one, two, or three body problem is concerned. Since in general relativity, a point mass produces a singularity problem (in fact, this has some physical difficulties, for example, a δ-function distribution cannot be introduced arbitrarily), we must consider the centre of mass. However, as mentioned earlier this concept is not well understood in general relativity. Dixon's two-point tensor formulation (see also Carmeli, 1982; Damour, 1987) may prove to be useful, but cannot readily be used and is too detailed to consider here. If this difficulty were solved, then equation (7.6) should give conceptually an analogous result as in Newtonian mechanics, which might be of little practical value. This is because that usually the matter contribution is bigger than that from gravitation, so that the above procedure has weakened the original relation by dropping the significant term.

Consider now a more complete study with the energy-momentum tensor of the matter explicitly given. In addition to the above difficulty there arises another general difficulty of general relativity. Consider a perfect fluid model, then we have the standard expression for the energy-mementum tensor,

$$T^{\mu\nu} = (\rho + p) V^{\mu} V^{\nu} + p g^{\mu\nu} \ , \qquad (7.7)$$

$$\overline{V} \ \equiv \ V^{\mu}\partial/\partial x^{\mu} \text{ is the velocity vector}$$

where ρ and p are the total energy density and the pressure which are, by definition, measured in a momentarily comoving local Lorentz frame of the fluid element. Applying equation (B8) to equations (7.3) and (7.7), we obtain

$$\begin{aligned} C^2 &\equiv 2 \sum_{j,k} \left(\int (\rho + p) V^0 x^{[j} V^{k]} + p x^{[j} g^{k]0} + x^{[j} t^{k]0} d^3x \right)^2 \\ &\leq \left\{ \int [(\rho + p)(V^0)^2 + p^2 + 1] r^2 d^3 x \right\} \cdot \\ &\quad \left\{ \int \sum_j (\rho + p)(V^j)^2 d^3x + \int \sum_j (g^{j0})^2 + (t^{j0})^2 d^3x \right\} \qquad (7.8) \end{aligned}$$

Here, the difficulty is again how to replace the term involving $\Sigma(V^i)^2$. In order to do this we must invoke the conservation of energy, viz.

$$P^0 = \int [(\rho + p) V^0 V^0 + p g^{00} + t^{00}] d^3x \qquad (7.9)$$

and carry out two coordinate transformations: one between the coordinate frame and an orthonormal frame attached to it, the other between this attached Lorentz frame and the locally comoving Lorentz frame. Another difficulty is the property of the t term. In order to obtain some forbidden motions like those found in classical mechanics, this term is required to increase as the bodies come closer. It is not immediately obvious that this is the case. However, the following demonstration in special relativity suggests that there might be a way to overcome these difficulties.

Consider the same problem in special relativity, with again a perfect fluid system. Because we now have a global Lorentz frame, we only need the simpler one of the above transformations (or equivalently use the normalisation condition g(V,V)=-1 and the special metric g=η). It turns out that the well known difficulty of defining kinetic energy in special relativity is not a problem here, and the difficulty of utilising equation (7.9) can be easily resolved. Thus equations (7.3) and (7.8) can be written

$$P^0 = \int \left(\frac{\rho + p}{1 - v^2} - p \right) d^3x \quad , \quad J^{jk} = 2 \int \frac{(\rho + p) x^{[j} v^{k]}}{1 - v^2} d^3x$$

$$
\begin{aligned}
C^2 &\equiv 2 \sum_{j,k} \int \frac{(\rho + p)x^{[j}v^{k]}}{1 - v^2} d^3x \\
&\leq \int \frac{(\rho + p)r^2}{1 - v^2} d^3x \cdot \int \frac{(\rho + p)v^2}{1 - v^2} d^3x \\
&= \int \frac{(\rho + p)\, r^2}{1 - v^2} d^3x \cdot (P^0 - \int \rho\, d^3x)
\end{aligned}
\qquad (7.10)
$$

and we see that we have succeeded in determining reasonable forbidden motion for a relativistic system. Here we emphasise that the **weak energy condition** (Synge, 1960; Hawking & Ellis, 1973) plays a significant role in obtaining both equation (7.8) and (7.10). This condition guarantees that p and ρ+p are nonnegative. We also observe that the factor (1-v^2) does not present a serious difficulty, since this simply means that when defining the 'moment of inertia' the proper distance should be used instead of the slightly different coordinate distance. This is also the reason behind the choice, in equation (7.8), of associating the factor V^0 with the coordinate r rather than with V^i. (This is possible with no loss of generality.)

A comment on how to define 'moment of inertia' and 'kinetic energy' in relativity is in order here. The difficulty of such classical concepts is not that they cannot be defined; it is rather why define them, what is their use, what is the most useful definition? Of course, a quantity can be regarded as a counterpart in relativity if it reduces to the classical one in the Newtonian order. This is similar to what happens in the 'comma goes to semicolon' rule, so such counterparts in relativity are not unique. We can thus impose another restriction: formal similarity. A more physically useful restriction is to study the relation of such quantities to conservation laws -- in physics a useful concept is one which is related to conservation laws.

The success of equation (7.10) in special relativity is encouraging and it is possible that a system of charged particles can be studied in this way to obtain ordered motion for some finite time. It is hoped that this will be studied at a later date.

Discussion and Conclusion

In this section we discussed the relation between forbidden motion and asymptotic symmetry. Despite the difficulties in general relativity of defining such concepts as energy and angular momentum, a mathematical method was developed which, although restricted in some cases, provided an ideal platform from which to study the possibility of forbidden motion within the context of geometrical spacetimes. This method is proposed as a generalisation of Sundman's inequality method of defining hierarchy in classical systems and proved to be particularly successful (see section 6.5) when we considered the effects of energy and angular momentum constraints on geodesic motion in a Schwarzschild spacetime and in the case where the spacetime under investigation was described by a Minkowski metric. Further application of this method to less restricted systems suffers from many difficulties. However, the analysis carried out in this section suggests that the presence of an asymptotic symmetry in a system of gravitationally interacting bodies must impose some restriction on possible motion. Therefore, it should be possible to discuss ordered motion for a general class of solutions displaying the property of asymptotic symmetry (i.e. Killing vectors can be defined at large distances from the local system).

It is also argued, by analogy to the classical problem, that although the inclusion of gravitational radiation, which may carry away both energy and angular momentum, prevents even the 2-body problem from being adequately discussed in general relativity, it does not prove to be a serious difficulty when considering forbidden motion, if the quantities involved are slowly varying. The main problem encountered is due to the fact that the spacetime is curved and that there is no a priori spacetime geometry in general relativity. Thus, in general, it is not possible to assume the existence of Killing vectors (symmetries). It is also found that when attempting to define a centre of mass, the nonlinear aspect of general relativity prevents us from carrying out a rigorous analysis except in some special cases. Because of this we hope to apply the method considered here to the theory developed by Dixon (1979), where the centre of mass is defined.

The inequality method described in this paper allows us to consider in more detail the conditions for the existence of forbidden, or ordered, motions in systems outwith the realm of Newtonian mechanics. By replacing the classical emphasis on 'integrals of motion' with one of 'symmetries of the spacetime' we are able, for some systems at least, to develop useful results regarding the formation of forbidden regions in a general relativistic context. In order to do this a coordinate framework was required. However, it should be possible, although not trivial, to express the relations usefully in a coordinate-free formalism. This is hoped to be carried out as a future work. In this respect, a use of the tetrad formalism may be a useful alternative approach.

A typical feature of the inequality method is its flexibility and this will be applied to a post-Newtonian formalism in the next section. It is also hoped that the application to stationary axially-symmetric systems may provide some useful results for the spacetime surrounding pulsars and Kerr black holes (see Chandrasekhar, 1983). Much work is required, however, in order to improve the applicability and the generality of this method. Any results in the field of conservation laws in general relativity would be invaluable to this analysis. One great benefit which may be obtained from the approach outlined in this section is in the study of mass transfer between the two stars of a binary system. By considering the gravitational field of the system in a similar way to that of Contopoulos (1990) and Chandrasekhar (1989) it may be possible to utilise the flexibility of the inequality method to obtain the relativistic limits on the motion of the transferring mass. This would be an interesting alternative to the classical **Roche lobe** analysis for a binary system.

7.3 Bounded Motion of the Post-Newtonian N-Body Problem

In the last chapter and previous section, the possibility of establishing bounded motion in the relativistic, gravitational few-body system was investigated. It was shown that the existence of an asymptotic symmetry in the system should result in the motion being bounded. In the present section we will apply the method developed to the post-Newtonian approximation. Although the post-Newtonian approximation has been criticised for the adverse side-effect of introducing implicitly a 'neo-Newtonian' interpretation of general relativity and in general such an approximation method may not give conceptually useful results (Carmeli, 1982; Damour, 1987), this investigation is carried out for the following reasons.

Firstly, post-Newtonian approximation is a simpler and easily manageable case, in which concrete results may be achieved. Secondly, it may serve as a further example with which to test the method: if bounded motion cannot be established to the post-Newtonian order then we may have to conclude that the method proposed in section 7.2 is not applicable to the full relativistic case. However, if we can obtain bounded motion then this may at least suggest that the method captures some of the essential mechanism of bounded motion in general relativity (Schutz, 1990). Finally, in a more practical sense, the investigation of a gravitational system to post-Newtonian order is in itself a useful exercise. Such investigations have been carried out extensively since the general relativistic theory was established (see Will, 1981). Nevertheless, it has been conventional to restrict the study to the relativistic corrections of the orbital elements. To the author's knowledge, no bounded motion has yet been established.

The investigation of bounded motion to post-Newtonian order is the main interest of the present section. We first give the energy and (angular) momentum integrals of the post-Newtonian equations of motion. Then the inequality method is applied to the first order post-Newtonian N-body problem. Some inequalities are obtained, based on which bounded motion may be easily established for the (post-Newtonian) 2- and 3-body problems. The standard post-Newtonian formulation is used (Will, 1981; Weinberg, 1972; Fock, 1959; Misner et al, 1973) and the units are chosen such that the speed of light and the gravitational constant are both unity. The notation of this section is the same as the chapters on classical studies.

Conserved Quantities of the Post-Newtonian Equations of Motion

The Post-Newtonian equations of motion have been obtained and studied many times since the foundation of general relativity. In spite of the divergence of approaches in achieving them, the results are essentially the same. These equations of motion can be interpreted as those of the centres of inertial mass of perfect fluid extended bodies (averaged over some interior time scales) (Will, 1981), or as those of the point masses (Weinberg, 1972). In fact, in the first order post-Newtonian approximation, the motion of the centre of inertial mass (or the point mass) follows the geodesics of the post-Newtonian metric generated by other bodies (see Will, 1981), and this approximation allows the concept of point mass to be admitted.

It can be shown straightforwardly that the post-Newtonian equations of motion can be (with no additional approximation) described in a Lagrangian form (Fock, 1959). This means that we can introduce a Lagrangian from general relativity and then study the Lagrangian system in the framework of Newtonian mechanics in a rigorous manner. Therefore all the classical techniques such as the Hamiltonian formulation, canonical transformation and vector analysis can be used. It follows immediately that such a post-Newtonian Lagrangian system possesses similar classical conservation laws since the Lagrangian is invariant with respect to a time translation, spatial translations and rotations. However, these conservation laws might result from the isometry of the background Minkowski space-time rather than the asymptotic symmetries of the original relativistic problem. Therefore the forbidden motion obtained to post-Newtonian order may not be useful as a reliable support for the general question of the relation between forbidden motion and asymptotic symmetries.

To obtain the (exact) conservation laws of the (first order post-Newtonian) equations of motion (or Lagrangian), let us write the Lagrangian (see Will, 1981) in a more convenient form

$$\begin{aligned}\mathscr{L}(\mathbf{R}(t), \mathbf{V}(t)) = -\sum_b m_b + \sum_b \left(\frac{1}{2} m_b V_b^2 + \frac{1}{8} m_b V_b^4\right) \\ + \frac{1}{2} \sum_{(b,d)} \frac{m_b m_d}{R_{bd}} [(3V_b^2 + 3V_d^2) - 7\mathbf{V}_b \cdot \mathbf{V}_d \\ -(\mathbf{V}_b \cdot \mathbf{n}_{bd})(\mathbf{V}_d \cdot \mathbf{n}_{bd})] \\ - \sum_{(b,d)} \left(-\frac{m_b m_d}{R_{bd}} + \frac{m_b m_d (m_b + m_d)}{2R_{bd}^2}\right).\end{aligned} \qquad (7.11)$$

where the notation is the same as the standard classical approach: m_b stands for the constant mass of the b^th^ body, the vectors $\mathbf{R}_b$, $\mathbf{V}_b$ denote radial and velocity vectors in Euclidean 3-space, t being the universal time and $\mathbf{n}_{bd}=\mathbf{R}_{bd}/R_{bd}$, $\mathbf{R}_{bd}=\mathbf{R}_b-\mathbf{R}_d$. The sum over (b,d) denotes the summation over all possible pairs without repetition.

