Scott Tremaine

DYNAMICS OF PLANETARY SYSTEMS

PRINCETON SERIES IN ASTROPHYSICS

Dynamics of Planetary Systems

PRINCETON SERIES IN ASTROPHYSICS

EDITED BY DAVID N. SPERGEL

Theory of Rotating Stars, *by Jean-Louis Tassoul*

Theory of Stellar Pulsation, *by John P. Cox*

Galactic Dynamics, Second Edition, *by James Binney and Scott Tremaine*

Dynamical Evolution of Globular Clusters, *by Lyman S. Spitzer, Jr.*

Supernovae and Nucleosynthesis: An Investigation of the History of Matter, from the Big Bang to the Present, *by David Arnett*

Unsolved Problems in Astrophysics, *edited by John N. Bahcall and Jeremiah P. Ostriker*

Galactic Astronomy, *by James Binney and Michael Merrifield*

Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment, *by Julian H. Krolik*

Plasma Physics for Astrophysics, *by Russell M. Kulsrud*

Electromagnetic Processes, *by Robert J. Gould*

Conversations on Electric and Magnetic Fields in the Cosmos, *by Eugene N. Parker*

High-Energy Astrophysics, *by Fulvio Melia*

Stellar Spectral Classification, *by Richard O. Gray and Christopher J. Corbally*

Exoplanet Atmospheres: Physical Processes, *by Sara Seager*

Physics of the Interstellar and Intergalactic Medium, *by Bruce T. Draine*

The First Galaxies in the Universe, *by Abraham Loeb and Steven R. Furlanetto*

Exoplanetary Atmospheres: Theoretical Concepts and Foundations, *by Kevin Heng*

Magnetic Reconnection: A Modern Synthesis of Theory, Experiment, and Observations, *by Masaaki Yamada*

Physics of Binary Star Evolution: From Stars to X-ray Binaries and Gravitational Wave Sources, *by Thomas M. Tauris and Edward P.J. van den Heuvel*

Dynamics of Planetary Systems, *by Scott Tremaine*

Dynamics of Planetary Systems

Scott Tremaine

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
99 Banbury Road, Oxford OX2 6JX

press.princeton.edu

ISBN 9780691207124
ISBN (pbk.) 9780691207117
ISBN (e-book) 9780691244228

British Library Cataloging-in-Publication Data is available

Editorial: Ingrid Gnerlich & Whitney Rauenhorst
Production Editorial: Ali Parrington
Jacket/Cover Design: Wanda España
Production: Jacqueline Poirier
Publicity: Matthew Taylor & Charlotte Coyne

Cover image: The innermost Galilean satellite, Io, casts its shadow onto the cloud decks of Jupiter in this Voyager 1 mosaic, taken 4 March 1979, at a planet-spacecraft distance of 1 million km. Photograph by Ian Regan (source images courtesy of NASA/JPL, via the PDS Ring-Moon Systems Node's OPUS service).

This book has been composed in LaTeX. The publisher would like to acknowledge the author of this volume for providing the print-ready files from which this book was printed.

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

The subject of this book, traditionally called celestial mechanics, is the oldest branch of theoretical physics. The publication in 1687 of the *Principia*, Newton's masterpiece on celestial mechanics, is widely regarded as the capstone of the Scientific Revolution. Since then, celestial mechanics has attracted the attention of many of the greatest physicists and mathematicians of the past several centuries, including Lagrange, Laplace, Gauss, Poincaré, Kolmogorov, and others. Concepts first explored in celestial mechanics are central to many if not most branches of physics, and its successful high-precision predictions of the motions of the planets have impacted disciplines as diverse as navigation and philosophy.

Celestial mechanics experienced a renaissance in the second half of the twentieth century. Starting in 1957, space flight created a demand for accurate and rapid orbit calculations as well as a need to understand the qualitative behavior of a wide variety of orbits. The development of high-speed digital computation enabled the study of classic problems in celestial mechanics with new tools. Advances in nonlinear dynamics and chaos theory provided new insights into the long-term behavior of orbits. Spacecraft visited every planet in the solar system and sent back data that dramatically expanded our understanding of the rich dynamics of their orbits, spins, and satellites. Finally, we have discovered thousands of planets outside the solar system, and celestial mechanics plays a central role in the analysis of the observations and the interpretation of their implications for the formation and evolution of planetary systems.

The primary goal of this book is to provide an introduction to celestial mechanics that reflects these developments. The reader is assumed to have an undergraduate background in classical mechanics and methods of mathematical physics (vectors, matrices, special functions, and so on), and much of what is needed is summarized in Appendixes B, C and D. The book contains most of the material that is needed for the reader to carry out research in the dynamics of planetary systems.

A book is defined in large part by what is left out, and a lot has been left out of this one. There is no analysis of spacecraft dynamics, except for a few examples and problems. There is almost no discussion of planet formation, since the tools that are needed to study this subject are mostly different from those of celestial mechanics. For similar reasons there is no discussion of planetary rings. Although general relativity offers a more accurate description of planetary motions than does Newtonian mechanics, its main use is in compiling high-accuracy

planetary ephemerides and so it is only described briefly, in Appendix J. Perturbation theory for planets and satellites on nearly circular, nearly coplanar orbits was the main focus of celestial mechanics in the nineteenth and early twentieth centuries, but many of the problems for which this theory was needed can now be solved using computer algebra or numerical orbit integration; thus the topic is described in much less detail than in earlier books at this level. There is only limited discussion of the rich phenomenology of extrasolar planets, since this is a large and rapidly growing subject that deserves a book of its own.

There are problems at the end of the book, many of which are intended to elaborate on topics that are not covered fully in the main text. Some of the problems are more easily done using a computer algebra system.

The notation in the book is mostly standard. We regularly use the notation $f(x) = \mathrm{O}(x)$ to indicate that $|f(x)/x|$ is no larger than a constant value as $|x| \to \infty$. We assume that $0^0 = 1$, although most mathematical and scientific software treats it as undefined. The symbols $\simeq$ and $\sim$ are used to indicate approximate equality, with $\simeq$ suggesting higher accuracy than $\sim$. Vectors and matrices are denoted by boldface type ($\mathbf{a}$, $\mathbf{A}$) and operators by boldface sans-serif type (A). We usually do not distinguish row vectors from column vectors; thus we write $\mathbf{a} = (a_1, a_2, a_3)$, in which $\mathbf{a}$ is a row vector, as well as $\mathbf{A}\mathbf{a}$, in which the matrix $\mathbf{A}$ multiplies the column vector $\mathbf{a}$.

We are all indebted to the Smithsonian/NASA Astrophysics Data System, https://ui.adsabs.harvard.edu, and the arXiv e-print service, https://arxiv.org, which have revolutionized access to the astronomy literature. In large part thanks to their efforts, most of the literature referenced here can easily and freely be accessed on the web.

All of the plots were prepared using Matplotlib (Hunter 2007), and most of the orbit integrations were done using REBOUND (Rein & Liu 2012).

I have learned this subject largely through my colleagues, collaborators and students, including Eugene Chiang, Luke Dones, Subo Dong, Martin Duncan, Wyn Evans, Dan Fabrycky, Eric Ford, Jean-Baptiste Fouvry, Adrian Hamers, Julia Heisler, Kevin Heng, Matthew Holman, Mario Jurić, Boaz Katz, Jacques Laskar, Renu Malhotra, Norman Murray, Fathi Namouni, Annika Peter, Cristobal Petrovich, Gerald Quinlan, Thomas Quinn, Roman Rafikov, Nicole Rappaport, Hanno Rein, Prasenjit Saha, Kedron Silsbee, Aristotle Socrates, Serge Tabachnik, Dan Tamayo, Alar Toomre, Jihad Touma, Paul Wiegert, Jack Wisdom, Qingjuan Yu and Nadia Zakamska. I thank Hanno Rein, Renu Malhotra and her students, and especially Alar Toomre, who read and commented on large parts of the manuscript, as well as Alysa Obertas and David Vokrouhlický, who contributed their research results for the figures. Above all, I am indebted to Peter Goldreich, who introduced me to this subject. My long collaboration with him was one of the highlights of my research career.

Much of this book was completed at home during the pandemic that began in 2020. I am grateful to my wife Marilyn for her unswerving support for this project, without which it would neither have been started nor finished.

Dynamics of Planetary Systems

Chapter 1

The two-body problem

1.1 Introduction

The roots of celestial mechanics are two fundamental discoveries by Isaac Newton. First, in any inertial frame the acceleration of a body of mass m subjected to a force $\mathbf{F}$ is

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \frac{\mathbf{F}}{m}. \tag{1.1}$$

Second, the gravitational force exerted by a point mass m_1 at position $\mathbf{r}_1$ on a point mass m_0 at $\mathbf{r}_0$ is

$$\mathbf{F} = \frac{\mathbb{G}m_0m_1(\mathbf{r}_1 - \mathbf{r}_0)}{|\mathbf{r}_1 - \mathbf{r}_0|^3}, \tag{1.2}$$

with $\mathbb{G}$ the gravitational constant.[1] Newton's laws have now been superseded by the equations of general relativity but remain accurate enough to describe all observable phenomena in planetary systems when they are supplemented by small relativistic corrections. A summary of the relevant effects of general relativity is given in Appendix J.

The simplest problem in celestial mechanics, solved by Newton but known as the **two-body problem** or the **Kepler problem**, is to determine

[1] For values of this and other constants, see Appendix A.

the orbits of two point masses ("particles") under the influence of their mutual gravitational attraction. This is the subject of the current chapter.[2]

The equations of motion for the particles labeled 0 and 1 are found by combining (1.1) and (1.2),

$$\frac{\mathrm{d}^2\mathbf{r}_0}{\mathrm{d}t^2} = \frac{\mathbb{G}m_1(\mathbf{r}_1 - \mathbf{r}_0)}{|\mathbf{r}_1 - \mathbf{r}_0|^3}, \quad \frac{\mathrm{d}^2\mathbf{r}_1}{\mathrm{d}t^2} = \frac{\mathbb{G}m_0(\mathbf{r}_0 - \mathbf{r}_1)}{|\mathbf{r}_0 - \mathbf{r}_1|^3}. \tag{1.3}$$

The total energy and angular momentum of the particles are

$$\begin{aligned} E_{\mathrm{tot}} &= \tfrac{1}{2}m_0|\dot{\mathbf{r}}_0|^2 + \tfrac{1}{2}m_1|\dot{\mathbf{r}}_1|^2 - \frac{\mathbb{G}m_0m_1}{|\mathbf{r}_1 - \mathbf{r}_0|}, \\ \mathbf{L}_{\mathrm{tot}} &= m_0\mathbf{r}_0 \times \dot{\mathbf{r}}_0 + m_1\mathbf{r}_1 \times \dot{\mathbf{r}}_1, \end{aligned} \tag{1.4}$$

in which we have introduced the notation $\dot{\mathbf{r}} = \mathrm{d}\mathbf{r}/\mathrm{d}t$. Using equations (1.3) it is straightforward to show that the total energy and angular momentum are conserved, that is, $\mathrm{d}E_{\mathrm{tot}}/\mathrm{d}t = 0$ and $\mathrm{d}\mathbf{L}_{\mathrm{tot}}/\mathrm{d}t = 0$.

We now change variables from $\mathbf{r}_0$ and $\mathbf{r}_1$ to

$$\mathbf{r}_{\mathrm{cm}} \equiv \frac{m_0\mathbf{r}_0 + m_1\mathbf{r}_1}{m_0 + m_1}, \quad \mathbf{r} \equiv \mathbf{r}_1 - \mathbf{r}_0; \tag{1.5}$$

here $\mathbf{r}_{\mathrm{cm}}$ is the **center of mass** or **barycenter** of the two particles and $\mathbf{r}$ is the **relative position**. These equations can be solved for $\mathbf{r}_0$ and $\mathbf{r}_1$ to yield

$$\mathbf{r}_0 = \mathbf{r}_{\mathrm{cm}} - \frac{m_1}{m_0 + m_1}\mathbf{r}, \quad \mathbf{r}_1 = \mathbf{r}_{\mathrm{cm}} + \frac{m_0}{m_0 + m_1}\mathbf{r}. \tag{1.6}$$

Taking two time derivatives of the first of equations (1.5) and using equations (1.3), we obtain

$$\frac{\mathrm{d}^2\mathbf{r}_{\mathrm{cm}}}{\mathrm{d}t^2} = \mathbf{0}; \tag{1.7}$$

[2] Most of the basic material in the first part of this chapter can be found in textbooks on classical mechanics. The more advanced material in later sections and chapters has been treated in many books over more than two centuries. The most influential of these include Laplace (1799–1825), Tisserand (1889–1896), Poincaré (1892–1897), Plummer (1918), Brouwer & Clemence (1961) and Murray & Dermott (1999).

thus the center of mass travels at uniform velocity, a consequence of the absence of any external forces.

In these variables the total energy and angular momentum can be written

$$E_{\rm tot} = E_{\rm cm} + E_{\rm rel}, \quad \mathbf{L}_{\rm tot} = \mathbf{L}_{\rm cm} + \mathbf{L}_{\rm rel}, \tag{1.8}$$

where

$$\begin{aligned} E_{\rm cm} &= \tfrac{1}{2} M |\dot{\mathbf{r}}_{\rm cm}|^2, & \mathbf{L}_{\rm cm} &= M \mathbf{r}_{\rm cm} \times \dot{\mathbf{r}}_{\rm cm}, \\ E_{\rm rel} &= \tfrac{1}{2} \mu |\dot{\mathbf{r}}|^2 - \frac{\mathbb{G}\mu M}{|\mathbf{r}|}, & \mathbf{L}_{\rm rel} &= \mu\, \mathbf{r} \times \dot{\mathbf{r}}; \end{aligned} \tag{1.9}$$

here we have introduced the **reduced mass** and **total mass**

$$\mu \equiv \frac{m_0 m_1}{m_0 + m_1}, \quad M \equiv m_0 + m_1. \tag{1.10}$$

The terms $E_{\rm cm}$ and $\mathbf{L}_{\rm cm}$ are the kinetic energy and angular momentum associated with the motion of the center of mass. These are zero if we choose a reference frame in which the velocity of the center of mass $\dot{\mathbf{r}}_{\rm cm} = \mathbf{0}$. The terms $E_{\rm rel}$ and $\mathbf{L}_{\rm rel}$ are the energy and angular momentum associated with the relative motion of the two particles around the center of mass. These are the same as the energy and angular momentum of a particle of mass μ orbiting around a mass M (the "central body") that is fixed at the origin of the vector $\mathbf{r}$.

Taking two time derivatives of the second of equations (1.5) yields

$$\frac{{\rm d}^2 \mathbf{r}}{{\rm d}t^2} = -\frac{\mathbb{G}M}{r^3}\mathbf{r} = -\frac{\mathbb{G}M}{r^2}\hat{\mathbf{r}}, \tag{1.11}$$

where $r = |\mathbf{r}|$ and the unit vector $\hat{\mathbf{r}} = \mathbf{r}/r$. Equation (1.11) describes any one of the following:

(i) the motion of a particle of arbitrary mass subject to the gravitational attraction of a central body of mass M that is fixed at the origin;

(ii) the motion of a particle of negligible mass (a **test particle**) under the influence of a freely moving central body of mass M;

(iii) the motion of a particle with mass equal to the reduced mass μ around a fixed central body that attracts it with the force $\mathbf{F}$ of equation (1.2).

