INSTITUTE OF PHYSICS
SERIES IN HIGH ENERGY PHYSICS, COSMOLOGY AND GRAVITATION

GRAVITATION: FROM THE HUBBLE LENGTH TO THE PLANCK LENGTH

EDITED BY
I CIUFOLINI, E COCCIA,
V GORINI, R PERON
AND N VITTORIO

IoP

GRAVITATION: FROM THE HUBBLE LENGTH TO THE PLANCK LENGTH

Edited by

Ignazio Ciufolini

University of Lecce, Italy

Eugenio Coccia

University of Rome 'Tor Vergata', Italy

Vittorio Gorini

University of Insubria, Como, Italy

Roberto Peron

University of Lecce, Italy

Nicola Vittorio

University of Rome 'Tor Vergata', Italy

IoP

INSTITUTE OF PHYSICS PUBLISHING
BRISTOL AND PHILADELPHIA

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN 0 7503 0948 2

Library of Congress Cataloging-in-Publication Data are available

Commissioning Editor: John Navas
Production Editor: Simon Laurenson
Production Control: Sarah Plenty and Leah Fielding
Cover Design: Victoria Le Billon
Marketing: Nicola Newey, Louise Higham and Ben Thomas

Published by Institute of Physics Publishing, wholly owned by The Institute of Physics, London

Institute of Physics Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

US Office: Institute of Physics Publishing, The Public Ledger Building, Suite 929, 150 South Independence Mall West, Philadelphia, PA 19106, USA

Typeset in LaTeX 2ε by Text 2 Text Limited, Torquay, Devon
Printed in the UK by MPG Books Ltd, Bodmin, Cornwall

Contents

Preface

This volume brings together the contents of the courses given at the doctoral school on 'Gravitation: from the Hubble Length to the Planck Length' which took place in September 2002 in the beautiful environment of the historic Villa Mondragone, near Frascati, Italy. The school was sponsored and financed by SIGRAV (the Italian Society of Relativity and Gravitation), the Italian National Institute of Nuclear Physics and the University of Rome 'Tor Vergata'.

The main actor on the stage was gravitation: though the weakest among the fundamental interactions that drive the universe, it is nevertheless, in various respects, the most encompassing and pervasive one. As stressed by the title of the school, one can see that, whenever large concentrations of matter and energy are involved, gravitation works at all scales, from the microscopic domain (such as the interior of black holes and at the very birth of the universe, where quantum effects are crucially relevant as well) up to the huge clusters and superclusters of galaxies which form the large-scale texture of the present-day cosmos.

Gravity is not just the familiar mutually attractive force, affecting all types of matter–energy but a peculiar manifestation of spacetime itself. Indeed, as Einstein has taught us, spacetime is not a rigid arena—a simple ground for the play of others—but, due to the equivalence of gravitation and inertia, it is a flexible and dynamic part of the whole machinery. This renders the intimate behaviour of gravitation not only much more complex than was once thought but also much more interesting. Theory predicts a whole host of new phenomena, most of them giving rise to tiny effects save under extreme conditions, and a challenge for experimentalists. In turn, experiments and observations do as usual supply insights extending our overall understanding and providing the stimulus to develop new viewpoints and new theories.

Each chapter covers a particular feature, ranging from refined experimental techniques in gravitational physics all the way to cosmology and to the 'quantum frontier'. The authors have tried to be as clear and as pedagogical as possible, while, at the same time, bringing the reader to the edge of current research topics. This renders the volume much more than a simple 'proceedings book'.

Of course, only a selection of topics could be treated here. Nevertheless, we hope that these chapters will provide the reader with the flavour of current

research on spacetime and gravitation and with the feeling of fascination that such frontier investigations are able to transmit to our human perception.

Eugenio Coccia, Vittorio Gorini and Roberto Peron
9 June 2004

Chapter 1

Introduction

Roberto Peron and Amedeo Balbi

Almost a century after its development by Albert Einstein, the general theory of relativity is living in a new golden age. Being a beautiful theory both on the mathematical side and from the point of view of its clear physical insight, it is a continuous source of experimental predictions: it turns out that the picture of the world we get from its equations is tightly bound to what we call 'reality'.

It is not by chance that most of the greatest minds of scientific thought—Galilei, Newton, Einstein—found in gravitation the key to unveiling so many secrets of nature. In fact, Einstein (following a path opened by Gauss and Riemann) discovered that gravitation is nothing but the behaviour of spacetime itself. The path of a test particle in a gravitational field is simply geodetic motion in a curved spacetime; in turn, spacetime is curved by the presence of matter and energy in it. We can speak, following Wheeler, about *geometrodynamics*.

It is interesting to note that Einstein started from the desire to extend the principle of relativity from the class of inertial reference frames but ended up obtaining so much more. It was a conceptual jump, not a simple evolution. It is once more amazing to see the number and variety of physical consequences one may obtain from the relatively simple assumptions upon which general relativity is based.

The lectures in this book cover a wide spectrum of topics in the field of gravitational physics. All of them are written by leading scientists: their main scope is to give the reader a general view of their respective fields of research, focusing on foundations, state-of-the-art, open problems, either from a theoretical or an experimental point of view. We find unsolved problems in general relativity and cosmology: even if we restrict ourselves to solar system science, year after year the complex dynamics of systems driven by the 'force' of gravity is revealed.

From the theoretical point of view, the appearance of general relativity brought onto the scene new mathematical methods, capable of dealing better with the geometrical character of this theory. This process had its peak in the formulation of the singularity theorems by Penrose and Hawking. These theorems

helped us to obtain a better understanding of the limits of general relativity as a physical theory: at some level, there should be a much more refined theory with quantum effects included in the geometry of spacetime. So, in recent decades, a number of attempts in this direction have flourished, namely covariant and canonical quantum gravity and, more recently, string theory and loop quantum gravity. This is only mentioning the mainstream.

The problem of singularity is directly linked to a description of the birth and expansion of the universe. This is usually done in the framework of the Friedmann–Robertson–Walker relativistic models but several reasons have led to an extension of this—called inflationary cosmology—which adds new assumptions about spacetime dynamics (i.e. the existence of an inflaton field coupled to gravity) going beyond standard geometrodynamics.

Considering the dynamics of spacetime on a very large scale introduces some issues at the border between physics and philosophy: in particular, the link between local and 'global' reference frames could open the way to some hypothesis on the global distribution and flow of matter–energy. This line of thought—known as 'Mach's principle'—had played a fundamental rôle in the formulation of general relativity by Einstein. Far from being an isolated question, it opens the way to a deeper view of the mathematical structure of the theory and finds its place in studying some peculiar predictions of the theory, like gravitomagnetism.

The experimental side of gravitation science is as varied as the theoretical one. First of all, geometrodynamics leads to a number of direct experimental predictions that show the various ways in which relativistic gravitation could differ from the Newtonian one. So light is bent, test particles trajectories are different, clocks near masses behave differently. These features are better handled in a particular formalism called Parametrized Post-Newtonian (PPN), in which deviations from Newtonian physics are taken into account using an expansion of metric and stress–energy tensors around the 'Newtonian ones'. In this way, it is possible to describe many different gravity theories, each with its own set of coefficients. Comparison with the experiment leads to constraints on these values, excluding sets of theories which are not compatible with observed behaviour. In fact, there exists a number of 'alternative theories', usually obtained by relaxing some of the hypotheses at the basis of general relativity or introducing new fields coupled to the metric one. In this respect, the PPN formalism is not representative of a particular class of physical theories but it is to be seen as a classification device.

Side by side with the 'classical tests', following the physical consequences of geometrodynamics leads to qualitatively new features of spacetime itself. The existence of phenomena such as gravitomagnetism and gravitational waves in a sense completes the characterization of geometrodynamics as a true field theory, showing the different ways in which spacetime is a main actor in the physical scene.

A separate place must be reserved for the Equivalence Principle. Its formulation is at the heart of geometrodynamics—as Einstein conceived it—and constitutes perhaps its main distinctive character with respect to the other theories of gravitation. The need for a field theory in order to describe gravitational phenomena was more or less clear but the treatment of it in an 'apparent force way' was indeed revolutionary. This point has been widely criticized (together with general covariance) in a number of ways but what remains once again shows its power. The importance of an experimental verification of this principle with the greatest possible accuracy is evident and there is active work today focused on a number of laboratory-based or spaceborne experiments.

Returning to Newtonian gravitation, could it be that, even on a local scale, we find something different from what we expected? Some claim that, on a small scale, the 'gravitational force' does not scale as $1/r^2$; others predict secular variations of the gravitational constant G. All these topics are worth testing experimentally, each of them a possible source of insight and ideas of valuable interest even outside gravitation physics.

1.1 Gravitation in the solar system and beyond

The Earth and the solar system environment are perhaps the main places where gravitational phenomena can be studied. Due to the relative closeness of solar system objects, their motion can be tracked with relatively high accuracy, providing thus a great deal of information about gravitational dynamics. This field of study has grown in importance over the years, since improvement in knowledge about the 'gravitational environment' around Earth and the other objects in the solar system means improvements in space navigation techniques. This is much more important for space techniques applied to Earth sciences, where this knowledge has important applications in remote sensing of the 'Earth system'.

Chapter 2 'Probing spacetime in the solar system' by B Bertotti presents the state-of-the-art methods regarding measurements in the solar system by focusing on three fundamental physical quantities: transit times, angles and frequencies. While a lot of other measurements could be performed in this context, the three ones given here retain their status of 'fundamental'; they constitute a framework to which others techniques must relate to, and a source for continuously improved experimental data. Due to their relative conceptual simplicity, they allow us to see very clearly the progress achieved. From Lunar Laser Ranging to GAIA to Cassini, current or forthcoming missions are extending the accuracy to the point in which care must be taken of previously ignored effects. It is in that small frontier between known and unknown (modelled and unmodelled) that new science is done.

The phenomenon of gravitomagnetism—an interaction of gravitational origin caused by currents of mass–energy—is a peculiar prediction of general relativity. Chapter 3 'Frame-dragging and its measurement' by I Ciufolini

contains a detailed description of the solar system and astrophysical implications of gravitomagnetism. Together with recent advances in studying these phenomena using test gyroscopes, test particles, clocks and photons, the most recent results in measuring Lense–Thirring effect (gravitomagnetic precession caused by Earth angular momentum) in LAGEOS and LAGEOS II orbits are presented. The Equivalence Principle lies at the basis of the general relativity theory, and indeed its testing remains fundamental for the experimental confirmation of the theory. From this principle Einstein directly deduced light deflection and changing of clock rates near a mass. In chapter 4 'The special relativistic Equivalence Principle' by K Nordtvedt an extended version of this principle is introduced, fully exploiting special relativity. It is shown that in this way one can predict a number of further effects, including geodetic and gravitomagnetic precession. These effects do not include all the possible consequences of general relativity theory, but are present in all locally Lorentz-invariant, complete metric theories of gravity.

Among the various techniques developed for space measurements, Lunar Laser Ranging shines as one of the most precise. It has the particular honour of having been started in conjunction with the first lunar manned landing, and this raised its fascination. Chapter 5 'Lunar laser ranging; a comprehensive probe of post-Newtonian gravity' by K Nordtvedt describes its use for studies of post-Newtonian effects in the Sun–Earth–Moon system. The order $1/c^2$ equations of motion reveal effects that have no counterpart in Newtonian dynamics, and could be in principle different also with respect to Einstein general relativity. Cosmological consequences (related to scalar–tensor theories) may be tested too. Analysis of Lunar Laser Ranging data can therefore improve constraints on alternative theories of gravitation, and its expected improvements will render this as useful a tool as in the past.

1.2 Cosmological issues

The connection between the large-scale properties of the Universe and the extremely small scales investigated by fundamental physics becomes evident when one explores the evolution of the Universe in its early stages. Gravitational instability governs the growth of the cosmic structure, seeded by primordial fluctuations in the spacetime metric. The emergence of these fluctuations is directly related to physical processes taking place in the Universe when the energy is of the order of the Planck scale. Phenomena that are not testable in laboratories on Earth can then be probed by the imprint they have left on the cosmic evolution. Chapter 6 'The Early Universe and the Cosmic Microwave Background' by A Balbi, outlines the interplay between fundamental physics and cosmological observations and describes the revolutionary progress in our understanding of the physical Universe that has taken place over the past decade. Some of the questions that are investigated by modern cosmology are: What is the nature of the scalar

fields that govern inflation? What are the different contributions to the energy density of the Universe? What is the nature of the quantum vacuum, whose energy seems to dominate the cosmic budget today?

1.3 The other side: gravitation in the quantum regime

The standard model describing the unification of electromagnetic, weak and strong interactions into a single gauge theory is a beautiful and, in many respects, very successful description of reality. Unfortunately, gravity displays a peculiar behaviour with respect to the other fundamental forces and cannot be incorporated into the framework provided by the standard model: for example, gravity is purely attractive, and it is so weak that it basically plays no rôle at the atomic and sub-atomic level. Furthermore, the effective coupling of the gravitational interaction between point-like particles becomes extremely strong at the Planck scale $E_{\rm Pl} \approx 10^{19}$ GeV, resulting in divergences in the quantization of general relativity. The most promising way of connecting gravity to the other interactions is provided by string theory. Chapter 7, 'Strings, Gravity and Particle Physics' by A Sagnotti, reviews some of the key aspects of string theory, including extra-dimensions and branes, with applications to particle physics, black hole thermodynamics and color-flux strings.

1.4 Gravitation as a universal phenomenon

The contributions to this volume demonstrate how the study of gravitation can be both interesting and a source of precious information about the 'machinery' of our world. For many years the smallness of the gravitational interaction—compared to the other ones—permitted only a kinematical study (motion of the Heavens). More accurate observations and new theories have permitted a deeper insight into the dynamics of most gravitational systems, opening problems still unsolved (chaotic dynamics, for example). General relativity added a new, fundamental piece of information, showing how the fall of an apple is a consequence of the fundamental properties of spacetime. This, after all, showed how a fundamental theory can be very simple. We hope such clarity will be achieved by the quantum theory of gravity, whatever it will be.

The simplicity we believe is a characteristic of a fundamental theory has its counterpart in the overwhelming complexity of natural phenomena as we see them. The experimental procedures employed add a further degree of complexity to our view. In the midst of all this 'chaos', we feel easy—even when studying 'this' or 'that'—by staying in touch with something that is everywhere, thereby confirming the unity of our universe.