It will become clear, after the conserved Hamiltonian is obtained, that the last term may be interpreted as 'potential energy'. The generalised linear momentum of each body $\mathbf{P}_b$, corresponding to the coordinates $\mathbf{R}_b$, is determined in the standard way, namely,

$$\begin{aligned}
\mathbf{P}_b \equiv\ & \{\partial\mathcal{L} / \partial\dot{X}_b, \partial\mathcal{L} / \partial\dot{Y}_b, \partial\mathcal{L} / \partial\dot{Z}_b\} \\
=\ & (m_b + \tfrac{1}{2} m_b V_b^2)\,\mathbf{V}_b + \left(\sum_{d\neq b} \frac{m_b m_d}{R_{bd}}\right) 3\,\mathbf{V}_b \\
& - \frac{7}{2}\sum_{d\neq b}\left(\frac{m_b m_d}{R_{bd}}\mathbf{V}_d\right) - \frac{1}{2}\sum_{d\neq b}\left(\frac{m_b m_d}{R_{bd}}(\mathbf{V}_d \cdot \mathbf{n}_{bd})\mathbf{n}_{bd}\right) \\
=\ & (m_b + \tfrac{1}{2} m_b V_b^2 - \sum_{d\neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}})\,\mathbf{V}_b \\
& + \sum_{d\neq b} \frac{7}{2}\left(\frac{m_b m_d}{R_{bd}}(\mathbf{V}_b - \mathbf{V}_d)\right) - \sum_{d\neq b} \frac{1}{2}\left(\frac{m_b m_d}{R_{bd}}(\mathbf{V}_d \cdot \mathbf{n}_{bd})\,\mathbf{n}_{bd}\right). \qquad (7.12)
\end{aligned}$$

The Hamiltonian is then given (the velocities have not been replaced by momenta),

$$\begin{aligned}
\mathcal{H}(\mathbf{R}(t), \mathbf{V}(t)) =\ & \sum_b (\mathbf{V}_b \bullet \mathbf{P}_b) - \mathcal{L} \\
=\ & \sum_b m_b + \sum_b (\tfrac{1}{2} m_b V_b^2 + \tfrac{3}{8} m_b V_b^4) \\
& + \frac{1}{2}\sum_{(b,d)} \frac{m_b m_d}{R_{bd}}[(3V_b^2 + 3V_d^2) - 7\,\mathbf{V}_b \cdot \mathbf{V}_d \\
& \qquad -(\mathbf{V}_b \cdot \mathbf{n}_{bd})(\mathbf{V}_d \cdot \mathbf{n}_{bd})] \\
& + \sum_{(b,d)}\left(-\frac{m_b m_d}{R_{bd}} + \frac{m_b m_d (m_b + m_d)}{2R_{bd}^2}\right). \qquad (7.13)
\end{aligned}$$

This corresponds to the conserved 'total energy' of the system. This energy can be split into a 'potential energy', a 'kinetic energy', and the 'total inertial mass', (Σm_b). The last term in this equation can be realised as the 'potential energy' (denoted by U) because it is not only a function of coordinates, but also the part which has opposite signs in the Lagrangian and Hamiltonian respectively. The remaining terms, involving velocity, of this total energy are regarded as the 'kinetic energy'. However, attention must be paid to the coefficient difference of the $(V_b)^4$ terms in the Lagrangian and Hamiltonian, and to the fact that the total inertial mass has different signs in the

Lagrangian and Hamiltonian. Because of this the Lagrangian cannot be written as the difference between the 'kinetic' and 'potential' energies. As in the classical N-body problem, the total linear and angular momenta are also conserved, viz.

$$\mathbf{P} \equiv \sum_b \mathbf{P}_b = \mathrm{const}\,; \quad \mathbf{J} \equiv \sum_b (\mathbf{R}_b \times \mathbf{P}_b) = \mathrm{const}. \qquad (7.14)$$

The conservation of the total linear momentum can be shown from the Euler-Lagrange equations of motion. For example,

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{X}_b}\right) = \frac{\partial \mathcal{L}}{\partial X_b} \quad \Rightarrow \quad \frac{d}{dt}\left(\sum_b \frac{\partial \mathcal{L}}{\partial \dot{X}_b}\right) = \sum_b \frac{\partial \mathcal{L}}{\partial X_b} \;,$$

$$\sum_b \frac{\partial \mathcal{L}}{\partial X_b} = 0 \quad \Rightarrow \quad \frac{d}{dt}\left(\sum_b \frac{\partial \mathcal{L}}{\partial \dot{X}_b}\right) = 0 \quad \Rightarrow \quad \sum_b \frac{\partial \mathcal{L}}{\partial \dot{X}_b} = \mathrm{const}$$

The conservation of the angular momentum can be shown in the following outlined scheme,

$$\frac{d\mathbf{J}}{dt} = \frac{d}{dt}\left(\sum_b \mathbf{R}_b \times \mathbf{P}_b\right) = \sum_b \mathbf{V}_b \times \mathbf{P}_b + \sum_b \mathbf{R}_b \times \frac{d\mathbf{P}_b}{dt}$$
$$= \sum_b \mathbf{V}_b \times \mathbf{P}_b + \sum_b \mathbf{R}_b \times \nabla_{\mathbf{R}_b} \mathcal{L} = 0 \;.$$

This proof is essentially similar to the classical proof: the Euler-Lagrange equations of motion must be used and the manipulation of the gradient of position vectors must again be invoked.

Although it is straightforward to write out the total linear and angular momentum, we give their explicit expressions because some of the terms occurred in $\mathbf{P}_b$ cancel out to yield simpler results, viz.

$$\mathbf{P} = \sum_b \Big(m_b + \frac{1}{2} m_b V_b^2 - \sum_{d \neq b} \frac{1}{2} \frac{m_b m_d}{R_{bd}}\Big) \mathbf{V}_b$$
$$- \sum_b \sum_{d \neq b} \frac{1}{2}\left(\frac{m_b m_d}{R_{bd}} (\mathbf{V}_d \cdot \mathbf{n}_{bd})\, \mathbf{n}_{bd}\right)$$
$$= \sum_b \Big(m_b + \frac{1}{2} m_b V_b^2\Big) \mathbf{V}_b - \sum_{(b,\,d)} \frac{1}{2} \frac{m_b m_d}{R_{bd}} (\mathbf{V}_b + \mathbf{V}_d)$$
$$- \sum_{(b,\,d)} \frac{1}{2}\left(\frac{m_b m_d}{R_{bd}} [(\mathbf{V}_b + \mathbf{V}_d) \cdot \mathbf{n}_{bd}]\, \mathbf{n}_{bd}\right). \qquad (7.15)$$

$$\mathbf{J} = \sum_{b}(m_b + \frac{1}{2}m_b V_b^2 - \sum_{d \neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}})\, \mathbf{R}_b \times \mathbf{V}_b$$

$$+ \sum_{b}\sum_{d \neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}} [7\mathbf{R}_b \times (\mathbf{V}_b - \mathbf{V}_d)$$

$$-(\mathbf{V}_d \cdot \mathbf{n}_{bd})\, \mathbf{R}_b \times \mathbf{n}_{bd}]$$

$$= \sum_{b}(m_b + \frac{1}{2}m_b V_b^2 - \sum_{d \neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}})\, \mathbf{R}_b \times \mathbf{V}_b$$

$$+ \sum_{(b,\,d)} \frac{1}{2}\frac{m_b m_d}{R_{bd}} \{7\mathbf{R}_{bd} \times (\mathbf{V}_b - \mathbf{V}_d)$$

$$- [(\mathbf{V}_b + \mathbf{V}_d) \cdot \mathbf{n}_{bd}]\, \mathbf{R}_b \times \mathbf{n}_{bd}\} \; . \qquad (7.16)$$

Finally, we note that there is another useful expression for the total linear momentum relevant to defining the centre of mass, namely,

$$M_b \equiv m_b + \frac{1}{2}m_b V_b^2 - \sum_{d \neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}} \; , \qquad (7.17)$$

$$\mathbf{P} = \frac{d}{dt}(\sum_{b} M_b \mathbf{R}_b)$$

$$= \frac{d}{dt}\left\{\sum_{b}\left(m_b + \frac{1}{2}m_b V_b^2 - \sum_{d \neq b} \frac{1}{2}\frac{m_b m_d}{R_{bd}}\right)\mathbf{R}_b\right\} \; . \qquad (7.18)$$

From this equation a centre of mass can be defined as

$$\mathbf{R}_c \equiv \frac{\sum M_b \mathbf{R}_b}{\sum M_b} \; , \qquad (\sum M_b)\mathbf{R}_c = \sum M_b \mathbf{R}_b = 0. \qquad (7.19)$$

In this section we have summarised the conservation laws of the first order post-Newtonian N-body problem. These equations are all exact, except equations (7.18) and (7.19), in which further approximation are involved.

Sundman's Inequality of the Post-Newtonian N-body problem

As already noted in the previous section, the inequality method is extremely flexible. In general, many inequalities may be constructed for a given problem. This is also true in the post-Newtonian limit considered here; different interpretations of the parameters such as mass result in different constructions. On the other hand we can construct inequalities in which the 'energy' and 'angular momentum' are changing slowly. With these points in mind we give the following three simplest relations.

If we apply the inequality to the simplest and most important terms, then we have

$$
\begin{aligned}
H_1 &\equiv \sum_b m_b + \sum_b \left(\frac{1}{2} m_b V_b^2 + \frac{3}{8} m_b V_b^4\right) + U \\
\mathbf{J}_1 &\equiv \sum_b M_b \mathbf{R}_b \times \mathbf{V}_b \\
&= \sum_b \left(m_b + \frac{1}{2} m_b V_b^2 - \sum_{d \neq b} \frac{1}{2} \frac{m_b m_d}{R_{bd}}\right) \mathbf{R}_b \times \mathbf{V}_b
\end{aligned}
$$

$$
\begin{aligned}
(\mathbf{J}_1)^2 &\leq \left(\sum_b M_b R_b^2\right)\left(\sum_b M_b V_b^2\right) \\
&\leq 2\left(\sum_b M_b R_b^2\right)\left(H_1 - \sum_b m_b - U\right) \qquad (7.20)
\end{aligned}
$$

where H_1 and $(J_1)^2$ are slowly changing integrals according to the conservation of energy and angular momentum. If the centre of mass is located at the origin of the coordinate system, the 'moment of inertia' can be rewritten in another form, as is also true in the classical study,

$$
\sum_b M_b R_b^2 = \left(\sum_{(b,\,d)} M_b M_d R_{bd}^2\right) / \left(\sum_b M_b\right) \qquad (7.21)
$$

where ΣM_b is constant to the approximation taken. The advantage of this expression of the 'moment of inertia' is that only relative distances are involved. It therefore helps reduce the (configurational) dimension of the system by one, since we can now chose the separation of one pair as the unit of length. We finally obtain an inequality which is similar to the classical Sundman's inequality

$$
\begin{aligned}
(\mathbf{J}_1)^2 \leq \frac{2}{\sum_b M_b} & \left\{\sum_{(b,\,d)} M_b M_d R_{bd}^2\right\} \\
& \left\{H_1 - \sum_b m_b - \sum_{(b,\,d)} \left(-\frac{m_b m_d}{R_{bd}} + \frac{m_b m_d (m_b + m_d)}{2R_{bd}^2}\right)\right\} \qquad (7.22)
\end{aligned}
$$

From this equation, bounded motion can be determined for a finite time scale. The problem that the mass, M_b, is not constant can be eliminated by a use of the Virial theorem, which gives $M_b \leq m_b$. Thus,

$$
\begin{aligned}
(\mathbf{J}_1)^2 \leq \frac{2}{\sum_b M_b} & \left\{\sum_{(b,\,d)} m_b m_d R_{bd}^2\right\} \\
& \left\{H_1 - \sum_b m_b - \sum_{(b,\,d)} \left(-\frac{m_b m_d}{R_{bd}} + \frac{m_b m_d (m_b + m_d)}{2R_{bd}^2}\right)\right\} \qquad (7.23)
\end{aligned}
$$

The above analysis is the simplest generalisation of the classical result. Let us now try to modify it by including more of the energy and angular momentum terms, and thus increase the time scale of validity. If we define

$$H_2 \equiv \sum_b m_b + \sum_b \left(\frac{1}{2} m_b V_b^2 + \frac{3}{8} m_b V_b^4\right) + \frac{1}{2} \sum_{(b,d)} \frac{m_b m_d}{R_{bd}} [(3V_b^2 + 3V_d^2) - 7\mathbf{V}_b \cdot \mathbf{V}_d]$$

$$\mathbf{J}_2 \equiv \sum_b \left(m_b + \frac{1}{2} m_b V_b^2 - \sum_{d \neq b} \frac{1}{2} \frac{m_b m_d}{R_{bd}}\right) \mathbf{R}_b \times \mathbf{V}_b + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} \{\mathbf{R}_{bd} \times (\mathbf{V}_b - \mathbf{V}_d)\}$$

then a use of the inequalities developed in Appendix B gives

$$(\mathbf{J}_2)^2 \leq \left\{\sum_b M_b R_b^2 + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} R_{bd}^2\right\} \left\{\sum_b M_b V_b^2 + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} (\mathbf{V}_b - \mathbf{V}_d)^2\right\}$$

where the two factors on the right side may be further simplified, viz.

$$\{\text{First Factor}\} = \left\{\sum_b M_b R_b^2 + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} R_{bd}^2\right\}$$

$$= \left\{\left(\sum_{(b,d)} M_b M_d R_{bd}^2\right) / \left(\sum_b M_b\right) + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} R_{bd}^2\right\}$$

$$\{\text{Second Factor}\} = \left\{\sum_b M_b V_b^2 + \sum_{(b,d)} \frac{7}{2} \frac{m_b m_d}{R_{bd}} (\mathbf{V}_b - \mathbf{V}_d)^2\right\}$$

$$= \left\{\sum_b \left(m_b + \frac{1}{2} m_b V_b^2 - \sum_{d \neq b} \frac{1}{2} \frac{m_b m_d}{R_{bd}}\right) V_b^2 + \sum_{(b,d)} \frac{m_b m_d}{R_{bd}} \left[\frac{7}{2}(V_b^2 + V_d^2) - 7\mathbf{V}_b \cdot \mathbf{V}_d\right]\right\}$$

$$= \left\{\sum_b \left(m_b + \frac{1}{2} m_b V_b^2\right) V_b^2 + \sum_{(b,d)} \frac{m_b m_d}{R_{bd}} [3(V_b^2 + V_d^2) - 7\mathbf{V}_b \cdot \mathbf{V}_d]\right\}$$

$$= \left\{\sum_b \left(\frac{1}{2} m_b + \frac{2}{8} m_b V_b^2\right) V_b^2 + \frac{1}{2} \sum_{(b,d)} \frac{m_b m_d}{R_{bd}} [3(V_b^2 + V_d^2) - 7\mathbf{V}_b \cdot \mathbf{V}_d]\right\} \times 2$$

$$\leq 2\{H_2 - \sum_b m_b - U\}$$

and we finally obtain the inequality

$$(J_2)^2 \le 2\left\{\frac{\sum\limits_{(b,\,d)} m_b m_d R_{bd}^2}{\sum\limits_b M_b} + \sum_{(b\,,\,d)} (\frac{7}{2} m_b m_d R_{bd})\right\} \{H_2 - \sum_b m_b - U\} \qquad (7.24)$$

This relation is more satisfactory than equations (7.20, 22, 23), since the changes of H_2 and $(J_2)^2$ are slower than H_1 and $(J_1)^2$ respectively. For the (post-Newtonian) 2-body problem, J_2=**J**, since the neglected terms in **J** vanish if the centre of mass is set at the origin of the coordinate system. On the other hand, since the velocity vectors of the two bodies are almost anti-parallel (with centre of mass at the origin of the coordinate system), the term neglected in $\mathcal{H}$ is non-negative (except when these two velocity vectors are almost perpendicular to the connecting vector; however, when this happens the neglected term is vanishingly small). Therefore equation (7.24) is rigorous for a 2-body problem, that is,

$$(J)^2 \le 2\left\{\frac{m_b m_d R_{bd}^2}{M_b + M_d} + \frac{7}{2}\frac{m_b m_d R_{bd}^2}{R_{bd}}\right\}$$
$$\left\{\mathcal{H} - \sum_b m_b + \frac{m_b m_d}{R_{bd}} - \frac{m_b m_d (m_b + m_d)}{2 R_{bd}^2}\right\} \qquad (7.25)$$

where b and d are the indices of the two bodies.