Whatever the interpretation, the two-body problem has been reduced to a one-body problem.

Equation (1.11) can be derived from a Hamiltonian, as described in §1.4. It can also be written

$$\ddot{\mathbf{r}} = -\nabla\Phi_{\mathrm{K}}, \tag{1.12}$$

where we have introduced the notation $\nabla f(\mathbf{r})$ for the gradient of the scalar function $f(\mathbf{r})$ (see §B.3 for a review of vector calculus). The function $\Phi_{\mathrm{K}}(r) = -\mathbb{G}M/r$ is the **Kepler potential**. The solution of equations (1.11) or (1.12) is known as the **Kepler orbit**.

We begin the solution of equation (1.11) by evaluating the rate of change of the relative angular momentum $\mathbf{L}_{\mathrm{rel}}$ from equation (1.9):

$$\frac{1}{\mu}\frac{\mathrm{d}\mathbf{L}_{\mathrm{rel}}}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\times\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} + \mathbf{r}\times\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = -\frac{\mathbb{G}M}{r^2}\mathbf{r}\times\hat{\mathbf{r}} = \mathbf{0}. \tag{1.13}$$

Thus the relative angular momentum is conserved. Moreover, since $\mathbf{L}_{\mathrm{rel}}$ is normal to the plane containing the test particle's position and velocity vectors, the position vector must remain in a fixed plane, the **orbital plane**. The plane of the Earth's orbit around the Sun is called the **ecliptic**, and the directions perpendicular to this plane are called the north and south ecliptic poles.

We now simplify our notation. Since we can always choose an inertial reference frame in which the center-of-mass angular momentum $\mathbf{L}_{\mathrm{cm}} = \mathbf{0}$ for all time, we usually shorten "relative angular momentum" to "angular momentum." Similarly the "relative energy" E_{rel} is shortened to "energy." We usually work with the angular momentum per unit mass $\mathbf{L}_{\mathrm{rel}}/\mu = \mathbf{r}\times\dot{\mathbf{r}}$ and the energy per unit mass $\frac{1}{2}|\dot{\mathbf{r}}|^2 - \mathbb{G}M/|\mathbf{r}|$. These may be called "specific angular momentum" and "specific energy," but we shall just write "angular momentum" or "energy" when the intended meaning is clear. Moreover we typically use the same symbol—$\mathbf{L}$ for angular momentum and E for energy—whether we are referring to the total quantity or the quantity per unit mass. This casual use of the same notation for two different physical

quantities is less dangerous than it may seem, because the intended meaning can always be deduced from the units of the equations.

1.2 The shape of the Kepler orbit

We let (r, ψ) denote polar coordinates in the orbital plane, with ψ increasing in the direction of motion of the orbit. If $\mathbf{r}$ is a vector in the orbital plane, then $\mathbf{r} = r\hat{\mathbf{r}}$ where $(\hat{\mathbf{r}}, \hat{\boldsymbol{\psi}})$ are unit vectors in the radial and azimuthal directions. The acceleration vector lies in the orbital plane and is given by equation (B.18),

$$\ddot{\mathbf{r}} = (\ddot{r} - r\dot{\psi}^2)\hat{\mathbf{r}} + (2\dot{r}\dot{\psi} + r\ddot{\psi})\hat{\boldsymbol{\psi}}, \tag{1.14}$$

so the equations of motion (1.12) become

$$\ddot{r} - r\dot{\psi}^2 = -\frac{\mathrm{d}\Phi_{\mathrm{K}}(r)}{\mathrm{d}r}, \quad 2\dot{r}\dot{\psi} + r\ddot{\psi} = 0. \tag{1.15}$$

The second equation may be multiplied by r and integrated to yield

$$r^2\dot{\psi} = \text{constant} = L, \tag{1.16}$$

where $L = |\mathbf{L}|$. This is just a restatement of the conservation of angular momentum, equation (1.13).

We may use equation (1.16) to eliminate $\dot{\psi}$ from the first of equations (1.15),

$$\ddot{r} - \frac{L^2}{r^3} = -\frac{\mathrm{d}\Phi_{\mathrm{K}}}{\mathrm{d}r}. \tag{1.17}$$

Multiplying by $\dot{r}$ and integrating yields

$$\tfrac{1}{2}\dot{r}^2 + \frac{L^2}{2r^2} + \Phi_{\mathrm{K}}(r) = E, \tag{1.18}$$

where E is a constant that is equal to the energy per unit mass of the test particle. Equation (1.18) can be rewritten as

$$\tfrac{1}{2}v^2 - \frac{\mathbb{G}M}{r} = E, \tag{1.19}$$

where $v = (\dot{r}^2 + r^2\dot{\psi}^2)^{1/2}$ is the speed of the test particle.

Equation (1.18) implies that

$$\dot{r}^2 = 2E + \frac{2\,\mathbb{G}M}{r} - \frac{L^2}{r^2}. \tag{1.20}$$

As $r \to 0$, the right side approaches $-L^2/r^2$, which is negative, while the left side is positive. Thus there must be a point of closest approach of the test particle to the central body, known as the **periapsis** or **pericenter**.[3] In the opposite limit, $r \to \infty$, the right side of equation (1.20) approaches $2E$. Since the left side is positive, when $E < 0$ there is a maximum distance that the particle can achieve, known as the **apoapsis** or **apocenter**. Orbits with $E < 0$ are referred to as **bound** orbits since there is an upper limit to their distance from the central body. Orbits with $E > 0$ are **unbound** or **escape** orbits; they have no apoapsis, and particles on such orbits eventually travel arbitrarily far from the central body, never to return.[4]

The periapsis distance q and apoapsis distance Q of an orbit are determined by setting $\dot{r} = 0$ in equation (1.20), which yields the quadratic equation

$$2Er^2 + 2\,\mathbb{G}Mr - L^2 = 0. \tag{1.22}$$

For bound orbits, $E < 0$, there are two roots on the positive real axis,

$$q = \frac{\mathbb{G}M - \left[(\mathbb{G}M)^2 + 2EL^2\right]^{1/2}}{2|E|}, \quad Q = \frac{\mathbb{G}M + \left[(\mathbb{G}M)^2 + 2EL^2\right]^{1/2}}{2|E|}. \tag{1.23}$$

For unbound orbits, $E > 0$, there is only one root on the positive real axis,

$$q = \frac{\left[(\mathbb{G}M)^2 + 2EL^2\right]^{1/2} - \mathbb{G}M}{2E}. \tag{1.24}$$

[3] For specific central bodies other names are used, such as perihelion (Sun), perigee (Earth), periastron (a star), and so forth. "Periapse" is incorrect—an apse is not an apsis.

[4] The **escape speed** $v_{\rm esc}$ from an object is the minimum speed needed for a test particle to escape from its surface; if the object is spherical, with mass M and radius R, equation (1.19) implies that

$$v_{\rm esc} = \left(\frac{2\,\mathbb{G}M}{R}\right)^{1/2}. \tag{1.21}$$

To solve the differential equation (1.17) we introduce the variable $u \equiv 1/r$, and change the independent variable from t to ψ using the relation

$$\frac{\mathrm{d}}{\mathrm{d}t} = \dot{\psi}\frac{\mathrm{d}}{\mathrm{d}\psi} = Lu^2\frac{\mathrm{d}}{\mathrm{d}\psi}. \tag{1.25}$$

With these substitutions, $\dot{r} = -L\mathrm{d}u/\mathrm{d}\psi$ and $\ddot{r} = -L^2u^2\mathrm{d}^2u/\mathrm{d}\psi^2$, so equation (1.17) becomes

$$\frac{\mathrm{d}^2u}{\mathrm{d}\psi^2} + u = -\frac{1}{L^2}\frac{\mathrm{d}\Phi_{\mathrm{K}}}{\mathrm{d}u}. \tag{1.26}$$

Since $\Phi_{\mathrm{K}}(r) = -\mathbb{G}M/r = -\mathbb{G}Mu$ the right side is equal to a constant, $\mathbb{G}M/L^2$, and the equation is easily solved to yield

$$u = \frac{1}{r} = \frac{\mathbb{G}M}{L^2}[1 + e\cos(\psi - \varpi)], \tag{1.27}$$

where $e \geq 0$ and ϖ are constants of integration.[5] We replace the angular momentum L by another constant of integration, a, defined by the relation

$$L^2 = \mathbb{G}Ma(1 - e^2), \tag{1.28}$$

so the shape of the orbit is given by

$$r = \frac{a(1 - e^2)}{1 + e\cos f}, \tag{1.29}$$

where $f = \psi - \varpi$ is known as the **true anomaly**.[6]

The closest approach of the two bodies occurs at $f = 0$ or azimuth $\psi = \varpi$ and hence ϖ is known as the **longitude of periapsis**. The periapsis distance is $r(f = 0)$ or

$$q = a(1 - e). \tag{1.30}$$

[5] The symbol ϖ is a variant of the symbol for the Greek letter π, even though it looks more like the symbol for the letter ω; hence it is sometimes informally called "pomega."

[6] In a subject as old as this, there is a rich specialized vocabulary. The term "anomaly" refers to any angular variable that is zero at periapsis and increases by 2π as the particle travels from periapsis to apoapsis and back. There are also several old terms we shall not use: "semilatus rectum" for the combination $a(1 - e^2)$, "vis viva" for kinetic energy, and so on.

When the eccentricity is zero, the longitude of periapsis ϖ is undefined. This indeterminacy can drastically slow or halt numerical calculations that follow the evolution of the orbital elements, and can be avoided by replacing e and ϖ by two new elements, the **eccentricity components** or h **and** k **variables**

$$k \equiv e\cos\varpi, \quad h \equiv e\sin\varpi, \tag{1.31}$$

which are well defined even for $e = 0$. The generalization to nonzero inclination is given in equations (1.71).

Substituting q for r in equation (1.22) and replacing L^2 using equation (1.28) reveals that the energy per unit mass is simply related to the constant a:

$$E = -\frac{\mathbb{G}M}{2a}. \tag{1.32}$$

First consider bound orbits, which have $E < 0$. Then $a > 0$ by equation (1.32) and hence $e < 1$ by equation (1.28). A circular orbit has $e = 0$ and angular momentum per unit mass $L = (\mathbb{G}Ma)^{1/2}$. The circular orbit has the largest possible angular momentum for a given semimajor axis or energy, so we sometimes write

$$\mathbf{j} \equiv \frac{\mathbf{L}}{(\mathbb{G}Ma)^{1/2}}, \quad \text{where} \quad j = |\mathbf{j}| = (1-e^2)^{1/2} \tag{1.33}$$

ranges from 0 to 1 and represents a dimensionless angular momentum at a given semimajor axis.

The apoapsis distance, obtained from equation (1.29) with $f = \pi$, is

$$Q = a(1+e). \tag{1.34}$$

The periapsis and the apoapsis are joined by a straight line known as the **line of apsides**. Equation (1.29) describes an ellipse with one focus at the origin (**Kepler's first law**). Its major axis is the line of apsides and has length $q + Q = 2a$; thus the constant a is known as the **semimajor axis**. The **semiminor axis** of the ellipse is the maximum perpendicular distance of the orbit from the line of apsides, $b = \max_f[a(1-e^2)\sin f/(1+e\cos f)] = a(1-e^2)^{1/2}$. The **eccentricity** of the ellipse, $(1-b^2/a^2)^{1/2}$, is therefore equal to the constant e.

Box 1.1: The eccentricity vector

The **eccentricity vector** offers a more elegant but less transparent derivation of the equation for the shape of a Kepler orbit. Take the cross product of $\mathbf{L}$ with equation (1.11),

$$\mathbf{L} \times \ddot{\mathbf{r}} = -\frac{\mathbb{G}M}{r^3}\mathbf{L} \times \mathbf{r}. \tag{a}$$

Using the vector identity (B.9b), $\mathbf{L} \times \mathbf{r} = -\mathbf{r} \times \mathbf{L} = -\mathbf{r} \times (\mathbf{r} \times \dot{\mathbf{r}}) = r^2\dot{\mathbf{r}} - (\mathbf{r} \cdot \dot{\mathbf{r}})\mathbf{r}$, which is equal to $r^3 \mathrm{d}\hat{\mathbf{r}}/\mathrm{d}t$. Thus

$$\mathbf{L} \times \ddot{\mathbf{r}} = -\mathbb{G}M\frac{\mathrm{d}\hat{\mathbf{r}}}{\mathrm{d}t}. \tag{b}$$

Since $\mathbf{L}$ is constant, we may integrate to obtain

$$\mathbf{L} \times \dot{\mathbf{r}} = -\mathbb{G}M(\hat{\mathbf{r}} + \mathbf{e}), \tag{c}$$

where $\mathbf{e}$ is a vector constant of motion, the **eccentricity vector**. Rearranging equation (c), we have

$$\mathbf{e} = \frac{\dot{\mathbf{r}} \times (\mathbf{r} \times \dot{\mathbf{r}})}{\mathbb{G}M} - \frac{\mathbf{r}}{r}. \tag{d}$$

To derive the shape of the orbit, we take the dot product of (c) with $\hat{\mathbf{r}}$ and use the vector identity (B.9a) to show that $\hat{\mathbf{r}} \cdot (\mathbf{L} \times \dot{\mathbf{r}}) = -L^2/r$. The resulting formula is

$$r = \frac{L^2}{\mathbb{G}M}\frac{1}{1 + \mathbf{e} \cdot \hat{\mathbf{r}}} = \frac{a(1 - e^2)}{1 + \mathbf{e} \cdot \hat{\mathbf{r}}}; \tag{e}$$

in the last equation we have eliminated L^2 using equation (1.28). This result is the same as equation (1.29) if the magnitude of the eccentricity vector equals the eccentricity, $|\mathbf{e}| = e$, and the eccentricity vector points toward periapsis.

The eccentricity vector is often called the **Runge–Lenz vector**, although its history can be traced back at least to Laplace (Goldstein 1975–1976). Runge and Lenz appear to have taken their derivation from Gibbs & Wilson (1901), the classic text that introduced modern vector notation.

Unbound orbits have $E > 0$, $a < 0$ and $e > 1$. In this case equation (1.29) describes a hyperbola with focus at the origin and asymptotes at azimuth

$$\psi = \varpi \pm f_\infty, \quad \text{where} \quad f_\infty \equiv \cos^{-1}(-1/e) \tag{1.35}$$

is the **asymptotic true anomaly**, which varies between π (for $e = 1$) and $\frac{1}{2}\pi$ (for $e \to \infty$). The constants a and e are still commonly referred to as semimajor axis and eccentricity even though these terms have no direct geometric interpretation.

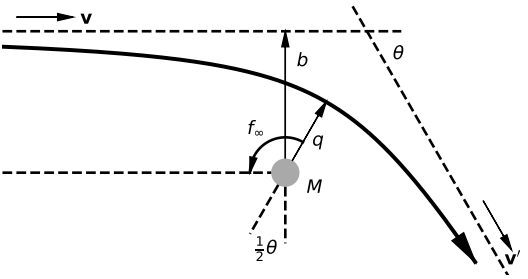

Figure 1.1: The geometry of an unbound or hyperbolic orbit around mass M. The impact parameter is b, the deflection angle is θ, the asymptotic true anomaly is f_∞, and the periapsis is located at the tip of the vector $\mathbf{q}$.