Chapter 2

Probing spacetime in the solar system

Bruno Bertotti
Dipartimento di Fisica Nucleare e Teorica, Università di Pavia, Pavia

2.1 Introduction

This very selective (in particular in the bibliography) and synthetic exposé on experimental gravitation [20] in space—somewhat different from the original presentation—is organized around three physical quantities: transit times, angles and frequencies. For each of these quantities, this chapter, reviews the fundamental instrumental concepts, together with the driving errors and a paradigmatic experiment. Its purpose is to stress and exemplify the importance of the current and outstanding instrumental improvements for the understanding of the structure of spacetime: it may be useful for theoreticians who wish to design new experiments and for experimentalists who may find unforeseen applications and implications for their techniques. For more details, see [7]. At the fundamental level, a 'moral' is the need to formulate an experiment in an invariant way: coordinates are only a computational tool, a ladder with which to climb the geometrical wall [3]. For example, the distinction between the gravitational and transversal Doppler shifts is coordinate-dependent: to avoid pitfalls, the full expression (2.8) should be used. The initial planning of the Global Positioning System by the American military was marred by this confusion and, as a result, civilian physicists had to intervene to define the correct software [1].

From the instrumental point of view, the recent impressive improvement in accuracy in measurements of distances, angles and frequencies in the solar system does not mean that the corresponding errors, as usually expressed in terms of 'standard deviations', should be taken as a mantra, without a careful and often critical analysis. I only mention the fact that, for these three quantities, the dynamic range in relation to the sought signal is huge: although this hindrance

depends on the respective time scales and cannot be discussed in general terms, let it be enough to point out the crude ratio between the observable and the error:

(i) For Lunar Laser Ranging, the error in the distance D is $\sigma_D = 1$ cm $= 2.5 \times 10^{-11} D$.
(ii) In GAIA's astrometric project, the goal in angular accuracy is 10 μarcsec $= 5 \times 10^{-11}$.
(iii) In Cassini's Doppler experiments, the observable is the fractional frequency shift y: $\sigma_y = 3 \times 10^{-15} = 3 \times 10^{-11} y_{\rm E}$, where $y_{\rm E} = 10^{-4}$ corresponds to the Earth's orbital speed, the main contribution to y.

The data analysis must dig 10 or 11 orders of magnitude into the record before being able to deal with the signal at the accuracy determined by the instrumental errors. Many different contributions, all different in nature, larger than this must be eliminated or estimated before coming to the gist of the experiment. Very often, they are not known well enough *a priori*, and must be determined simultaneously with the target signal. Among the difficulties which may arise in this process, I can mention the following ones.

- Under- and over-parametrization: a good physical understanding of the physics and the relevance of all the contributions other than the main signal is necessary. Adding unnecessary parameters dilutes the information content.
- Correlations between the target parameter p and another parameter, say p': in this case, the experiment only provides, in the (p, p') plane, a very elongated error ellipse, which undermines the accuracy in p if p' is not known by other means.
- Gaps in the record, especially if they also have time scales in the target signal, are dangerous. For example, since the Moon is, *de facto*, laser tracked preferably between its quarter and full phase, the data distribution itself is modulated with the lunar synodic phase $\lambda_{\rm s}$. Since the Equivalence Principle violating signal has the same signature (equation (2.6)), this results in a deterioration in the accuracy [17].
- When the signal is constant or varies on a time scale longer than the record, a red spectrum for the noise of the observable quantity may be serious. Because of the Wiener–Khinchin theorem, such a spectrum is equivalent to a correlation in the observables and, hence, to an effectively smaller data set: more dangerously, it may be evidence of systematic errors (like thermal drift) and non-stationary behaviour.

Space offers several advantages, including:

- freedom from the Earth's gravity and the related very strong dynamic anisotropy;
- small (but not vanishing!) non-gravitational forces;
- very little residual gas; and

- the space around a spacecraft is optically thin—communication by electromagnetic waves is easy.

The main disadvantage is the launch cost, essentially due to the fact that the escape velocity

$$v_{\rm esc} = \sqrt{2\frac{GM_\oplus}{R_\oplus}} \simeq 11.2\ {\rm km\ s^{-1}} \tag{2.1}$$

is much larger (about three times) than the typical exhaust velocity of the gas, of the order of its thermal speed. With chemical fuel, it is not possible to put a body in orbit directly—multiple stages and large, expendable fuel tanks must be used.

2.2 Distance

2.2.1 Fundamentals

Traditionally the unit of length—the centimetre—was defined, through interferometric measurements, as a given multiple of the wavelength λ_0 of a stable optical spectral line. Two rods have the same length if they correspond to the same number of wavelengths. The unit of time was independent and provided by a microwave resonator based on an atomic system. By transferring a frequency standard from the microwave to the optical band, it was possible to measure the frequency c/λ_0 and to obtain the velocity of light c (with the dimension cm $\rm s^{-1}$). This transfer over a frequency range of about seven orders of magnitude, however, is subject to relevant errors; moreover, the stability of lasers used in interferometric techniques is much worse than that of atomic frequency standards. As a consequence, *the standard of length has recently been foregone* and the velocity of light c is now a conventionally fixed quantity. If $c = G = 1$ (adopted here), lengths are measured in light seconds and masses are lengths: $m_\odot = 1.48$ km, $m_\oplus = 0.44$ cm. Note that short-term relativistic effects on a Keplerian orbit are of order m, therefore ≈ 1 km in the solar system and ≈ 1 cm around the Earth.

This point of view is quite appropriate to space physics, where rigid rods cannot be used: absolute distances are obtained from the transit times of short light or radio pulses, timed with atomic clocks. It is also fully consistent with general relativity, in which three-dimensional rigid bodies cannot be defined in general: only the proper time is needed to construct the four-dimensional manifold (the *chronometric point of view*; see the illuminating discussion in [19], especially chapters II and III). Figure 2.1 shows how an invariant measurement of distance is accomplished. Using several nearby, freely falling objects and the equation of geodesic deviation, an invariant and operational way to measure the curvature can be defined [3].

However, in the solar system, the round-trip light-time is available only for very few bodies: the distances of the other ones are obtained *dynamically*. In

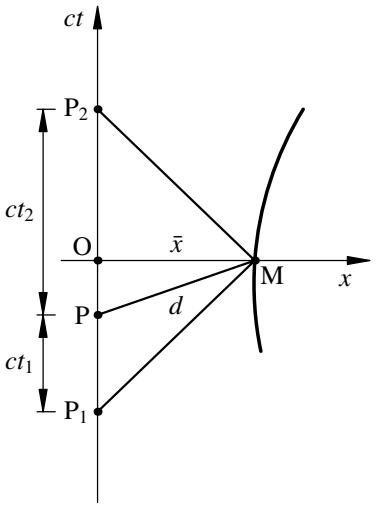

Figure 2.1. The chronometric measurement of distance. In the spacetime frame (ct, x), we have $M = (0, \bar{x})$, $P = (ct_1 - \bar{x}, 0)$. In special relativity, the proper length of the hypotenuse MP is the square root of the *difference* of the squares of the time-like and the space-like sides: $d^2 = (\bar{x} - ct_1)^2 - \bar{x}^2 = c^2\, t_1 t_2$. This reduces to the elementary expression $d = ct_1$ when the events M and P are simultaneous, so that $t_1 = t_2$.

the neighbourhood of the Earth, the orbital period of a gravitationally bound body determines, by the third Kepler law, the ratio $GM_\oplus/a^3$: unless the Earth's parameter $GM_\oplus$ is known, all the semimajor axes a are determined to within a constant scaling $a \to a' = ka$. The measurement of a single semimajor axis (e.g. the Moon's) fixes the scale and all lengths. Similarly, for bodies orbiting around the Sun, unless $GM_\odot$ is known, the semimajor axes are determined to within a change of scale. Range measurements of a single quantity, for instance those carried out with the Viking spacecraft on Mars for the distance from the Earth, fix all interplanetary distances. At present, $GM_\oplus$ and $GM_\odot$ are known with the fractional accuracy 2×10^{-9} and 1.2×10^{-11}, respectively. Correspondingly, the Astronomical Unit (AU) has an error of ≈ 6 m. Note also that, in space physics, the mass M and the gravitational constant G never appear separately.

2.2.2 Techniques

The electromagnetic measurement of distance is, of course, the basis of *radar*, an impressive technological development, which was started in Great Britain for military reasons and played an essential role in the Battle of Britain against the German airforce in 1940. The first suggestion of using radio signal was put forward by R A Watson Watt in 1935 and, later, the British Government, in

particular through the work of Sir H Tizard, brought it to an operational stage [8]. At present, the military use of radar is still paramount but new civilian applications, in particular *Synthetic Aperture Radar*, were developed for all-weather mapping.

However, an optical radar, with a wavelength λ several orders of magnitude smaller achieves a hugely larger gain (proportional to $1/\lambda^2$) and allows much shorter pulses and, hence, much more accurate ranging. A Q-switched laser produces regular trains of very short (even 40 ps) pulses, which are fed into the focal plane of a large reflecting telescope and sent to the target. On the target, special optical systems—called *retroreflectors*—send the pulse back in the same direction from which it comes. An elementary, two-dimensional realization of such a device is just two orthogonal mirrors. The same telescope receives the returned pulse and the delay $\Delta t = 2D/c$ is measured electronically. Several Earth satellites equipped with retroreflectors are routinely tracked in the context of space geodesy (*Satellite Laser Ranging*, SLR); in particular, the two LAGEOS (LAser GEOdynamic Satellites) suffer little atmospheric drag and provide a very good realization of gravitational motion: they have achieved remarkable accuracies ($\sigma_D < 0.5$ cm, table 2.1). These measurements are also routinely accomplished for the Moon (*Lunar Laser Ranging*, [LLR]) (currently with $\sigma_D < 2$ cm), using four retroreflectors placed there by the NASA Apollo missions and a Soviet spacecraft.

The basis of radar measurements is the *link budget*, relating the emitted (P) to the received (P') power in terms of the wavelength λ, the distance D and the gains G and G' of the main transmitter and the mirror on the target, respectively:

$$P' = PG^2G'^2\left(\frac{\lambda}{4\pi D}\right)^4. \tag{2.2}$$

The *gain* G of a parabolic antenna or reflector depends on the angular position of the source and is the ratio of the power flux in that direction and the isotropic flux. On the axis,

$$G_0 = \frac{4\pi A_e}{\lambda^2}$$

where A_e is the *effective area*, somewhat less than the geometrical area of the dish. The gain measures the lack of perfect collimation due to diffraction.

A laser tracking telescope is capable of transmitting a laser pulse of very short duration τ_t and large energy $P\tau_t$. If $N_t = P\tau_t/h\nu$ is the number of photons in a single pulse, with the radar equation we obtain the number of photons N_r received through the same aperture and detected with a photomultiplier in the primary focus: N_r decreases very fast with the distance, as $1/D^4$. The received pulse has a duration τ_r greater than τ_t due to dispersion in the atmosphere and the superposition of several sub-pulses from different retroreflectors encompassed by the beam. The accuracy with which the round-trip transit time $2D/c$ can be measured results in an accuracy σ_D for the distance. It is important to note that,

Table 2.1. The two LAGEOS satellites: mean orbital elements, mean secular rates and rotation period. For the latter, we give the initial value (roughly 0.5 s for both satellites) and an estimate of the rotation period at the end of 2001.

	LAGEOS 1	LAGEOS 2
Launch	4 May 1976	22 October 1992
Semimajor axis	12 270 km	12 167 km
Revolution period	13 500 s	13 380 s
Rotation period	0.5/ $\approx$ 2500 s	0.5/ $\approx$ 30 s
Eccentricity	0.0039	0.0133
Inclination	109.80°	52.65°
Perigee rate	−0.213°/d	0.438°/d
Node rate	0.342°/d	−0.632°/d

for the Moon, N_r is much less than unity, so that an actual measurement requires averaging over many successive shots. Moreover, the number N_r has a Poissonian distribution and a standard deviation $\sqrt{N_r}$: this places a fundamental limitation on the accuracy with which one can measure the arrival time of the centroid of the returned pulse.

The assumption of a straight path is far from correct, due to atmospheric refraction, which increases the optical path by as much as 2 m, depending on atmospheric conditions and the elevation over the horizon. The bulk of this correction is evaluated with meteorological measurements and a model of the atmosphere but, for a higher accuracy, another harmonic of the laser line is used simultaneously. Since air is a dispersive medium, there is a difference between the transit times in the two frequencies, which can be used to obtain the geometrical distance D.

2.2.3 Lunar Laser Ranging

This section, a complement to Professor Nordtvedt's contribution (chapter 5), briefly introduces the outstanding Lunar Laser Ranging (LLR) experimental programme. This programme is the main and, to some degree, unexpected scientific result of the expensive and largely forgotten NASA Apollo missions for the human exploration of the Moon: three retroreflectors were placed there by the astronauts.

If the ratio $M_g/M_i = 1 + \eta$ between the inertial and gravitational mass of the Earth (at $\boldsymbol{r}_1$) and the Moon (at $\boldsymbol{r}_2 = \boldsymbol{r}_1 + \boldsymbol{r}$) are not the same, their motion in an external (the Sun's) gravitational potential per unit mass $U(\boldsymbol{r}_1)$ fulfils

$$\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}t^2} + \frac{M_E + m_M}{r^3}\boldsymbol{r} = -(\eta_2 - \eta_1)\nabla U(\boldsymbol{r}_1) - \boldsymbol{r}\cdot\nabla\nabla U(\boldsymbol{r}_1) + \mathrm{O}(r^2). \qquad (2.3)$$

The last but one term is related to the solar *tide* in the quadrupole approximation: a violation of the Weak Equivalence Principle (EP) brings in a new solar term, a relative acceleration, *independent of the distance* r. To see what its effect is, note that it corresponds to the potential (per unit mass)

$$U_{\mathrm{EP}} = -H_1 = -(\eta_2 - \eta_1)\frac{m_\odot}{R^3}\boldsymbol{R}\cdot\boldsymbol{r} = -(\eta_2 - \eta_1)n_\odot^2\boldsymbol{R}\cdot\boldsymbol{r} \tag{2.4}$$

where $\boldsymbol{R}$ is the vector from the Earth to the Sun and $n_\odot$ the mean motion of the Earth. H_1 is the corresponding perturbation in the Hamiltonian function. Neglecting the eccentricity of both the Moon and the Earth, H_1 depends on time as $\cos\lambda_{\mathrm{s}} = \boldsymbol{R}\cdot\boldsymbol{r}/(Rr)$ where λ_{s} is the *synodic longitude* of the Moon. The effect on the distance r is nearly the same as the effect on the osculating semimajor axis a, governed by the first Lagrange equation

$$\frac{\mathrm{d}a}{a\,\mathrm{d}t} = -\frac{2}{na^2}\frac{\partial H_1}{\partial M} \tag{2.5}$$

in terms of the mean anomaly $M = n(t - t_0)$ and the mean motion n of the Moon around the Earth. Since M and λ_{s} only differ by their origin (the perigee and the Sun, respectively), this directly integrates to

$$\frac{\delta r}{r} = -\frac{2}{n^2r^2}H_1 = 2(\eta_2 - \eta_1)\frac{Rn_\odot^2}{rn^2}\cos\lambda_{\mathrm{s}}. \tag{2.6}$$

At an elementary level and when the eccentricity of the Moon is neglected, this equation is a simple consequence of the energy theorem: the rate of change in the *osculating* orbital energy (per unit mass) $-m_{\mathrm{E}}/(2a)$ equals the power

$$(\eta_2 - \eta_1)\frac{m_\odot}{R^3}\boldsymbol{R}\cdot\boldsymbol{v}$$

of the new force, from which (2.5) is easily derived.