A satisfactory inequality has not yet been found for the full energy and angular momentum of systems with more than two bodies. However, equation (7.25) is possibly a general relation for systems with more than two bodies. We now demonstrate the difficulty by the following construction,

$$\mathbf{J} = \sum_b (m_b + \frac{1}{2} m_b V_b^2 - \sum_{d \ne b} \frac{1}{2}\frac{m_b m_d}{R_{bd}})\, \mathbf{R}_b \times \mathbf{V}_b$$
$$+ \sum_b \sum_{d \ne b} \frac{1}{2}\frac{m_b m_d}{R_{bd}} \sqrt{7}\, \mathbf{R}_b \times [\sqrt{7}(\mathbf{V}_b - \mathbf{V}_d) - \frac{1}{\sqrt{7}}(\mathbf{V}_d \cdot \mathbf{n}_{bd})\, \mathbf{n}_{bd}]$$

$$J^2 \le \left\{\sum_b M_b R_b^2 + \sum_b \sum_{d \ne b} \frac{m_b m_d}{R_{bd}} 7 R_b^2\right\}$$
$$\left\{\sum_b M_b V_b^2 + \sum_b \sum_{d \ne b} \frac{1}{4}\frac{m_b m_d}{R_{bd}} [\sqrt{7}(\mathbf{V}_b - \mathbf{V}_d) - \frac{1}{\sqrt{7}}(\mathbf{V}_d \cdot \mathbf{n}_{bd})\, \mathbf{n}_{bd}]^2\right\}$$

where

{Second Factor} =

$$=\sum_{b} M_b V_b^2+\sum_{b}\sum_{d\neq b}\frac{1}{4}\frac{m_b m_d}{R_{bd}}[\sqrt{7}(\mathbf{V}_b-\mathbf{V}_d)-\frac{1}{\sqrt{7}}(\mathbf{V}_d\cdot\mathbf{n}_{bd})\mathbf{n}_{bd}]^2$$

$$=\sum_{b} M_b V_b^2+\sum_{b}\sum_{d\neq b}\frac{1}{4}\frac{m_b m_d}{R_{bd}}[7(\mathbf{V}_b-\mathbf{V}_d)^2$$

$$-2(\mathbf{V}_b\cdot\mathbf{n}_{bd})(\mathbf{V}_d\cdot\mathbf{n}_{bd})+\frac{15}{7}(\mathbf{V}_d\cdot\mathbf{n}_{bd})^2]$$

$$=\sum_{b} M_b V_b^2+\sum_{(b,d)}\frac{1}{2}\frac{m_b m_d}{R_{bd}}[7(\mathbf{V}_b-\mathbf{V}_d)^2-2(\mathbf{V}_b\cdot\mathbf{n}_{bd})(\mathbf{V}_d\cdot\mathbf{n}_{bd})]$$

$$+\sum_{b}\sum_{d\neq b}\frac{1}{4}\frac{m_b m_d}{R_{bd}}\frac{15}{7}(\mathbf{V}_d\cdot\mathbf{n}_{bd})^2$$

$$=\sum_{b}(m_b+\frac{1}{2}m_b V_b^2)V_b^2+\sum_{(b,d)}\frac{m_b m_d}{R_{bd}}[3(V_b^2+V_d^2)-7\mathbf{V}_b\cdot\mathbf{V}_d$$

$$-(\mathbf{V}_b\cdot\mathbf{n}_{bd})(\mathbf{V}_d\cdot\mathbf{n}_{bd})]+\sum_{b}\sum_{d\neq b}\frac{1}{4}\frac{m_b m_d}{R_{bd}}\frac{15}{7}(\mathbf{V}_d\cdot\mathbf{n}_{bd})^2$$

$$=2\left\{\mathcal{H}-\sum_{b} m_b-U-\frac{1}{8}\sum_{b} m_b V_b^4+\frac{1}{2}\sum_{b}\sum_{d\neq b}\frac{1}{4}\frac{m_b m_d}{R_{bd}}\frac{15}{7}(\mathbf{V}_d\cdot\mathbf{n}_{bd})^2\right\}$$

$$\leq 2\left\{\mathcal{H}-\sum_{b} m_b-U+\frac{15}{56}\sum_{b}\left(\sum_{d\neq b}\frac{m_d}{R_{bd}}\right)^2\right\}$$

In this deduction we have used the Virial theorem in the last step. Now we have the inequality

$$\mathbf{J}^2\leq 2\left\{\sum_{b} M_b R_b^2+\sum_{b}\sum_{d\neq b}\frac{m_b m_d}{R_{bd}}7R_b^2\right\}$$

$$\left\{\mathcal{H}-\sum_{b} m_b-U+\frac{15}{56}\sum_{b}\left(\sum_{d\neq b}\frac{m_d}{R_{bd}}\right)^2\right\} \qquad (7.26)$$

From this relation we see that bounded motion also exists in the more complicated (post-Newtonian) 3-body problem since the RHS is a function of coordinates only. However, the result has been achieved because the Virial theorem was used in the last step. Because of the positive residual term, applying this relation to the (post-Newtonian) 2-body problem leads to a relation weaker than equation (7.25). We are therefore of the opinion that this term will vanish if we carry out a better construction of the inequality from the beginning.

An improvement in the construction may be made by utilising the correspondence between the terms in the energy and those in the momentum. This is very useful in the construction of the terms on the right hand side of the inequality from the angular momentum and by using the generalised Sundman's inequalities, so that these terms may be replaced by the 'kinetic energy' terms in the total energy.

The shapes of the possible and forbidden regions have not yet been plotted. This is straightforward, because from an observation of the classical analysis (cf. chapter 4) we see that the inequalities obtained suffice for an investigation of bounded motion. However, the calculation of the critical configurations and the condition for bounded motion is more complicated and will be the subject of a future work.

Discussion and Conclusion

In this section we have generalised the Sundman's inequality study to the post-Newtonian approximation of the gravitational N-body problem, which suffices to establish the existence of bounded motion for the gravitational 2- and 3-body system in the same limit. The result is especially satisfactory for the 2-body problem. It will also be interesting for practical consideration if we can apply this approach to the post-Minkowskian approximation.

The results for systems with more than two bodies are not completely satisfactory. The difficulty encountered may be a reflection of the same problem in the full relativistic case; it, however, may also be due to the truncation of the complete problem. A more complete treatment of the post-Newtonian approximation may be related to the following question defined in the framework of Newtonian mechanics. If an autonomous Lagrangian system is defined in Newtonian space-time and the total energy and angular momentum are conserved (with the Cartesian coordinates as generalised coordinates), then does the generalised Sundman inequality, of the form $J^2 \leq 2I(H-U)$, hold for any such Lagrangian system? If not, can we determine the class of Lagrangians which satisfies such relation? We hope that the answer to this question can shed some light on improving the result in our study of the post-Newtonian N-body problem.

7.4 Summary

In this chapter and Appendix B we developed an inequality method to investigate restrictions imposed by symmetries on possible motion of relativistic few-body problems. Such restrictions may lead to interesting bounded motion or hierarchical orbital motion which is important in, say, studying mass transfer of a binary system for it may provide a possible relativistic alternative to the classical Roche lobe. Using this method we were able to establish some inequality relations for relativistic systems.

The power of the method has only been touched in this chapter. We hope to apply this approach to the theory of Dixon (1979) and establish some relations for the motion of the centre of mass the author defined. The post-Minkowskian approximation is also an interesting field to apply our approach.

I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, while the great ocean of truth lay all undiscovered before me. --- Sir Isaac Newton

CHAPTER 8 Future Work

The purpose of this thesis was to investigate hierarchical orbital motion and chaotic motion in both classical and relativistic gravitational few-body systems. The classical problem has been a well-known historical problem which cannot be solved analytically; whereas the relativistic problem concerned in this thesis is the first attempt at such an investigation. Therefore, it is inevitable that in a research such as this there remain many unanswered questions. Some (in particular the classical part) have existed for many years previously, others (relativistic part) have arisen during the course of this work. This final chapter is concerned with looking at some of the more important questions which may be solved in the near future following the methods suggested.

8.1 Classical Gravitational 3-Body Problems

In chapter 4, we have obtained inequalities stronger than Sundman's for the 3-body problem. Because of the role played by the moment of inertia tensor and ellipsoid in the deduction and Saari's (1984, 1987) result of the same problem in a flat N-body system, we feel strongly that similar relations exist for the general N-body problem. This study has therefore left us with several open questions which may only require some straightforward attack. However, the answer to these questions would greatly benefit the study of chapters 6 and 7. We propose the following two lines of future research.

On the one hand, we may extend the canonical transformation approach to the general N-body system. As is pointed out in the main body of the thesis, such an approach is direct and no relation of the system would be weakened in the course of deduction; however, difficulty may arise due to the large amount of algebra.

On the other hand, we may look for an alternative proof of the stronger inequalities, a general inequality proof similar to those given in appendix B. One may first look at the 3-body problem and then generalise the approach to systems with more bodies. In fact, the inequalities of appendix B was obtained by the present author in this way, based on the belief that the number '3' is nothing special in this context.

Once inequalities are obtained in this way, then they can be applied to relativistic systems because of the generality of the proof.

The research of chapter 5 has also left us with some future computational work and theoretical investigation. Firstly, we need to carry out longer time-scale numerical experiment to reach a more definite answer to the question of the importance of the tunnel-shaped stability surface, and to investigate larger sample of systems to test Roy's statistical stability conjecture. Secondly, it is desirable to carry out the systematic numerical experiments using the $(\mu\mu_3)$ instead of the $(\varepsilon_{23}\varepsilon_{32})$ parameters, because the distortion produced due to the transformation from the former to the latter parameters prevent us from looking at some of the numerical results with respect to the Hill-type stability. With his scheme, it will also be interesting to investigate the behaviour of the spatial motion of the 3-body system.

Following the numerical work of McKenzie & Szebehely (1981) on the circular restricted 3-body problem, it is worth investigating the stability and instability for the motion in the neighbourhood of the equilateral triangle point.

Finally, it will be useful if we could find a theoretical explanation for the phenomena observed from the experiments.

8.2 Relativistic Gravitational Few-Body Problems

In chapters 6 and 7, we have generalised the classical inequality approach on possible and forbidden motions into the framework of general relativity. Since this is the first attempt at such, many open questions are left for future researches. In addition to improving the relationships of appendix B in the light of the classical approach, we shall mention the following.

It is desirable to apply the method developed in this thesis to Dixon's (1979) formulation of the gravitational systems and tensorial conservation laws. The advantage of this theory is that the centre of mass was defined and the relationships obtained are coordinate-free. On more practical grounds, we hope to apply our approach to the post-Minkowskian approximation of general relativity. Compared with the post-Newtonian approximation, this is applicable to compact objects, and therefore of astrophysical importance.

Finally, we shall mention a particularly interesting problem, geodesic motion in the Kerr geometry, which may provide many important theoretical results. Since there is standard bounded motion in this problem, it will be interesting to apply our inequality method to this problem to test its validity. Due to the existence of independent extra integrals (cf. Toda Hamiltonian) nonlinear in the 4-momentum, the motion is integrable. It is well-known that such integrals do not correspond to any obvious symmetry (Killing vector), but are related to Killing tensors. However, if an approximation is made for this problem using the Kerr coordinates, then the Killing tensors may be lost, and thus the extra integrals may disappear (like the Toda Hamiltonian). In this way relativistic chaos would occur. Such an approach would also shed some light on the validity of approximation methods in general, which is often used in the study of relativistic problems.

APPENDIX A.

Elementary Number-Theoretic Results

In this Appendix we summarise some basic results of number theory relevant to the KAM theorem (see chapter 2) and a comprehensive understanding of chaos in general. These results are included in most introductory books to theory of numbers. We shall mention in particular two books by Baker (1975 and 1984), and the original works of Arnold (1963) and Moser (1962).

It is well-known that real numbers can be divided into rational and irrational numbers; in fact, even the ancient Greeks knew that $\sqrt{2}$ cannot be expressed as a fraction of two integers. However, it was not until 1844 that the theory of transcendental numbers was originated by Liouville, who showed that a class of numbers satisfies no algebraic equation with integer coefficients. The theory of numbers was not perfected until the end of the nineteenth century by Cantor.

Although the structure of quasi-periodic and chaotic solutions to nonintegrable systems is often said to be like that of rational and irrational numbers; it is the property of algebraic and transcendental numbers that is important to chaos and the KAM theorem. Briefly, a number is said to be algebraic if it is a zero of a polynomial with integer coefficients (eg. 0.3, $\sqrt{2}$, the golden section g); otherwise it is termed a transcendental number (eg. e, π, 0.1010010001...). The degree of the irreducible polynomial is called the degree of the algebraic number. We shall adopt the convention that by a rational p/q, we mean that p and q are relatively prime integers. We have the following results for numbers:

(1). If (p_1, q_1) and (p_2, q_2) are two pairs of relatively prime integral numbers respectively, then so are $((p_1q_2 \pm q_1p_2), (q_1q_2))$ and $((kq_1 \pm p_1), q_1)$, where k is an integer.

(2). The Dirichlet theorem (found in 1842) says that for any **real** number α and any integer Q there exist **some** integers p, q (not necessarily relatively prime) with $0<q<Q$ such that

$$|q\alpha - p| \le \frac{1}{Q} \ , \qquad \left|\alpha - \frac{p}{q}\right| \le \frac{1}{qQ}$$

and the two expressions are equivalent.

(2'). The Dirichlet theorem is usually interpreted in the following way, as a corollary of the theorem, showing that irrational numbers can be approximated arbitrarily closely by rationals. For any **real** number α, there exist **some** rationals p/q such that

$$|q\alpha - p| \to 0 \quad , \qquad \left|\alpha - \frac{p}{q}\right| \to 0 \qquad \text{as } q \to \infty.$$

This statement is true; however, it is valid in a broader sense (eg. q<Q is not needed), thus should be taken as an independent result. Moreover, the two expressions are no longer equivalent; in fact, more rationals satisfy the second expression. For example, one can allow q to be arbitrarily large such that the second expression is arbitrarily small, but the first expression is finite. An example was given in chapter 1 when α is a rational number.

(2"). There is another very important corollary of the Dirichlet theorem, showing how good the approximation of irrationals by rationals is. For any **irrational** number α, there exists **infinitely** many rationals p/q such that

$$|q\alpha - p| < \frac{1}{q} \quad , \qquad \left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}$$

and the two expressions are equivalent. We will see that the continued fraction algorithm gives the best possible construction of such a rational approximation.

Note that this corollary is not true for **rationals**; there are only **finitely** many rationals. In fact if α=a/b, then either p/q=a/b or p/q≠a/b; in the latter case, it is easy to show that

$$|q\alpha - p| \geq \frac{1}{b} \quad , \qquad \left|\alpha - \frac{p}{q}\right| \geq \frac{1}{qb} .$$

(3). On the other hand, Liouville established in 1844 a lower limit for the approximation of irrationals by rationals. Liouville's theorem says that, for any **algebraic** number α with degree n>1, there exists c=c(α) such that

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^n}$$

for **all** rationals p/q. This theorem led to the first construction of transcendental numbers. The above inequality is still an exceptional case, because it was shown by Cantor in 1874 that almost all numbers (in the sense of Lebesgue measure) are transcendental and the set of all algebraic numbers is countable.

Liouville's theorem was later improved as

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^{\lambda}}\ , \quad \lambda > \sqrt{2n}$$

and $\lambda>2$ is the best possible.

However, when we have a set of numbers a similar inequality is satisfied by almost all sets of numbers. These are the conditions closely related to the KAM theorem. Almost all vectors $\alpha=(\alpha_1, ..., \alpha_k)$ satisfy the inequality

$$|q_1\alpha_1 + \ldots + q_k\alpha_k| \geq \frac{c}{Q^{k+1}}\ , \quad Q = |q_1| + \ldots + |q_k|.$$

for a constant $c=c(\alpha)$ and **all** integral vectors $q=(q_1, ..., q_k)\neq 0$.