Suppose that a particle is on an unbound orbit around a mass M. Long before the particle approaches M, it travels at a constant velocity which we denote by $\mathbf{v}$ (Figure 1.1). If there were no gravitational forces, the particle would continue to travel in a straight line that makes its closest approach to M at a point $\mathbf{b}$ called the **impact parameter vector**. Long after the particle passes M, it again travels at a constant velocity $\mathbf{v}'$, where $v \equiv |\mathbf{v}| = |\mathbf{v}'|$ because of energy conservation. The deflection angle θ is the angle between $\mathbf{v}$ and $\mathbf{v}'$, given by $\cos\theta = \mathbf{v} \cdot \mathbf{v}'/v^2$. The deflection angle is related to the asymptotic true anomaly f_∞ by $\theta = 2f_\infty - \pi$; then

$$\tan\tfrac{1}{2}\theta = -\frac{\cos f_\infty}{\sin f_\infty} = \frac{1}{(e^2-1)^{1/2}}. \tag{1.36}$$

The relation between the pre- and post-encounter velocities can be written

$$\mathbf{v}' = \mathbf{v}\cos\theta - \hat{\mathbf{b}} v \sin\theta. \tag{1.37}$$

In many cases the properties of unbound orbits are best described by the asymptotic speed v and the impact parameter $b = |\mathbf{b}|$, rather than by orbital elements such as a and e. It is straightforward to show that the angular momentum and energy of the orbit per unit mass are $L = bv$ and $E = \frac{1}{2}v^2$. From equations (1.28) and (1.32) it follows that

$$a = -\frac{\mathbb{G}M}{v^2}, \quad e^2 = 1 + \frac{b^2v^4}{(\mathbb{G}M)^2}. \tag{1.38}$$

Then from equation (1.36),

$$\tan\tfrac{1}{2}\theta = \frac{\mathbb{G}M}{bv^2}. \tag{1.39}$$

The periapsis distance $q = a(1-e)$ is related to the impact parameter b by

$$q = \frac{\mathbb{G}M}{v^2}\left[\left(1 + \frac{b^2v^4}{\mathbb{G}^2M^2}\right)^{1/2} - 1\right] \quad \text{or} \quad b^2 = q^2 + \frac{2\,\mathbb{G}Mq}{v^2}. \tag{1.40}$$

Thus, for example, if the central body has radius R, the particle will collide with it if

$$b^2 \le b_{\rm coll}^2 \equiv R^2 + \frac{2\,\mathbb{G}MR}{v^2}. \tag{1.41}$$

The corresponding cross section is $\pi b_{\rm coll}^2$. If the central body has zero mass the cross section is just πR^2; the enhancement arising from the second term in equation (1.41) is said to be due to **gravitational focusing**.

In the special case $E = 0$, a is infinite and $e = 1$, so equation (1.29) is undefined; however, in this case equation (1.22) implies that the periapsis distance $q = L^2/(2\,\mathbb{G}M)$, so equation (1.27) implies

$$r = \frac{2q}{1+\cos f}, \tag{1.42}$$

which describes a parabola. This result can also be derived from equation (1.29) by replacing $a(1-e^2)$ by $q(1+e)$ and letting $e \to 1$.

1.3 Motion in the Kepler orbit

The **period** P of a bound orbit is the time taken to travel from periapsis to apoapsis and back. Since $\mathrm{d}\psi/\mathrm{d}t = L/r^2$, we have $\int_{t_1}^{t_2} \mathrm{d}t = L^{-1}\int_{\psi_1}^{\psi_2} r^2 \mathrm{d}\psi$; the integral on the right side is twice the area contained in the ellipse between azimuths ψ_1 and ψ_2, so the radius vector to the particle sweeps out equal areas in equal times (**Kepler's second law**). Thus $P = 2A/L$, where the area of the ellipse is $A = \pi ab$ with a and $b = a(1-e^2)^{1/2}$ the semimajor and semiminor axes of the ellipse. Combining these results with equation (1.28), we find

$$P = 2\pi\left(\frac{a^3}{\mathbb{G}M}\right)^{1/2}. \tag{1.43}$$

The period, like the energy, depends only on the semimajor axis. The **mean motion** or mean rate of change of azimuth, usually written n and equal to $2\pi/P$, thus satisfies[7]

$$n^2 a^3 = \mathbb{G}M, \tag{1.44}$$

which is **Kepler's third law** or simply **Kepler's law**. If the particle passes through periapsis at $t = t_0$, the dimensionless variable

$$\ell = 2\pi\frac{t-t_0}{P} = n(t-t_0) \tag{1.45}$$

is called the **mean anomaly**. Notice that the mean anomaly equals the true anomaly f when $f = 0, \pi, 2\pi, \ldots$ but not at other phases unless the orbit is circular; similarly, ℓ and f always lie in the same semicircle (0 to π, π to 2π, and so on).

[7] The relation $n = 2\pi/P$ holds because Kepler orbits are closed—that is, they return to the same point once per orbit. In more general spherical potentials we must distinguish the **radial period**, the time between successive periapsis passages, from the **azimuthal period** $2\pi/n$. For example, in a harmonic potential $\Phi(r) = \frac{1}{2}\omega^2 r^2$ the radial period is π/ω but the azimuthal period is $2\pi/\omega$. Smaller differences between the radial and azimuthal period arise in perturbed Kepler systems such as multi-planet systems or satellites orbiting a flattened planet (§1.8.3). For the Earth the radial period is called the **anomalistic year**, while the azimuthal period of 365.256 363 d is the **sidereal year**. The anomalistic year is longer than the sidereal year by 0.003 27 d. When we use the term "year" in this book, we refer to the Julian year of exactly 365.25 d (§1.5).

The position and velocity of a particle in the orbital plane at a given time are determined by four **orbital elements**: two specify the size and shape of the orbit, which we can take to be e and a (or e and n, q and Q, L and E, and so forth); one specifies the orientation of the line of apsides (ϖ); and one specifies the location or phase of the particle in its orbit (f, ℓ, or t_0).

The trajectory $[r(t), \psi(t)]$ can be derived by solving the differential equation (1.20) for $r(t)$, then (1.16) for $\psi(t)$. However, there is a simpler method.

First consider bound orbits. Since the radius of a bound orbit oscillates between $a(1-e)$ and $a(1+e)$, it is natural to define a variable $u(t)$, the **eccentric anomaly**, by

$$r = a(1 - e\cos u); \tag{1.46}$$

since the cosine is multivalued, we must add the supplemental condition that u and f always lie in the same semicircle (0 to π, π to 2π, and so on). Thus u increases from 0 to 2π as the particle travels from periapsis to apoapsis and back. The true, eccentric and mean anomalies f, u and ℓ are all equal for circular orbits, and for any bound orbit $f = u = \ell = 0$ at periapsis and π at apoapsis.

Substituting equation (1.46) into the energy equation (1.20) and using equations (1.28) and (1.32) for L^2 and E, we obtain

$$\dot r^2 = a^2 e^2 \sin^2 u\, \dot u^2 = -\frac{\mathbb{G}M}{a} + \frac{2\,\mathbb{G}M}{a(1-e\cos u)} - \frac{\mathbb{G}M(1-e^2)}{a(1-e\cos u)^2}, \tag{1.47}$$

which simplifies to

$$(1 - e\cos u)^2 \dot u^2 = \frac{\mathbb{G}M}{a^3} = n^2 = \dot\ell^2. \tag{1.48}$$

Since $\dot u, \dot\ell > 0$ and $u = \ell = 0$ at periapsis, we may take the square root of this equation and then integrate to obtain **Kepler's equation**

$$\ell = u - e\sin u. \tag{1.49}$$

Kepler's equation cannot be solved analytically for u, but many efficient numerical methods of solution are available.

The relation between the true and eccentric anomalies is found by eliminating r from equations (1.29) and (1.46):

$$\cos f = \frac{\cos u - e}{1 - e\cos u}, \quad \cos u = \frac{\cos f + e}{1 + e\cos f}, \tag{1.50}$$

with the understanding that f and u always lie in the same semicircle. Similarly,

$$\sin f = \frac{(1-e^2)^{1/2}\sin u}{1 - e\cos u}, \quad \sin u = \frac{(1-e^2)^{1/2}\sin f}{1 + e\cos f}, \tag{1.51a}$$

$$\tan \tfrac{1}{2} f = \left(\frac{1+e}{1-e}\right)^{1/2} \tan \tfrac{1}{2} u, \tag{1.51b}$$

$$\exp(\mathrm{i}f) = \frac{\exp(\mathrm{i}u) - \beta}{1 - \beta\exp(\mathrm{i}u)}, \quad \exp(\mathrm{i}u) = \frac{\exp(\mathrm{i}f) + \beta}{1 + \beta\exp(\mathrm{i}f)}, \tag{1.51c}$$

where

$$\beta \equiv \frac{1 - (1-e^2)^{1/2}}{e}. \tag{1.52}$$

If we assume that the periapsis lies on the x-axis of a rectangular coordinate system in the orbital plane, the coordinates of the particle are

$$x = r\cos f = a(\cos u - e), \quad y = r\sin f = a(1-e^2)^{1/2}\sin u. \tag{1.53}$$

The position and velocity of a bound particle at a given time t can be determined from the orbital elements a, e, ϖ and t_0 by the following steps. First compute the mean motion n from Kepler's third law (1.44), then find the mean anomaly ℓ from (1.45). Solve Kepler's equation for the eccentric anomaly u. The radius r is then given by equation (1.46); the true anomaly f is given by equation (1.50); and the azimuth $\psi = f + \varpi$. The radial velocity is

$$v_r = \dot{r} = n\frac{\mathrm{d}r}{\mathrm{d}\ell} = n\frac{\mathrm{d}r/\mathrm{d}u}{\mathrm{d}\ell/\mathrm{d}u} = \frac{nae\sin u}{1 - e\cos u} = \frac{nae\sin f}{(1-e^2)^{1/2}}, \tag{1.54}$$

and the azimuthal velocity is

$$v_\psi = r\dot{\psi} = \frac{L}{r} = na\frac{(1-e^2)^{1/2}}{1 - e\cos u} = na\frac{1 + e\cos f}{(1-e^2)^{1/2}}, \tag{1.55}$$

in which we have used equation (1.28).

For unbound particles, recall that $a < 0$, $e > 1$, and the period is undefined since the particle escapes to infinity. The physical interpretations of the mean anomaly ℓ and mean motion n that led to equations (1.44) and (1.45) no longer apply, but we may *define* these quantities by the relations

$$\ell = n(t - t_0), \quad n^2|a|^3 = \mathbb{G}M. \tag{1.56}$$

Similarly, we define the eccentric anomaly u by

$$r = |a|(e\cosh u - 1). \tag{1.57}$$

The eccentric and mean anomalies increase from 0 to ∞ as the true anomaly increases from 0 to $\cos^{-1}(-1/e)$ (eq. 1.35).

By following the chain of argument in equations (1.47)–(1.49), we may derive the analog of Kepler's equation for unbound orbits,

$$\ell = e\sinh u - u. \tag{1.58}$$

The relation between the true and eccentric anomalies is

$$\cos f = \frac{e - \cosh u}{e\cosh u - 1}, \quad \cosh u = \frac{e + \cos f}{1 + e\cos f}, \tag{1.59a}$$

$$\sin f = \frac{(e^2-1)^{1/2}\sinh u}{e\cosh u - 1}, \quad \sinh u = \frac{(e^2-1)^{1/2}\sin f}{1 + e\cos f}, \tag{1.59b}$$

$$\tan\tfrac{1}{2}f = \left(\frac{e+1}{e-1}\right)^{1/2}\tanh\tfrac{1}{2}u. \tag{1.59c}$$

A more direct but less physical approach to deriving these results is to substitute $u \to \mathrm{i}u$, $\ell \to -\mathrm{i}\ell$ in the analogous expressions for bound orbits.

For parabolic orbits we do not need the eccentric anomaly since the relation between time from periapsis and true anomaly can be determined analytically. Since $\dot{f} = L/r^2$, we can use equation (1.42) to write

$$t - t_0 = \int_0^f \frac{\mathrm{d}f\, r^2}{L} = \left(\frac{8q^3}{\mathbb{G}M}\right)^{1/2}\int_0^f \frac{\mathrm{d}f}{(1+\cos f)^2}. \tag{1.60}$$

In the last equation we have used the relation $L^2 = 2\mathbb{G}Mq$ for parabolic orbits. Evaluating the integral, we obtain

$$\left(\frac{\mathbb{G}M}{2q^3}\right)^{1/2}(t-t_0) = \tan\tfrac{1}{2}f + \tfrac{1}{3}\tan^3\tfrac{1}{2}f. \tag{1.61}$$

This is a cubic equation for $\tan\frac{1}{2}f$ that can be solved analytically.

1.3.1 Orbit averages

Many applications require the time average of some quantity $X(\mathbf{r},\mathbf{v})$ over one period of a bound Kepler orbit of semimajor axis a and eccentricity e. We call this the **orbit average** of X and use the notation

$$\langle X\rangle = \int_0^{2\pi}\frac{\mathrm{d}\ell}{2\pi}X = \int_0^{2\pi}\frac{\mathrm{d}u}{2\pi}(1-e\cos u)X, \tag{1.62}$$

in which we have used Kepler's equation (1.49) to derive the second integral. An alternative is to write

$$\langle X\rangle = \int_0^{P}\frac{\mathrm{d}t}{P}X = \int_0^{2\pi}\frac{\mathrm{d}f}{P\dot{f}}X = \frac{1}{PL}\int_0^{2\pi}\mathrm{d}f\,r^2X; \tag{1.63}$$

here P and $L = r^2\dot{f}$ are the orbital period and angular momentum. Substituting equations (1.28), (1.29) and (1.43) for the angular momentum, orbit shape and period, the last equation can be rewritten as

$$\langle X\rangle = (1-e^2)^{3/2}\int_0^{2\pi}\frac{\mathrm{d}f}{2\pi}\frac{X}{(1+e\cos f)^2}. \tag{1.64}$$

Equation (1.62) provides the simplest route to derive such results as

$$\langle a/r\rangle = 1, \tag{1.65a}$$
$$\langle r/a\rangle = 1+\tfrac{1}{2}e^2, \tag{1.65b}$$
$$\langle (r/a)^2\rangle = 1+\tfrac{3}{2}e^2, \tag{1.65c}$$
$$\langle (r/a)^2\cos^2 f\rangle = \tfrac{1}{2}+2e^2, \tag{1.65d}$$

$$\langle (r/a)^2 \sin^2 f \rangle = \tfrac{1}{2} - \tfrac{1}{2}e^2, \tag{1.65e}$$

$$\langle (r/a)^2 \cos f \sin f \rangle = 0. \tag{1.65f}$$

Equation (1.64) gives

$$\langle (a/r)^2 \rangle = (1-e^2)^{-1/2}, \tag{1.66a}$$

$$\langle (a/r)^3 \rangle = (1-e^2)^{-3/2}, \tag{1.66b}$$

$$\langle (a/r)^3 \cos^2 f \rangle = \tfrac{1}{2}(1-e^2)^{-3/2}, \tag{1.66c}$$

$$\langle (a/r)^3 \sin^2 f \rangle = \tfrac{1}{2}(1-e^2)^{-3/2}, \tag{1.66d}$$

$$\langle (a/r)^3 \sin f \cos f \rangle = 0. \tag{1.66e}$$

Additional orbit averages are given in Problems 1.2 and 1.3.

1.3.2 Motion in three dimensions

So far we have described the motion of a particle in its orbital plane. To characterize the orbit fully we must also specify the spatial orientation of the orbital plane, as shown in Figure 1.2.