The much larger force $-\boldsymbol{r}\cdot\boldsymbol{\nabla}\boldsymbol{\nabla}U(\boldsymbol{r}_1)$ is due to the solar tides and, since it corresponds to a potential function quadratic in $\boldsymbol{r}$, it gives a correction in the distance r with the period $2\lambda_{\mathrm{s}}$, easily distinguishable from the Equivalence Principle signal. An error of 10^{-10} in $\delta r/r$ (equation (2.2.3)) corresponds to an error in $\eta_2 - \eta_1$ of about 2×10^{-11}: after about 30 yr of data and a large number of lunar months, this test, in effect, has attained the accuracy 5×10^{-13}. From the fundamental point of view, a non-vanishing value for the parameter η is a complex and overall effect, resulting from all terms which appear in the mass of a body, in particular its binding energies. While for small bodies, such as those employed in laboratory tests, the main binding energies are microscopic, for the solar system bodies we are testing possible differences in the contribution of the gravitational binding energy—the main one—to the inertial and gravitational mass. The Lunar Laser Ranging test of the Equivalence Principle, therefore, is essentially different from laboratory tests. As amply discussed by K Nortdvedt, LLR is now an outstanding tool, not only for testing gravitational theories but also for investigating the dynamics and interior of the Moon.

2.3 Angle

2.3.1 Fundamentals

If two light rays with null vectors p_1 and p_2 arrive at an event $x(s)$ of an observer at its proper time s, their angular separation δ is invariantly defined as follows: construct, for each of them, the projection $p_\perp = p - v(p \cdot v)$ orthogonal to the four-velocity $v = \mathrm{d}x/\mathrm{d}s$; δ is the angle between the space vectors $p_{1\perp}$ and $p_{2\perp}$. This shows that an angle depends on the motion of the (not necessarily inertial) observer who measures it.

In classical physics, a goniometer is a *rigid* graduated circle, to which the directions of distant objects (for instance, sighted through a telescope) are referred; but even in special relativity, rigorously rigid bodies do not exist generically and, contrasting with distances and frequencies, there is no 'fundamental' absolute goniometer. Indeed, a rigid body requires that the material properties and the distances of its points be the same at all times; and their world lines must be parallel, a requirement which can be fulfilled only if it is inertially at rest. One must make do with actual bodies, with a finite elasticity and estimate and monitor their lack of rigidity and its effect on the measurement. Note also that the direction of a light beam at the instrument is affected not only by the position of the source but also by gravitational effects on the propagation; for this reason accurate astrometry is inextricably linked with the gravitational deflection. Since for a generic star in the sky, the photon trajectory is displaced by the Sun by approximately $m_\odot$, at 1 AU the deflection $\approx 10^{-8} = 2 \times 10^{-3''}$ must be fully integrated into GAIA's analysis; considering its very large data base, a very good measurement of γ is expected.

2.3.2 Techniques

Measurements of angles has always had a crucial role in traditional astronomy, beginning with Erathostenes' estimate of the radius of the Earth [12]. The optical resolution of the naked eye, $\approx 1' = 0.3$ mrad, has been drastically overcome in the seventeenth and eighteenth centuries with *graduated circles*, built by clever and patient craftsmen, especially in England. J Horrox was able to divide a 3 ft staff into 10 000 parts, each about 0.1 mm wide. By about 1820, graduated circles were achieving accuracies better than a second of arc, to be compared with the daily parallax of the Sun of $16.12''$ [9]. The measurement of the yearly parallax of 61 Cygni, with a parallax of $0.292''$, was announced by F W Bessel in 1838: this was a milestone in three-dimensional astronomy and truly made astrophysics possible. At a distance of 1 pc, the parallax—the angle subtended by 1 AU—is $1''$; inversely, with an accuracy of 10 μas $= 10^{-5''}$ parallactic distances up to 10^5 pc—further than the whole galaxy—will be available.

The main astronomical realization on the ground of angular measurements is *Very Long Baseline Interferometry* (VLBI), based upon a 'rigid' body—the Earth itself—which rotates with a 'constant' angular velocity $\boldsymbol{\omega}_\mathrm{E}$. Very schematically,

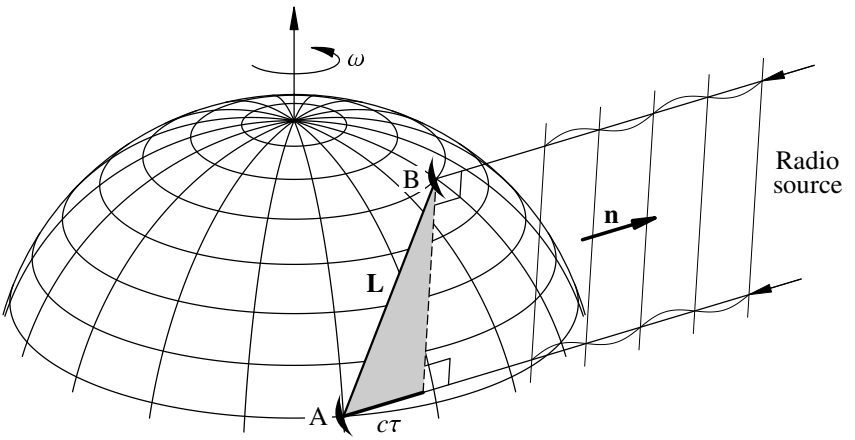

Figure 2.2. In VLBI, two radiotelescopes A and B, separated by the baseline $\boldsymbol{L}$, measure the phase delay $\tau = \boldsymbol{L} \cdot \boldsymbol{n}/c$ for a source in the direction $\boldsymbol{n}$. In the body-fixed frame, $\boldsymbol{n} = (\boldsymbol{n} \cdot \hat{\boldsymbol{\omega}})\hat{\boldsymbol{\omega}} + \boldsymbol{n}_\perp$ is the sum of a constant component along the rotation axis and a part $\boldsymbol{n}_\perp$ which varies sinusoidally with the rotation period. Accordingly, the delay $\tau = \tau_0 + \tau_1 \cos(\omega t + \phi)$ is the sum of a constant and a periodic contribution.

two (or more) large radio antennas at the far-away points A and B separated by a vector $\boldsymbol{L}$—the *baseline*—point to a radio source in the direction $\boldsymbol{n}$. Within a limited bandwidth and a wavelength $\simeq \lambda$, electromagnetic wavefronts are received at the two stations with a delay τ of the order of L, consisting of a constant part and a part with the period of a day (figure 2.2). With sophisticated software, the two trains are correlated and $\tau(t)$ is determined. From the delays corresponding to the two sources, their angular separation can be determined. A phase error of the order of unity implies an error in the delay $\sigma_\tau \approx \lambda$, corresponding, with $L = 10\,000$ km and at 30 GHz, to an angular accuracy of 10^{-9} rad $= 0.2$ mas. The error is proportional to $1/L$: for better performance and coverage, VLBI antennas in space have been planned and built. In differential VLBI, the angular separation between two sources is measured, realizing an accurate goniometer.

Clearly, a VLBI astronomical 'goniometer' requires accurate knowledge of (a) the rotation vector $\boldsymbol{\omega}_\mathrm{E}$ and (b) the effect of the relative motion of the two tectonic plates on which the stations A and B stand. The latter is, even conceptually, a delicate task, since it requires establishing a 'rigidly' rotating Cartesian coordinate system $\mathcal{S}$ with the origin at the centre of mass of the Earth, with respect to which the motion of each plate is known. This is usually done by requiring that, in $\mathcal{S}$, the overall mean tectonic motion vanishes. A large, world-wide effort has allowed the routine realization of these requirements, resulting in errors in the rotation rate ω_E smaller than 1 mas y^{-1}. The details of this procedure

Table 2.2. Three projects of space astrometry: past (HIPPARCOS) and future. The average accuracy in angular position for objects up to a given visual magnitude are given in the third and fourth columns, the last column gives limiting visual magnitude.

Mission	Launch	Number of stars	Accuracy (in mas)	At Mag.	Limiting magnitude.
HIPPARCOS (ESA)	1989	1.2×10^5	1	9	12
SIM (NASA)	2009	$\approx 10^4$	0.004	13	20
GAIA (ESA)	2012	$\approx 10^9$	0.010	15	20

are complex, highly technical and dependent on the required accuracy: let it suffice to say here that VLBI inextricably links three different areas: positional astronomy of the radio sources, the rotational motion of the Earth and plate tectonics. VLBI is an important part of *space geodesy*. Angular resolutions of ≈ 10 μas in the angular separation can be achieved.

2.3.3 Space astrometry: GAIA

In the much more important optical band, due to the atmospheric scintillation near the ground, accurate astrometry needs instruments in space: they have achieved huge improvements over ground astrometric telescopes. After the milestone achievements of the Hipparcos mission of the European Space Agency (ESA), several projects are being planned, with the purpose of plotting a map of the sky with a very large number of optical astronomical sources and a very high accuracy (table 2.2). Note also that with a simple homogeneous distribution of sources, an increase by a factor 10 in angular accuracy implies an increase by a factor 1000 in the number of objects, with a huge increase in complexity and computational load.

Figure 2.3 shows the optical bench of GAIA—a follow-up of Hipparcos—a cornerstone of the ESA. It collects light with two large rectangular mirrors (ASTRO1 and ASTRO2) 'fixed' in the equatorial plane of a spacecraft, spinning at the angular velocity $60'' \ s^{-1}$ (a period of 6 h). Their longer size (along the equator) is $d = 1400$ cm. A precessional motion allows covering the whole sky. Their axes are separated by an angle $\Gamma = 106°$: the sources in their fields of view are brought onto the same focal plane, mapping angles into distances. As the spacecraft rotates, a large CCD array detects the motion of all the images across the plane and their distances [13]. At the end of the mission, all angular measurements are integrated into a single sky map.

I confine myself to two general remarks, just to give an idea of the fundamental limits of such an instrument and the integrated character of the mission. First, the required accuracy of 10 μas $= 5 \times 10^{-11}$ rad places severe

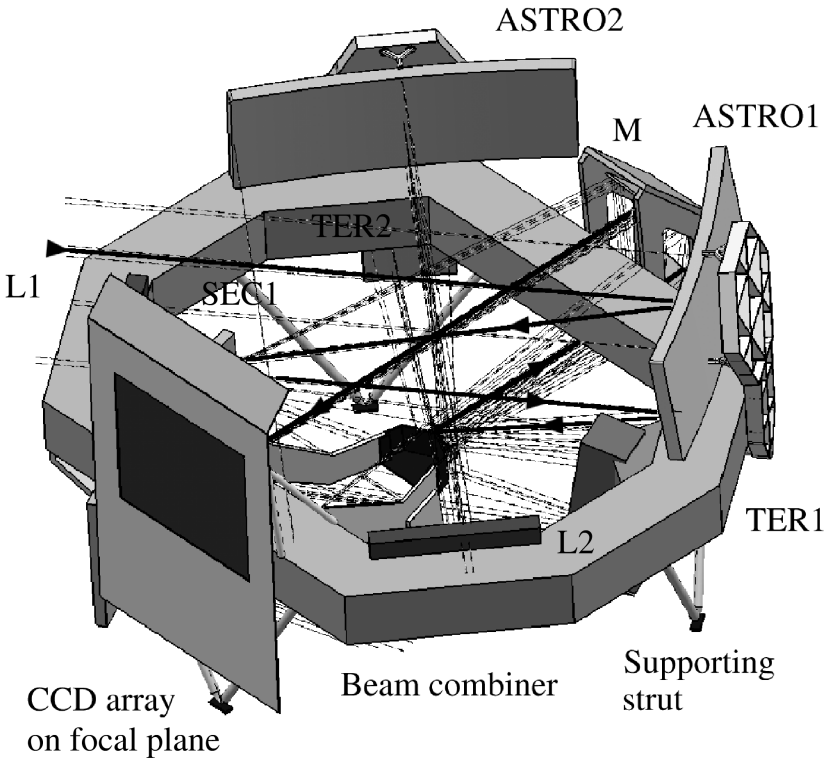

Figure 2.3. The concept of GAIA's astrometric optical bench. The photons collected by the two main rectangular mirrors ASTRO1 and ASTRO2—$d = 1400$ cm wide and $\Gamma = 106°$ apart—are brought onto the beam combiner (by the secondary (SEC) and the tertiary (TER) mirrors) which, after a reflection on the beam combiner and the mirror M, sends them onto the focal plane for detection by the CCD array. For simplicity, only rays of the first beam are given in bold: they are reflected on the secondary SEC1 and the tertiary TER1 mirrors below ASTRO1. Opposite the main mirrors are the two main components L_1, L_2 of the laser metrological system to measure Γ; the three supporting struts are also shown. Figure kindly provided by F Mignard.

constraints on the rigidity of the optical bench. At time scales shorter than the rotation period, the fundamental angle Γ must be constant to 1 μas, which in turn, imposes a high thermal stability: an interferometric laser system on board is needed to provide a measurement of Γ. For longer times, similarly to VLBI, the rotation frequency is included, together with the angular data, in the set of parameters to be determined. Second, the diffraction pattern of a point source, with width $\lambda/d \simeq 100$ mas, is 10 000 times larger than the required angular

accuracy 10 μas. This impressive hindrance is overcome by scanning each source many times during the mission and relying on the stability and a good knowledge of its profile. The resulting attainable fraction of λ/d is essentially determined by the shot noise, i.e. by the *total number of photons N* received from a single source; in turn, this sets requirements on the duration of the mission and the width of the two fields of view.

2.4 Frequency

2.4.1 Fundamentals

An atomic frequency standard is based upon a microscopic quantum system capable of sharp energy states, which can control a generator of electromagnetic waves. At a fundamental level, its stability is ensured by the *Strong Equivalence Principle*, according to which the 'constants' of local physics do not depend upon time and place. A violation of this principle could show up in a disagreement between frequency standards of different nature—none has been found so far.