(4). A corollary of Minkowski's theorem may be regarded as a generalisation of the Dirichlet theorem. It says that if $(\alpha_1, ..., \alpha_k)$ is a set of any real numbers and if $Q>0$ then there exist integers p and $(q_1, ..., q_k)$, not all zero, such that $|q_i| < Q$ $(i=1, ..., k)$ and

$$|q_1\alpha_1 + \ldots + q_k\alpha_k - p| \leq \frac{1}{Q^k}\ .$$

(5). The continued fraction algorithm sets up an 1-1 correspondence between all irrational α and all infinite sets of integers $(a_0, a_1, \ldots)$ with $(a_1, \ldots)$ all positive. It also sets up an 1-1 correspondence between all rationals α and all finite sets of integers $(a_0, a_1, ..., a_k)$ with $(a_1, ..., a_{k-1})$ all positive and $a_k\geq 2$. We shall use the following notations for a continued fraction

$$\alpha = [a_0, a_1, a_2, \ldots\ ] \equiv a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}} \qquad \alpha = [a_0, a_1, \ldots, a_k] \equiv a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_k}}}\ .$$

It is also conventional to call the integers $(a_0, a_1, \ldots)$ the partial quotients of α; whereas the numbers $\alpha_n=[a_0, a_1, ..., a_n]=p_n/q_n$ are known as the complete quotients of α. We also define α_{n+1} by $\alpha=[a_0, a_1, ..., a_n, \alpha_{n+1}]$.

A continued fraction represents a quadratic irrational iff it is ultimately periodic, that is, iff the partial quotients satisfy $a_{m+n}=a_n$ for some positive integer m and for sufficiently large n.

If we define $p_0=a_0$, $q_0=1$ and $p_1=a_1a_0+1$, $q_1=a_1$, then the complete quotients can be generated recursively by the equations

$$\begin{cases} p_n = a_n p_{n-1} + p_{n-2} \\ q_n = a_n q_{n-1} + q_{n-2}. \end{cases}$$

Based on this equation, one can show that α lies between p_n/q_n and p_{n+1}/q_{n+1}. We also have

$$p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1}$$

$$\left| \frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}} \right| = \frac{1}{q_n q_{n+1}}.$$

It follows that

$$\left| \alpha - \frac{p_n}{q_n} \right| \leq \frac{1}{q_n q_{n+1}}$$

and so certainly p_n/q_n converges to α as $n \to \infty$; and p_n/q_n is called the convergent of α.

In fact the following stronger inequality hold for any convergent of α

$$\frac{1}{(a_{n+1}+2)q_n^2} < \left| \alpha - \frac{p_n}{q_n} \right| < \frac{1}{a_{n+1} q_n^2},$$

and convergents are indeed the **best** approximations to α in the sense that, if (p, q) are integers with $0<q< q_n$ then

$$|q\alpha - p| \geq |q_n \alpha - p_n| \quad \Rightarrow \quad \left| \alpha - \frac{p}{q} \right| \geq \left| \alpha - \frac{p_n}{q_n} \right| .$$

One can also show that if a rational p/q satisfies

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{2q^2},$$

then it is a convergent to α.

(6). We see from above that the continued fraction algorithm constructs the convergents p/q of α, each of them satisfies

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^2} .$$

In fact we have a series of similar results. For example, at least one of any two consecutive convergents, say p_n/q_n and p_{n+1}/q_{n+1}, satisfies

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2} .$$

Moreover, at least one of any three consecutive convergents satisfies

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2} .$$

Therefore, for any irrational number there are always infinitely many rationals satisfying each of the three inequalities. In fact the constant $1/\sqrt{5}$ of the third inequality is the **best possible** as can be verified by taking the golden section number $g = (1+ \sqrt{5})/2 = [1, 1, 1, ...]$. This number is the most irrational number in the sense that the approximation of it by rationals is the worst out of all irrational numbers.

If one excludes all irrationals whose continued fractions have all but finitely many partial quotients equal to 1, then the best possible constant becomes $1/\sqrt{8}$. There is an infinite sequence of such constants which tends to 1/3.

(7). Using the above property of continued fractions and convergents, it is easy to prove that a periodic function of a single variable can only have one principal period. This proof was first given by Jacobi (see Forsyth, 1893, P200).

(8).

The Diophantine equation $x^2=2y^4-1$ has exactly two solutions in positive integers, namely, (1, 1) and (239, 13). However, the other Diophantine equation $x^2-2y^2=1$ has infinitely many positive integer solutions. It is easy to verify that (3, 2) is one such solution to this equation, and if (x, y) is a solution then new solutions (x^*, y^*) may be constructed using the following formulae

$$\begin{cases} x^* = 3x + 4y \\ y^* = 2x + 3y. \end{cases}$$

This has some similarity with the concept local integrals of dynamical systems. In the space of positive integers, the original equation may be regarded as a local integral of the linear mapping which is valid only for some points of the space.

APPENDIX B.

Generalised Cauchy's Inequality and Sundman's Inequality

This Appendix includes the most general proof of Cauchy's and Sundman's inequalities used in chapter 4, and some generalisation of them. The inequalities may seem basic; but they were constructed and proved completely by the author, because they are needed in his inequality approach proposed in chapters 6 and 7 to the investigation of bounded motion in general relativity. The classical book on inequality by Hardy *et al* (1934) may be a good reference for further generalisation of the inequalities given here. The main inequalities used in chapter 4, 6 and 7 are

$$2\sum_{i,k}(A^{[i}B^{k]})^2 = \frac{1}{2}\sum_{i,k}(A^iB^k - A^kB^i)^2$$

$$= \left(\sum_i (A^i)^2\right)\left(\sum_i (B^i)^2\right) - \left(\sum_i A^iB^i\right)^2 \qquad \text{(B1)}$$

$$\left(\sum_i A^iB^i\right)^2 \le \left(\sum_i (A^i)^2\right)\left(\sum_i (B^i)^2\right) \qquad \text{(B2)}$$

$$2\sum_{i,k}(A^{[i}B^{k]})^2 = \frac{1}{2}\sum_{i,k}(A^iB^k - A^kB^i)^2$$

$$\le \left(\sum_i (A^i)^2\right)\left(\sum_i (B^i)^2\right) \qquad \text{(B3)}$$

$$\left(\sum_{i,b} A^i_bB^i_b\right)^2 \le \left(\sum_{i,b} (A^i_b)^2\right)\left(\sum_{i,b} (B^i_b)^2\right) \qquad \text{(B4)}$$

$$2\sum_{i,k}\left(\sum_b A^{[i}_bB^{k]}_b\right)^2 = \frac{1}{2}\sum_{i,k}\left(\sum_b A^i_bB^k_b - A^k_bB^i_b\right)^2$$

$$\le \left(\sum_{i,b} (A^i_b)^2\right)\left(\sum_{i,b} (B^i_b)^2\right) \qquad \text{(B5)}$$

where $C^{[ik]} \equiv \frac{1}{2}(C^{ik} - C^{ki})$; $i, k = 1, \ldots, m$; $b, d = 1, \ldots, n$.

Equation (B2) is usually called Cauchy's inequality, while (B3) is a generalisation of Sundman's inequality. These two equations follow immediately from the basic relation (B1), whose proof is straightforward. Equations (B4) and (B5) are further generalisation from quantities at one point to quantities at many points. (B4) follows immediately from (B2), but equation (B5) needs a more involved proof.

[Proof of (B5)]: The proof of equation (B5) is divided into three steps. We prove two basic inequalities in the first two steps.

$$(\,*\,)\quad \sum_i\left(\sum_b C^i_b\right)^2 \le \left(\sum_b \sqrt{\sum_i (C^i_b)^2}\right)^2$$

$$\text{LHS} = \sum_i\left(\sum_b C^i_b\right)\left(\sum_d C^i_d\right) = \sum_{b,d}\sum_i C^i_b C^i_d \le \sum_{b,d}\left|\sum_i C^i_b C^i_d\right|$$

$$\text{RHS} = \sum_b\sqrt{\sum_i (C^i_b)^2}\,\sum_d\sqrt{\sum_i (C^i_d)^2} = \sum_{b,d}\sqrt{\left(\sum_i (C^i_b)^2\right)\left(\sum_i (C^i_d)^2\right)}$$

$$\text{but (B2)} \Leftrightarrow \left|\sum_i C^i_b C^i_d\right| \le \sqrt{\left(\sum_i (C^i_b)^2\right)\left(\sum_i (C^i_d)^2\right)}$$

$$(\,**\,)\quad \sum_{i,k}\left(\sum_b C^{ik}_b\right)^2 \le \left(\sum_b \sqrt{\sum_{i,k} (C^{ik}_b)^2}\right)^2$$

Equation (**) follows immediately from equation (*) if we substitute the index i by the indices (i,k). Now we can give the proof of equation (B5) itself.

$$\text{If } C^{ik}_b = A^{[i}_b B^{k]}_b \text{ , then}$$

$$2\sum_{i,k}\left(\sum_b A^{[i}_b B^{k]}_b\right)^2 \le 2\left(\sum_b\sqrt{\sum_{i,k}(A^{[i}_b B^{k]}_b)^2}\right)^2 \qquad \text{(use **)}$$

$$= \left(\sum_b\sqrt{2\sum_{i,k}(A^{[i}_b B^{k]}_b)^2}\right)^2$$

$$\le \left(\sum_b\sqrt{\left(\sum_i (A^i_b)^2\right)\left(\sum_i (B^i_b)^2\right)}\right)^2 \qquad \text{(use B3)}$$

$$= \left(\sum_b\sqrt{\sum_i (A^i_b)^2}\sqrt{\sum_i (B^i_b)^2}\right)^2$$

$$\le \left\{\sum_b\left(\sqrt{\sum_i (A^i_b)^2}\right)^2\right\}\left\{\sum_b\left(\sqrt{\sum_i (B^i_b)^2}\right)^2\right\} \qquad \text{(use B2)}$$

$$= \left(\sum_{i,b}(A^i_b)^2\right)\left(\sum_{i,b}(B^i_b)^2\right)$$

Finally, we note that a positive function different from point to point can be put into the equations without changing them. The sum over points (a or b) can be replaced by an integral without changing the relations, but the measure element must be positive everywhere, which is usually the case. The last restriction is from the right hand side of the relation, which is not required by the left hand side. Also note that the use of the energy condition in chapter 7, that is the positiveness of (p+ρ) is not required by the inequalities but rather the requirement to collaborate the splitting of potential and kinetic energy.

There are many other possibilities in addition to the above five equations, which are the most immediate generalisations. Here we give some more supplementary ones used in chapter 7, their validity should be readily seen by the appropriate understanding of the above five basic equations.

$$\left.\begin{array}{l} [g(\bar{A},\bar{B})]^2 \leq [g(\bar{A},\bar{A})]\cdot[g(\bar{B},\bar{B})] \\ \left(\sum_i \pm A^iB^i\right)^2 \leq \left(\sum_i (A^i)^2\right)\left(\sum_i (B^i)^2\right) \end{array}\right\} \qquad \text{(B6)}$$

The first equation is valid for any positive definite metric g, while the second, although it may not be very useful by itself, is important when combined with the following relations. All of the inequalities given in this appendix still hold no matter whether it is a '+' or '-' preceding any term on the left hand side.

$$\left(\sum_{i,b} \pm A^i_bB^i_b \pm C^i_bD^i_b\right)^2 \leq \left(\sum_{i,b}(A^i_b)^2 + (C^i_b)^2\right)\left(\sum_{i,b}(B^i_b)^2 + (D^i_b)^2\right) \qquad \text{(B7)}$$

$$2\sum_{i,k}\left(\sum_b \pm A^{[i}_bB^{k]}_b \pm C^{[i}_bD^{k]}_b\right)^2 \leq \left(\sum_{i,b}(A^i_b)^2 + (C^i_b)^2\right)\left(\sum_{i,b}(B^i_b)^2 + (D^i_b)^2\right) \qquad \text{(B8)}$$

These two equations are useful because the angular momentum in relativity involve the sum of several skew-symmetric terms. We also note that one source of the flexibility (in addition to the fact that stronger equations exist) of these inequalities is that we can interchange the positions of A and B, or C and D on the right hand side of the above two equations. We must choose this from the physical content carried by them, that is, to put the same kind of quantities together. Another flexibility is possible in these two equations because, for example, we can always multiply C with a big number and divided D by the same number, hence finally omit the term involving D. Though the final relation is weaker, this is a very useful technique when there are some complicated but less important quantities involved in the angular momentum. The validity of these two equations is more apparent if we put them in a more general form, viz.

$$\left(\sum_{i,b} \pm A^i_b B^i_b \pm \sum_{k,d} C^k_d D^k_d\right)^2 \leq \left(\sum_{i,b} (A^i_b)^2 + \sum_{k,d} (C^k_d)^2\right)\left(\sum_{i,b} (B^i_b)^2 + \sum_{k,d} (D^k_d)^2\right)$$

$$2\sum_{i,k}\left(\sum_{b} \pm A^{[i}_b B^{k]}_b \pm \sum_{d} C^{[i}_d D^{k]}_d\right)^2 \leq \left(\sum_{i,b} (A^i_b)^2 + \sum_{k,d} (C^k_d)^2\right)\left(\sum_{i,b} (B^i_b)^2 + \sum_{k,d} (D^k_d)^2\right)$$

Finally, in the attempt to construct better inequalities, an equality for the many-point problem like that of the one-point problem, equation (B1), should be very interesting and useful.

APPENDIX C.

Transformation of the Critical Stability Surfaces from ($\mu\mu_3$) to (ε_{23} ε_{32}) Space and Attractors

The critical stability surface (cf. chapter 5) is usually calculated first in the O-$\alpha\mu\mu_3$ space by solving a set of algebraic equations, then transformed into the O-$\alpha\varepsilon_{23}\varepsilon_{32}$ space (for more details see Walker et al, 1980). The transformation from the former space to the latter one is carried out by an iterative procedure, which in theory defines an iterative discrete mapping like what is discussed in chapter 2. The purpose of this appendix is to view the converging and diverging process of the iterative procedure as the generic behaviour of mapping, which is nowadays a hot topic due to the progress made in understanding chaos.

Since the function α_c=F(μ, μ_3) is not explicit, the transformation to the (ε_{23} ε_{32}) parameter space cannot be carried out explicitly either. Instead an iterative procedure must be used (see Walker et al , 1980), which can be viewed as a 3-dimensional mapping (non-volume preserving) with two parameters, viz.

$$\begin{cases} \mu^{\cdot} = 0.5 - 0.5\sqrt{1 - \dfrac{4\,\varepsilon_{23}}{\alpha_{23}}} \\ \mu^{\cdot}_3 = \dfrac{\varepsilon_{32}}{(\alpha_{23})^3} \\ \alpha^{\cdot}_{23} = f(\mu^{\cdot}, \mu^{\cdot}_3) \end{cases} \qquad \varepsilon_{23} \text{ and } \varepsilon_{32} \text{ are parameters}$$

where α and α_{23} are related to each other by the very simple equation (5.3). To calculate α_c for a given (ε_{23} ε_{32}) pair, we choose an estimated value for it, say 0.5, and then carry out the above mapping until when a substitution of the calculated (μ μ_3 α) into equation (5.7) gives values of (ε_{23} ε_{32}) sufficiently close to the given values. The error in α_c of this calculation cannot be controlled directly, but can only be controlled through the ε's.