We work with the usual Cartesian coordinates (x, y, z) and spherical coordinates (r, θ, ϕ) (see Appendix B.2). We call the plane $z = 0$, corresponding to $\theta = \frac{1}{2}\pi$, the **reference plane**. The **inclination** of the orbital plane to the reference plane is denoted I, with $0 \le I \le \pi$; thus $\cos I = \hat{\mathbf{z}} \cdot \hat{\mathbf{L}}$, where $\hat{\mathbf{z}}$ and $\hat{\mathbf{L}}$ are unit vectors in the direction of the z-axis and the angular-momentum vector. Orbits with $0 \le I \le \frac{1}{2}\pi$ are **direct** or **prograde**; orbits with $\frac{1}{2}\pi < I < \pi$ are **retrograde**.

Any bound Kepler orbit pierces the reference plane at two points known as the **nodes** of the orbit. The particle travels upward ($\dot{z} > 0$) at the **ascending node** and downward at the **descending node**. The azimuthal angle ϕ of the ascending node is denoted Ω and is called the **longitude of the ascending node**. The angle from ascending node to periapsis, measured in the direction of motion of the particle in the orbital plane, is denoted ω and is called the **argument of periapsis**.

An unfortunate feature of these elements is that neither ω nor Ω is defined for orbits in the reference plane ($I = 0$). Partly for this reason, the

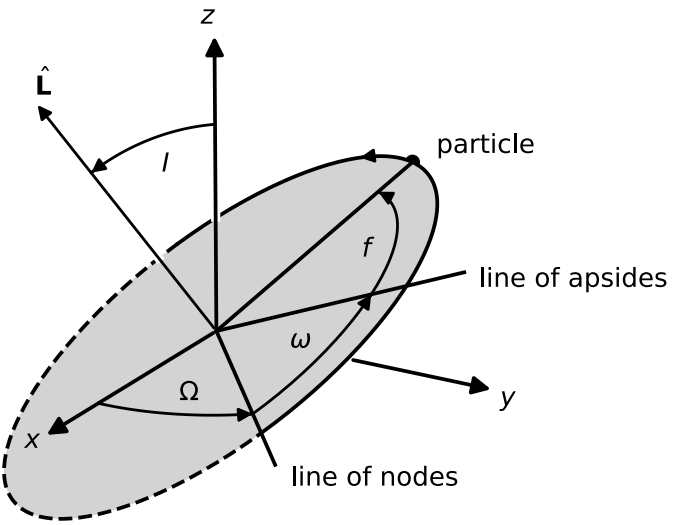

Figure 1.2: The angular elements of a Kepler orbit. The usual Cartesian coordinate axes are denoted by (x, y, z), the reference plane is $z = 0$, and the orbital plane is denoted by a solid curve above the equatorial plane ($z > 0$) and a dashed curve below. The plot shows the inclination I, the longitude of the ascending node Ω, the argument of periapsis ω and the true anomaly f.

argument of periapsis is often replaced by a variable called the **longitude of periapsis** which is defined as

$$\varpi \equiv \Omega + \omega. \tag{1.67}$$

For orbits with zero inclination, the longitude of periapsis has a simple interpretation—it is the azimuthal angle between the x-axis and the periapsis, consistent with our earlier definition of the same symbol following equation (1.29)—but if the inclination is nonzero, it is the sum of two angles

measured in different planes (the reference plane and the orbital plane).[8] Despite this awkwardness, for most purposes the three elements (Ω, ϖ, I) provide the most convenient way to specify the orientation of a Kepler orbit.

The **mean longitude** is

$$\lambda \equiv \varpi + \ell = \Omega + \omega + \ell, \tag{1.68}$$

where ℓ is the mean anomaly; like the longitude of perihelion, the mean longitude is the sum of angles measured in the reference plane (Ω) and the orbital plane ($\omega + \ell$).

Some of these elements are closely related to the Euler angles that describe the rotation of one coordinate frame into another (Appendix B.6). Let (x', y', z') be Cartesian coordinates in the **orbital reference frame**, defined such that the z'-axis points along the angular-momentum vector $\mathbf{L}$ and the x'-axis points toward periapsis, along the eccentricity vector $\mathbf{e}$. Then the rotation from the (x, y, z) reference frame to the orbital reference frame is described by the Euler angles

$$(\alpha, \beta, \gamma) = (\Omega, I, \omega). \tag{1.69}$$

The position and velocity of a particle in space at a given time t are specified by six orbital elements: two specify the size and shape of the orbit (e and a); three specify the orientation of the orbit (I, Ω and ω), and one specifies the location of the particle in the orbit (f, u, ℓ, λ, or t_0). Thus, for example, to find the Cartesian coordinates (x, y, z) in terms of the orbital elements, we write the position in the orbital reference frame as $(x', y', z') = r(\cos f, \sin f, 0)$ and use equation (1.69) and the rotation matrix for the transformation from primed to unprimed coordinates (eq. B.61):

$$\begin{aligned}
\frac{x}{r} &= \cos\Omega\cos(f+\omega) - \cos I \sin\Omega \sin(f+\omega), \\
\frac{y}{r} &= \sin\Omega\cos(f+\omega) + \cos I \cos\Omega \sin(f+\omega), \\
\frac{z}{r} &= \sin I \sin(f+\omega);
\end{aligned} \tag{1.70}$$

[8] Thus "longitude of periapsis" is a misnomer, since ϖ is *not* equal to the azimuthal angle of the eccentricity vector, except for orbits of zero inclination.

here r is given in terms of the orbital elements by equation (1.29).

When the eccentricity or inclination is small, the polar coordinate pairs e–ϖ and I–Ω are sometimes replaced by the eccentricity and inclination components[9]

$$k \equiv e\cos\varpi, \quad h \equiv e\sin\varpi, \quad q \equiv \tan I\cos\Omega, \quad p \equiv \tan I\sin\Omega. \tag{1.71}$$

The first two equations are the same as equations (1.31).

For some purposes the shape, size and orientation of an orbit can be described most efficiently using the angular-momentum and eccentricity vectors, $\mathbf{L}$ and $\mathbf{e}$. The two vectors describe five of the six orbital elements: the missing element is the one specifying the location of the particle in its orbit, f, u, ℓ, λ or t_0 (the six components of the two vectors determine only five elements, because $\mathbf{e}$ is restricted to the plane normal to $\mathbf{L}$).

Note that ω and Ω are undefined for orbits with zero inclination; ω and ϖ are undefined for circular orbits; and ϖ, Ω and I are undefined for radial orbits ($e \to 1$). In contrast the angular-momentum and eccentricity vectors are well defined for *all* orbits. The cost of avoiding indeterminacy is redundancy: instead of five orbital elements we need six vector components.

1.3.3 Gauss's f and g functions

A common task is to determine the position and velocity, $\mathbf{r}(t)$ and $\mathbf{v}(t)$, of a particle in a Kepler orbit given its position and velocity $\mathbf{r}_0$ and $\mathbf{v}_0$ at some initial time t_0. This can be done by converting $\mathbf{r}_0$ and $\mathbf{v}_0$ into the orbital elements $a, e, I, \omega, \Omega, \ell_0$, replacing ℓ_0 by $\ell = \ell_0 + n(t - t_0)$ and then reversing the conversion to determine the position and velocity from the new orbital elements. But there is a simpler method, due to Gauss.

Since the particle is confined to the orbital plane, and $\mathbf{r}_0, \mathbf{v}_0$ are vectors lying in this plane, we can write

$$\mathbf{r}(t) = f(t, t_0)\mathbf{r}_0 + g(t, t_0)\mathbf{v}_0, \tag{1.72}$$

[9] The function $\tan I$ in the elements q and p can be replaced by any function that is proportional to I as $I \to 0$. Various authors use I, $\sin\frac{1}{2}I$, and so forth. The function $\sin I$ is not used because it has the same value for I and $\pi - I$.

which defines **Gauss's f and g functions**. This expression also gives the velocity of the particle,

$$\mathbf{v}(t) = \frac{\partial f(t,t_0)}{\partial t}\mathbf{r}_0 + \frac{\partial g(t,t_0)}{\partial t}\mathbf{v}_0. \tag{1.73}$$

To evaluate f and g for bound orbits we use polar coordinates (r,ψ) and Cartesian coordinates (x,y) in the orbital plane, and assume that $\mathbf{r}_0$ lies along the positive x-axis ($\psi_0 = 0$). Then the components of equation (1.72) along the x- and y-axes are:

$$\begin{aligned} r(t)\cos\psi(t) &= f(t,t_0)r_0 + g(t,t_0)v_r(t_0), \\ r(t)\sin\psi(t) &= g(t,t_0)v_\psi(t_0), \end{aligned} \tag{1.74}$$

where v_r and v_ψ are the radial and azimuthal velocities. These equations can be solved for f and g:

$$\begin{aligned} f(t,t_0) &= \frac{r(t)}{r_0}\left[\cos\psi(t) - \frac{v_r(t_0)}{v_\psi(t_0)}\sin\psi(t)\right], \\ g(t,t_0) &= \frac{r(t)}{v_\psi(t_0)}\sin\psi(t). \end{aligned} \tag{1.75}$$

We use equations (1.16), (1.28), (1.29), (1.54) and the relation $\psi = f - f_0$ to replace the quantities on the right sides by orbital elements (unfortunately f is used to denote both true anomaly and one of Gauss's functions). The result is

$$\begin{aligned} f(t,t_0) &= \frac{\cos(f-f_0) + e\cos f}{1 + e\cos f}, \\ g(t,t_0) &= \frac{(1-e^2)^{3/2}\sin(f-f_0)}{n(1+e\cos f)(1+e\cos f_0)}. \end{aligned} \tag{1.76}$$

Since these expressions contain only the orbital elements n, e and f, they are valid in any coordinate system, not just the one we used for the derivation. For deriving velocities from equation (1.73), we need

$$\frac{\partial f(t,t_0)}{\partial t} = n\frac{e\sin f_0 - e\sin f - \sin(f-f_0)}{(1-e^2)^{3/2}},$$

$$\frac{\partial g(t,t_0)}{\partial t} = \frac{e\cos f_0 + \cos(f - f_0)}{1 + e\cos f_0}. \tag{1.77}$$

The f and g functions can also be expressed in terms of the eccentric anomaly, using equations (1.50) and (1.51a):

$$\begin{aligned} f(t,t_0) &= \frac{\cos(u-u_0) - e\cos u_0}{1 - e\cos u_0}, \\ g(t,t_0) &= \frac{1}{n}[\sin(u-u_0) - e\sin u + e\sin u_0], \\ \frac{\partial f(t,t_0)}{\partial t} &= -\frac{n\sin(u-u_0)}{(1-e\cos u)(1-e\cos u_0)}, \\ \frac{\partial g(t,t_0)}{\partial t} &= \frac{\cos(u-u_0) - e\cos u}{1 - e\cos u}. \end{aligned} \tag{1.78}$$

To compute $\mathbf{r}(t)$, $\mathbf{v}(t)$ from $\mathbf{r}_0 \equiv \mathbf{r}(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0)$ we use the following procedure. From equations (1.19) and (1.32) we have

$$\frac{1}{a} = \frac{2}{r} - \frac{v^2}{\mathbb{G}M}; \tag{1.79}$$

so we can compute the semimajor axis a from $r_0 = |\mathbf{r}_0|$ and $v_0 = |\mathbf{v}_0|$. Then Kepler's law (1.44) yields the mean motion n. The total angular momentum is $L = |\mathbf{r}_0 \times \mathbf{v}_0|$ and this yields the eccentricity e through equation (1.28). To determine the eccentric anomaly at t_0, we use equation (1.46) which determines $\cos u_0$, and then determine the quadrant of u_0 by observing that the radial velocity $\dot{r}$ is positive when $0 < u_0 < \pi$ and negative when $\pi < u_0 < 2\pi$. From Kepler's equation (1.49) we then find the mean anomaly ℓ_0 at $t = t_0$.

The mean anomaly at t is then $\ell = \ell_0 + n(t - t_0)$. By solving Kepler's equation numerically we can find the eccentric anomaly u. We may then evaluate the f and g functions using equations (1.78) and the position and velocity at t from equations (1.72) and (1.73).

1.4 Canonical orbital elements

The powerful tools of Lagrangian and Hamiltonian dynamics are essential for solving many of the problems addressed later in this book. A summary of the relevant aspects of this subject is given in Appendix D. In this section we show how Hamiltonian methods are applied to the two-body problem.

The Hamiltonian that describes the trajectory of a test particle around a point mass M at the origin is

$$H_{\rm K}(\mathbf{r},\mathbf{v}) = \tfrac{1}{2}\mathbf{v}^2 - \frac{\mathbb{G}M}{|\mathbf{r}|}. \tag{1.80}$$

Here $\mathbf{r}$ and $\mathbf{v}$ are the position and velocity, which together determine the position of the test particle in 6-dimensional phase space. The vectors $\mathbf{r}$ and $\mathbf{v}$ are a canonical coordinate-momentum pair.[10] Hamilton's equations read

$$\frac{{\rm d}\mathbf{r}}{{\rm d}t} = \frac{\partial H_{\rm K}}{\partial \mathbf{v}} = \mathbf{v}, \quad \frac{{\rm d}\mathbf{v}}{{\rm d}t} = -\frac{\partial H_{\rm K}}{\partial \mathbf{r}} = -\frac{\mathbb{G}M}{|\mathbf{r}|^3}\mathbf{r}. \tag{1.81}$$

These are equivalent to the usual equations of motion (1.11).

The advantage of Hamiltonian methods is that the equations of motion are the same in any set of phase-space coordinates $\mathbf{z} = (\mathbf{q},\mathbf{p})$ that are obtained from $(\mathbf{r},\mathbf{v})$ by a canonical transformation (Appendix D.6). For example, suppose that the test particle is also subject to an additional potential $\Phi(\mathbf{r},t)$ arising from some external mass distribution, such as another planet. Then the Hamiltonian and the equations of motion in the original variables are

$$H(\mathbf{r},\mathbf{v},t) = H_{\rm K}(\mathbf{r},\mathbf{v}) + \Phi(\mathbf{r},t), \quad \frac{{\rm d}\mathbf{r}}{{\rm d}t} = \frac{\partial H}{\partial \mathbf{v}}, \quad \frac{{\rm d}\mathbf{v}}{{\rm d}t} = -\frac{\partial H}{\partial \mathbf{r}}. \tag{1.82}$$

[10] We usually—but not always—adopt the convention that the canonical momentum $\mathbf{p}$ that is conjugate to the position $\mathbf{r}$ is velocity $\mathbf{v}$ rather than Newtonian momentum $m\mathbf{v}$. Velocity is often more convenient than Newtonian momentum in gravitational dynamics since the acceleration of a body in a gravitational potential is independent of mass. If necessary, the convention used in a particular set of equations can be verified by dimensional analysis.