The shift detected as an electromagnetic signal transfers a frequency from one standard $_1$ to another $_2$ at a different place, has a special-relativistic component

$$\frac{\omega_2}{\omega_1} = \left[\frac{1-v_1^2}{1-v_2^2}\right]^{1/2} \frac{1-\boldsymbol{v}_2\cdot\hat{\boldsymbol{k}}}{1-\boldsymbol{v}_1\cdot\hat{\boldsymbol{k}}} \tag{2.7}$$

where $\hat{\boldsymbol{k}}$ is the propagation unit wavevector. In addition to the ordinary Doppler effect $O(v)$ (second factor), there is the transversal effect $O(v^2)$ (first factor), present even if the distance between $_1$ and $_2$ does not change. There is also a gravitational shift; e.g. photons detected on the Earth at an altitude h above the source are redshifted by $\Delta\omega/\omega = hg$. In the solar system, with a gravitational potential per unit mass U, the fractional Doppler shift is of order $\Delta U \approx m_\odot/r \approx v^2$. Note that the distinction between the gravitational shift and the transversal Doppler effect is not invariant: in a freely falling frame, there is no gravity acceleration but the velocities of the source and the detector are different. In interplanetary propagation, where the potential and the kinetic energy are generally of the same order of magnitude, the two effects are comparable. The proper way to obtain the frequency shift in the solar system is a single and exact description, involving the two four-velocities v^μ of the transmitter and the receiver, and the null four-vector k^μ of the parallely propagated photon; then

$$\frac{\omega_1}{\omega_2} = \frac{[g_{\mu\nu}k^\mu v^\nu]_1}{[g_{\mu\nu}k^\mu v^\nu]_2}. \tag{2.8}$$

This expression can be proved on the basis of the definition of null geodesics [18] and has an obvious physical interpretation: $g_{\mu\nu}k^\mu v^\nu$ is proportional to the photon energy, as measured by an observer with four-velocity v^μ.

At the cost of looking naïve, I wish to remind that, in relativity, physical time is a local quantity and global clock synchronization is, in principle, impossible. There is, however an interesting case of approximate synchronization in the neighbourhood of the Earth. The *geoid* is defined as a surface where the total gravitational potential (per unit mass)

$$U_{\mathrm{T}} = U - \tfrac{1}{2}\omega_{\mathrm{E}}^2 r^2 \sin^2\theta \tag{2.9}$$

is constant. The last term is the centrifugal potential, a function of the distance r from the centre and the colatitude θ; U is the gravitational potential per unit mass of the Earth, *including the oblateness contribution*. For all clocks at rest on one such surface (in the rotating system), the relation between the proper time s and the coordinate time t

$$\mathrm{d}s^2 = \mathrm{d}t^2(1 - 2U_{\mathrm{T}} + \cdots) \tag{2.10}$$

is the same; hence, the global coordinate t is synchronous with their proper times.

2.4.2 Techniques

The electromagnetic Doppler effect in the radio band has, of course, many industrial applications, in particular to measure fluid velocities. To obtain the frequency displacement $\omega(t) - \omega_0$ of the received signal $s(t)$ from the main oscillator frequency ω_0, a phase-locked receiver is needed. In the traditional version, it is based upon a *Voltage Controlled Oscillator* (VCO) which, under an input $z(t)$, produces an output $r(t)$ with frequency proportional to its amplitude. A *closed-loop* receiver is obtained when the input $z(t)$ is the low-frequency part of the beat signal between $r(t)$ and $s(t)$. For a better performance, especially with a large dynamic range of the incoming frequency, an *open loop* is used, which continuously records the electric field of the incoming wave with a time resolution better than a period and evaluates its instantaneous frequency, referenced to the standard. In the acquisition phase, when the incoming frequency is not known well, the oscillator frequency is swept up and down in a limited bandwidth, until the incoming carrier is found and lock is achieved; then the 'tracking' phase follows [16].

The use of Doppler measurements as a tool for fundamental physics in space has been made possible by the developments of extremely accurate frequency standards (figure 2.4). The main one, which is currently available and operational for radio astronomy and space, is the *hydrogen maser*, a microwave cavity tuned to the hyperfine frequency splitting $\nu_0 = 1,420,405,751.68$ Hz of the ground level of atomic hydrogen due to the spin interaction between the electron and the nucleus. Better performances are offered by laboratory devices, like the Superconducting Cavity Stabilized Oscillators (SCSO). Fractional changes are measured by

$$y(t) = \frac{\omega(t) - \omega_0}{\omega_0} \qquad \left(\omega = 2\pi\nu = \frac{\mathrm{d}\phi}{\mathrm{d}t}\right). \tag{2.11}$$

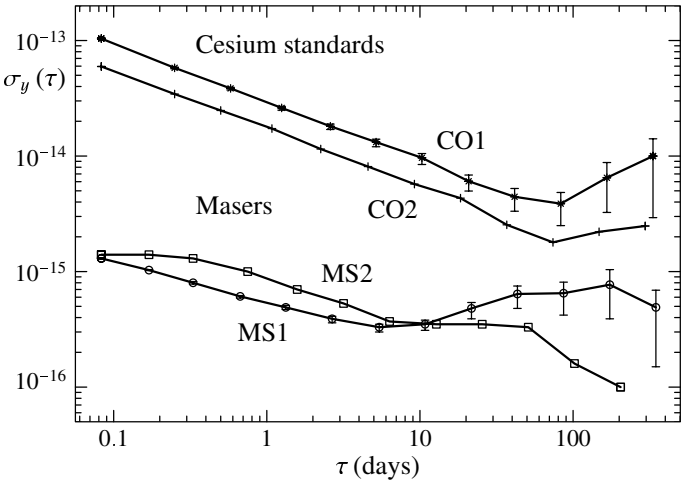

Figure 2.4. Measured Allan deviation $\sigma_y(\tau)$, as a function of the integration time τ, of frequency standards based upon hydrogen masers (MS1 and MS2) and cesium oscillators (CO1 and CO2), relative to the main frequency standard kept at the National Bureau of Standards (Boulder, USA). Long-term drifts have been fitted out for the hydrogen maser signals.

To define its accuracy, the averaging time τ must be specified: the relevant quantity is the running average

$$y_\tau(t) = \frac{1}{\tau}\int_{t-\tau/2}^{t+\tau/2} \mathrm{d}t'\, y(t') = \frac{\phi(t+\tau/2) - \phi(t-\tau/2)}{\omega_0 \tau} \tag{2.12}$$

also expressed in terms of the phase ϕ. The measure of its statistical fluctuations must take into account the fact that we can only measure frequency changes, not absolute frequencies: this is accomplished with the *Allan variance* [2]

$$\sigma_y^2(\tau) = \tfrac{1}{2}\langle [y_\tau(t+\tau) - y_\tau(t)]^2 \rangle. \tag{2.13}$$

Commercial hydrogen masers reach a stability better than a part in 10^{15} for averaging times of the order of 1 h: in the laboratory, using different standards, longer averaging times can be attained (figure 2.4). Of course, a frequency standard can also be used as a clock, with an accuracy for the measurement of an interval T equal to $\sigma_\mathrm{T} = T\sigma_y(T)$. It should be noted that the accuracy of a clock depends in an essential way on the length of the measured interval.

The scientific use of Doppler measurements involving microwave links to interplanetary spacecraft has been made possible by two great technological advances. First, continuous improvements in NASA's Deep Space Network (DSN), which operates several large dish antennas at three locations widely

Table 2.3. The Deep Space Network frequencies (in MHz) of coherent uplink (↑) and downlink (↓) carriers and their conversion ratios R (a rational number). Two nearby downlinks in K_a-band are used in the Cassini mission.

	S-band	X-band	K_a-band	
↑		7175	34 316	
↓	2299	8430	32 034	32 029
R		880/749	3344/749	14/15

spaced in longitude (Goldstone, California; Madrid, Spain; Canberra, Australia); in particular, at Goldstone the new, advanced DSS25 station is only devoted to science. Second, higher-frequency bands are being used (see table 2.3): at DSS25 the new K_a band (table 2.3) has been successfully implemented with a very sophisticated instrumentation. The Cassini spacecraft is the first interplanetary probe to use this band: the Italian Space Agency has provided on board the complex high-gain antenna (4 m in diameter) and the frequency transponder. In the current, two-way configuration, a Doppler measurement uses a stable and very narrow spectral line controlled on the ground by a frequency standard: this is the carrier for transmission. On board, a coherent carrier is locked to the arriving beam and sent back to the ground, where the total frequency shift $y(t)$ (including both the up- and down-link) is measured as a function of time.

2.4.3 The Cassini conjunction experiment

Cassini is a huge interplanetary probe launched in October 1997, due to arrive at Saturn in July 2004 for an exploration of the Saturnian system, which will last 4 yr (at least). In normal conditions, Cassini's frequency stability requirement for the new K_a link is $\sigma_y(\tau) = 3 \times 10^{-15}$ for $1000 < \tau < 10\,000$ s, including all disturbances, both at the station and the spacecraft, and those due to the traversed media (the atmosphere, the ionosphere and interplanetary space). This corresponds to an accuracy in velocity $\sigma_v = 10^{-4}$ cm s^{-1} and, over 1000 s, an accuracy of 1 mm in the change of distance. For the acceleration the accuracy is $\sigma_y(\tau)c/\tau = 10^{-7}$ cm s^{-2}. It is important to note that, contrary to the radar technique, with this method *absolute distances are not accessible*. Nothwithstanding the novelty, after some instrumental problems, in good working conditions (i.e. in absence of manoeuvres which produce unknown displacements in the centre of phase of the antenna, and when the weather at the ground station is only slightly perturbed) Cassini's system works fairly well and this specification is often fulfilled.

According to general relativity, the deflection of electromagnetic waves by the Sun is twice the Newtonian value

$$\delta_{\mathrm{N}} = 2\frac{m_{\odot}}{b} = 4 \times 10^{-6}\frac{R_{\odot}}{b}. \tag{2.14}$$

Writing

$$\delta = (1+\gamma)\,\delta_{\mathrm{N}} \tag{2.15}$$

the general relativity value corresponds to $\gamma = 1$. In both cases, the deflection is inversely proportional to the impact parameter b of the beam, bounded below by the solar radius $R_{\odot}$. Therefore, this fundamental test can discriminate between a scalar and a tensor theory of gravity: a small scalar field, a remnant of the dilaton field φ of primordial cosmology, may contribute at the present time. The value of $\gamma - 1$ is an indication of the mixture between a tensor field of rank 2 and other fields which determines gravity; since in the Newtonian case $\gamma = 0$, it is not a surprise that in the scalar case $\gamma < 1$. Theoreticians have not yet been able to construct a clear-cut and computable fundamental theory for this interaction: in a simple version, where φ is coupled to matter through a potential $V(\varphi)$, one finds (see Nordtvedt's chapter (5) in this volume, formula (5.11)):

$$\gamma - 1 = -\frac{1}{2}\left(\frac{\mathrm{d}\ln V(\varphi)}{\mathrm{d}\varphi}\right)^2. \tag{2.16}$$

As the Universe expands, φ tends to the minimum of $V(\varphi)$ and $\gamma - 1$ becomes small, but remains negative. The deficiency may be of order 10^{-5}–10^{-7} ([10] and cited papers). Measuring this tiny discrepancy, therefore, is of fundamental importance for the understanding of the nature of gravitation. The same remnant scalar field produces also a change in the other PPN parameter β, a violation of the Weak EP and a change in time of the gravitational constant at the cosmological scale.

Traditionally, the deflection parameter γ can be obtained directly, by comparing the angular distance between two celestial sources in the sky with and without the Sun nearby (as done for the first time during the solar eclipse of 1919 [11]). However, one can also use the fact that a deflected path implies that the transit time Δt between the emitting and receiving stations is larger than the geometric value and changes with time; this is called the *Shapiro effect*. Both methods have been actively pursued and have confirmed the predictions of general relativity with an accuracy $\sigma_{\gamma} \approx 10^{-3}$ [15, 20]. But there is a third way, which relies on the obvious fact that a deflection changes the angle between the direction of propagation of the photons and the velocities, of order $v \simeq 10^{-4}$, of the emitting and receiving stations (details in [5]). A time-dependent Doppler shift y_{GR} of order $v\delta$ is produced: numerically and, in general, (figure 2.3),

$$y_{\mathrm{GR}} \approx \frac{1+\gamma}{2} 8 \times 10^{-10}\left(\frac{R_{\odot}}{b}\right). \tag{2.17}$$

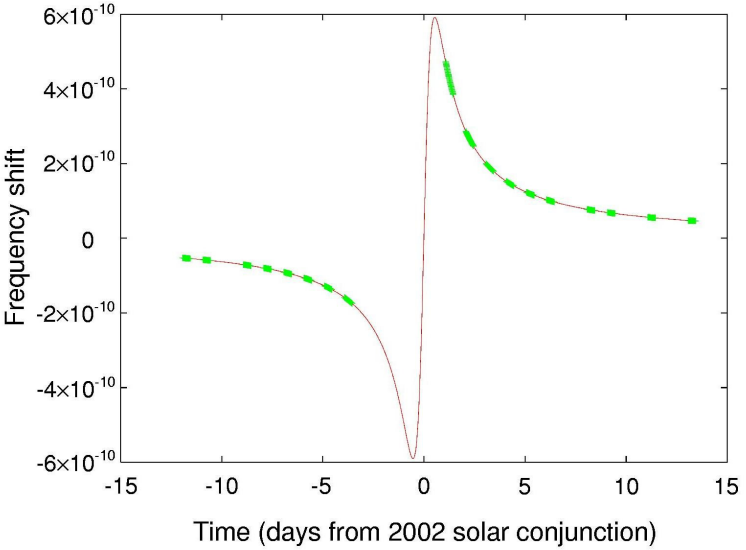

Figure 2.5. The gravitational signal $y_{\rm GR}$ and the available 18 passages during Cassini's cruise to Saturn, as a function of days from solar conjunction, which occurred on 21 June 2002.

Naïvely equating this signal to the Cassini stability requirement $\sigma_y = 3 \times 10^{-15}$, at grazing incidence, one gets an error $\sigma_\gamma \approx 10^{-5}$.