The procedure was successful in their (Walker et al , 1980) calculation of the circular case, since the critical stability surface possesses a monotonic property (see Fig. 5.6a). But in the elliptical case the procedure diverges around many points, because of the complexity of the 'tunnel-shaped' critical stability surface due to the eccentricities (see Fig. 5.6b). This can be explained by looking at the process of the procedure if we note that

$$\left.\begin{array}{l} d\varepsilon_{23} = 2\,\mu(1-\mu)\,\alpha_{23}d\alpha_{23} + (1-2\mu)\,\alpha_{23}^2\,d\mu \\ d\varepsilon_{32} = 3\,\mu_3\,\alpha_{23}^2\,d\alpha_{23} \quad + \alpha_{23}^3\,d\mu_3 \end{array}\right\} \xrightarrow{d\varepsilon=0} \left\{\begin{array}{l} d\alpha_{23}\cdot d\mu < 0 \\ d\alpha_{23}\cdot d\mu_3 < 0 \end{array}\right.$$

which produces the typical iterative route shown in Fig. C1.

In fact this divergence is a very general phenomenon concerning mappings. The procedure may also be viewed as a 1-dimensional mapping with two parameters,

$$\alpha_{23}^{\cdot} = f(\mu^{\cdot},\, \mu_3^{\cdot}) = f\left(0.5 - 0.5\sqrt{1 - \frac{4\ \varepsilon_{23}}{\alpha_{23}}}\ ,\ \frac{\varepsilon_{32}}{(\alpha_{23})^3}\right) = g(\alpha_{23}\ ;\ \varepsilon_{23}\,,\,\varepsilon_{32})\,.$$

In order that the mapping converges, it must possess 'attractors' dense everywhere in the space. However, it is well known that even the property of a simple quadratic mapping can be very complicated. There is no guarantee for convergence without carefully studying the mapping, which is defined implicitly here. In the study of chapter 5, the analytical property of the mapping is not of importance, because we are aiming only at the calculation of α_c, for which purpose an interpolation method can always be used as supplementary.

This is another example which shows the significance of eccentricity. As is well-known, in principle an arbitrarily small eccentricity in the orbit of the primaries will change completely the nature of the restricted 3-body problem, viz., the existence/nonexistence of the Jacobian integral. Eccentricity is also the most important parameter characterising regular and irregular motion, as was evident from the numerical investigation of chapter 5.

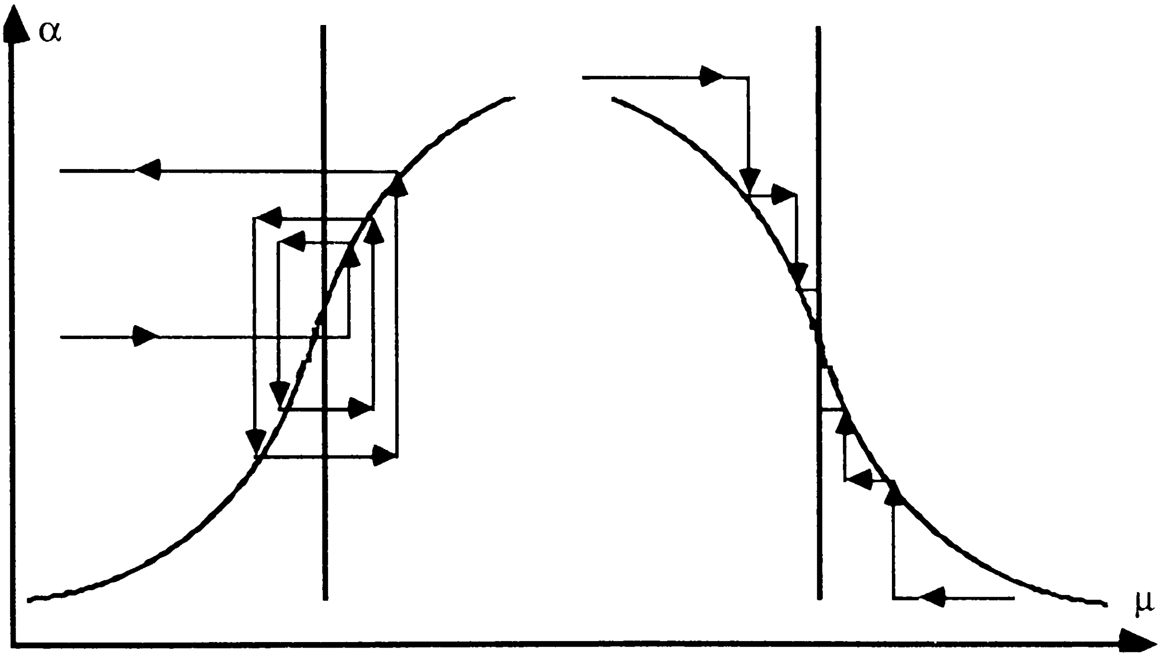


Fig. C1 Convergence and divergence of the iterative procedure of calculating α_c for a given (ε_{23} ε_{32}) pair. Only a section with μ_3=const. is shown in the diagram. The procedure converges in a region of the critical stability surface like that on the right side; it diverges in a region like that on the left side.

APPENDIX D.
Proof of Equations (6.2) and (6.3)

[Proof of (6.2a, b)]: We only need to prove the first relation of equation (6.2a); all the others follow immediately from this relation by a use of equation (6.1). The following proof is based on the most general properties of scalar, vector, and tensor; Lie and covariant derivatives, and the Leibniz rule these derivatives obey; symmetry feature of connection, compatibility of connection and metric. In contrast to a conventional component proof, we can see the more basic assumptions on which the the results really depend.

$(\mathcal{L}_{\bar\xi}\, g)(\bar A, \bar B)$	
$= \mathcal{L}_{\bar\xi}[g(\bar A, \bar B)] - g(\mathcal{L}_{\bar\xi}\bar A, \bar B) - g(\bar A, \mathcal{L}_{\bar\xi}\bar B)$	Leibniz rule for $\mathcal{L}_{\bar\xi}$
$= \nabla_{\bar\xi}[g(\bar A, \bar B)] - g(\mathcal{L}_{\bar\xi}\bar A, \bar B) - g(\bar A, \mathcal{L}_{\bar\xi}\bar B)$	$\mathcal{L}_{\bar\xi} f = \nabla_{\bar\xi} f$
$= \quad g(\nabla_{\bar\xi}\bar A, \bar B) + g(\bar A, \nabla_{\bar\xi}\bar B)$	$\left\{\begin{array}{l}\text{Leibniz rule for } \nabla_{\bar\xi} \\ \nabla g \equiv 0 \text{ (compatibility)}\end{array}\right.$
$\quad - g(\mathcal{L}_{\bar\xi}\bar A, \bar B) - g(\bar A, \mathcal{L}_{\bar\xi}\bar B)$	
$= g(\nabla_{\bar A}\bar\xi, \bar B) + g(\bar A, \nabla_{\bar B}\bar\xi)$	Symmetry of ∇
$= (\nabla_{\bar A}\tilde\xi)(\bar B) + (\nabla_{\bar B}\tilde\xi)(\bar A)$	$g(\nabla_{\bar A}\bar\xi, \quad) \equiv \nabla_{\bar A}\tilde\xi$
$= (\nabla_{\bar A}\bar\xi)(\tilde B) + (\nabla_{\bar B}\bar\xi)(\tilde A)$	

[**Proof of (6.3a, b)**]: Only equation (6.3a) needs a non-trivial proof. The following proof is independent of coordinates. Using the compatibility condition between connection and metric, $\nabla g=0$, we obtain

$$\nabla_{\bar{P}}[g(\bar{P},\bar{\xi})] = g(\nabla_{\bar{P}}\bar{P},\bar{\xi}) + g(\bar{P},\nabla_{\bar{P}}\bar{\xi}) \quad \Leftarrow \nabla g \equiv 0 \text{ (compatibility)}$$

where only the second term need further calculation,

$$\begin{array}{l||l}
g(\bar{P},\nabla_{\bar{P}}\bar{\xi}) = g(\bar{P},\nabla_{\bar{\xi}}\bar{P} + \mathcal{L}_{\bar{P}}\bar{\xi}) & \nabla_{\bar{P}}\bar{\xi} - \nabla_{\bar{\xi}}\bar{P} = \mathcal{L}_{\bar{P}}\bar{\xi} \text{ (symmetric } \nabla)\\
= g(\bar{P},\nabla_{\bar{\xi}}\bar{P}) + g(\bar{P},\mathcal{L}_{\bar{P}}\bar{\xi}) & \\
= g(\bar{P},\nabla_{\bar{\xi}}\bar{P}) - g(\bar{P},\mathcal{L}_{\bar{\xi}}\bar{P}) & \mathcal{L}_{\bar{P}}\bar{\xi} = -\mathcal{L}_{\bar{\xi}}\bar{P}\\
\begin{array}{l} = \frac{1}{2}\{\nabla_{\bar{\xi}}\, g(\bar{P},\bar{P}) - (\nabla_{\bar{\xi}}\, g)(\bar{P},\bar{P})\}\\ \quad -\frac{1}{2}\{\mathcal{L}_{\bar{\xi}}\, g(\bar{P},\bar{P}) - (\mathcal{L}_{\bar{\xi}}\, g)(\bar{P},\bar{P})\}\end{array} & \left\{\begin{array}{l}\text{Leibniz rule for } \nabla_{\bar{\xi}}\\ \text{Leibniz rule for } \mathcal{L}_{\bar{\xi}}\end{array}\right.\\
= \frac{1}{2}(\mathcal{L}_{\bar{\xi}}\, g)(\bar{P},\bar{P}) & \nabla g \equiv 0 \;;\; \nabla_{\bar{\xi}} f \equiv \mathcal{L}_{\bar{\xi}} f
\end{array}$$

REFERENCES AND BIBLIOGRAPHY

Ablowitz, M.J. & Ramani, A. & Segur, H. (1980): *J. Math. Phys.*, **21** (No. 4), 715-721.

Abraham, Ralph & Marsden, J.E. (1966, 1978): Foundations of Mechanics, Benjamin/Cummings.

Aksenov, E.P. (1979): The Doubly Averaged Elliptical Restricted 3-Body Problem, *Sov. Astron.*, **23** (No. 2), 236-243.

Alfriend, K.T. & Richardson, D.L. (1973): Third and Fourth Order Resonances in Hamiltonian Systems, *Celest. Mech.*, **7**, 408-420.

Armenti, A. (1972): A Classification of Particle Motions in Newtonian Mechanics and General Relativity, *Celest. Mech.*, **6**, 383-415.

Arnold, Vladimir Igorevich (1963): Small Denominators, II. Proof of Kolmogorov's Theorem on the Conservation of Quasi-Periodic Motions under Small Perturbations of the Hamiltonian, *Russ. Math. Surveys*, **18** (No. 5), 9-36.

--- (1963): Small Denominators and Problems of Stability of Motion in Classical and Celestial Mechanics, *Russ. Math. Surveys*, **18** (No. 6), 85-191.

--- (1978, 1989): Mathematical Methods of Classical Mechanics, Springer-Verlag.

--- (1983, 1988): Geometrical Methods in the Theory of Ordinary Differential Equations, Springer-Verlag.

Ash, M.E. (1976): Doubly Averaged Effect of the Moon and Sun on a High Altitude Earth Satellite Orbit, *Celest. Mech.*, **14**, 209-238.

Bailey, I. & Israel, W. (1980): *Ann. of Phys.*, **130**, 188.

Baker, Alan (1975): Transcendental Number Theory, Cambridge University Press.

--- (1984, 1986): A Concise Introduction to the theory of Numbers, Cambridge University Press.

Barnett, W.A. & Berndt, E.R. & White, H. (1988): Dynamic Econometric Modelling, proceedings of the Third International Symposium in Economic Theory and Econometrics, Cambridge University Press.

Barrar, R.B. (1970): Convergence of the Von Zeipel Procedure, *Celest. Mech.*, **2**, 494-504.

Barrow, John D. (1988, 1990): The World Within The World, Oxford University Press.

Baugher, J.F. (1988): The Space-Age Solar System, John Wiley & Sons.

Baumgardt, C. (1952): Johannes Kepler: Life and Letters, Victor Gollancz.

Baumgarte, J. (1973): Stabilisation of the Differential Equations of Keplerian Motion, in

Tapley & Szebehely (ed. 1973), 38-44.

--- & Stiefel, E.L. (1972): in Lecture Notes in Mathematics, Springer-Verlag, 362, 207-236.

--- & Stiefel, E.L. (1974): Stabilization by Manipulation of the Hamiltonian, *Celest. Mech.*, **10**, 71-85.

Bell, E.T. (1937): Men of Mathematics, Victor Gollancz.

Berge, P. & Pomean, Y. & Vidal, C. (1984): Order within Chaos, John Wiley & Sons.

Berry, M.V. & Percival, I.C. & Weiss, N.O. (ed. 1987): Dynamical Chaos, *Proc. Roy. Soc. L.*, **A413**, 1-199.

--- & Tabor, M. (1976): *Proc. Roy. Soc. L.*, **A349**, 101.

--- (1978): Regular and Irregular Motion, in Jorna (ed.) (1978), 16-120.

Binney, J. & Tremaine, S. (1987): Galactic Dynamics, Princeton University Press.

Birkhoff, George D. (1927): Dynamical Systems, American Mathematical Society, New York.

--- (1913): Proof of Poincare's Geometric Theorem, *Trans. Amer. Math. Soc.*, **14**, 14-22.

--- (1917): Dynamical Systems with Two Degrees of Freedom, *Trans. Amer. Math. Soc.*, **18**, 199-300.

--- (1922): Surface Transformations and Their Dynamical Applications, *Acta. Math.*, **43**, 1-119.

--- (1925): An Extension of Poincare's Last Geometric Theorem, *Acta. Math.*, **47**, 297-311.

Birrell, N.D. & Davies, P.C.W. (1982): Quantum Fields in Curved Space, Cambridge University Press.

Bishop, R.L. & Goldberg, S.I. (1968): Tensor Analysis on Manifolds, Macmillan.

Blagg, M.A. (1913): On a Suggested Substitute for Bode's Law, *Mon. Not. Roy. Ast. Soc.*, **73**, 414-422.

Born, M. (1924, 1927): The Mechanics of the Atom, Bell, London.

--- (1956): Physics in My Generation, Pergamon Press.

--- (1971): The Born-Einstein Letters, MacMillan.

Bountis, T. & Segur, H. & Vivaldi, F. (1982): Integrable Hamiltonian Systems and the Painleve Property, *Phys. Rev.*, **A25** (No. 3), 1257-1264.

Bozis, G. (1976): Zero Velocity Surfaces for the General Planar 3-Body Problem, *Astrophys. Space Sci.*, **43**, 355-368.

Brillouin, Leon (1964): Scientific Uncertainty and Information, Academic Press.

Broucke, R. (1971): Periodic Collision Orbits in the Elliptic Restricted 3-Body Problem, *Celest. Mech.*, **3**, 461-477.