In the new canonical variables,[11]

$$H(\mathbf{z},t) = H_{\rm K}(\mathbf{z}) + \Phi(\mathbf{z},t), \quad \frac{{\rm d}\mathbf{q}}{{\rm d}t} = \frac{\partial H}{\partial \mathbf{p}}, \quad \frac{{\rm d}\mathbf{p}}{{\rm d}t} = -\frac{\partial H}{\partial \mathbf{q}}. \tag{1.83}$$

If the additional potential is small compared to the Kepler potential, $|\phi(\mathbf{r},t)| \ll \mathbb{G}M/r$, then the trajectory will be close to a Kepler ellipse. Therefore the analysis can be much easier if we use new coordinates and momenta $\mathbf{z}$ in which Kepler motion is simple.[12] The six orbital elements—semimajor axis a, eccentricity e, inclination I, longitude of the ascending node Ω, argument of periapsis ω and mean anomaly ℓ—satisfy this requirement as all of the elements are constant except for ℓ, which increases linearly with time. This set of orbital elements is not canonical, but they can be rearranged to form a canonical set called the **Delaunay variables**, in which the coordinate-momentum pairs are:

$$\begin{aligned} &\ell, && \Lambda \equiv (\mathbb{G}Ma)^{1/2}, \\ &\omega, && L = [\mathbb{G}Ma(1-e^2)]^{1/2}, \\ &\Omega, && L_z = L\cos I. \end{aligned} \tag{1.84}$$

Here L_z is the z-component of the angular-momentum vector $\mathbf{L}$ (see Figure 1.2); $L = |\mathbf{L}|$ (eq. 1.28); and Λ is sometimes called the **circular angular momentum** since it equals the angular momentum for a circular orbit. The proof that the Delaunay variables are canonical is given in Appendix E.

The Kepler Hamiltonian (1.80) is equal to the energy per unit mass, which is related to the semimajor axis by equation (1.32); thus

$$H_{\rm K} = -\frac{\mathbb{G}M}{2a} = -\frac{(\mathbb{G}M)^2}{2\Lambda^2}. \tag{1.85}$$

[11] For notational simplicity, we usually adopt the convention that the Hamiltonian and the potential are functions of position, velocity, or position in phase space rather than functions of the coordinates. Thus $H(\mathbf{r},\mathbf{v},t)$ and $H(\mathbf{z},t)$ have the same value if $(\mathbf{r},\mathbf{v})$ and $\mathbf{z}$ are coordinates of the same phase-space point in different coordinate systems.

[12] However, the additional potential $\Phi(\mathbf{z},t)$ is often much more complicated in the new variables; for a start, it generally depends on all six phase-space coordinates rather than just the three components of $\mathbf{r}$. Since dynamics is more difficult than potential theory, the tradeoff—simpler dynamics at the cost of more complicated potential theory—is generally worthwhile.

Since the Kepler Hamiltonian is independent of the coordinates, the momenta Λ, L and L_z are all constants along a trajectory in the absence of additional forces; such variables are called **integrals of motion**. Because the Hamiltonian is independent of the momenta L and L_z their conjugate coordinates ω and Ω are also constant, and $\mathrm{d}\ell/\mathrm{d}t = \partial H_{\mathrm{K}}/\partial\Lambda = (\mathbb{G}M)^2\Lambda^{-3} = (\mathbb{G}M/a^3)^{1/2} = n$, where n is the mean motion defined by Kepler's law (1.44). Of course, all of these conclusions are consistent with what we already know about Kepler orbits.

Because the momenta are integrals of motion in the Kepler Hamiltonian and the coordinates are angular variables that range from 0 to 2π, the Delaunay variables are also angle-action variables for the Kepler Hamiltonian (Appendix D.7). For an application of this property, see Box 1.2.

One shortcoming of the Delaunay variables is that they have coordinate singularities at zero eccentricity, where ω is ill-defined, and zero inclination, where Ω and ω are ill-defined. Even if the eccentricity or inclination of an orbit is small but nonzero, these elements can vary rapidly in the presence of small perturbing forces, so numerical integrations that follow the evolution of the Delaunay variables can grind to a near-halt.

To address this problem we introduce other sets of canonical variables derived from the Delaunay variables. We write $\mathbf{q} = (\ell, \omega, \Omega)$, $\mathbf{p} = (\Lambda, L, L_z)$ and introduce a generating function $S_2(\mathbf{q}, \mathbf{P})$ as described in Appendix D.6.1. From equations (D.63)

$$\mathbf{p} = \frac{\partial S_2}{\partial \mathbf{q}}, \quad \mathbf{Q} = \frac{\partial S_2}{\partial \mathbf{P}}, \tag{1.86}$$

and these equations can be solved for the new variables $\mathbf{Q}$ and $\mathbf{P}$. For example, if $S_2(\mathbf{q}, \mathbf{P}) = (\ell + \omega + \Omega)P_1 + (\omega + \Omega)P_2 + \Omega P_3$ then the new coordinate-momentum pairs are

$$\begin{aligned} \lambda = \ell + \omega + \Omega, &\quad \Lambda, \\ \varpi = \omega + \Omega, &\quad L - \Lambda = (\mathbb{G}Ma)^{1/2}\left[(1-e^2)^{1/2} - 1\right], \\ \Omega, &\quad L_z - L = (\mathbb{G}Ma)^{1/2}(1-e^2)^{1/2}(\cos I - 1). \end{aligned} \tag{1.87}$$

Here we have reintroduced the mean longitude λ (eq. 1.68) and the longitude of periapsis ϖ (eq. 1.67). Since λ and ϖ are well defined for orbits of

zero inclination, these variables are better suited for describing nearly equatorial prograde orbits. The longitude of the node Ω is still ill-defined when the inclination is zero, although if the motion is known or assumed to be restricted to the equatorial plane the first two coordinate-momentum pairs are sufficient to describe the motion completely.

With the variables (1.87) two of the momenta $L - \Lambda$ and $L_z - L$ are always negative. For this reason some authors prefer to use the generating function $S_2(\mathbf{q}, \mathbf{P}) = (\ell + \omega + \Omega)P_1 - (\omega + \Omega)P_2 - \Omega P_3$, which yields new coordinates and momenta

$$\begin{aligned} \lambda = \ell + \omega + \Omega, &\qquad \Lambda, \\ -\varpi = -\omega - \Omega, &\qquad \Lambda - L = (\mathbb{G}Ma)^{1/2}\bigl[1 - (1-e^2)^{1/2}\bigr], \\ -\Omega, &\qquad L - L_z = (\mathbb{G}Ma)^{1/2}(1-e^2)^{1/2}(1-\cos I). \end{aligned} \tag{1.88}$$

Another set is given by the generating function $S_2(\mathbf{q}, \mathbf{P}) = \ell P_1 + (\ell + \omega)P_2 + \Omega P_3$, which yields coordinates and momenta

$$\begin{aligned} \ell, &\qquad \Lambda - L = (\mathbb{G}Ma)^{1/2}\bigl[1 - (1-e^2)^{1/2}\bigr], \\ \ell + \omega, &\qquad L = (\mathbb{G}Ma)^{1/2}(1-e^2)^{1/2}, \\ \Omega, &\qquad L_z = (\mathbb{G}Ma)^{1/2}(1-e^2)^{1/2}\cos I. \end{aligned} \tag{1.89}$$

The action $\Lambda - L$ that appears in (1.88) and (1.89) has a simple physical interpretation. At a given angular momentum L, the radial motion in the Kepler orbit is governed by the Hamiltonian $H(r, p_r) = \frac{1}{2}p_r^2 + \frac{1}{2}L^2/r^2 - \mathbb{G}M/r$ (cf. eq. 1.18). The corresponding action is $J_r = \oint \mathrm{d}r\, p_r/(2\pi)$ (eq. D.72). The radial momentum $p_r = \dot{r}$ by Hamilton's equations; writing r and $\dot{r}$ in terms of the eccentric anomaly u using equations (1.46) and (1.54) gives

$$J_r = \frac{na^2e^2}{2\pi}\int_0^{2\pi} \mathrm{d}u\, \frac{\sin^2 u}{1 - e\cos u} = na^2\bigl[1 - (1-e^2)^{1/2}\bigr] = \Lambda - L. \tag{1.90}$$

Thus $\Lambda - L$ is the action associated with the radial coordinate, sometimes called the **radial action**. The radial action is zero for circular orbits and equal to $\frac{1}{2}(\mathbb{G}Ma)^{1/2}e^2$ when $e \ll 1$.

Box 1.2: The effect of slow mass loss on a Kepler orbit

If the mass of the central object is changing, the constant M in equations like (1.11) must be replaced by a variable $M(t)$. We assume that the evolution of the mass is (i) due to some spherically symmetric process (e.g., a spherical wind from the surface of a star), so there is no recoil force on the central object; (ii) slow, in the sense that $|\mathrm{d}M/\mathrm{d}t| \ll M/P$, where P is the orbital period of a planet.

Since the gravitational potential remains spherically symmetric, the angular momentum $L = (\mathbb{G}Ma)^{1/2}(1-e^2)^{1/2}$ (eq. 1.28) is conserved.

Moreover, actions are adiabatic invariants (Appendix D.10), so during slow mass loss the actions remain almost constant. The Delaunay variable $\Lambda = (\mathbb{G}Ma)^{1/2}$ (eq. 1.84) is an action. Since Λ and L are distinct functions of Ma and e, and both are conserved—one adiabatically and one exactly—then both Ma and e are also conserved. In words, during slow mass loss the orbit expands, with $a(t) \propto 1/M(t)$, but its eccentricity remains constant. The accuracy of this approximate conservation law is explored in Problem 2.8.

At present the Sun is losing mass at a rate $\dot{M}/M = -(1.1\pm0.3)\times10^{-13}\ \mathrm{yr}^{-1}$ (Pitjeva et al. 2021). Near the end of its life, the Sun will become a red-giant star and expand dramatically in radius and luminosity. At the tip of the red-giant branch, about 7.6 Gyr from now, the solar radius will be about 250 times its present value or 1.2 au and its luminosity will be 2 700 times its current value (Schröder & Connon Smith 2008). During its evolution up the red-giant branch the Sun will lose about 30% of its mass, and according to the arguments above the Earth's orbit will expand by the same fraction. Whether or not the Earth escapes being engulfed by the Sun depends on the uncertain relative rates of the Sun's future expansion and its mass loss.

Finally, consider the generating function $S_2(\mathbf{q},\mathbf{P}) = P_1(\ell+\omega+\Omega) + \frac{1}{2}{P_2}^2\cot(\omega+\Omega) + \frac{1}{2}{P_3}^2\cot\Omega$, which yields the **Poincaré variables**

$$\begin{aligned} \lambda &= \ell+\omega+\Omega, & \Lambda,\\ [2(\Lambda-L)]^{1/2}\cos\varpi, & & [2(\Lambda-L)]^{1/2}\sin\varpi,\\ [2(L-L_z)]^{1/2}\cos\Omega, & & [2(L-L_z)]^{1/2}\sin\Omega. \end{aligned} \tag{1.91}$$

These are well defined even when $e = 0$ or $I = 0$. In particular, in the limit

of small eccentricity and inclination the Poincaré variables simplify to

$$\begin{aligned} \lambda, \qquad & \Lambda, \\ (\mathbb{G}Ma)^{1/4} e\cos\varpi, \qquad & (\mathbb{G}Ma)^{1/4} e\sin\varpi, \\ (\mathbb{G}Ma)^{1/4} I\cos\Omega, \qquad & (\mathbb{G}Ma)^{1/4} I\sin\Omega. \end{aligned} \tag{1.92}$$

Apart from the constant of proportionality $(\mathbb{G}Ma)^{1/4}$ these are just the Cartesian elements defined in equations (1.71).

All of these sets of orbital elements remain ill-defined when the inclination $I = \pi$ (retrograde orbits in the reference plane) or $e = 1$ (orbits with zero angular momentum); however, such orbits are relatively rare in planetary systems.[13]

1.5 Units and reference frames

Measurements of the trajectories of solar-system bodies are some of the most accurate in any science, and provide exquisitely precise tests of physical theories such as general relativity. Precision of this kind demands careful definitions of units and reference frames. These will only be treated briefly in this book, since our focus is on understanding rather than measuring the behavior of celestial bodies.

Tables of physical, astronomical and solar-system constants are given in Appendix A.

1.5.1 Time

The unit of time is the Système Internationale or SI second (s), which is defined by a fixed value for the frequency of a particular transition of cesium atoms. Measurements from several cesium frequency standards are combined to form a timescale known as **International Atomic Time** (TAI).

[13] A set of canonical coordinates and momenta that is well defined for orbits with zero angular momentum is described by Tremaine (2001). Alternatively, the orbit can be described using the angular-momentum and eccentricity vectors, which are well defined for any Kepler orbit; see §5.3 or Allan & Ward (1963).

In contrast, **Universal Time** (UT) employs the Earth's rotation on its axis as a clock. UT is not tied precisely to this clock because the Earth's angular speed is not constant. The most important nonuniformity is that the length of the day increases by about 2 milliseconds per century because of the combined effects of tidal friction and post-glacial rebound. There are also annual and semiannual variations of a few tenths of a millisecond. Despite these irregularities, a timescale based approximately on the Earth's rotation is essential for everyday life: for example, we would like noon to occur close to the middle of the day. Therefore all civil timekeeping is based on **Coordinated Universal Time** (UTC), which is an atomic timescale that is kept in close agreement with UT by adding extra seconds ("leap seconds") at regular intervals.[14] Thus UTC is a discontinuous timescale composed of segments that follow TAI apart from a constant offset.

An inconvenient feature of TAI for high-precision work is that it measures the rate of clocks at sea level on the Earth; general relativity implies that the clock rate depends on the gravitational potential and hence the rate of TAI is different from the rate measured by the same clock outside the solar system. For example, the rate of TAI varies with a period of one year and an amplitude of 1.7 milliseconds because of the eccentricity of the Earth's orbit. **Barycentric Coordinate Time** (TCB) measures the proper time experienced by a clock that co-moves with the center of mass of the solar system but is far outside it. TCB ticks faster than TAI by 0.49 seconds per year, corresponding to a fractional speedup of 1.55×10^{-8}.

The times of astronomical events are usually measured by the **Julian date**, denoted by the prefix JD. The Julian date is expressed in days and decimals of a day. Each day has $86\,400$ seconds. The Julian year consists of exactly 365.25 days and is denoted by the prefix J. For example, the initial conditions of orbits are often specified at a standard epoch, such as

$$\text{J2000.0} = \text{JD}\ 2\,451\,545.0, \tag{1.93}$$

which corresponds roughly to noon in England on January 1, 2000. The modified Julian day is defined as

$$\text{MJD} = \text{JD} - 2\,400\,000.5; \tag{1.94}$$

[14] The utility of leap seconds is controversial, and their future is uncertain.

the integer offset reduces the length of the number specifying relatively recent dates, and the half-integer offset ensures that the MJD begins at midnight rather than noon.

In contrast to SI seconds (s) and days (1 d = 86 400 s) there is no unique definition of "year": most astronomers use the Julian year but there is also the anomalistic year, sidereal year, and the like (see footnote 7). For this reason the use of "year" as a precise unit of time is deprecated. However, we shall occasionally use years, megayears and gigayears (abbreviated yr, Myr, Gyr) to denote 1, 10^6 and 10^9 Julian years. The age of the solar system is 4.567 Gyr and the age of the Universe is 13.79 Gyr. The future lifetime of the solar system as we know it is about 7.6 Gyr (see Box 1.2).