The main hindrance to attain an accuracy of this order and the reason why the experiment has not been done earlier is the fact that the beam must traverse the solar corona, a dense, unmodellable, unpredictable and fast-varying plasma: it produces an *outward* deflection, at a small impact parameters even larger than the gravitational effect (see [4], in particular figure 2.1). It can be easily shown that, if $N_{\rm e}(t) = \int {\rm d}s\, n_{\rm e}$ is the total electron plasma columnar content encountered along the beam (up and down) by a photon emitted at the time t from the ground station, the Doppler observable $y(t)$ is affected by a term $\propto {\rm d}N_{\rm e}/(\omega^2\, {\rm d}t)$, inversely proportional to the square of the carrier's frequency ω. I recall that no frequency measurement is carried out on board; moreover, the time scale of the experiment is comparable with the round-trip light-time, so that the electron columnar contents (mainly localized near the Sun) in the uplink ($N_{{\rm e}\uparrow}(t)$) and the downlink ($N_{{\rm e}\downarrow}(t)$) are independent, and generally different. The observable $y(t)$ is, therefore, the sum of three contributions: the non-dispersive part $y_{\rm nd}(t)$, which includes the gravitational signal, $y_\uparrow(t)$ and $y_\downarrow(t)$. In Cassini's experiment, for the first time, the elimination of all the plasma contributions has been made possible with the use of a multi-frequency link. As indicated in table 2.2, two carriers, in the X- and

K_a-bands, respectively, are transmitted to the spacecraft; the X uplink, besides the normal X downlink, controls on board also a side channel in the K_a band; with the complete up and down K_a link we have three independent observables, which can be conveniently labelled as $y_{XX}(t)$, $y_{XK}(t)$, $y_{KK}(t)$. They are linear combinations of y_{nd}, the uplink and the downlink plasma contribution. With a simple linear system the latter ones can be eliminated. What is left, $y_{nd}(t)$, contains not only the signal $y_{GR}(t)$ but also other contributions, in particular from the troposphere and the orbital dynamics.

Details of the error budget are given in [14]. A 30-day experiment was carried out from 6 June to 7 July 2002 and has confirmed that the plasma compensation system works. In [6], other aspects of the experiment are discussed, in particular the important role of the dynamical model: the non-gravitational forces acting on the spacecraft perturb its motion quite appreciably and must be suitably modelled and determined. At the level 1σ, the result [6]

$$\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5} \tag{2.18}$$

does not show any violation of general relativity but with an accuracy greater by about a factor 50 over previously published experiments.

Acknowledgments

This material is in part taken from [7]: I am very grateful to D Vokrouhlický for his insight and collaboration.

References

[1] Ashby N 1998 Relativistic effects in the Global Positioning System *Gravitation and Relativity: At the turn of the Millennium, Proc. GR-15 Conference* ed N Dadich and J Narlikar (Pune: IUCAA) p 231
[2] Barnes J A *et al* 1971 *IEEE Trans. Instrum. Meas.* **IM-20** 105
[3] Bertotti B 1962 The theory of measurement in general relativity *Evidence for Gravitational Theories* ed C Møller (New York: Academic) p 174
[4] Bertotti B 1998 *Gen. Rel. Grav.* **30** 209
[5] Bertotti B and Giampieri G 1992 *Class. Quantum Grav.* **9** 777
[6] Bertotti B, Iess L and Tortora P 2003 A test of general relativity with radio links with the Cassini spacecraft *Nature* **425** 374
[7] Bertotti B, Farinella P and Vokrouhlický D 2003 *Physics of the Solar System* (Dordrecht: Kluwer)
[8] Bowe E G 1987 *Radar Days* (Bristol: Hilger)
[9] Chapman A 1990 *Dividing the Circle; The Development of Critical Angular Measurements in Astronomy 1500–1850* (New York: Horwood)
[10] Damour T, Piazza F and Veneziano G 2002 *Phys. Rev.* D **66** 046007
[11] Dyson F W, Eddington A S and Davidson C 1920 *Phil. Trans. R. Soc.* A **220** 291
[12] Dragoni G 1975 Introduzione allo studio e alle opere di Eratostene *Physics* **17** 41

[13] ESA 2000 *GAIA. Composition, Formation and Evolution of the Galaxy* ESA-SCI 4
[14] Iess L, Giampieri G, Anderson J D and Bertotti B 1999 *Class. Quantum Grav.* **16** 1487
[15] Lebach D E *et al* 1995 *Phys. Rev. Lett.* **75** 1439
[16] Lindsey W C and Simon M K 1973 *Telecommunication Systems Engineering* (New York: Dover)
[17] Nordtvedt K 1998 *Class. Quantum Grav.* **15** 3363
[18] Schrödinger E 1955 *Expanding Universes* (Cambridge: Cambridge University Press)
[19] Synge J L 1960 *Relativity: the General Theory* (Amsterdam: North-Holland)
[20] Will C M 2001 *Living Reviews* **4** 4

Chapter 3

Frame-dragging and its measurement

Ignazio Ciufolini
Dip. Ingegneria dell'Innovazione, Università di Lecce, Via Monteroni, 73100 Lecce, Italy

3.1 Some historical background on the measurement of gravitomagnetism and the gravitational field inside a rotating shell

This is just a brief introduction to past and present experiments to measure gravitomagnetism and the problem of the gravitational field inside a rotating shell: for a more exhaustive introduction we refer readers to Ciufolini and Wheeler [1].

In 1915, Einstein published his theory of general relativity. Among the sources of inspiration was Mach's idea on the origin of inertia and inertial forces [1, 2]. Mach thought that centrifugal and inertial forces were the result of rotation and accelerations with respect to the masses in the universe.

Influenced by Mach, several investigators studied the problem of the gravitational field inside a rotating shell. In a seminal paper of 1918, Thirring published a solution of the Einstein field equation representing the metric inside a rotating shell to first order in M/R (mass over shell radius) and to first order in ω, the angular velocity of shell [3]. In 1966, Brill and Cohen derived the metric inside a shell with an arbitrary mass and to lowest order in angular velocity [4]. An extension of the Brill–Cohen results to higher orders in ω was then published in 1985 by Pfister and Braun [5, 6].

Nevertheless, the exact solution representing the spacetime geometry inside a shell with an arbitrary mass and rotating with an arbitrary angular velocity is still unknown. An exact solution inside a rotating shell would give us insight into the role of the 'Mach principle' in general relativity.

Indeed, the level at which general relativity satisfies Mach's idea on the origin of inertia has been discussed in a large number of books and papers (see, for example, [1, 2]). However, general relativity satisfies at least a 'weak

manifestation' of Mach's ideas: the dragging of inertial frames. Indeed, in Einstein's gravitational theory, the concept of an inertial frame has only a local meaning and a local inertial frame is 'rotationally dragged' by mass–energy currents because moving masses influence and change the orientation of the axes of a local inertial frame, i.e. of the gyroscopes; thus, a current of mass such as the spinning Earth 'drags' and changes the orientation of the gyroscopes with respect to the distant stars.

It might be surprising to know that the first experiments to detect the gravitational influence of the rotation of a mass and to measure the dragging of a gyroscope by a rotating body were performed well before the development of Einstein's theory of general relativity [1].

In 1896, Benedikt and Immanuel Friedländer [8] tried to measure the dragging effect due to a rapidly rotating, heavy fly-wheel on a torsion balance. Immanuel Friëdlander wrote:

> In the same way as centrifugal force is acting on a static wheel due to the rotation of the heavy earth and the cosmos, there should, I thought, appear on accordingly smaller scale a centrifugal force action on bodies near moving heavy fly-wheels. Would this phenomenon be detectable....

In 1904, August Föppl [9] tried to measure the dragging effect on a gyroscope due to the rotation of Earth: he reached an accuracy of about 2% of the Earth's angular velocity. However, the general relativistic dragging effect on a gyroscope at the surface of the Earth (at a European or US latitude) is about 2×10^{-10} of its rotation rate! These experiments, performed before the development of general relativity, were inspired by Mach's ideas on inertia.

In 1916, de Sitter [10] calculated the tiny shift in the perihelion of Mercury due to the rotation of the Sun, a particular case of the shift in the pericentre of an orbiting test particle due to the angular momentum of the central body, see section 2. This shift of about $-0.002''$/century is about 5×10^{-5} times smaller than the standard general relativistic Mercury precession of $\sim 43''$/century and is too small to be measured.

In their well-known 1918 paper, Lense and Thirring [7] calculated the gravitomagnetic secular perturbations of the moons of various planets. In particular, the V moon of Jupiter has a considerable gravitomagnetic secular precession; however, the observations do not yet allow this effect to be separated out and measured.

In 1959, Yilmaz [11] proposed using polar satellites to detect the gravitomagnetic field, thus avoiding the effects due to the non-sphericity of the Earth's gravity field. In 1976, Van Patten and Everitt [12] proposed measuring the Lense–Thirring nodal precession using two drag-free, guided satellites, counter-rotating in the same polar plane. The reason for proposing two counter-rotating satellites was to avoid the error associated with the determination of the inclination.

Between 1959 and 1960 [13, 14], G E Pugh and Leonard Schiff independently proposed an experiment using orbiting gyroscopes: this became the well-known Gravity Probe-B experiment, or GP-B (launched on 20 April 2004). The Stanford University group has been working for more than 30 years to make and fly superconducting gyroscopes in an Earth-orbiting satellite [15]. At an altitude of about 650 km, the axis of a gyroscope is predicted to undergo a gravitomagnetic frame-dragging of about 42 milliarcsec per year. GP-B should detect the gravitomagnetism of the Earth and measure it to an accuracy of about 1% or less.

Several other experiments have been proposed for measuring the gravitomagnetic field; for a review, see Ciufolini and Wheeler [1]. Here we only mention the Foucault pendulum at the South Pole [16], the ring laser gyroscopes [17] and the orbiting gradiometers. Between 1980 and 1989, the use of gravity gradient resonant detectors orbiting the Earth [18] and superconducting gravity gradiometers in a polar orbit [19] was proposed to measure magnetic components of the Riemann tensor, with the accuracy needed to detect the gravitomagnetic field. Some indirect astrophysical evidence of frame-dragging was obtained in 1988 by the periastron precession rate of the binary pulsar PSR $1913 + 16$ [20].

In 1984 and 1988, we proposed [21] the detection of the gravitomagnetic field by measuring the orbital drag on non-polar, passive, laser-ranged satellites. The fundamental idea [21, 22] of this experiment, called the LAGEOS III experiment, is based on two considerations:

(a) position measurements of laser–ranged satellites, of the LAGEOS (1976) type (see later), are accurate enough to detect the very tiny effect due to the gravitomagnetic field—the Lense–Thirring precession; and
(b) to 'cancel out' the enormous perturbations due to the non-sphericity of Earth's gravity field, we need a new satellite with an inclination supplementary to that of LAGEOS, and with the other orbital parameters, a and e, nearly equal to those of LAGEOS.

The accuracy of this experiment was estimated, by several studies and papers [22], to be, *in* 1988, of the order of 10% of the Lense–Thirring effect.

In 1998, the LARES experiment [23] was proposed and selected as a phase-A study by ASI, the Italian Space Agency. This space mission would allow the Lense–Thirring effect to be measured at the 1% level and is briefly described in section 3.5.

Between 1995 and 2001, using two laser-ranged satellites, we measured [24–27,73] the Lense–Thirring effect several times with an accuracy ranging from about 50% to about 30%. These are described in section 3.5. Between 1998 and 2001, this method provided a direct measurement of the Earth's gravitomagnetism with an accuracy of the order of 30%. Indeed, we also report here the latest measurement of the Lense–Thirring effect, obtained in 2001 with the LAGEOS satellites using nearly 8 years of data. This 2001 result fully confirms and improves our previous measurements of the Earth's frame-dragging.

On 20 April 2004, the Gravity Probe-B experiment was launched in order to try to measure the Earth's frame-dragging with 1% accuracy or better.

In 2004, accurate measurements of the Earth's Lense–Thirring effect have been obtained using the recently released Earth's gravity field models generated by the space missions CHAMP and GRACE and *only the nodes*, of the LAGEOS satellites by analysing about 10 years of data. The accuracy of this recent measurement has been less than 20% [73] (see section 3.5.3). From this experiment we have concluded that the Lense–Thirring effect exists and its experimental value is within $\sim$ 20% of that predicted by Einstein's theory of general relativity. However, the most recent measurement, in full agreement with the prediction of general relativity, has an accuracy of *only* about 5% [75].

3.2 Frame-dragging, the weak-field slow-motion analogy: an invariant characterization of gravitomagnetism

In the weak-field slow-motion approximation, a formal analogy with electrodynamics has been developed by using the Einstein field equation and the geodesic equation. In geometrodynamics [1, 2], in the weak-field slow-motion approximation for a stationary, localized, mass–energy distribution, the $(0i)$ components of the Einstein field equation can be written in the Lorentz gauge: $\Delta h_{0i} \cong 16\pi\rho v^i$. This is formally analogous to the Maxwell–Ampère equation for the vector potential of electrodynamics in the Coulomb gauge: $\Delta A^i = -4\pi\rho_e v^i$. $\boldsymbol{h}$ has been called the gravitomagnetic potential. The *gravitomagnetic field* [1, 28] has then been defined as $\boldsymbol{H} = \boldsymbol{\nabla} \times \boldsymbol{h}$. Furthermore, by the geodesic equation for a test particle of mass m in the weak-field slow-motion limit, one has then:

$$m\frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}t^2} \cong m\left(\boldsymbol{G} + \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \times \boldsymbol{H}\right).$$

This is formally analogous to the Lorentz force where $\boldsymbol{G} \cong -M/|\boldsymbol{x}|^2\hat{\boldsymbol{x}}$ is the standard Newtonian acceleration and $\boldsymbol{H}$ is the gravitomagnetic field (see [1]).

This is the weak-field slow-motion analogy of the gravitomagnetic field in geometrodynamics with the magnetic field of electrodynamics. However, a characterization of gravitomagnetism independent of any approximation has also been proposed [1, 28] and is described later.

One might then describe gravitomagnetism as all those phenomena generated by mass–energy currents and, acting as a source on the right-hand side of the Einstein field equation. However, in the presence of any mass, one can always observe a mass current in a boosted frame. Therefore, a more rigorous definition of gravitomagnetism has been proposed.

This characterization of gravitomagnetism is independent of the frame and the coordinate system used and is based on spacetime curvature invariants built using the Riemann curvature tensor (see later). It is also independent of any approximation.

In electromagnetism, in the frame in which an electric charge is at rest, we only have a non-zero electric field but no magnetic field. However, if we consider an observer moving relative to the charge, in this new frame we have a magnetic field. Similarly, in general relativity, in the frame where a mass is at rest, the gravitomagnetic potential $\boldsymbol{h}$ is zero. However, if we consider an observer moving relative to the mass, in the local frame of the observer we have a non-zero gravitomagnetic potential $\boldsymbol{h}$.

Therefore, following the method for characterizing curvature singularities and classifying different spacetimes [30], one should inspect the invariants of the spacetime. However, in a vacuum, as a consequence of the Einstein field equation, the Ricci curvature scalar $R = R^{\alpha}_{\alpha}$ is identically equal to zero. Another scalar invariant is the Kretschmann invariant $R_{\alpha\beta\mu\nu}R^{\alpha\beta\mu\nu}$. However, in the case of a metric characterized by mass and angular momentum, such as the Kerr metric, the Kretschmann invariant is a function of M/r^3 and J/r^4, with the leading term $\sim M/r^3$, therefore, this invariant is different from zero in the presence of a mass M, whether or not there is any angular momentum.