--- (1979): Simple Non-integrable Systems with Two Degrees of Freedom, in

Szebehely, (ed.) (1979).

--- (1980): Kepler Equation and Strange Attractor, *J. Astronautical Sci.*, **28**, 255-265.

Brouwer, D. (1937): On the Accumulation of Errors in Numerical Integration, *Astron. J.*, **46**, 149-153.

--- & Clemence, G.M. (1961): Methods of Celestial Mechanics, Academic Press.

Brumberg, V.A. & Kovalevsky, J. (1986): Unsolved Problems of Celestial Mechanics, *Celest. Mech.*, **39**, 133-140.

Buchler, J.R. & Perdang, J.M. & Spiegel, E.A. (ed. 1985): Chaos in Astrophysics, Reidel.

Burd, A.B. & Buric, N. & Ellis, G.F.R. (1990): A Numerical Analysis of Chaotic Behaviour in Bianchi IX Models, *General Relativity and Gravitation*, **22** (3), 349-363.

Buric, N. & Percival, I.C. & Vivaldi, F. (1990): Critical Function and Modular Smoothing, *Nonlinearity*, **3**, 21-37.

Campbell, D.K. (1989): Nonlinear Science, in Cooper (ed. 1989), 218-262.

Campbell, J.A. & Jefferys, W.H. (1970): Equivalence of the Perturbation Theories of Hori and Deprit, *Celest. Mech.*, **2**, 467-473.

Caranicolas, N. & Vozikis, C.H. (1987): Chaos in a Quartic Dynamical Model, *Celest. Mech.*, **40**, 35-49.

Carmeli, M. (1982): Classical Fields: General Relativity and Gauge Theory, Wiley-Interscience, New York.

Carter, B. (1968): Global Structure of the Kerr family of Gravitational Fields, *Phys. Rev.*, 174, 1559-1571.

--- (1968): Hamilton-Jacobi and Schrodinger Separable Solutions of Einstein's Equations, *Commun. Math. Phys.*, **10**, 280-310.

--- (1969): Killing Horizons and Orthogonally Transitive Groups in space-Time, *J. Math. Phys.* **10**, 70-81.

--- (1970): The Commutation Property of a Stationary Axisymmetric System, *Commun. Math. Phys.*, **17**, 233-238.

--- (1971): Axisymmetric Black Hole Has Only Two Degrees of Freedom, *Phys. Rev. Lett.*, **26**, 331-332.

Casasayas, J. & Jorba, A. & Nunes, A. (1988): Qualitative Study of Motion Under the Potentials $V(r)=-ar^{-\alpha}$, *Celest. Mech.*, **42**, 129-139.

Chandrasekhar, S. (1983): The Mathematical Theory of Black Holes, Oxford University Press.

--- (1989): The 2-Centre Problem in General Relativity, *Proc. Roy. Soc. Lond.*, **A421**, 227-258.

Channell, P.J. & Scovel, C. (1990): Symplectic Integration of Hamiltonian Systems, *Nonlinearity*, **3**, 231-259.

Chapsiadis, A. & Michalodimitrakis (1988): *Celest. Mech.*, **41**, 53-63.

Chirikov, Boris V. (1979): A Universal Instability of Many-Dimensional Oscillator Systems, *Phys. Reports*, 52, 265-379.

Choi, J.S. & Tapley, B.D. (1973): An Extended Canonical Perturbation Method, *Celest. Mech.*, **7**, 77-90.

Choquet-Bruhat, Y. & DeWitt-Morette, C. & Dillard-Bleick, M. (1977): Analysis, Manifolds, and Physics, North-Holland.

Chow, S.N. & Hale, J.K. (1982): Methods of Bifurcation Theory, Springer-Verlag.

Cohen, C.J. & Hubbard, B.C. (1973): Planetary Elements for 10 000 000 Years, *Celest. Mech.*, **7**, 438-448.

Contopoulos, G. (1963): *Astrophys. J.*, **138**, 1297-1305.

--- (1971): *Astron. J.*, **76**, 147-156.

--- (1973): Problems of Stellar Dynamics, in Tapley & Szebehely (ed. 1973), 177-191.

--- (1986): Bifurcations in Systems of Three Degrees of Freedom, *Celest. Mech.*, **38**, 1-22.

--- (1987): Large Degree Stochasticity in a Galactic Model, *Astron. Astrophys.*, **172**, 55-66.

--- (1988): Nonuniqueness of Families of Periodic Solutions in a Four Dimensional Mapping, *Celest. Mech.*, **44**, 393-409.

--- (1990): Chaos in Two-Fixed-Black-Hole Problem, preprint.

Conway, B.A. (1986): *Celest. Mech.*, **39**, 199-211.

--- (1988): in Roy (ed. 1988).

Cooper, N.G. (ed.) (1989): From Cardinals to Chaos, Reflections on the Life and Legacy of Stanislaw Ulam, Cambridge University Press.

Courant, R. & Hilbert, D. (1924, 1953): Methods of Mathematical Physics, Interscience.

Cvitanovic, D. (1984): Universality in Chaos, Adam Hilger, Bristol.

Damour, T. (1987): The Problem of Motion in Newtonian and Einstein Gravity, in Hawking & Israel (ed. 1987).

Danby, J.M.A. (1962): Fundamentals of Celestial Mechanics, Macmillan.

--- (1964): Stability of the Triangular Points in the Elliptic Restricted Problem of Three Bodies, *Astron. J.* **69**, 165-172.

--- (1964): The Stability of the Triangular Lagrangian Points in the General Problem of Three Bodies, *Astron. J.* **69**, 294-296.

--- (1987): *Celest. Mech.*, **40**, 303-312.

Davies, Paul C.W. (ed.) (1989): The New Physics, Cambridge University Press.

Davies, T.V. & James E.M. (1966): Nonlinear Differential Equations, Addison-Wesley.

Dean, R.A. (1966): Elements of Abstract Algebra, John Wiley & Sons.

Deprit, A. (1969): Canonical Transformations Depending on a Small Parameter, *Celest. Mech.*, **1**, 12-30.

--- & Henrard, J. (1969): Birkhoff's Normalization, *Celest. Mech.*, **1**, 222-251.

Dermott, S.F. & Malhotra, R. & Murray, C.D. (1988): Dynamics of the Uranian and Saturnian Satellite Systems: A Chaotic Route to Melting Miranda?, *Icarus*, **76**, 295-334.

Devaney, R.L. (1987): Chaotic Dynamical Systems, Addison-Wesley.

DeWitt, B.S. & Brehme, R.W. (1960): *Ann. of Phys.*, **9**, 220.

DeWitt, M.C. (1979): Celestial Mechanics, Quantum Mechanics and Path Integration, in Szebehely, (ed.) (1979).

Dixon, W.G. (1978): Special Relativity: the Foundation of Macroscopic Physics, Cambridge University Press.

--- (1979): Extended Bodies in General Relativity: Their Description and Motion, In Ehlers (ed. 1979) 156-219.

Dodson, M.M. & Vickers, J.A.G. (ed.) (1989): Number Theory and Dynamical Systems, London Mathematical Society Lecture Note Series 134, Cambridge University Press.

Dolan, P. & Kladouchou, A. & Card, C. (1989): On the Significance of Killing Tensors, *General Relativity and Gravitation*, **21**, 427-437.

Dormand, J.R. & Woolfson, M.M. (1989): The Origin of the Solar System: the Capture Theory, Ellis Horwood.

Dray, T. & Padmanabhan, T. (1989): Conserved Quantities from Piecewise Killing Vectors, *General Relativity and Gravitation*, **21**, 741-745.

Dysthe, K.B. & Gudmestad, O.T. (1975): On Resonance and Stability of Conservative Systems, *J. Math. Phys.*, **16** (No. 1), 56-64.

Easton, R. (1971): Some Topology of the 3-Body Problem, *J. Diff. Equations*, **10**, 371-377.

--- (1975): Some Topology of the n-Body Problems, *J. Diff. Equations*, **19**, 258-269.

--- (1979): Perturbed Twist Maps, Homoclinic Points and Ergodic Zones, in Szebehely, (ed.) (1979).

Ehlers, J. (ed.) (1979): Isolated Gravitating Systems in General Relativity, North-Holland.

Eminhizer, C.R. & Helleman, R.H.G. & Montroll, E.W. (1976): On a Convergent Nonlinear Perturbation Theory Without Small Denominators or Secular Terms, *J. Math. Phys.*, **17** (No. 1), 121-140.

Evans, D.S. (1968): *Quarterly J. Roy. Astr. Soc.*, **9**, 388-400.

Falconer, K. (1990): Fractal Geometry, John Wiley & Sons.

Feder, J. (1988): Fractals, Plenum Press.

Feigenbaum, M.J. (1983): *Physica*, **7D**, 16-39.

Fermi, E. (1923): in Enrico Fermi, Collected Papers, I., University of Chicago Press, 79-87.

--- & Pasta, J. & Ulam, S. (1955): Studies of Nonlinear Problems, in Enrico Fermi, Collected Papers, II., 978-988.

Ferraz-Mello, S. (1985): Ideal Resonance in Regular Variables, *Celest. Mech.*, **35**, 209-220.

Fetter, A.L. & Walecka, J.D. (1980): Theoretical Mechanics of Particles and Continua, McGraw-Hill.

Fock, V. (1959): The Theory of Space Time and Gravitation, Pergamon.

Ford, J. (1983): How Random is a Coin Toss? *Phys. Today*, **36** (No. 4), 40-47.

Forsyth, A.R. (1893): Theory of Functions of a Complex Variable, Cambridge.

Friedrich, H. & Wintgen, D. (1989): The Hydrogen Atom in a Uniform Magnetic Field - An Example of Chaos, *Phys. Rep.*, **183** (No. 2), 37-79.

Froeschle, C. (1971): On the Number of Isolating Integrals in Systems with three Degrees of Freedom, *Astrophys. Space Sci.*, **14**, 110-117.

--- (1984): The Lyapounov Characteristic Exponents - Applications to Celestial Mechanics, *Celest. Mech.*, **34**, 95-115.

--- & Rickman, H. (1989): Chaotic Dynamics and Monte Carlo Modelling, *Celest. Mech.*, **45**, 93-98.

--- & Scheidecker, J.P. (1973): On the Disappearance of Isolating Integrals in Dynamical Systems with More Than Two Degrees of Freedom, *Astrophys. Space Sci.*, **25**, 373-386.

--- & Scheidecker, J.P. (1975): Stochasticity of Dynamical Systems with Increasing Number of Degrees of Freedom, *Phys. Rev. A*, **12** (No. 5), 2137-2143.

--- & Scholl, H. (1976): On the Dynamical topology of the Kirkwood Gaps, *Astron. Astrophys.*, **48**, 389-393.

--- & Scholl, H. (1977): A Qualitative Comparison Between the Circular and Elliptic Sun-Jupiter-Asteroid Problem at Commensurabilities, *Astron. Astrophys.*, **57**, 33-39.

Fukushima, T. (1988): *Celest. Mech.*, **44**, 61-75.

Galgani, L. (1985): in Buchler et al (ed.) (1985), 245-257.

Gallavotti, G. (1983): The Elements of Mechanics, Springer-Verlag.

Garfinkel, B. (1966): *Astron. J.*, **71**, 657-669.

Giovanni, B. & Valsecchi & Carusi, A. & Roy, A.E. (1984): The Effect of Orbital Eccentricities on the Shape of the Hill-Type Analytical Stability Surfaces in the

General 3-Body Problem, *Celest. Mech.*, **32**, 217-230.

Goldreich, P. (1965): An explanation of the Frequent Occurrence of Commensurable Mean Motions in the Solar System, *Mon. Not. Roy. Astr. Soc.*, **130**, 159-181.

Goldstein, H. (1950, 1980): Classical Mechanics, Addison-Wesley.

Golubev, V.G. (1967): Regions Where Motion Is Impossible in the Three-Body Problem, *Soviet Phys. Doklady*, **12** (No. 6), 529-531.

--- (1968): Hill Stability in the Unrestricted Three-Body Problem, *Soviet Phys. Doklady*, **13** (No. 5), 373-375.

Grebenikov, E.A. (1965): Methods of Averaging Equations in Celestial Mechanics, *Soviet Physics-Astron.*, **9** (No. 1), 146-148.

Green, Robin M. (1985): Spherical Astronomy, Cambridge University Press.

Greene, J.M. (1968): Two-Dimensional Measure-Preserving Mappings, *J. Math. Phys.*, **9** (No. 5), 760-768.

--- (1979): A Method for Determining a Stochastic Transition, *J. Math. Phys.*, **20** (No. 6), 1183-1201.

--- & Percival, I.C. (1981): Hamiltonian Maps in the Complex Plane, *Physica*, **3D**, 530-548.

Gutzwiller, M.C. (1971): Periodic Orbits and Classical Quantization Conditions, *J. Math. Phys.*, **12** (No. 3), 343-358.

Hadamard, J.S. (1898): *J. Math. pures appl.* **4**, 27-73.

Hadjidemetriou, J.D. (1981): The Present Status of Periodic Orbits, *Celest. Mech.*, **23**, 277-286.

--- (1984): Periodic Orbits, *Celest. Mech.*, **34**, 379-393.

Hagihara, Y. (1957): Stability in Celestial Mechanics, Tokyo.

Hale, J.K. & LaSalle, J.P. (ed.) (1967): Differential Equations and Dynamical Systems, Academic Press.

Hardy, G.H. & Littlewood, J.E. & Polya, G. (1934, 1988): Inequalities, Cambridge University Press.

Harrington, R.S. (1972): Stability Criteria for Triple Stars, *Celest. Mech.*, **6**, 322-327.

Hasegawa, H. & Robnik, M. & Wunner, G. (1989): Classical and Quantal Chaos in the Diamagnetic Kepler Problem, *Theoretical Phys. Prog. Jap. Supp.*, **98**, 189-286.

Hawking, S.W. & Ellis, G.F.R. (1973): The Large-Scale Structure of Space-Time, Cambridge University Press.

Hawking, S.W. & Israel, W. (ed.) (1987): Three Hundred Years of Gravitation, Cambridge University Press.

Heggie, Douglas C. (1974): A Global Regularization of the N-Body Problem, *Celest. Mech.*, **10**, 217-241.

--- (1975): Binary Evolution in Stellar Dynamics, *Mon. Not. Roy. Astr. Soc.*, **173**, 729-787.

--- (1976): Redundant Variables for 'Global' Regularization of the Three-Body Problem, *Celest. Mech.*, **14**, 69-71.

Heisenberg, Werner (1958): Physics and Philosophy, George Allen and Unwin.

--- (1967): Nonlinear Problems in Physics, *Phys. Today*, **20** (No. 5), 27-33.

Henon, Michel. (1965-70): Exploration Numerique Du Probleme Restreint

--- (1965): I. Masses Egales, Orbites Periodiques, *Ann. Astr.*, **28**, 499-511.

--- (1965): II. Masses Egales, Stabilite Des Orbites Periodiques, *Ann. Astr.*, **28**, 992-1007.

--- (1966): III. Masses Egales, Orbites Non Periodiques, *Bull. Astr. Paris*, **1**(fasc 1), 57-79.