The SI unit of length is defined in terms of the second, such that the speed of light is exactly

$$c \equiv 299\,792\,458\,\mathrm{m\,s^{-1}}. \tag{1.95}$$

1.5.2 Units for the solar system

The history of the determination of the scale of the solar system and the mass of the Sun is worth a brief description. Until the mid-twentieth century virtually all of our data on the orbits of the Sun and planets came from tracking their positions on the sky as functions of time. This information could be combined with the theory of Kepler orbits developed earlier in this chapter (plus small corrections arising from mutual interactions between the planets, which are handled by the methods of Chapter 4) to determine all of the orbital elements of the planets including the Earth, except for the overall scale of the system. Thus, for example, the ratio of semimajor axes of any two planets was known to high accuracy, but the values of the semimajor axes in meters were not.[15] To reflect this uncertainty, astronomers introduced the concept of the **astronomical unit** (abbreviated au), which was originally defined to be the semimajor axis of the Earth's orbit. Thus the semimajor axes of the planets were known in astronomical units long before the value of the astronomical unit was known to comparable accuracy.

[15] This indeterminacy follows from dimensional analysis: measurements of angles and times cannot be combined to find a quantity with dimensions of length.

Since Kepler's third law (1.44) is $\mathbb{G}M = 4\pi^2 a^3/P^2$, and orbital periods P can be determined so long as we have accurate clocks, any fractional uncertainty ϵ in the astronomical unit implies a fractional uncertainty of 3ϵ in the **solar mass parameter** $\mathbb{G}M_\odot$.

Over the centuries, the astronomical unit was measured by many different techniques, including transits of Venus, parallaxes of nearby solar-system objects over Earth-sized baselines and stellar aberration. Nevertheless, even in the 1950s the astronomical unit was only known with a fractional accuracy of about 10^{-3}. Soon after, radar observations of Venus and Mars and ranging data from interplanetary spacecraft reduced the uncertainty by several orders of magnitude. The current uncertainty is much smaller than variations in the Earth's semimajor axis due to perturbations from the other planets, so in 2012 the International Astronomical Union (IAU) re-defined the astronomical unit to be an exact unit of length,

$$1\,\text{au} \equiv 149\,597\,870\,700\,\text{m}. \tag{1.96}$$

Distances to other stars are measured in units of **parsecs** (abbreviated pc), the distance at which 1 au subtends one second of arc. Thus the parsec is also an exact unit of length, though an irrational number of meters:

$$1\,\text{pc} = \frac{648\,000}{\pi}\,\text{au} \simeq 3.085\,677\,6 \times 10^{16}\,\text{m}. \tag{1.97}$$

The determination of the scale of the solar system allowed the determination of $\mathbb{G}M_\odot$ to comparable accuracy. In contrast, the gravitational constant $\mathbb{G}$, determined by laboratory experiments, is only known to a fractional accuracy of 2×10^{-5} (see Appendix A). Therefore the masses of the Sun and solar-system planets are much less well known than $\mathbb{G}$ times the masses, and for accurate work they should always be quoted along with the assumed value of $\mathbb{G}$.

In 2015 the IAU recommended that orbit calculations should be based on the nominal value of the solar mass parameter

$$\mathbb{G}M_\odot \equiv 1.327\,124\,4 \times 10^{20}\,\text{m}^3\,\text{s}^{-2}. \tag{1.98}$$

The adjective "nominal" means that this should be understood as a standard conversion factor that is close to the "actual" value (probably with a fractional error of less than 1×10^{-9}). For most dynamical problems it is better to use a consistent set of constants that is common to the whole community rather than the best current estimate of each constant.

1.5.3 The solar system reference frame

The **Barycentric Celestial Reference System** (BCRS) is a coordinate system created in 2000 by the IAU. The system uses harmonic coordinates (eq. J.6), with origin at the solar system barycenter and time given by TCB. This is the reference system appropriate for solving the equations of motion of solar system bodies. The orientation of the BCRS coordinate system coincides with that of the International Celestial Reference System (ICRS), which is defined by the adopted angular coordinates of a set of extragalactic radio sources. For more detail see Kaplan (2005) and Urban & Seidelmann (2013).

These definitions are based on the assumption that the local inertial reference frame (the BCRS) is not rotating relative to the distant universe (the ICRS), sometimes called **Mach's principle**. This assumption is testable: the relative rotation of these frames is consistent with zero and less than 4×10^{-5} arcsec yr^{-1} (Folkner 2010).

1.6 Orbital elements for exoplanets

The orbital elements of extrasolar planets ("exoplanets") are much more difficult to determine accurately than the elements of solar-system bodies. In most cases we only know some of the six orbital elements, depending on the detection method.

Here we describe three methods of planet detection based on the classical observational techniques of spectroscopy, photometry, astrometry and imaging. We do not discuss a further important technique, gravitational microlensing, because it measures only the mass of the planet and its projected separation from the host star and thus provides only limited constraints on the orbital elements and dynamics (Gaudi 2011).

1.6.1 Radial-velocity planets

One of the most powerful methods to detect and characterize exoplanets is through periodic variations in the velocity of their host star, which arise as both star and planet orbit around their common center of mass.[16] These variations can be detected through small Doppler shifts in the stellar spectrum.[17]

To illustrate the analysis, we consider a system containing a single planet. The star is at $\mathbf{r}_0$ and the planet is at $\mathbf{r}_1$. The velocity of the star is given by the time derivative of equation (1.6),

$$\mathbf{v}_0 = \mathbf{v}_{\rm cm} - \frac{m_1}{m_0 + m_1}\mathbf{v}, \tag{1.99}$$

where $\mathbf{v}$ is the velocity of the planet relative to the star. The velocity of the center of mass $\mathbf{v}_{\rm cm}$ is constant (eq. 1.7). We may choose our coordinates such that the positive z-axis is parallel to the line of sight from the observer to the system and pointing away from the observer; thus edge-on orbits have $I = 90°$, face-on orbits have $I = 0$, and positive line-of-sight velocity implies that the star is receding from us. Then the line-of-sight velocity of the star relative to the center of mass is

$$v_{\rm los} \equiv (\mathbf{v}_0 - \mathbf{v}_{\rm cm}) \cdot \hat{\mathbf{z}} = -\frac{m_1}{m_0 + m_1}\mathbf{v} \cdot \hat{\mathbf{z}}. \tag{1.100}$$

From equation (1.70), $\mathbf{v} \cdot \hat{\mathbf{z}} = \dot{z} = \sin I[\dot{r}\sin(f + \omega) + r\cos(f + \omega)\dot{f}] = \sin I[v_r \sin(f+\omega) + v_\psi \cos(f+\omega)]$. Then using equations (1.54) and (1.55),

$$v_{\rm los} = -\frac{m_1}{m_0 + m_1}\frac{na}{(1 - e^2)^{1/2}}\sin I\big[\cos(f + \omega) + e\cos\omega\big]. \tag{1.101}$$

[16] The possibility of detecting planets by radial-velocity variations and by transits was first discussed in a prescient short paper by Struve (1952).

[17] Unfortunately the term "radial velocity" is commonly used to denote two different quantities: (i) the component of the planet's velocity relative to the host star along the line joining them, and (ii) the component of the star's velocity relative to the observer along the line joining them. In practice the meaning is usually clear from the context, but when there is the possibility of confusion we shall use the term "line-of-sight velocity" as an unambiguous replacement for (ii).

Since the orbital period $P = 2\pi a^{3/2}/[\,\mathbb{G}(m_0+m_1)\,]^{1/2}$ is directly observable while the semimajor axis is not, we eliminate a in favor of P to obtain

$$v_{\rm los} = -\frac{m_1}{m_0+m_1}\left[\frac{2\pi\,\mathbb{G}(m_0+m_1)}{P}\right]^{1/3}\frac{\sin I}{(1-e^2)^{1/2}}\big[\,\cos(f+\omega)+e\cos\omega\big]. \tag{1.102}$$

Using equations (1.50) and (1.51a), this result can also be expressed in terms of the eccentric anomaly,

$$v_{\rm los} = -\frac{m_1}{m_0+m_1}\left[\frac{2\pi\,\mathbb{G}(m_0+m_1)}{P}\right]^{1/3}\sin I \times\frac{(1-e^2)^{1/2}\cos u\cos\omega-\sin u\sin\omega}{1-e\cos u}. \tag{1.103}$$

To obtain $v_{\rm los}(t)$, the line-of-sight velocity as a function of time (the **velocity curve**), we write the mean anomaly as $\ell = 2\pi(t-t_0)/P$ where t_0 is the time of periapsis passage, solve Kepler's equation (1.49) for u, and then substitute u into equation (1.103). The velocity curve is not sinusoidal unless the orbit is circular, but it is still useful to define the **semi-amplitude** K as half the difference between the maximum and minimum velocity. From equation (1.102) the extrema of $v_{\rm los}$ occur at $f=-\omega$ and $f=\pi-\omega$, so

$$K = \frac{m_1}{m_0+m_1}\left[\frac{2\pi\,\mathbb{G}(m_0+m_1)}{P}\right]^{1/3}\frac{\sin I}{(1-e^2)^{1/2}}. \tag{1.104}$$

These results tell us what can and cannot be determined from the velocity curve. The orbital period P is equal to the period of the velocity curve, and the eccentricity e and argument of periapsis ω can be determined from the shape of the curve. The longitude of the node Ω is not constrained. The masses of the star and planet, m_0 and m_1, and the inclination I cannot be individually determined, only the combination

$$\mu \equiv \frac{m_1^3\sin^3 I}{(m_0+m_1)^2}, \tag{1.105}$$

known as the **mass function**. The mass function is related to the semi-amplitude and period by

$$\mu = \frac{P}{2\pi\,\mathbb{G}} K^3 (1 - e^2)^{3/2}. \tag{1.106}$$

Since exoplanet masses are usually much smaller than the mass of their host star, and the mass of the host star can usually be determined from its spectral properties, the mass function yields a combination of the planetary mass and orbital inclination, $m_1 \sin I$.

The semi-amplitude K varies as $a^{-1/2}$ for planets of a given mass, so radial-velocity searches are most sensitive to planets orbiting close to the host star. Planets whose orbital periods are much larger than the survey duration will contribute a constant acceleration or linear drift to the line-of-sight velocity of the host star, and this signal provides evidence for the existence of a distant planet but only weak constraints on its properties.

The most precisely measured radial-velocity planets are found orbiting pulsars. The pulsar emits pulsed radio signals at regular intervals Δt. The pulse emitted at $t_n = n\Delta t + \text{const}$ arrives at $t'_n = t_n + r(t_n)/c$ where $r(t_n)$ is the distance of the pulsar at t_n and c is the speed of light. Now write $r(t) = \text{const} + v_{\rm los} t$ where $v_{\rm los}$ is the line-of-sight velocity of the pulsar, and we have $\Delta t'_n = t'_{n+1} - t'_n = \Delta t(1 + v_{\rm los}/c)$. Thus measuring the intervals between pulses yields the line-of-sight velocity (up to an undetermined constant, since the rest-frame pulse interval Δt is unknown), and as usual periodic variations in the line-of-sight velocity are the signature of a planet.

Pulsar planets are rare, presumably because planets cannot survive the supernova explosion that creates the pulsar, and only a handful are known. The prototype is the system of three planets discovered around the pulsar PSR B1257+12 (Wolszczan & Frail 1992).

1.6.2 Transiting planets

In a small fraction of cases, a planetary system is oriented such that one or more of its planets crosses the face of the host star as seen from Earth, an

event known as a **transit**.[18] During the transit, there is a characteristic dip in the stellar flux, which repeats with a period equal to the planet's orbital period.

Suppose that the planet has radius $R_{\rm p}$ and the star has radius R_*. In most cases $R_{\rm p} \ll R_*$; for example, the radii of Earth and Jupiter relative to the Sun are $R_\oplus/R_\odot = 0.009\,153$ and $R_{\rm J}/R_\odot = 0.099\,37$.[19] During a transit the visible area of the stellar disk is reduced to a fraction $1 - f$ of its unobscured value, where

$$f = \frac{R_{\rm p}^2}{R_*^2}, \tag{1.107}$$

and the flux from the star is reduced by a similar amount (depending on limb darkening, to be discussed later in this subsection). An observer watching Earth or Jupiter transit the Sun would find $f = 8.377\times10^{-5}$ and $f = 0.009\,88$ respectively. With current technology, Jupiter-like transits can be detected from the ground but Earth-like transits can only be detected by space-based observatories.

The probability that a planet will transit depends strongly on its semimajor axis. To determine this probability, we again use a coordinate system in which the z-axis is parallel to the line of sight. Then the planet transits if and only if the minimum value of $x^2 + y^2$ is less than $(R_* + R_{\rm p})^2$. From equations (1.70), $x^2 + y^2 = r^2 - z^2 = r^2[1 - \sin^2 I \sin^2(f + \omega)]$ so the minimum value of $x^2 + y^2$ is $r^2 \cos^2 I$. Therefore if the planet is on a circular orbit with semimajor axis a, it transits if and only if $|\cos I| < (R_* + R_{\rm p})/a$. If the distribution of orientations of the planetary orbits is random—an untested

[18] Transits and occultations are usually distinguished from eclipses. In an eclipse (e.g., an eclipse of the Sun by the Moon) both bodies have similar angular size. In a transit (e.g., a transit of Venus across the Sun) a small body passes in front of a large one, and in an occultation a small body passes behind a large one.

[19] Planets are not perfect spheres: in general, the polar radius $R_{\rm pol}$ of a rotating planet is smaller than its equatorial radius $R_{\rm eq}$, and the planet is said to have an **equatorial bulge** (Box 1.3). If we assume that the spin and orbital axes of the planet are aligned, then both are normal to the line of sight if the planet transits the star. Approximating the shape of the planet as an ellipse, its area on the plane normal to the line of sight is $\pi R_{\rm eq} R_{\rm pol}$ so the effective radius for computing the transit depth is $R_{\rm eff} = (R_{\rm eq} R_{\rm pol})^{1/2}$. For the Earth and Jupiter the effective radii are $R_{\oplus,{\rm eff}} = 6\,367.4$ km and $R_{\rm J,eff} = 69\,134$ km. In contrast the Sun is nearly spherical, with a fractional difference in the polar and equatorial radii $\lesssim 10^{-5}$.

but extremely plausible assumption—then $|\cos I|$ is uniformly distributed between 0 and 1, so the probability of transit is

$$p = \frac{R_* + R_{\rm p}}{a}. \tag{1.108}$$

A useful reference time for the duration of the transit is

$$\tau_0 = \frac{2R_*}{v} = 2R_* \left(\frac{a}{\mathbb{G}M_*}\right)^{1/2} = 12.98\,\text{hours}\,\frac{R_*}{R_\odot}\left(\frac{a}{\text{au}}\frac{M_\odot}{M_*}\right)^{1/2}. \tag{1.109}$$

Here v is the planet's orbital velocity, M_* is the stellar mass, and a is the planet's semimajor axis; in deriving these equations we have assumed that the planet's orbit is circular. The reference time equals the actual transit time only if the planet radius $R_{\rm p} \ll R_*$, the stellar radius $R_* \ll a$, and the transit passes through the center of the star. The actual transit time is usually shorter than τ_0 since the planet travels along a chord across the star rather than through its center.