Let us then again use the formal analogy between electromagnetism and weak-field general relativity [31]. In electromagnetism, to characterize the electromagnetic field, one can calculate the scalar invariant $-\frac{1}{2}F_{\alpha\beta}F^{\alpha\beta} = E^2 - B^2$, which is analogous to the Kretschmann invariant

$$R_{\alpha\beta\mu\nu}R^{\alpha\beta\mu\nu} \sim \left(\frac{M}{r^3}\right)^2 + C\left(\frac{J}{r^4}\right)^2$$

(see later). However, in electrodynamics, one can also construct the scalar pseudoinvariant $\frac{1}{4}F_{\alpha\beta}{}^{*}F^{\alpha\beta} = \boldsymbol{E}\cdot\boldsymbol{B}$ where * is the dual operation: ${}^{*}F^{\alpha\beta} = \frac{1}{2}\varepsilon^{\alpha\beta\mu\nu}F_{\mu\nu}$. We observe that if we have only a charge q, in its rest frame we have only an electric field, and the invariant $F_{\alpha\beta}{}^{*}F^{\alpha\beta}$ is zero, therefore, even in the frames where $\boldsymbol{B} \neq \boldsymbol{0}$ and $\boldsymbol{E} \neq \boldsymbol{0}$, this invariant will be zero. However, if in the rest frame, we have a charge q and a magnetic dipole $\boldsymbol{m}$, in this frame we have, in general, $F_{\alpha\beta}{}^{*}F^{\alpha\beta} \neq 0$ and this invariant will, of course, be different from zero in any other frame.

Therefore, to characterize the spacetime geometry and curvature generated by the mass–energy currents or by the intrinsic angular momentum, J, of a central body (in [1] it is shown that, in the weak-field limit, the angular momentum generated by the mass–energy currents plays a role in general relativity analogous to the magnetic dipole moment generated by a loop of charge current in electromagnetism), we should look for an analogous spacetime invariant.

This invariant should, therefore, be built out of the dual of the Riemann tensor ${}^{*}R^{\alpha\beta\mu\nu} \equiv \frac{1}{2}\varepsilon^{\alpha\beta\sigma\rho}R^{\mu\nu}_{\sigma\rho}$, 'squared' or 'multiplied' by $R_{\alpha\beta\mu\nu}$. This pseudo invariant is of the type [1, 28] $\frac{1}{2}\varepsilon^{\alpha\beta\sigma\rho}R^{\mu\nu}_{\sigma\rho}R_{\alpha\beta\mu\nu}$. Because of the formal analogy with electromagnetism and since this pseudo invariant, ${}^{*}\boldsymbol{R}\cdot\boldsymbol{R}$, is built using the Levi Civita pseudo-tensor, it should change sign for time reflections ($t \to -t$)

and, therefore, it should be proportional to J. A list of all the possible spacetime invariants built out of the Riemann tensor and its dual is given in [30].

The invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ in general relativity is especially meaningful. In fact, whereas in classical electrodynamics ${}^*\boldsymbol{F}\cdot\boldsymbol{F}$ characterizes the electromagnetic field, but not the spacetime geometry $\eta_{\alpha\beta}$, in geometrodynamics the invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ characterizes the gravitational field and, therefore, the spacetime geometry.

Indeed, by calculating ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$, the result is (for simplicity we just give here the weak-field lowest-order, result; for the general, exact, expressions of $\boldsymbol{R}\cdot\boldsymbol{R}$ and ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$, and related discussions see [1, 28, 29]):

$$ {}^*\boldsymbol{R}\cdot\boldsymbol{R} \simeq 288\frac{JM}{r^7}\cos\theta + \cdots \tag{3.1} $$

whereas the Kretschmann invariant $\boldsymbol{R}\cdot\boldsymbol{R}$, for the Kerr metric, is in the weak-field limit,

$$ \boldsymbol{R}\cdot\boldsymbol{R} \simeq 48\left(\frac{M^2}{r^6} - 21\frac{J^2}{r^8}\cos^2\theta\right) + \cdots. \tag{3.2} $$

Since the external gravitational field of a stationary black hole is determined by its mass M, charge Q and intrinsic angular momentum J and since, for the Kerr–Newman metric, the invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ is [29] still proportional to J, the previous result is quite general in the case of black holes and is valid, asymptotically, in the weak-field limit for any quasistationary solution. Furthermore, the previous result, which was obtained in Einstein theory, is generally valid in any metric theory of gravity (with no prior geometry) *not* necessarily described at the post-Newtonian order by the Parametrized Post-Newtonian (PPN) formalism [1, 28, 29]. This can be seen in two ways. Let us first write the weak-field, slow-motion expression of an asymptotically flat metric of a metric theory of gravity in the form:

$$ g_{00} \simeq -1 + 2U + \text{higher-order terms} $$
$$ g_{ik} \simeq \delta_{ik}(1 + 2U) + \text{higher-order terms} $$

and

$$ \boldsymbol{g} \equiv (g_{01}, g_{02}, g_{03}). $$

Then, the pseudo invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ can be easily calculated at the lowest order to be

$$ {}^*\boldsymbol{R}\cdot\boldsymbol{R} \simeq \nabla^2[\nabla U\cdot(\nabla\times\boldsymbol{g})] + \text{higher-order terms}. $$

In the case of a static distribution of matter and a corresponding static metric, with $\boldsymbol{g} = \boldsymbol{0}$, we then have ${}^*\boldsymbol{R}\cdot\boldsymbol{R} = 0$. However, for a stationary distribution of matter and a corresponding stationary metric, with $\boldsymbol{g} \neq \boldsymbol{0}$, for example with $\boldsymbol{g} \sim J$, we have ${}^*\boldsymbol{R}\cdot\boldsymbol{R} \neq 0$. In the first case, for a static distribution of matter and a static metric, with a boost with velocity $\boldsymbol{v}$, we have $\boldsymbol{g} \sim \boldsymbol{v}U$; however, ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ is, of course, still zero: ${}^*\boldsymbol{R}\cdot\boldsymbol{R} \sim \nabla^2[\nabla U\cdot(\nabla\times\boldsymbol{v}U)] = 0$. A second argument confirms

the validity of this result in any metric theory of gravity not necessarily described by the PPN formalism. For a generic source, in any metric theory of gravity (with no prior geometry), the full expression for the scalar ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ must be dependent on some of the intrinsic physical quantities characterizing the source, such as the total mass–energy of the source, its intrinsic angular momentum, its multipole mass moments etc. i.e. it must be dependent on some integral of the mass–energy density, ε, of the mass–energy currents, εu^i, etc. In particular, since ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ must change sign for time reflections, its full expression for a generic source must be proportional to some odd function of the intrinsic mass–energy currents $\varepsilon u^i, \ldots$ (which cannot be eliminated by a change of origin or a Lorentz transformation), characterizing the system, such as the intrinsic angular momentum of the source.

Therefore, independently from the field equations of a particular metric theory, the pseudo invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ may be used to determine the existence and presence of 'intrinsic' gravitomagnetism in that metric theory of gravity. Indeed, using this invariant ${}^*\boldsymbol{R}\cdot\boldsymbol{R}\sim(JM/r^7)\cos\theta$, we can determine whether or not there is a gravitomagnetic contribution to the spacetime geometry and curvature. We just need to calculate ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$; if it is different from zero, we have a mass–energy current contribution to the spacetime curvature; if it is zero, there is no mass–energy current contribution. We do not need to concern ourselves with the local Lorentz transformation or any other frame and coordinate transformations on a static background, either ${}^*\boldsymbol{R}\cdot\boldsymbol{R}$ is zero, as in the Schwarzschild case, or it is different from zero, as in the Kerr case. Of course, a spacetime with ${}^*\boldsymbol{R}\cdot\boldsymbol{R}\neq 0$ is qualitatively different from a spacetime with ${}^*\boldsymbol{R}\cdot\boldsymbol{R}=0$, whatever the frame and coordinate transformations.

In conclusion, we may say that gravitomagnetism [1, 28, 29] is that phenomenon in which the spacetime structure and curvature are determined and affected not only by mass–energy but also by mass–energy currents relative to other matter, i.e. mass–energy currents not generable or eliminable with a Lorentz transformation (for example the intrinsic angular momentum of a body that cannot be generated or eliminated by a Lorentz transformation). This characterization of gravitomagnetism is independent of the frame and coordinate system used and is only based on spacetime curvature invariants.

3.3 Gravitomagnetic phenomena in test gyroscopes, test particles, clocks and photons

Einstein's theory of general relativity [1, 2] predicts the occurrence of peculiar phenomena in the vicinity of a spinning body, caused by its rotation. The period of a particle orbiting around a spinning body in the same direction as the rotation of the body, i.e. 'co-rotating' with the central object, is longer than the period of a particle orbiting at the same distance but in the opposite direction i.e. 'counter-rotating' with respect to the spin of the central object. The difference between the

co-rotating and counter-rotating orbital periods is

$$\Delta\tau = 4\pi\frac{J}{M}. \tag{3.3}$$

Furthermore, a particle orbiting around a spinning body has its orbital plane 'dragged' around the spinning body in the same sense as the rotation of the body. Small gyroscopes that determine the axes of a local free-falling inertial frame, where 'locally' means that the gravitational field is 'unobservable', rotate with respect to 'distant stars' due to the rotation of the body.

Thus, an external current of mass, such as the spinning Earth, 'drags' and changes the orientation of gyroscopes. Indeed, a test gyroscope has a precession $\dot{\boldsymbol{\Omega}}$ with respect to 'an asymptotic inertial frame', in a weak field with angular velocity:

$$\dot{\boldsymbol{\Omega}} = -\frac{1}{2}\boldsymbol{H} = \frac{[-\boldsymbol{J} + 3(\boldsymbol{J}\cdot\hat{\boldsymbol{x}})\hat{\boldsymbol{x}}]}{|\boldsymbol{x}|^3} \tag{3.4}$$

where $\boldsymbol{J}$ is the angular momentum of the central object and $\boldsymbol{H}$ its gravitomagnetic field generated by $\boldsymbol{J}$ (see section 3.2). This is the 'rotational dragging of inertial frames' or 'frame-dragging' ('dragging' as Einstein called it).

The whole orbital plane of a test particle is itself a type of enormous gyroscope (for motion under a central force) dragged by the gravitomagnetic field. Indeed, the orbit of a test particle around a central body with angular momentum $\boldsymbol{J}$ has a secular rate of change in the longitude of the line of the nodes (intersection between the orbital plane of the test particle and the equatorial plane of the central object), discovered by Lense–Thirring (1918) [7], in a weak field given by:

$$\dot{\boldsymbol{\Omega}}^{\text{Lense–Thirring}} = \frac{2\boldsymbol{J}}{[a^3(1-e^2)^{3/2}]} \tag{3.5}$$

where a is the semimajor axis of the test particle and e its orbital eccentricity. The pericentre of an orbiting test particle is also a type of enormous gyroscope (for motion under a central force $\sim 1/r^2$). Indeed, the orbit of a test particle has a secular rate of change in the mean longitude in the orbit L_0 (i.e. $L_0 = n\cdot\Delta t + \tilde{\omega}$, where $n = 2\pi/P$ is the satellite's mean motion, P its orbital period, Δt the interval of time from passage of the satellite through the pericentre, $\tilde{\omega} = \Omega + \omega$ the longitude of the pericentre and ω is the argument of the pericentre, i.e. the angle from the equatorial plane to the pericentre) and of the longitude of the pericentre $\dot{\tilde{\boldsymbol{\omega}}}$, (defining the Runge–Lenz vector):

$$\dot{\tilde{\boldsymbol{\omega}}}^{\text{Lense–Thirring}} = \frac{2J(\hat{\boldsymbol{J}} - 3\cos I\hat{\boldsymbol{l}})}{[a^3(1-e^2)^{3/2}]} \tag{3.6}$$

where $\hat{\boldsymbol{l}}$ is the orbital angular momentum, a unit vector, of the test particle, and I its orbital inclination (angle between the orbital plane and the equatorial plane of the central object).

Between 1995 and 2001, the Lense–Thirring effect was measured with about 30% accuracy using the LAGEOS and LAGEOS II satellites [21, 22, 24–28], see section 3.5.2.

However, test particles and gyroscopes are not the only objects affected by the spin of the central object: photons and clocks are also affected. A photon co-rotating around a spinning body takes less time to return to a 'fixed point' (with respect to distant stars) than a photon rotating in the opposite direction. In the Kerr metric [1] characterized by the mass and angular momentum of the central object, a fixed point can be determined by constant Boyer–Lindquist spatial coordinates, i.e. by the constant spatial coordinates of the weak-field slow-motion metric (see (3.7)). Operationally a fixed point can be realized by a small telescope always pointing toward the same distant star, always oriented toward the centre of the spinning body and at the same distance from it by using gradiometers and rockets attached to the telescope. For example, around the spinning Earth, the difference between the travel time of two pulses of electromagnetic radiation counter-propagating in the same circuit would be

$$\oint \frac{g_{0i}}{g_{00}}\,\mathrm{d}x^i \sim \frac{8\pi J_\oplus}{r}$$

or $\sim 10^{-16}$ [1, 32]. Since light rays are used to synchronize clocks, the difference in the travel time of co-rotating and counter-rotating photons implies the impossibility of synchronizing clocks all around a closed path around a spinning body. The behaviour of light rays, analysed in this chapter, and the behaviour of clocks around a spinning body are intimately connected. Let us then briefly analyse the behaviour of clocks around a spinning object. In several papers, the 'frame-dragging clock effect' around a spinning body has been estimated and space experiments have been proposed to test it [32–36]. We observe first that to synchronize clocks around a path in a stationary field, we can use light rays or even very slowly moving clocks, so that the special relativistic time dilation is a higher-order effect, always at the same distance from the central spinning body, so that the mass time dilation is equal for both clocks. Thus, when a clock co-rotating very slowly (using rockets) around a spinning body and at a constant distance from it returns to its starting point, it finds itself advanced relative to a clock kept there at 'rest' (with respect to 'distant stars', see earlier). Similarly a clock, counter-rotating arbitrarily slowly and at a constant distance around the spinning body finds itself retarded relative to the clock at rest at its starting point [1, 32]. For example, when a clock that co-rotates very slowly around the spinning Earth, at $\sim$ 6000 km altitude, returns to its starting point, it finds itself advanced relative to a clock kept there at 'rest' (with respect to 'distant stars') by

$$\oint \frac{g_{0i}}{g_{00}}\,\mathrm{d}x^i \sim \frac{4\pi J_\oplus}{r} \sim 5 \times 10^{-17}\ \mathrm{s}$$

where $g_{0i} \sim 2J_\oplus/r^2$ is the Earth's gravitomagnetic field and $J_\oplus \cong 145\ \mathrm{cm}^2$ is the Earth's angular momentum. Similarly, a clock that counter-rotates very slowly

around the spinning Earth finds itself retarded relative to a clock kept at 'rest' there by the same amount. However, a larger clock effect, of the order of 10^{-7}, has been estimated in [33, 34]. Let us explain this apparent disagreement. The orbital period of a particle or clock freely co-orbiting (along a geodesic) around a spinning body is longer than the orbital period of a particle or clock freely counter-orbiting on the same path [33, 34], see formula (3.7). The difference between the two orbital periods, i.e. the difference between the two times read by a clock at a fixed point (with respect to 'distant stars', see earlier) when the two counter-rotating particles come back to this point after one revolution is $\sim 4\pi J/M$, i.e. around the spinning Earth, is $\sim 1.4 \times 10^{-7}$ s [33–36]: this is basically the effect derived in [33, 34]. Nevertheless, the difference between the time read by the two clocks when they meet again after a whole revolution is still $\sim 10^{-16}$ [32, 35, 36].