--- (1966): IV. Masses Egales, Orbites Non Periodiques (Fin), *Bull. Astr. Paris*, **1**(fasc 2), 49-66.

Numerical Exploration of the Restricted Problem

--- (1969): V. Hill's Case: Periodic Orbits and Their Stability, *Astron. Astrophys.*, **1**, 223-238.

--- (1970): VI. Hill's Case: Non-Periodic Orbits, *Astron. Astrophys.*, **9**, 24-36.

--- (1976): Stability of the Interplay Motions, *Celest. Mech.*, **15**, 243-261.

Henon, M. & Heiles, C. (1964): The Applicability of the Third Integral of Motion: Some Numerical Experiments, *Astron. J.*, **69** (No. 1), 73-79.

Henon, M. (1969): Numerical Study of Quadratic Area-Preserving Mappings, *Quart. Apl. Math.*, **27**, 291-312.

--- (1970): *Astron. & Astrophys.*, **9**, 24 -36.

--- (1974): Integrals of the Toda Lattice, Phys. Rev., B9, 1921-1923.

--- (1976): A Two-Dimensional Mapping with a Strange Attractor, *Comm. Math. Phys.*, **50**, 69-77.

Henon, M. & Petit, J.M. (1986): Series Expansions for Encounter-Type Solution of Hill's Problem, *Celest. Mech.*, **38**, 67-100.

Henrici, P. (1963): Discrete Variable Methods in Ordinary Differential Equations, John Wiley & Sons.

Hietarinta, J. (1987): Direct Methods for the Search of the Second Invariant, *Phys. Rep.*, **147** (No. 2), 87-154.

Hill, G.W. (1878): Researches in the Lunar Theory, *Am. J. Math.*, **1** (No. 5), 5-26, 129-147, 245-260.

Holden, A.V. (ed.) (1986): Chaos, Manchester University Press.

Holmes, P. (1990): *Phys. Rep.*, **193** (No. 3), 137-163.

Horton, C.W. & Reichl, L.E. & Szebehely, V. (ed.) (1983): Long-Time Prediction in Dynamics, Wiley-Interscience.

Hoveijn, I. & Verhulst, F.(1990): Chaos in the 1:2:3 Hamiltonian Normal Form, *Physica*, **D44**, 397-406.

Infeld, L. (ed.) (1962): Relativistic Theories of Gravitation, Proceeding of a Conference Held in Warsaw and Jablonna, Pergamon Press.

Islam, J.N. (1985): Rotating Fields in General Relativity, Cambridge University Press.

Jacobi, C.G. (1836): *Compt. Rend.*, **3**, 59.

Janich, K. (1984): Topology, Springer-Verlag.

Jensen, R.V. & Susskind, S.M. & Sanders, M.M. (1991): Chaotic Ionization of Highly excited Hydrogen Atoms, Comparison of Classical and Quantum Theory with Experiments, *Phys. Rep.*, **201**, 1-56.

Jorna, S. (ed.) (1978): Topics in Nonlinear Dynamics, American Institute of Physics, New York.

Jordan, D.W. & Smith, P. (1977, 1990): Nonlinear Ordinary Differential Equations, Oxford University Press.

Jupp, A.H. (1972): Ideal Resonance by Lie Series, *Celest. Mech.*, **5**, 8-26.

Kamel, A.A. (1970): Perturbation Method in the Theory of Nonlinear Oscillations, *Celest. Mech.*, **3**, 90-106.

Kaneko, K. (1986): Collapse of Tori and Genesis of Chaos in Dissipative Systems, World Scientific.

Karttunen, H. & Kroger, P. et al (1987): Fundamental Astronomy, Springer-Verlag.

Kinoshita, H. & Nakai, H. (1984): Motions of the Perihelions of Neptune and Pluto, *Celest. Mech.*, **34**, 203-217.

Kirchgraber, U. (1976): Error Bounds for Perturbation Methods, *Celest. Mech.*, **14**, 351-362..

Kolmogorov, Andrei Nikolaevich (1954): The Conservation of Conditionally Periodic Motions with a Small Variation in the Hamiltonian, *Dokl. Akad. Nauk. USSR.*, **98**, 527-530. (sov phys doklady)

--- (1954): The General Theory of Dynamical Systems and Classical Mechanics, address to the 1954 International Congress of Mathematicians in Amsterdam, also in R. Abraham & J.E. Marsden (1978).

--- (1957):

Kopejkin, S.M. (1988): *Celest. Mech.*, **44**, 87-115.

Korner, T.W. (1988): Fourier Analysis, Cambridge University Press.

Kovalevsky, J. & Brumberg, V.A. (ed.) (1986): Relativity in Celestial Mechanics and Astrometry, Springer-Verlag.

Kozlov, V.V. (1983): Integrability and Non-integrability in Hamiltonian Mechanics, *Russ. Math. Surveys*, **38** (No. 1), 3-76.

Landau, L.D. & Lifshitz, E.M. (1962): The Classical Theory of Fields, Pergamon.

Lang, S. (1972): Differential Manifolds, Addison-Wesley.

Laskar, J. & Marchal, C. (1984): Triple Close Approach in the 3-Body Problem: A Limit for the Bounded Orbits, *Celest. Mech.*, **32**, 15-28.

Lear, J. (1965): Kepler's Dream, California University Press.

Leimanis, E. & Minorsky, N. (1958): Dynamics and Nonlinear Mechanics, John Wiley & Sons.

Levi-Civita, Tullio (1964): The N-Body Problem in General Relativity, Dordrecht.

--- (1906): *Acta. Math.*, **30**, 305-327.

Liapunov, A.M. (1889): On the Stability of Motion in a Particular Case of the Problem of Three Bodies, Kharkof. *Bull. Astron.* **VI**, 481.

Liapunov, A.M. (1907): Reprinted in *Ann. Math. Studies*, **17**, Princeton, 1947.

Lichtenberg, A.J. & Lieberman, M.A. (1983): Regular and Stochastic Motion, Springer-Verlag.

Lidov, M.L. (1963): On the Approximation Analysis of the Orbit Evolution of Artificial Satellites, in Roy, M. (ed.) (1963).

Lighthill, J. (1986): The Recently Recognized Failure of Predictability in Newtonian Dynamics, *Proc. Roy. Soc. London*, **A407**, 35-50.

Liouville, J. (1837): *J. Math. Pures Appl.*, **2**, 16.

Lorenz, E.N. (1963): Deterministic Nonperiodic Flow, *J. Atmos. Sci.*, **20**, 130-141.

Lundberg, J. & Szebehely, V. et al (1985): *Celest. Mech.*, **36**, 191.

Lynden-Bell, D. (1962): Stellar Dynamics, Only Isolating Integrals Should Be Used in Jean's Theorem, *Mon. Not. Roy. Astr. Soc.*, **124** (No. 1), 1-9.

--- (1962): Stellar Dynamics, Potentials with Isolating Integrals, *Mon. Not. Roy. Astr. Soc.*, **124** (No. 2), 95-123.

Mackay, R.S. & Meiss, J.D. (ed.) (1987): Hamiltonian Dynamical systems, Adam Hilger.

--- & Meiss, J.D. & Stark, J. (1989): Converse KAM Theory for Symplectic Twist Maps, *Nonlinearity*, **2**, 555-570.

--- & Percival, I.C. (1985): Converse KAM: Theory and Practice, *Comm. Math. Phys.*, **98**, 469-512.

Mandelbrot, Benoit B. (1977, 1983): The Fractal Geometry of Nature, Freeman.

--- (1983): *Physica*, **7D**, 224-239.

Marchal, C. (1990): The Three-Body Problem, Elsevier.

--- (1971): Qualitative Study of an N-Body System: A New Condition of Complete Scattering, *Astron. Astrophys.*, **10**, 278-289.

--- & Saari, D.G. (1975): Hill Regions For the General 3-Body Problem, *Celest. Mech.*, **12**, 115-129.

--- & Bozis, G. (1982): Hill Stability and Distance Curves for the General 3-Body

Problem, *Celest. Mech.*, **26**, 311-333.

--- & Yoshida, J. & Sun, Y.S. (1984): A Test of Escape Valid for Very Small Mutual Distances, *Celest. Mech.*, **33**, 193-207.

--- & Yoshida, J. & Sun, Y.S. (1984): Three-Body Problem, *Celest. Mech.*, **34**, 65-93.

--- (1986): The Two-Center Problem, *Celest. Mech.*, **38**, 377-387.

Markellos, V.V. & Roy, A.E. (1981): Hill Stability of Satellite Orbits, *Celest. Mech.*, **23**, 269-275.

Maslov, V.P. & Fedoriuk, M.V. (1981): Semi-Classical Approximation in Quantum Mechanics, translated from Russian by J. Niederle & J. Tolar, Reidel.

Mcdonald, A.J.C. (1986): *Celest. Mech.*, **38**, 139.

McKelvey, J.P. (1990): The Case of The Curious Curl, *Am. J. Phys.*, **58** (4), 306-310.

McKenzie, R. & Szebehely, V. (1981): Nonlinear Stability Around the Triangular Libration Points, *Celest. Mech.*, **23**, 223-229.

McLachlan, N.W. (1947): Theory and Application of Mathieu Functions, Oxford.

--- (1950): Ordinary Nonlinear Differential Equations, Oxford.

McVittie, G.L. (1963): The General Relativity "Force" on a Satellite, in Roy, M. (ed.) (1963).

Mehra, J. (1975): The Solvay Conferences on Physics, Reidel.

--- (1982): The Historical Development of Quantum Theory (Vol. 1), Springer-Verlag.

Mersman, W.A. (1970): A New Algorithm for the Lie Transformation, *Celest. Mech.*, **3**, 81-89.

--- (1970): Explicit Recursive Algorithms for the Construction of Equivalent Canonical Transformations, *Celest. Mech.*, **3**, 384-389.

Message, P.J. (1984): The Stability of Our Solar System, *Celest. Mech.*, **34**, 155-163.

Meyer, K.R. (1987): Bifurcation of a Central Configuration, *Celest. Mech.*, **40**, 273-282.

Mikkola, S. (1987): A Cubic Approximation for Kepler's Equation, *Celest. Mech.*, **40**, 329-334.

Milani, Andrea (1989): Planet Crosssing Asteroids and Parallel Computing: Project SPACEGUARD, *Celest. Mech.*, **45**, 111-118.

Milani, A. & Nobili, Anna M. (1983): On Topological Stability in the General 3-Body Problem, *Celest. Mech.*, **31**, 213-240.

--- (1983): On the Stabillity of Hierarhcical Four-Body Systems, *Celest. Mech.*, **31**, 241-291.

--- (1985): Resonat Structure of the Outer solar System, *Celest. Mech.*, **35**, 269-287.

--- (1988): Instability of the 2+2 Body Problem, *Celest. Mech.*, **41**, 153-160.

Miller, R.H. (1989): Applications to Stellar and Galactic Dynamics, *Celest. Mech.*, **45**, 19-26.

Mira, Christian (1987): Chaotic Dynamics, World Scientific.

Misner, C.W. & Thorne, K.S. & Wheeler, J.A. (1973): Gravitation, Freeman.

Moore, P. (1988): The Planet Neptune, Ellis Horwood.

Morbidelli, A. & Giorgilli, A. (1990): On the Dynamics in the Asteroids Belt,

--- I. General Theory, *Celest. Mech.*, **47**, 145-172.

--- II. Detailed Study of the Main Resonances, *Celest. Mech.*, **47**, 173-204.

Morse, Marston & Schaack, G.B.V. (1934): The Critical Point Theory Under General Boundary Conditions, *Ann. Math.*, **35** (No. 3), 545-571.

Moser, Jurgen K.(1973): Stable and Random Motions in Dynamical Systems, Princeton University Press.

Moser, Jurgen K.(1958): Stability of the Asteroids, *Astron. J.*, **63**, 439-443.

--- (1958): New Aspects in the Theory of Stability of Hamiltonian Systems, *Comm. on Pure and Appl. Math.*, **11**, 81-114.

--- (1961): A New Technique for the Construction of Solutions of Nonlinear Differential Equations, *Proc. Nat. Acad. Sci. USA.*, 47, 1824-1831.

--- (1962): On Invariant Curves of Area-Preserving Mappings of an Annulus, *Nachr. Akad. Wiss. Gottingen Math. Phys. Kl.*, **1**, 1-20.

--- (1978): Is the Solar System Stable? *Math. Intelligencer*, **1**, 65-71.

Moulton, F.R. (1902, 1959): An Introduction to Celestial Mechanics, Macmillan.

--- (1910): The Straight Line Solutions of the Problem of N Bodies, *Ann. Math.*, **12**, 1-17.

Murdoch, D.C. (1970): Linear Algebra, John Wiley & Sons.

Murison, M.A. (1989): The Fractal Dynamics of Satellite Capture in the Circular Restricted 3-Body Problem, *Astron. J.*, **98** (No. 6), 2346-2359.

Murray, C.D. & Fox, K. (1984): Structure of the 3:1 Jovian Resonance: A Comparison of Numerical Methods, *Icarus*, **59**, 221-233.

Noether, A. (1918): Invariante Variationsprobleme, *Nachr. Konig. Gesell. Wissen. Gottingen, Math. Phys. Kl.*, 235-257. (see *Transport Theory and Stat. Phys.*, **1** (1971), 186-207, for an English translation.)

Nugeyre, J.B. & Bouvier, P. (1981): Formal Aspects of Possible Hierarchies Within a Stellar System, *Celest. Mech.*, **29**, 51-64.

Olver, P.J. (1986): Applications of Lie Groups to Differential Equations, Springer-Verlag.

Oppenheim, A.V. & Shafer, R.W. (1975): Digital Signal Processing, Prentice Hall.

Ovenden, M.W. (1973): Planetary Distances and the Missing Planet, in Tapley & Szebehely (ed. 1973), 319-332.

Ozorio de Almeida, A.M. (1988): Hamiltonian Systems: Chaos and Quantization, Cambridge University Press.

--- (1989): On the Quantisation of Homoclinic Motion, *Nonlinearity*, **2**, 519-540.

Painleve, Paul (1897): *Compt. Rend.*, **124**, 173.

Palmore, J.I. (1973): Classfying Relative Equalibria I, *Bull. Amer. Math. Soc.*, **79** (No. 5), 904-908.

--- (1975): Classfying Relative Equalibria II, *Bull. Amer. Math. Soc.*, **81** (No. 2), 489-491.

--- (1976): Measure of Degenerate Relative Equilibria I, *Ann. Math.*, **104**,, 421-429.

--- (1979): Relative Equilibria and the Virial Theorem, *Celest. Mech.*, **19**, 167-171.

Papapetrou, A. (1966): *Ann. Inst. H. Poincare*, **A4**, 83.

--- (1974): Lectures on General Relativity, Dordrecht-Holland.

Patton, J.M. (1988): On the Dynamical Derivation of the Titius-Bode Law, *Celest. Mech.*, **44**, 365-391.

Percival, I.C. (1974): Variational Principles for the Invariant Toroids of Classical Dynamics, *J. Phys. A*, **7** (No. 7), 794-802.