The interval between transits equals the orbital period (eq. 1.43),

$$P = 2\pi\left(\frac{a^3}{\mathbb{G}M_*}\right)^{1/2}. \tag{1.110}$$

The shape and duration of the transit event can be described more accurately using Figure 1.3. The point of closest approach of the planet to the center of the star is bR_* where the **impact parameter** b is a dimensionless number in the range 0 to ~ 1. There are four milestones during the transit event: first contact, where the projected planetary disk first touches the edge of the star; second contact, where the entire planetary disk first obscures the star, third contact, the last time at which the entire planetary disk obscures the star, and fourth contact, when the transit ends. Between first and second contact the flux from the star is steadily decreasing as more and more of the stellar disk is obscured; between second and third contact the flux is constant; and between third and fourth contact the flux is steadily returning to its original value. If the closest approach to the center of the star is at $t = 0$, then straightforward trigonometry shows that the times associated

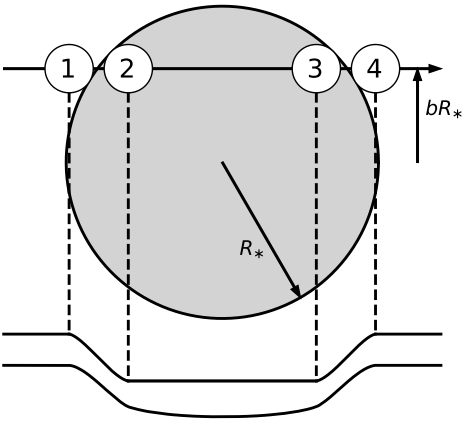

Figure 1.3: The geometry of a planetary transit. The large shaded circle of radius R_* shows the disk of the host star, and the unshaded circles of radius $R_{\rm p}$ show the position of the planetary disk at first, second, third and fourth contact. The minimum distance between the centers of the planet and the star is bR_*, where b is the impact parameter. In this image $b = 0.6$ and $R_{\rm p}/R_* = 0.15$. The curves at the bottom of the figure show the stellar flux as a function of time in two cases: no limb darkening (top), and solar limb darkening (bottom) as described in the paragraph containing equation (1.114). Analytic expressions for transit light curves are given by Sackett (1999), Mandel & Agol (2002) and Seager & Mallén-Ornelas (2003).

with these events are

$$t_4 = -t_1 = \frac{1}{v}\Big[(R_* + R_{\rm p})^2 - b^2 R_*^2\Big]^{1/2} = \tfrac{1}{2}\tau_0\big[(1 + R_{\rm p}/R_*)^2 - b^2\big]^{1/2},$$
$$t_3 = -t_2 = \frac{1}{v}\Big[(R_* - R_{\rm p})^2 - b^2 R_*^2\Big]^{1/2} = \tfrac{1}{2}\tau_0\big[(1 - R_{\rm p}/R_*)^2 - b^2\big]^{1/2}. \tag{1.111}$$

Here we have assumed that $R_* \ll a$ so the planet travels across the star at nearly constant velocity v; an equivalent constraint is that the transit duration is much less than the orbital period, $\tau_0 \ll P$. The total duration of the

transit is

$$t_4 - t_1 = \frac{2}{v}\Big[(R_* + R_{\rm p})^2 - b^2 R_*^2\Big]^{1/2} = \tau_0\Big[(1 + R_{\rm p}/R_*)^2 - b^2\Big]^{1/2}, \quad (1.112)$$

and the duration of the flat part of the transit, between second and third contact, is

$$t_3 - t_2 = \tau_0\Big[(1 - R_{\rm p}/R_*)^2 - b^2\Big]^{1/2}. \quad (1.113)$$

What can we measure from the transit depth, duration and shape? The fractional depth f of the transit determines the ratio of the planetary and stellar radii $R_{\rm p}/R_*$ through equation (1.107). Once this is known, the total duration $t_4 - t_1$ (eq. 1.112) and the duration of the flat part of the transit $t_3 - t_2$ (eq. 1.113) give two constraints on the impact parameter b and the reference time τ_0, so both can be determined. If the stellar mass M_* and radius R_* can be determined from the star's luminosity, colors and spectrum then equations (1.109) for the reference time and (1.43) for the orbital period give two constraints on the semimajor axis: if these agree then the planetary orbit is likely to be circular, and if not it must be eccentric.

This simple model predicts that the flux from the star is constant between second and third contact, which requires that the surface brightness of the star is uniform. In practice the surface brightness of the stellar disk is usually higher near the center, a phenomenon called **limb darkening**. One common parametrization of limb darkening is that the surface brightness at distance R from the center of the stellar disk of radius R_* is given by

$$\frac{I(R)}{I(0)} = 1 - a(1-\mu) - b(1-\mu)^2, \quad \text{where} \quad \mu = (1 - R^2/R_*^2)^{1/2}. \quad (1.114)$$

The limb-darkening coefficients a and b depend on the spectral type of the star and the wavelength range in which the surface brightness is measured. For a solar-type star measured in the Kepler wavelength band, $a \simeq 0.41$ and $b \simeq 0.26$.[20]

The depth of a transit (eq. 1.107) is independent of the semimajor axis a of the planet, but the probability that a planet will transit varies as a^{-1} (eq.

[20] Limb-darkening models for a wide range of stars are described in Claret & Bloemen (2011).

1.108), so transit searches are most sensitive to planets close to the host star. Planets whose orbital periods are larger than the survey duration are difficult to verify: a useful rule of thumb is that at least three transits are needed for a secure detection.

1.6.3 Astrometric planets

Planets can be detected by the periodic variations in the position of their host star as the star orbits around the center of mass of the star and planet.

The Kepler ellipse described by the star is projected onto an ellipse on the sky plane perpendicular to the line of sight. However, the semimajor axis and eccentricity of the projected ellipse are generally different from those of the original ellipse, and the focus of the projected ellipse differs from the projection of the focus of the original ellipse. Nevertheless all of the orbital elements, with some minor degeneracies, can be deduced from these measurements.

We consider a system containing a single planet of mass m_1 orbiting a star of mass m_0. We choose coordinates such that the positive z-axis is parallel to the line of sight from the observer to the system and pointing toward the observer.[21] The position of the star is $\mathbf{r}_0 = \mathbf{r}_{\rm cm} - m_1\mathbf{r}/(m_0+m_1)$ (eq. 1.6), where $\mathbf{r}_{\rm cm}$ is the position of the center of mass and $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_0$ is the vector from the star to the planet. Using equations (1.29) and (1.70) the position of the star on the sky, in the Cartesian coordinates x and y, is

$$\begin{aligned} x_0 &= x_{\rm cm} - \frac{1-e^2}{1+e\cos f}(A\cos f + F\sin f), \\ y_0 &= y_{\rm cm} - \frac{1-e^2}{1+e\cos f}(B\cos f + G\sin f), \end{aligned} \tag{1.115}$$

[21] Unfortunately this orientation is opposite to the orientation of the coordinate system in §1.6.1. The line-of-sight velocity is always defined to be positive if the star is receding from the observer, which implies that the positive z-axis points *away* from the observer. For astrometric binaries the x-y coordinate system on the sky is assumed to be right-handed (the positive y-axis is 90° counterclockwise from the positive x-axis), which requires that the positive z-axis points *toward* the observer.

where the **Thiele–Innes elements** are

$$\begin{aligned} A &= \frac{m_1 a}{m_0 + m_1}(\cos\Omega\cos\omega - \cos I\sin\Omega\sin\omega), \\ B &= \frac{m_1 a}{m_0 + m_1}(\sin\Omega\cos\omega + \cos I\cos\Omega\sin\omega), \\ F &= \frac{m_1 a}{m_0 + m_1}(-\cos\Omega\sin\omega - \cos I\sin\Omega\cos\omega), \\ G &= \frac{m_1 a}{m_0 + m_1}(-\sin\Omega\sin\omega + \cos I\cos\Omega\cos\omega); \end{aligned} \tag{1.116}$$

as usual a and e are the semimajor axis and eccentricity of the relative orbit, and f, I, ω and Ω are the true anomaly, inclination, argument of periapsis and longitude of the ascending node. The four Thiele-Innes elements replace a, I, Ω and ω; their advantage is that the positions are linear functions of these elements, which simplifies orbit fitting.

Equations (1.115) are simpler when written in terms of the eccentric anomaly, using equations (1.46), (1.50) and (1.51a):

$$\begin{aligned} x_0 &= x_{\rm cm} - A(\cos u - e) - F(1-e^2)^{1/2}\sin u, \\ y_0 &= y_{\rm cm} - B(\cos u - e) - G(1-e^2)^{1/2}\sin u. \end{aligned} \tag{1.117}$$

The eccentric anomaly is related to the time t through Kepler's equation (1.49), and with equation (1.45) this reads $n(t - t_0) = u - e\sin u$. Using these results we can fit the observations of x_0 and y_0 as a function of time to equations (1.117) to determine $x_{\rm cm}$, $y_{\rm cm}$, A, B, F and G, the eccentricity e, the mean motion n and the epoch of periapsis t_0.

The usual orbital elements are straightforward to derive from the Thiele–Innes elements. First,

$$\tan(\Omega + \omega) = \frac{B - F}{A + G}, \quad \tan(\Omega - \omega) = \frac{B + F}{A - G}, \tag{1.118}$$

and these equations can be solved for Ω and ω. If these are solutions then so are $\Omega + k_1\pi$ and $\omega + k_2\pi$, where k_1 and k_2 are integers. All but one of these solutions can be discarded because we also require that (i) $\sin(\Omega + \omega)$ has the same sign as $B - F$; (ii) $\sin(\Omega - \omega)$ has the same sign as $B + F$;

(iii) $0 \le \omega < 2\pi$; and (iv) $0 \le \Omega < \pi$. The last of these is a convention that is imposed because astrometric observations alone cannot distinguish the solutions (Ω, ω) and $(\Omega + \pi, \omega + \pi)$.

Next define

$$q_1 = \frac{A + G}{\cos(\Omega + \omega)}, \quad q_2 = \frac{A - G}{\cos(\Omega - \omega)}. \tag{1.119}$$

Then

$$I = 2\tan^{-1}(q_2/q_1)^{1/2}, \quad \frac{m_1 a}{m_0 + m_1} = \tfrac{1}{2}(q_1 + q_2). \tag{1.120}$$

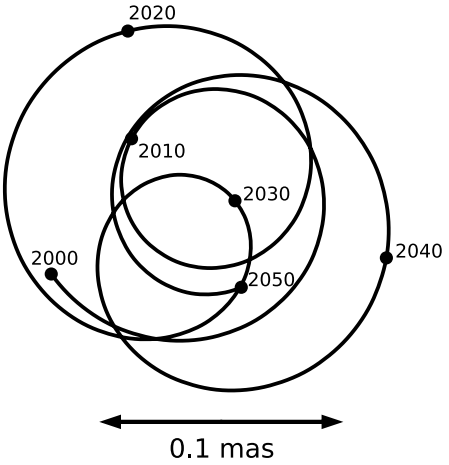

Figure 1.4: The astrometric signal from the solar system over the 50-year period from 2000 to 2050, as viewed from a star 100 parsecs away in the direction of the north ecliptic pole. The arrows mark an angular distance of 0.1 milliarcseconds.

The fit to the observations also yields the mean motion n, which constrains the semimajor axis and masses through Kepler's third law, $n^2 a^3 = \mathbb{G}(m_0 + m_1)$ (eq. 1.44). Combining this relation with the last of equations (1.120), we have

$$\frac{m_1^3}{(m_0 + m_1)^2} = \frac{(q_1 + q_2)^3 n^2}{8\,\mathbb{G}}; \tag{1.121}$$

the quantities on the right are observables and the left side is the **mass function** for astrometric planets. The mass m_0 of the host star can usually be determined from its spectral properties, so the mass function determines the planetary mass m_1.

The astrometric signal from a planet is proportional to its semimajor axis, so planets on larger orbits are easier to detect astrometrically. However, a reliable determination of the orbital elements usually requires data over at least one orbit, unless the data are extremely accurate. Thus the easiest planets to detect astrometrically are those with an orbital period smaller than the span of observations, but not by too much.

Astrometric data from multi-planet systems are hard to interpret if *any* of the massive planets in the system has an orbital period longer than the span of the observations. As an example, the astrometric signal arising from the motion of the Sun around the barycenter of the solar system is shown in Figure 1.4, as seen from a star 100 parsecs away. The figure shows that determining the masses and orbits of the giant planets in a planetary system like our own, even with an astrometric baseline of 1–2 decades, would be quite difficult.

1.6.4 Imaged planets

Imaging planets is difficult because the host star is so much brighter than the planet. For example, the luminosity of the Earth at visible wavelengths is only about 10^{-10} times the luminosity of the Sun. The contrast ratio is more favorable for young, massive planets at infrared wavelengths, in part because such planets are self-luminous, emitting thermal energy as they contract (Burrows et al. 1997). Even Jupiter emits roughly as much energy per unit time from contraction as it reflects from the Sun.

Most planets that have been successfully imaged are in orbits with large semimajor axes, where they are not swallowed in the glare from their host star: the median estimated semimajor axis of planets detected by direct imaging is well over 100 au. For a solar-mass host star the orbital period at 100 au is 1 000 yr, so the motion of most imaged planets relative to their host star has not been detected at all. What motion has been detected covers only a small fraction of the orbit, so the uncertainties in the orbital elements

are large. Nevertheless, it is worth examining briefly what elements can be detected in principle for imaged planets.

In contrast to astrometric planets, where the position of the host star relative to the center of mass is measured on the sky plane, we measure the position of an imaged planet at $\mathbf{r}_1$ relative to the host star at $\mathbf{r}_0$. By analogy with equations (1.117) we may write the Cartesian coordinates of this relative position on the sky plane as

$$\begin{aligned} x &= x_1 - x_0 = A'(\cos u - e) + F'(1-e^2)^{1/2} \sin u, \\ y &= y_1 - y_0 = B'(\cos u - e) + G'(1-e^2)^{1/2} \sin u, \end{aligned} \tag{1.122}$$

where u is the eccentric anomaly, e is the eccentricity, and the Thiele–Innes elements are

$$\begin{aligned} A' &= a(\cos\Omega\cos\omega - \cos I \sin\Omega\sin\omega), \\ B' &= a(\sin\Omega\cos\omega + \cos I \cos\Omega\sin\omega), \\ F' &= a(-\cos\Omega\sin\omega - \cos I \sin\Omega\cos\omega), \\ G' &= a(-\sin\Omega\sin\omega + \cos I \cos\Omega\cos\omega). \end{aligned} \tag{1.123}$$

As usual a, ω and Ω are the semimajor axis, argument of periapsis and longitude of the ascending node. The eccentric anomaly is related to the time t through Kepler's equation (1.49), which reads $n(t - t_0) = u - e\sin u$ where n is the mean motion. We can fit the observations of x and y as a function of time to equations (1.122) to determine A', B', F', G', e, n and t_0. Then we can follow the procedure in equations (1.118)–(1.120) to determine the other orbital elements. Like astrometry, imaging cannot distinguish the solutions (Ω, ω) and $(\Omega + \pi, \omega + \pi)$. A check of the results comes from Kepler's third law (1.44): this determines the mass of the host star from the mean motion and the semimajor axis, and this mass can be determined independently from the spectral properties of the star.

1.7 Multipole expansion of a potential

In most cases the distance between a planet and its host star, or a satellite and its host planet, is large enough that both can be treated as point masses.

However, accurate dynamical calculations must sometimes account for the distribution of mass within one or both of these bodies. Examples include tracking artificial satellites of the Earth, measuring the relativistic precession of Mercury's perihelion, or determining the precession rate of a planet's spin axis.