In Einstein's theory of general relativity, all these phenomena in test particles, gyroscopes, photons and clocks are the result of the rotation of the central mass.

3.4 Time delay due to the spin of a central body and inside a rotating shell

3.4.1 Spin time delay and gravitational lensing

Let us now study null geodesics around a rotating body; in particular, we apply our results to the behaviour of photons. Null geodesics in the Kerr metric, also in regard to gravitational lensing and the image's position, polarization and intensification, distortion and optical caustic, have been studied in several papers (see [37–39] and references therein; for gravitational lensing in a strong Schwarzschild field see [40]). However, here we derive the time delay in the arrival time of photons due to the angular momentum of the deflecting body. Using the weak-gravitational-field slow-motion approximation, we also derive and compare the light deflections caused by the angular momentum and quadrupole moment of the deflecting body.

By assuming a weak gravitational field and slow motion for the source, we can write the spacetime metric at the order beyond Newtonian theory, the post-Newtonian order, in terms of small classical potentials determined by the distribution and motion of the mass–energy via the solution of Poisson-like equations, obtained from the weak-field slow-motion limit of the Einstein equation. If the source is stationary, with mass density ρ and mass–current density ρv^i, in order to study null geodesics, we thus have the following metric [1, 47]:

$$\begin{aligned} g_{00} &= -1 + 2U \\ g_{0i} &= -4V_i \\ g_{ik} &= (1 + 2U)\delta_{ik} \end{aligned} \tag{3.7}$$

where δ_{ik} is the standard Kronecker delta and, by the Einstein equation, in the weak-field slow-motion approximation the classical potentials U and V_i satisfy

the Poisson equations: $\Delta U = -4\pi\rho$ and $\Delta V_i = -4\pi\rho v_i$. V_i, or $-4V_i \equiv g_{0i}$, is the gravitomagnetic potential. As usual, the Newtonian potential U for a central distribution of the mass can be written as a multipole expansion; i.e. if we only include the monopole, M, and the quadrupole contributions

$$U = \frac{M}{r} + \frac{1}{2} Q_{ij} \frac{x_i x_j}{r^5}$$

where $Q_{ij} = \int (3x'_i x'_j - r'^2 \delta_{ij}) \rho(\boldsymbol{x}') d^3x'$ is the standard quadrupole moment tensor.

For an equilibrium ellipsoid [41, 42], by assuming the outer surface to be equipotential, in spherical coordinates we have

$$U = \frac{M}{r}\left[1 - J_2 \left(\frac{R}{r}\right)^2 P_{20}(\cos\theta)\right]$$

where R is the equatorial radius of the ellipsoid and $P_{20} = \frac{1}{2}(3\cos^2\theta - 1)$ is the associated Legendre function. At the lowest order in the flattening $f \equiv (R - R_{\rm p})/R$, where $R_{\rm p}$ is the polar radius of the ellipsoid, the quadrupole coefficient, J_2, is $J_2 = \frac{2}{3} f + O(f^2)$. If $\rho =$ constant, the quadrupole coefficient is $J_2 = \frac{2}{5} f + O(f^2)$.

The gravitomagnetic potential g_{0i}, i.e. the non-diagonal part of the metric tensor, from $\Delta V_i = -4\pi\rho v_i$, can be written as

$$g_{0i}(\boldsymbol{X}) \cong -4 \int \frac{\rho(\boldsymbol{x}') v^i(\boldsymbol{x}')}{|\boldsymbol{X} - \boldsymbol{x}'|} \, \mathrm{d}^3x' \tag{3.8}$$

where $\boldsymbol{X}$ is the position vector. Far from a stationary source for a spheroidal rotating body with angular momentum $\boldsymbol{J}$, we then have

$$\boldsymbol{g}(\boldsymbol{X}) \cong \frac{-2\boldsymbol{J} \times \boldsymbol{X}}{|\boldsymbol{X}|^3} \tag{3.9}$$

and, when $\boldsymbol{J} = (0, 0, J)$, in spherical coordinates:

$$g_{0\phi} \cong -\frac{2J}{r} \sin^2\theta. \tag{3.10}$$

Let us derive the time delay and deflection in the electromagnetic waves due to the spin and quadrupole moment of the central body. The quasi-Cartesian coordinate system (x, y, z) is the standard isotropic PPN system such that the coordinate z goes through the observer at Earth, while (X, Y, Z) is a coordinate system attached to the deflecting body. To relate the coordinate systems (x, y, z) and (X, Y, Z), we use the Euler angles (ϕ, β, γ) (see figure 3.1). For simplicity, we assume the deflecting body to be axially symmetric and Z, by definition, is the symmetry axis of the body. In such a case, the shape of the body is invariant

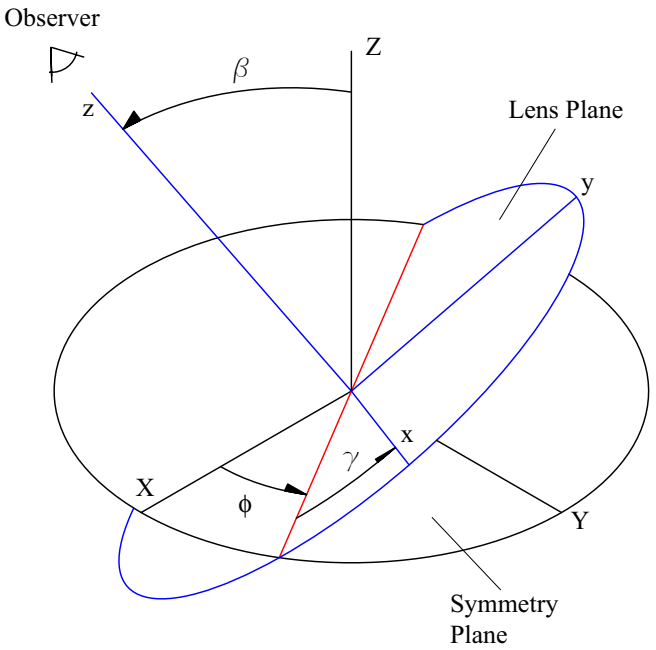

Figure 3.1. Euler's angles in our notation. β is the angle between the Z-axis of the body and the z-axis through the observer. γ is the angle between the line of nodes and the x-axis on the lens plane, and ϕ is the angle between the line of the nodes and the X-axis of the body. The origin of the coordinate systems is placed at the deflecting body.

for rotations of ϕ and we can thus choose $\phi = 0$. The rotation from (X, Y, Z) to (x, y, z) is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \cos\beta\sin\gamma & \cos\beta\cos\gamma & -\sin\beta \\ \sin\beta\sin\gamma & \sin\beta\cos\gamma & \cos\beta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{3.11}$$

We apply transformation (3.11) to the post-Newtonian metric (3.7) and, in the new coordinate system (x, y, z), we get

$$\begin{aligned} \mathrm{d}s^2 = {} & \overline{g}_{00}\,\mathrm{d}t^2 + \overline{g}_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j \\ & + \frac{4J}{r^3}(y\cos\beta - z\cos\gamma\sin\beta)\,\mathrm{d}x\,\mathrm{d}t \\ & - \frac{4J}{r^3}(x\cos\beta - z\sin\beta\sin\gamma)\,\mathrm{d}y\,\mathrm{d}t \\ & + \frac{4J}{r^3}(x\cos\gamma\sin\beta - y\sin\beta\sin\gamma)\,\mathrm{d}z\,\mathrm{d}t \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} \bar{g}_{00} &= (-1 + 2\bar{U}) \\ \bar{g}_{ij} &= (1 + 2\bar{U})\delta_{ij}, \end{aligned} \tag{3.13}$$

and

$$\bar{U} = \frac{M}{r}\left[1 - J_2\left(\frac{R}{r}\right)^2 P_{20}\left(\frac{z\cos\beta + y\cos\gamma\,\sin\beta + x\sin\beta\,\sin\gamma}{r}\right)\right]. \tag{3.14}$$

We can now easily derive the time delay using this metric element (3.12). In general, in a strong gravitational field, the time delays due to the gravitomagnetic field and to the non-sphericity of the matter distribution are nonlinearly coupled. However, in the weak-field slow-motion limit, i.e. $J/Mr \ll 1$ and $M/r \ll 1$, at the post-Newtonian order, we can analyse the two effects separately i.e. the gravitomagnetic and quadrupole moment time delays.

We have chosen the quasi-Cartesian coordinates such that the emitting and deflecting bodies have the same x and y coordinates (see figure 3.2) but a different z coordinate; i.e. the source, lens and observer are aligned. We have chosen this particularly simple configuration since here we are only interested in studying the time delay due to the gravitational field of the deflecting body (mass, quadrupole moment and gravitomagnetic time delay). However, there is an additional time delay, called the geometric time delay [43, page 143], due to the different geometrical path followed by different rays. Depending on the geometry of the system, this additional term can be very large and can be the main source of the time delay. However, when we compare the time delay of photons that follow the same geometrical path, we can neglect the geometric time delay, as in the case of two light rays with the same impact parameter but on different sides of the deflecting object. For of a small deflection angle with respect to the coordinate line $y = b\sin\alpha = \text{constant}$ and $x = b\cos\alpha = \text{constant}$ (see figure 3.2), the contribution to the travel time delay from the different path length due to the small deflection is of the order of $\sim \bar{U}^2$ [1] (to a small deflection angle of a photon path of the order of $\delta\phi \simeq 4M/r$ corresponds a change in the total distance l travelled by the photon of the order of $\delta l \simeq l(4M/r)^2$ and, depending on the geometrical configuration considered, this delay may need to be included in the total time delay. In a following paper, we shall analyse the higher-order time delays and compare then with the gravitomagnetic time delay. Here, for simplicity, we neglect any geometrical time delay. Now, as the speed of light equal to c in a local inertial frame: $\eta_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta = 0$, we have, in a general coordinate system, $\mathrm{d}s^2 = g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta = 0$. From this well-known condition of null arc length along the world line of photons, we then have

$$g_{00}\,\mathrm{d}t^2 \cong -g_{33}\,\mathrm{d}z^2 \tag{3.15}$$

or

$$\frac{\mathrm{d}z}{\mathrm{d}t} \cong \pm\sqrt{-\frac{g_{00}}{g_{33}}} \tag{3.16}$$

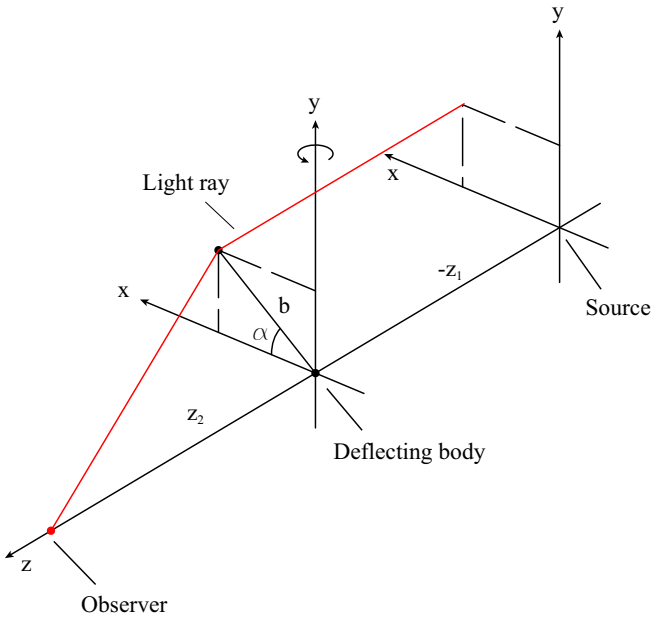

Figure 3.2. The observer, source and deflecting body have the same x and y coordinates: the source is at $z = -z_1$ ($z_1 > 0$), the observer at $z = z_2$ ($z_2 > 0$) and the deflecting body at the origin of the coordinate system. b is the impact parameter and α is the angular position of a light ray on the lens plane. The source is very far from the deflecting body, so that we assume all the light rays from the source parallel to the optical axis.

thus, from (3.16), we get, to first order in $\bar{U}$,

$$\mathrm{d}t = \sqrt{\frac{1+2\bar{U}}{1-2\bar{U}}}\,\mathrm{d}z \cong (1+2\bar{U})\,\mathrm{d}z. \tag{3.17}$$

Integrating this expression from $z = -z_1$ to $z = z_2$ corresponding, respectively, to the position of the source and observer, if $\overline{z} = z_1 \simeq z_2$ is much larger than the impact parameter b, we finally get

$$\Delta t_{J_2} = 2\bar{z} + 4M \ln\left(2\frac{\bar{z}}{b}\right) + \frac{2MR^2 J_2 \cos 2(\alpha+\gamma)\sin\beta^2}{b^2}. \tag{3.18}$$

In this expression the first term is the time taken for a radio pulse to travel from the source to Earth in the absence of a central mass: $M = 0$. The second term is the Shapiro time delay and the third one is the additional delay due to the quadrupole moment, J_2, of the deflecting body.