--- & Pomphrey, N. (1976): Vibrational Quantization of Polyatomic Molecules, *Molec. Phys.*, **31** (No. 1), 97-114.

--- & Pomphrey, N. (1978): Semiclassical Energy Levels for Linear Molecules, *Molec. Phys.*, **35** (No. 3), 649-663.

--- (1982): Chaotic Boundary of a Hamiltonian Map, *Physica*, **6D**, 67-77.

--- (1987): Chaos in Hamiltonian Systems, *Proc. Roy. Soc. L.*, **A413**, 131-144.

Petrosky, T.Y. & Broucke, R. (1988): Area-Preserving Mappings and Deterministic Chaos for Nearly Parabolic Motions, *Celest. Mech.*, **42**, 53-79.

Poincare, Henri. (1892): *Les Methods Nouvelles de la Mechanique Celeste*, Gauthier Villars. (Vol 3, Chap 26, Sect 301)

--- (1913, 195): Science and Method, translated by F. Maitland with a preface by B. Russell, Nelson.

Pollard, H. (1966): Mathematical Introduction to Celestial Mechanics, Prentice-Hall.

--- & Saari, D.G. (1970): Escape From a Gravitational System of Positive Energy, *Celest. Mech.*, **1**, 347-350.

Poston, T. & Stewart, I. (1978): Catastrophe Theory and Its Application, Pitman.

Pradeep, S. & Shrivastava (1990): Stability of Dynamical Systems: An Overview, *J. Guidance*, **13** (No. 3), 385-393.

Pregogine, Ilya. & Grecos, A. & George, C. (1977): On the Relation of Dynamics to Statistical Mechanics, *Celest. Mech.*, **16**, 489-507.

Pringle, J.E. & Wade, R.A. (ed.) (1985): Interacting Binary Stars, Cambridge University Press.

Ramani, A. & Grammaticos, B. & Bountis, T. (1989): The Painleve Property and Singularity Analysis of Integrable and Non-integrable Systems, *Phys. Rep.*, **180** (No. 3), 159-245.

Rand, R. & Podgorski, W. (1972): *Celest. Mech.*, **6**, 416-420.

Rannou, F. (1974): Numerical Study of Discrete Plane Area-Preserving Mappings, *Astron. Astrophs.*, **31**, 289-301.

Richardson, D.C. & Kelly, T.J. (1988): Two-Body Problem in the Post-Newtonian Approximation, *Celest. Mech.*, **43**, 193-210.

Roekaerts, D. & Schwartz, F. (1987): Direct Methods in the Search for Integrable Hamiltonians, *J. Phys.*, **A20**, L127-L133.

Roxin, E.O. (1972): Ordinary Differential Equations, Wadsworth.

Roy, Archie E. (1978, 1982, 1988): Orbital Motion, Adam Hilger.

--- (ed.) (1988): Long-Term Dynamical Behaviour of Natural and Artificial N-Body System, Kluwer Academic Publishers.

--- (1973): The Use of the Saros in Lunar Dynamical Studies, *The Mooon*, **7**, 6-13.

--- (1980): The Stability and Evolution of the Solar System, *The Moon and the Planets*, **22**, 67-81.

--- & Moran, P.E. & Black, W. (1972): Studies in the application of Recurrence Relations to Special Perturbation Methods, I., *Celest. Mech.*, **6**, 468-482.

--- & Ovenden, M.W. (1954): On the Occurrence of Commensurable Mean Motions in the Solar System, *M.N.R.A.S.*, **114**, 232-241.

--- & Ovenden, M.W. (1955): II. The Mirror Theorem, *M.N.R.A.S.*, **115**, 296-309.

--- & Steves, B.A. (1988): in Roy (ed.) (1988), 197-215.

--- & Walker, I.W. & Macdonald, A.J. et al (1987): Project LONGSTOP, *Vistas in Astron.*, **32**, 95-116.

Roy, M. (ed.) (1963): Dynamics of Satellites, Springer-Verlag.

Ruelle, D. (1990): Deterministic Chaos, the Science and Fiction, *Proc. Roy. Soc. L.*, **A427**, 241-248.

Russmann, H. (1976): On a New Proof of Moser's Twist Mapping Theorem, *Celest. Mech.*, **14**, 19-31.

Ryan, M.P. & Shapley, L.C. (1975): Homogeneous Relativistic Cosmologies, Princeton University Press.

Saari, D.G. (1976): The N-Body Problem of Celestial Mechanics, *Celest. Mech.*, **14**, 11-17.

--- (1980): On the Role and Properties of Central Configurations, *Celest. Mech.*, **21**, 9-20.

--- (1984): From Rotations and Inclinations to Zero Configurational Velocity Surfaces

I. A Natural Rotating Coordinate System, *Celest. Mech.*, **33**, 299-318.

--- (1987): From Rotations and Inclinations to Zero Configurational Velocity Surfaces II. The Best Possible Configurational Velocity Surfaces, *Celest. Mech.*, **40**, 197-223.

Sachs, R.K. (ed.) (1971): General Relativity and Cosmology, Academic Press.

Schiff, L.I. (1955): Quantum Mechanics, McGraw-Hill.

Schrodinger, Erwin (1944): What Is Life?, Cambridge University Press.

Schutz, B. (1980): Geometrical Methods of Mathematical Physics, Cambridge University Press.

--- (1985): A First Course in General Relativity, Cambridge University Press.

--- (1990): Personal discussion.

Schweber, S.S. (1961): An Introduction to Relativistic Quantum Field Theory, Harper & Row.

Seidelmann, P.K. (1986): Unsolved Problems of Celestial Mechanics - the Solar System, *Celest. Mech.*, **39**, 141-146.

Shapiro, I.I. (1963): The Prediction of Satellite Orbits, in Roy M. (ed. 1963), 257-312.

--- (1964): Fourth Test of General Relativity, *Phys. Rev. Let.*, **13**, 789-791.

Sidlichovsky, M. (1983): Double Averaged Three-Body Problem, *Celest. Mech.*, **29**, 295-305.

Siegel, Carl Ludwig (1941): Der Dreierstoss, *Ann. Math.*, **42**, 127-168.

--- (1941): On the Integrals of Canonical Systems, *Ann. Math.*, **42**, 806-822.

--- (1942): Iteration of Analytic Functions, *Ann. Math.*, **43**, 607-612.

Siegel, C.L. & Moser, J.K. (1971): Lectures on Celestial Mechanics, Springer-Verlag.

Simo, J.C. & Posbergh, T.A. & Marsden, J.E. (1990): *Phys. Rep.*, **193** (No. 6) 279-362.

Singh, V. (1976): The Relativistic Restricted Problem of Three Bodies, *Celest. Mech.*, **14**, 167-173.

Smale, S. (1970): Topology and Mechanics I, *Invent. Math.*, **10**, 305-331.

--- (1970): Topology and Mechanics II, *Invent. Math.*, **11**, 45-64.

--- (1970): Problems on the Nature of Relative Equilibria in Celestial Mechanics, in *Lecture Note in Math.*, Springer-Verlag, **197**, 194-201.

--- & Hirsch, M.W. (1974): Differential Equations, Dynamical Systems, and Linear Algebra, Academic Press.

--- (1980): The Mathematics of Time, Springer-Verlag.

Soffel, M.H. (1989): Relativity in Astometry, Celestial Mechanics and Geodesy, Springer-Verlag.

--- & Ruder, H. & Schneider, M. (1987): The Two-Body Problem in the PPN Theory, *Celest. Mech.*, **40**, 77-85.

--- & Wirrer, R. et al (1988): *Celest. Mech.*, **42**, 81-89.

Steen, L.A. (ed.) (1978): Mathematics Today, Springer-Verlag.

Stewart, Ian (1989): Does God Play Dice? The Mathematics of Chaos, Basil Blackwell.

Stiefel, Eduard L. & Scheifele, Gerhard (1971): Linear and Regular Celestial Mechanics, Springer-Verlag.

Stiefel, E.L. (1970): Remarks on Numerical Integration of Keplerian Orbits, *Celest. Mech.*, **2**, 274-281.

--- (1973): A Linear Theory of the Perturbed Two-Body Problem (Regularization), in Tapley & Szebehely (ed. 1973), 3-20.

--- (1976): Remarks on the Numerical Integration of Near-Parabolic Orbits, *Celest. Mech.*, **14**, 85-90.

Streater, R.F. & Wightman, A.S. (1964): PCT, Spin and Statistics, and All That, Benjamin.

Sundman, K.F. (1912): Memoire sur le probleme des trois corps, *Acta Mathematica*, **36**, 105-179.

Sussman, G.J. & Wisdom, J. (1988): Numerical Evidence That the Motion of Pluto Is Chaotic, *Science*, **241**, 433-437.

Sweet, P.A. (1990): Personal discussion.

Synge, J.L. (1960): Relativity: The General Theory, North-Holland.

Szebehely, V. (1967): Theory of Orbits, Academic Press.

--- (1973): Recent Advances in the Problem of three Bodies, in Tapley & Szebehely (ed.) (1973), 75-106.

--- (1977): Analytical Determination of the Measure of the Stability of Triple Stellar Systems, *Celest. Mech.*, **15**, 107-110.

--- (1978): Stability of Artificial and Natural Satellites, *Celest. Mech.*, **18**, 383-389.

--- (ed.) (1979): Instabilities in Dynamical Systems, Reidel.

--- (1984): Review of Concepts of Stability, *Celest. Mech.*, **34**, 49-64.

--- (1987): Celestial Mechanics Since Newton,*Vistas in Astronomy*, **30**, 313-318.

--- (1988): Limits of Predictability of Gravitational Systems, *Celest. Mech.*, **43**, 139-145.

--- & Zare, K. (1976): Stability of Classical Triplets and of Their Hierarchy, *Astron. Astrophys.*, **58**, 145-152.

Tanikawa, K. (1983): Impossibility of Capture of Retrograde Satellites in the Circular Restricted 3-Body Problem, *Celest. Mech.*, **29**, 367-402.

Tapley, B.D. & Szebehely, V. (ed.) (1973): Recent Advances in Dynamical Astronomy, Dordrecht-Holland.

Thorne, K.S. (1971): Relativistic Stars, Black Holes and Gravitational Waves (including an in-depth review of the theory of rotating, relativistic stars), In Sachs, R.K. (ed.

1971).

Tildesley, D.J. & Ball, R.C. (ed. 1989): Fractals in the Natural Science, *Proc. Roy. Soc. L.*, **A423**, 1-200.

Udry, S. & Martinet, L. (1990): Henon-Heiles and Toda Lattice Hamiltonian, *Physica*, **D44**, 61-74.

Valsecchi, G.B. & Carusi, A. & Roy, A.E. (1984): The Effect of Orbital Eccentricities on the Shape of the Hill-type Analytical Stability Surfaces, *Celest. Mech.*, **32**, 217-230.

Valtonen, M.J. & Mikkola, S. (1989): *Celest. Mech.*, **46**, 277-285.

van der Pol, B. (1922): On Oscillation Hysteresis in a Triode Generator with Two Degrees of Freedom, *Philos. Mag.*, **6** (No. 43), 700-719.

Veres, F. (1989): *Celest. Mech.*, **46**, 102-112.

Voinov, A.V. (1988): *Celest. Mech.*, **42**, 293-307.

Von Zeipel, H. (1916): *Ask. Astron. Mat. Fys.*, **11** (No. 1).

Voss, R.F. (1989): Random Fractals, *Physica*, **38D**, 362-371.

Waldvogel, J. (1972): A New Regularization of the Planar Problem of Three Bodies, *Celest. Mech.*, **6**, 221-231.

--- (1976): The Three-Body Problem Near Triple Collision, *Celest. Mech.*, **14**, 287-300.

Walker, I.W. & Emslie, A.G. & Roy, A.E. (1980): *Celest. Mech.* , **22**, 371-402.

--- & Roy, A.E. (1981): *Celest. Mech.* , **24**, 195- .

--- & Roy, A.E. (1983): *Celest. Mech.* , **29**, 117-148.

Walker, I.W. (1983): *Celest. Mech.* , **29**, 149-178.

--- (1983): *Celest. Mech.* , **29**, 267-294.

--- (1983): On the Stability of Close Binaries in Hierarchical 3-Body Systems, *Celest. Mech.* , **29**, 215-228.

Walker, M. & Penrose, R. (1970): *Commun. Math. Phys.*, **18**, 265-274.

Weinberg, Steven. (1972): Gravitation and Cosmology, Wiley.

Wheeler, J.A. & Zurek, W.H. (ed.) (1983): Quantum Theory and Measurement, Princeton.

Whittaker, E.T. (1904, 1964): A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, Cambridge University Press.

--- & Watson, G.N. (1902, 1988): A Course of Modern Analysis, Cambridge University Press.

--- (1953): A History of The Theory of Aether and Electricity, Thomas Nelson and Sons.

Will, C.M. (1981): Theory and Experiment in Gravitational Physics, Cambridge University Press.

Williams, C.A. (1984): The Problem of Small Divisors in Planetary Motion, *Celest.*

Mech., **34**, 395-410.

Wintner, Aurel. (1910, 1947): The Analytical Foundations of Celestial Mechanics, Princeton University Press.

Wisdom, J. (1980): The Resonance Overlap Criterion and the Onset of Stochastic Behavior in the Circular Restricted 3-Body Problem, *Astron. J.*, **85**, 1122-1133.

--- (1983): Chaotic Behavior and the Origin of the 3/1 Kirkwood Gap, *Icarus*, **56**, 51-74.

--- (1985): A Perturbative Treatment of Motion Near the 3/1 Commensurability, *Icarus*, **63**, 272-289.

--- (1987): Urey Prize Lecture: Chaotic Dynamics in the Solar System, *Icarus*, **72**, 241-275.

Woodhouse, N.M.J. (1975): Killing Tensors and Separation of the Hamilton-Jacobi Equation, *Comm. Math. Phys.*, **44** (No. 9), 9-38.

Yoshida, H. (1983): Necessary Condition for the Existence of Algebraic First Integrals, *Celest. Mech.*, **31**, 363-379; 381-399.

--- (1987): Anisotropic Kepler Problem, *Celest. Mech.*, **40**, 51-66.

--- (1988): A Note on Kowalevski Exponents and the Nonexistence of an Additional Analytic Integral, *Celest. Mech.*, **44**, 313-316.

--- (1988): Non-integrability of the Truncated Toda Lattice Hamiltonian at Any Order, *Commun. Math. Phys.*, **116**, 529-538.

Yoshikawa, M. (1987): A Simple Analytical Model for the Secular Resonance n_6 in the Asteroidal Belt, *Celest. Mech.*, **40**, 233-272.

Zare, K. (1974): A Regularization of Multiple Encounters in N-Body Problems, *Celest. Mech.*, **10**, 207-215.

--- (1976): The Effects of Integrals on the Totality of Solutions of Dynamical Systems, *Celest. Mech.*, **14**, 73-83.

--- (1977): Bifurcation Points in the Planar Problem of Three Bodies, *Celest. Mech.*, **16**, 35-38.

--- (1981): Properties of the Moment of Inertia in the Problem of Three Bodies, *Celest. Mech.*, **24**, 345-354.

(N.B. The first authors of the important reference books are underlined.)