Let $\rho(\mathbf{r})$ denote the density of a planet at position $\mathbf{r}$. The total mass of the planet is M and we assume that the origin is the center of mass of the planet. Then

$$\int \mathrm{d}\mathbf{r}\,\rho(\mathbf{r}) = M, \qquad \int \mathrm{d}\mathbf{r}\,\rho(\mathbf{r})\mathbf{r} = \mathbf{0}. \tag{1.124}$$

Using equations (C.44) and (C.55), the gravitational potential can be written in spherical coordinates $\mathbf{r} = (r, \theta, \phi)$ as

$$\begin{aligned}\Phi(r,\theta,\phi) &= -\mathbb{G}\int \frac{\mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \qquad (1.125)\\ &= -\mathbb{G}\sum_{l=0}^{\infty}\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')\frac{r_<^l}{r_>^{l+1}}\,\mathsf{P}_l(\cos\gamma)\\ &= -\sum_{l=0}^{\infty}\frac{4\pi\,\mathbb{G}}{2l+1}\sum_{m=-l}^{l}\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')\frac{r_<^l}{r_>^{l+1}}\,Y_{lm}^*(\theta',\phi')Y_{lm}(\theta,\phi).\end{aligned}$$

Here $\mathsf{P}_l(\cos\gamma)$ and $Y_{lm}(\theta,\phi)$ are a Legendre polynomial and a spherical harmonic (Appendices C.6 and C.7), $r_<$ and $r_>$ are the smaller and larger of r and r', $\cos\gamma = \mathbf{r}'\cdot\mathbf{r}/(r'r)$ is the cosine of the angle between the vectors $\mathbf{r}$ and $\mathbf{r}'$, and the asterisk denotes the complex conjugate. Any satellite must orbit outside all of the planetary mass, so the potential seen by the satellite simplifies to

$$\Phi(r,\theta,\phi) \equiv \sum_{l=0}^{\infty}\Phi_l(r,\theta,\phi), \tag{1.126}$$

where

$$\begin{aligned}\Phi_l(r,\theta,\phi) &= -\frac{\mathbb{G}}{r^{l+1}}\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')r'^l\mathsf{P}_l(\cos\gamma) \qquad (1.127)\\ &= -\frac{4\pi\,\mathbb{G}}{(2l+1)r^{l+1}}\sum_{m=-l}^{l}Y_{lm}(\theta,\phi)\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')r'^lY_{lm}^*(\theta',\phi').\end{aligned}$$

We examine the first three of these terms:

Monopole ($l = 0$) Since $\mathsf{P}_0(\cos\gamma) = 1$ (eq. C.45) and $\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}') = M$ (eq. 1.124), we have $\Phi_0(r,\theta,\phi) = -\mathbb{G}M/r$, the same as if all the mass of the planet were concentrated in a point at the origin.

Dipole ($l = 1$) Since $\mathsf{P}_1(\cos\gamma) = \cos\gamma = \mathbf{r}'\cdot\mathbf{r}/(r'r)$, the combination $r'\mathsf{P}_1(\cos\gamma)$ is a linear function of $\mathbf{r}'$ at fixed $\mathbf{r}$ and zero at $\mathbf{r}' = \mathbf{0}$. Then the second of equations (1.124) implies that the integral in the first line of equation (1.127) is zero. Thus $\Phi_1(r,\theta,\phi) = 0$.

Quadrupole ($l = 2$) Since $\mathsf{P}_2(\cos\gamma) = \frac{3}{2}\cos^2\gamma - \frac{1}{2}$, the combination $r'^2\mathsf{P}_2(\cos\gamma) = \frac{3}{2}(\mathbf{r}'\cdot\mathbf{r})^2/r^2 - \frac{1}{2}r'^2$. Therefore the quadrupole potential can be written

$$\Phi_2(r,\theta,\phi) = \frac{\mathbb{G}}{2r^5}\int d\mathbf{r}'\,\rho(\mathbf{r}')\big[r'^2r^2 - 3(\mathbf{r}'\cdot\mathbf{r})^2\big]. \tag{1.128}$$

When written in terms of the inertia tensor $\mathbf{I}$ of the planet (eq. D.85), this yields **MacCullagh's formula**

$$\Phi_2(r,\theta,\phi) = \frac{3\mathbb{G}}{2r^5}\sum_{ij=1}^{3} r_i I_{ij} r_j - \frac{\mathbb{G}}{2r^3}\sum_{i=1}^{3} I_{ii} = \frac{3\mathbb{G}}{2r^5}\mathbf{r}^{\mathrm{T}}\mathbf{I}\mathbf{r} - \frac{\mathbb{G}}{2r^3}\mathrm{Tr}(\mathbf{I}); \tag{1.129}$$

here $\mathbf{r}^{\mathrm{T}}$ is the row vector that is the transpose of the column vector $\mathbf{r}$, and $\mathrm{Tr}\,(\mathbf{I})$ is the trace of the inertia tensor.

Since $\Phi_l(r,\theta,\phi)$ in equation (1.127) falls off with distance $\propto r^{-l-1}$, at large distances from the host planet the potential is dominated by the monopole potential ($\propto r^{-1}$) and quadrupole potential ($\propto r^{-3}$).

1.7.1 The gravitational potential of rotating fluid bodies

Small bodies, such as rocks, comets and most asteroids, are irregularly shaped. Larger astronomical bodies are nearly spherical, because the forces due to gravity overwhelm the ability of any solid material to maintain other shapes (a brief quantitative discussion of this transition is given at the end of §8.6). Stars and planets are large enough that they can usually be treated as

a fluid. In this case the distribution of the matter is determined by a balance between gravity, pressure and centrifugal force due to rotation. Models of stellar and planetary interiors show that the resulting density distribution is always axisymmetric around the spin axis.[22]

Axisymmetry allows us to simplify the spherical-harmonic expansion (1.127) for the gravitational potential of the planet. If the axis of symmetry of the planet is chosen to be the polar axis ($\theta = 0$), the second line of equation (1.127) vanishes when $m \neq 0$ since $\int \mathrm{d}\phi' Y_{lm}(\theta', \phi') \propto \int \mathrm{d}\phi' \exp(\mathrm{i}m\phi') = 0$ when $m \neq 0$. Using the definition (C.46) of spherical harmonics in terms of associated Legendre functions, equations (1.126) and (1.127) can be rewritten as

$$\Phi(r,\theta) = -\frac{\mathbb{G}M}{r}\left[1 - \sum_{l=2}^{\infty} J_l \left(\frac{R_\mathrm{p}}{r}\right)^l \mathsf{P}_l(\cos\theta)\right], \tag{1.130}$$

where the dimensionless **multipole moments** J_l are given by

$$J_l \equiv -\frac{1}{MR_\mathrm{p}^l}\int \mathrm{d}\mathbf{r}'\,\rho(\mathbf{r}')\mathsf{P}_l(\cos\theta')r'^l. \tag{1.131}$$

The quantity R_p is an arbitrary reference radius that is introduced so that J_l is dimensionless; conventionally it is chosen to be close to the planetary radius.

Since $\mathsf{P}_2(\cos\theta) = \frac{1}{2}(3\cos^2\theta - 1)$ (eq. C.45), the **quadrupole moment** J_2 can be written in Cartesian coordinates as

$$J_2 = \frac{1}{MR_\mathrm{p}^2}\int \mathrm{d}\mathbf{r}\,\rho(\mathbf{r})(\tfrac{1}{2}x^2 + \tfrac{1}{2}y^2 - z^2). \tag{1.132}$$

For an axisymmetric body we define the moments of inertia of the planet around the equatorial and polar axes as (cf. eqs. D.87)

$$A = \int d\mathbf{r}\,\rho(\mathbf{r})(y^2 + z^2) = I_{xx} = I_{yy},$$

[22] Non-axisymmetric equilibrium bodies of self-gravitating fluid do exist. The first and most famous example is the sequence of Jacobi ellipsoids (Chandrasekhar 1969), which are uniformly rotating masses of homogeneous, incompressible fluid. However, only axisymmetric equilibria exist for typical planets, in which the material is compressible so the mass is concentrated toward the center.

$$C = \int d\mathbf{r}\, \rho(\mathbf{r})(x^2 + y^2) = I_{zz}, \tag{1.133}$$

which implies that

$$J_2 = \frac{C - A}{MR_{\rm p}^2}. \tag{1.134}$$

Then either MacCullagh's formula (1.129) or equation (1.130) yields[23]

$$\begin{aligned}\Phi(r,\theta) &= -\frac{\mathbb{G}M}{r} + \frac{\mathbb{G}MJ_2R_{\rm p}^2}{2r^3}(3\cos^2\theta - 1) + \mathrm{O}(r^{-4}) \\ &= -\frac{\mathbb{G}M}{r} + \mathbb{G}\frac{C - A}{2r^3}(3\cos^2\theta - 1) + \mathrm{O}(r^{-4}).\end{aligned} \tag{1.135}$$

Notice that measurements of the potential external to the planet allow us to determine the *difference* between the moments of inertia A and C but not the moments themselves. The rate of precession of the spin axis due to the torque from an external body, such as the Sun, yields the dynamical ellipticity $(C - A)/C$ (cf. eq. 7.10), so measurements of both the external gravitational field and the precession are needed to determine both moments of inertia C and A.

We also expect that rotating planets or stars are symmetric about the equatorial plane (the plane normal to the polar axis that passes through their center of mass),[24] so $\rho(r,\theta)$ is an even function of $\cos\theta$ if the center of mass coincides with the origin. Since $\mathsf{P}_l(-\cos\theta) = (-1)^l\mathsf{P}_l(\cos\theta)$ (eq. C.38), all multipole moments J_l with odd values of l vanish. In this case there is a sharper limit on the error in equation (1.135): $\mathrm{O}(r^{-5})$ rather than $\mathrm{O}(r^{-4})$.

Rotation flattens the density distribution of a planet (i.e., the planet becomes **oblate**), so the moment of inertia C around the polar axis is larger than the moment A around an equatorial axis, which in turn implies through equation (1.134) that the quadrupole moment J_2 is positive. In general the

[23] A function $f(r)$ is $\mathrm{O}(r^{-p})$ if $r^p f(r)$ is less than some constant when r is large enough.

[24] This result can be proved analytically in simple models of a planetary interior. In particular, if the planet is uniformly rotating (i.e., the fluid has zero velocity in a frame rotating at a constant angular speed $\mathbf{\Omega}$) and the equation of state is barotropic (i.e., the pressure is a function only of the density), then **Lichtenstein's theorem** states that in equilibrium the fluid has reflection symmetry around a plane perpendicular to $\mathbf{\Omega}$ (e.g., Lindblom 1992).

Box 1.3: Rotation, quadrupole moment and flattening

If the quadrupole moment J_2 is much larger than all of the J_n with $n > 2$, equation (1.135) implies that the gravitational potential outside the planet is

$$\Phi(r,\theta) = -\frac{\mathbb{G}M}{r}\left[1 - \frac{J_2 R_{\rm p}^2}{2r^2}(3\cos^2\theta - 1)\right]. \qquad \text{(a)}$$

We assume that the planet is rotating uniformly with angular speed Ω around its polar axis. Then the centrifugal potential is (eq. D.21)

$$\Phi_{\rm cent}(r,\theta) = -\tfrac{1}{2}\Omega^2(x^2+y^2) = -\tfrac{1}{2}\Omega^2 r^2 \sin^2\theta. \qquad \text{(b)}$$

If the surface of the planet can be treated as a fluid—that is, if it has an atmosphere or is large enough that the strength of the material at its surface is negligible—then the effective potential $\Phi_{\rm eff}(r,\theta) \equiv \Phi(r,\theta) + \Phi_{\rm cent}(r,\theta)$ must be constant on the surface.[a] Let the surface be $r = R_{\rm p} + \Delta R(\theta)$; we assume that the reference radius $R_{\rm p}$ is close enough to the mean radius of the surface that $|\Delta R(\theta)| \ll R_{\rm p}$. Then we may expand the effective potential to first order in $\Delta R(\theta)$, Ω^2 and J_2:

$$\Phi_{\rm eff}(R,\theta) = \text{constant} + \frac{\mathbb{G}M}{R_{\rm p}^2}\Delta R(\theta) + \frac{3\,\mathbb{G}M}{2R_{\rm p}} J_2 \cos^2\theta + \tfrac{1}{2}\Omega^2 R_{\rm p}^2 \cos^2\theta. \qquad \text{(c)}$$

If this is to be independent of the polar angle θ on the surface, we require

$$\frac{\Delta R(\theta)}{R_{\rm p}} = -\left(\tfrac{3}{2}J_2 + \frac{\Omega^2 R_{\rm p}^3}{2\,\mathbb{G}M}\right)\cos^2\theta + \text{constant}. \qquad \text{(d)}$$

Thus the difference between the equatorial radius $R_{\rm eq} = R_{\rm p} + \Delta R(\frac{1}{2}\pi)$ and the polar radius $R_{\rm pol} = R_{\rm p} + \Delta R(0)$ is

$$\frac{R_{\rm eq} - R_{\rm pol}}{R_{\rm p}} = \tfrac{3}{2}J_2 + \frac{\Omega^2 R_{\rm p}^3}{2\,\mathbb{G}M}. \qquad \text{(e)}$$

This simple relation connects three observables: the flattening or oblateness of the planet, the rotation rate and the quadrupole moment.

---

[a] Hydrostatic equilibrium in the rotating frame requires $\nabla p = -\rho\nabla\Phi_{\rm eff}$ where $p(\mathbf{r})$ is the pressure and $\rho(\mathbf{r})$ is the density. Since $\nabla \times \nabla p = \mathbf{0}$ for any scalar field $p(\mathbf{r})$ (eq. B.36a), we must have $\nabla\rho \times \nabla\Phi_{\rm eff} = \mathbf{0}$. This result implies that the gradient of the density must be parallel to the gradient of the effective potential, so surfaces of constant density and effective potential coincide.

multipole moments with even values of l decrease rapidly as l grows, so the non-spherical part of the potential is dominated by the quadrupole term even at the surface of the planet. Given this, there is a simple relation between the rotation rate, the quadrupole moment and the flattening of the planetary surface (Box 1.3).

1.8 Nearly circular orbits

1.8.1 Expansions for small eccentricity

Most planet and satellite orbits are nearly circular, so expansions of the trajectory in powers of the eccentricity e were an essential tool for studying orbits in the days when all algebra was done by hand. Such expansions continue to provide insight in many problems of celestial mechanics. Here we illustrate the derivations of these expansions, which are given to $\mathrm{O}(e^3)$. Expansions for other variables, or higher order expansions, can easily be derived by computer algebra.

(a) True anomaly in terms of eccentric anomaly Take the log of the first of equations (1.51c),

$$f = u - \mathrm{i}\log[1 - \beta\exp(-\mathrm{i}u)] + \mathrm{i}\log[1 - \beta\exp(\mathrm{i}u)], \tag{1.136}$$

and replace β by its expression (1.52) in terms of the eccentricity e. Then expand as a Taylor series in e:

$$f = u + e\sin u + \tfrac{1}{4}e^2\sin 2u + e^3(\tfrac{1}{4}\sin u + \tfrac{1}{12}\sin 3u) + \mathrm{O}(e^4). \tag{1.137}$$

(b) Eccentric anomaly in terms of true anomaly Similarly, using the second of equations (1.51c),

$$u = f - e\sin f + \tfrac{1}{4}e^2\sin 2f - e^3(\tfrac{1}{4}\sin f + \tfrac{1}{12}\sin 3f) + \mathrm{O}(e^4). \tag{1.138}$$

(c) Mean anomaly in terms of eccentric anomaly This is simply Kepler's equation (1.49),

$$\ell = u - e\sin u. \tag{1.139}$$