Similarly, in regard to the time delay due to the angular moment J of the deflecting body, from the condition $\mathrm{d}s^2 = 0$ and from $x =$ constant, $y =$ constant, solving (3.12) with respect to $\mathrm{d}t$, we have

$$\mathrm{d}t = -\frac{\bar{g}_{0z}}{\bar{g}_{00}}\,\mathrm{d}z \pm \frac{\sqrt{\bar{g}_{0z}^2\,\mathrm{d}z^2 - \bar{g}_{00}\bar{g}_{zz}\,\mathrm{d}z^2}}{\bar{g}_{00}} \tag{3.19}$$

where $\bar{g}_{0z} = (4J/r^3)(x\cos\gamma\sin\beta - y\sin\beta\sin\gamma)$. In the weak-field slow-motion approximation at the lowest order in $\bar{U}$ and $\bar{g}_{0z}$ expression (3.19), becomes

$$\mathrm{d}t \simeq \bar{g}_{0z}\,\mathrm{d}z \pm (1 + 2\bar{U})\,\mathrm{d}z. \tag{3.20}$$

The first term in (3.20) is the gravitomagnetic time delay, while the other terms have already been evaluated in the previous case (3.18): they are just the time it takes for a photon to travel from the source to the observer and the Shapiro time delay by a mass. Thus, let us integrate the first term of (3.20) from $z = -z_1$ to $z = z_2$ and assume that $\overline{z} = z_1 \simeq z_2 \gg b$. By setting, as before, $y =$ constant and $x =$ constant, we get the gravitomagnetic time delay:

$$\begin{aligned}\Delta t_J &= \lim_{\bar{z}b\to\infty}\int_{-z_1}^{z_2}\frac{4J}{r^3}(x\cos\gamma\sin\beta - y\sin\beta\sin\gamma)\,\mathrm{d}z \\ &= \frac{4J\cos(\alpha+\gamma)\sin\beta}{b}.\end{aligned} \tag{3.21}$$

From (3.18) and (3.21), we see that Δt_{J_2} (the time delay due to the quadrupole moment) is of order $\sim 1/b^2$ whereas Δt_J (the time delay due to the gravitomagnetic field) is of order $\sim 1/b$. This shows that there is a value of the impact parameter b such that $\Delta t_J > \Delta t_{J_2}$ and, if the angular momentum of the deflecting body is large enough, the 'spin time delay' may be a relevant effect.

To derive the deflection of electromagnetic waves due to the spin and quadrupole moment of the deflecting body, we use the geodesic equation in the weak-field approximation, we then have the deflection angles due to the quadrupole moment J_2 and the deflection angles due to the angular momentum J [32]:

$$\begin{aligned}\delta_x^{J_2} &= -\frac{4M\cos\alpha}{b} - \frac{4J_2MR^2\sin^2\beta\cos(3\alpha+2\gamma)}{b^3} \\ \delta_y^{J_2} &= -\frac{4M\sin\alpha}{b} - \frac{4J_2MR^2\sin^2\beta\sin(3\alpha+2\gamma)}{b^3}\end{aligned} \tag{3.22}$$

$$\begin{aligned}\delta_x^{J} &= -\frac{4J\sin\beta\cos(2\alpha+\gamma)}{b^2} \\ \delta_y^{J} &= -\frac{4J\sin\beta\sin(2\alpha+\gamma)}{b^2}.\end{aligned} \tag{3.23}$$

The first term in (3.22) is the standard deflection by a spherical object of mass M, whereas the second term is the additional deflection due to the quadrupole moment, J_2, of the central body.

Let us now study the possibility of determining the angular momentum J of the central deflecting body from time delay and deflection angle of different images of the source.

Let us consider three light rays emitted from a very far source, propagating parallel to the z-axis and with the same impact parameter b and let us assume that we are able to measure, determine or obtain [43] the following quantities: total time delay between the three rays, Δt_{12} and Δt_{13}; deflection angles δ_1, δ_2, δ_3; and the equatorial radius R of the deflecting body and distances from source and lens to the observer. In this way, we are able to determine the angle α for each light beam and the impact parameter b and we can write a system in which the only unknown quantities are: angular momentum, J, quadrupole moment, J_2, mass, M and Euler's angle β and γ. Solving this system we can, in principle, determine the time delay due to the angular moment J and the other unknown quantities [32].

We have chosen a special case in which $x = y = 0$ for the source, lens and observer and all the light rays have the same impact parameter. In this way, as we have already remarked, we do not need to consider the relative time delays in the arrival time due to the different geometry of the paths travelled by the photons [43] and due to the difference in the Shapiro time delays by the central mass: these delays are, in general, much larger than the spin and quadrupole moment time delays. Indeed, for other configurations in which the source is not exactly aligned with the lens and the observer, these effects—the different geometry of the path travelled and the difference in the standard Shapiro time delay—can be the main source of relative time delay. In these cases, we would then need to model and remove these delays between the different images on the basis of the geometry of the system [32]. In special cases, for example if we were to observe four images of the source and if the angle α of each deflected ray were to differ by π—the Einstein Cross has a configuration very similar to this—we could, at least in principle, eliminate the time delay due to the quadrupole moment (see [32]) and, thus, determine the spin time delay.

3.4.2 Some astrophysical sources and spin time delay

Let us now calculate the time delay due to the spin of some astrophysical sources: the Sun, the lensing galaxy of the Einstein Cross, Q2237 + 031; and a typical cluster of galaxies.

The Sun parameters are: $M_\odot = 1.477$ km, $R_\odot = 6.96 \times 10^5$ km, $J_{2\odot} \cong 1.7 \times 10^{-7}$ [1] and $a_\odot = J_\odot/M_\odot \cong 0.273$ km. Let us consider a co-rotating photon ($\frac{1}{2}\pi \geq \alpha \geq -\frac{1}{2}\pi$) and a counter-rotating one ($\frac{1}{2}\pi \leq \alpha \leq \frac{3}{2}\pi$), both coming from infinity and propagating near the Sun with an impact parameter b nearly equal to the Sun equatorial radius $b \cong R = 6.96 \times 10^5$ km. Let us

assume, for simplicity, that $\gamma = 0$ and $\beta = \frac{1}{2}\pi$: the maximum gravitomagnetic and quadrupole moment time delays, according to (3.18) and (3.21), are then, respectively,

$$\Delta t_{12}^{J} = \frac{8J}{b} = 1.54 \times 10^{-11}\ \mathrm{s} \tag{3.24}$$

$$\Delta t_{12}^{J_2} = 4\frac{J_2 M R^2}{b^2} = 3.35 \times 10^{-12}\ \mathrm{s}. \tag{3.25}$$

The time delay due to the Sun's spin could then, in principle, be detected using a laser interferometer around the Sun: this would consist of a source and a detector on opposite sides of the Sun, both at a distance of about 8×10^{10} km. The source, a laser, would, at the same time, emit photons toward the Sun but with slightly different angles so that they would travel on opposite sides of the Sun (i.e. photons co-rotating and counter-rotating with respect to the Sun's spin). Then, by gravitational lensing, they would be focused and observed by the detector on the opposite side. Thus, according to the previous calculation, there would be a relative time delay in the arrival time of the photons due to the Sun's spin. Of course all the other travel time delays should be modelled and removed from the observed delays, in particular the time delay due to the dispersion of electromagnetic waves by the solar plasma.

To derive the time delay due to the lensing galaxy of the Einstein Cross [44, 45], we assume a simple model for rotation and shape of the central object. Details about this model can be found in [46].

The angular separation between the four images is about $0.9''$, corresponding to a radius of closest approach of about $R \simeq 650 h_{75}^{-1}$ pc, and the mass inside a shell with $R \simeq 650 h_{75}^{-1}$ pc is $\sim 1.4 \times 10^{10} h_{75}^{-1} M_{\odot}$ [45]. Let us assume $J_2 \simeq 0.1$ and $J \simeq 10^{23}\ \mathrm{km}^2\ h_{75}^{-2}$, we then have

$$\Delta t_{12}^{J} = \frac{8J}{b} \simeq 4\ \mathrm{min} \tag{3.26}$$

$$\Delta t_{12}^{J_2} = 4\frac{J_2 M R^2}{b^2} \simeq 8\ \mathrm{h}. \tag{3.27}$$

Thus, at least in principle, one could measure the time delay due to the spin of the lensing galaxy by removing the large quadrupole moment time delay using the previously described method [32]: of course one should be able to model all other delays due to other physical effects accurately enough and remove them from the observed time delays between the four images.

As a third example, we consider the relative time delay of photons due to the spin of a typical cluster of galaxies: the precise calculations are shown elsewhere, nevertheless we give here some results. We consider a cluster of galaxies of mass $M_{\mathrm{C}} \cong 10^{14} M_{\odot}$, radius $R_{\mathrm{C}} \cong 5$ Mpc and angular velocity $\omega_{\mathrm{C}} \cong 10^{-18}\mathrm{s}^{-1}$. Depending on the geometry of the system and on the path followed by the photons, we then find relative time delays ranging from a few minutes to several days [32].

3.4.3 Spacetime geometry inside a rotating shell

Let us first assume that if a body or shell has a steady rotation, the spacetime is stationary (a spacetime is called stationary if it admits a time-like Killing vector field ξ_t^α, then there exists some coordinate system in which ξ_t^α can be written as $\xi_t^\alpha = (1, 0, 0, 0)$, from the Killing equation, in this coordinate system, the metric is thus time independent: $g_{\alpha\beta,0} = 0$). An external solution of this type is the stationary Kerr metric, characterized by the mass and angular momentum of the central body. The well-known Lense–Thirring metric [1,7] is, in Boyer–Lindquist coordinates, the weak-field slow-motion limit of the Kerr metric. The Brill–Cohen (1966) solution [4] describes the metric inside a shell with arbitrary mass and to lowest-order in the angular velocity. It is a lowest order series expansion in the angular velocity ω of the shell on the Scharzschild background of a spherical mass shell of arbitrary mass M, valid both inside and outside the shell. The Brill–Cohen metric is

$$\mathrm{d}s^2 = -H(r)\,\mathrm{d}t^2 + J(r)[\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta(\mathrm{d}\phi - \omega(r)\,\mathrm{d}t)^2] \qquad (3.28)$$

where

$$\begin{aligned} H(r) &= [(r-\alpha)/(r+\alpha)]^2 \qquad \text{for } r > R \\ H(r) &= [(R-\alpha)/(R+\alpha)]^2 \qquad \text{for } r \le R \\ J(r) &= (1+\alpha/r)^4 \qquad \text{for } r > R \\ J(r) &= (1+\alpha/R)^4 \qquad \text{for } r \le R \end{aligned}$$

and where $\alpha = 2M$, and R, ω, and M are, respectively, the radius, angular velocity and mass of the shell (for further details on this solution, see [6] and related papers in [48]).

However, in the following, we shall only consider Thirring's weak-field slow-motion solution for the metric inside a rotating shell and, for simplicity, we neglect the stresses of the rotating shell (see [4,48,49] for related discussions and references). Inside a hollow, static, spherically symmetric distribution of matter, in vacuum, we have the flat metric $\eta_{\alpha\beta}$ [1]. Thus, in the weak-field slow-motion limit, we assume that the metric inside a slowly rotating massive shell can be written as $g_{\alpha\beta} \cong \eta_{\alpha\beta} + h_{\alpha\beta}$, where $h_{00} = h_{ii} = 0$ and the $(0i)$ components of the Einstein field equation then satisfy in the Lorentz gauge:

$$\Delta h_{0i} \cong 16\pi\rho v^i \qquad (3.29)$$

with solution

$$h_{0i}(\boldsymbol{x}) \cong -4\int \frac{\rho(\boldsymbol{x}')v^i(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}\,\mathrm{d}^3x' \qquad (3.30)$$

where ρ is the shell's mass density and ρv^i is the mass–current density.

We can then apply this result to determine the metric inside a thin shell of total mass M and radius R, rotating with small angular velocity $\boldsymbol{\omega}$, by integrating

expression (3.30) inside the shell, where $\boldsymbol{v} = \boldsymbol{\omega} \times \boldsymbol{x}$. By a rotation of the spatial axes so that $\boldsymbol{\omega} = (0, 0, \omega)$ and by using the mass density of a thin spherical shell, $\rho(\boldsymbol{x}') = (M/4\pi R^2)\delta(R - r')$ (we neglect here the stresses of the rotating shell [4, 48, 49]), we have

$$\begin{aligned} \boldsymbol{h} &= -4 \int \frac{\rho(\boldsymbol{x}')(\boldsymbol{\omega} \times \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} r'^2 \, \mathrm{d}\Omega' \, \mathrm{d}r' \\ &= -\frac{M}{\pi} \boldsymbol{\omega} \times \int_0^{\pi} \sin\theta' \, \mathrm{d}\theta' \int_0^{2\pi} \frac{R\hat{\boldsymbol{x}}'}{|\boldsymbol{x} - R\hat{\boldsymbol{x}}'|} \, \mathrm{d}\varphi'. \end{aligned} \tag{3.31}$$

The integral over $\mathrm{d}\Omega \equiv \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$ if $r < R$ is

$$\int \frac{\hat{\boldsymbol{x}}'}{|\boldsymbol{x} - \hat{\boldsymbol{x}}'|} \, \mathrm{d}\Omega' = \frac{4\pi}{3R} \boldsymbol{x}.$$

Therefore, for any $\boldsymbol{x}$ inside the shell, we have

$$\begin{aligned} \boldsymbol{h} \equiv (h_{0x}, h_{0y}, h_{0z}) &= -\frac{4}{3} \frac{M}{R} \boldsymbol{\omega} \times \boldsymbol{x} \\ &= \left(\frac{4}{3} \frac{M}{R} \omega y, -\frac{4}{3} \frac{M}{R} \omega x, 0 \right). \end{aligned} \tag{3.32}$$

By substituting the components of $h_{\alpha\beta}$ inside the slowly rotating shell in the geodesic equation, we find that the acceleration of a test particle inside a rotating shell, due to the spin of the shell [1, 3, 48], is

$$\begin{aligned} \ddot{x} &= -\frac{8}{3} \frac{M}{R} \omega \dot{y} + \frac{4}{15} \frac{M}{R} \omega^2 x \\ \ddot{y} &= \frac{8}{3} \frac{M}{R} \omega \dot{x} + \frac{4}{15} \frac{M}{R} \omega^2 y \\ \ddot{z} &= -\frac{8}{15} \frac{M}{R} \omega^2 z \end{aligned} \tag{3.33}$$

where the ω^2 terms are due to the other components of $h_{\alpha\beta}$ and may be interpreted as a change in the inertial and gravitational mass-density of the shell due to the velocity ωR. Due to the rotation of the shell, the test particle is affected by forces *formally* similar to Coriolis and centrifugal forces. For discussions on the interpretation of the accelerations inside a rotating shell, we refer readers to [4, 48–50].

Finally, we have that the axes of the local inertial frames, i.e. the gyroscopes, are dragged by the rotating shell with constant angular velocity $\dot{\boldsymbol{\Omega}}^G$, according to [1, 3]:

$$\dot{\boldsymbol{\Omega}}^G \cong -\frac{1}{2} \boldsymbol{\nabla} \times \boldsymbol{h} = \frac{4}{3} \frac{M}{R} \boldsymbol{\omega}. \tag{3.34}$$

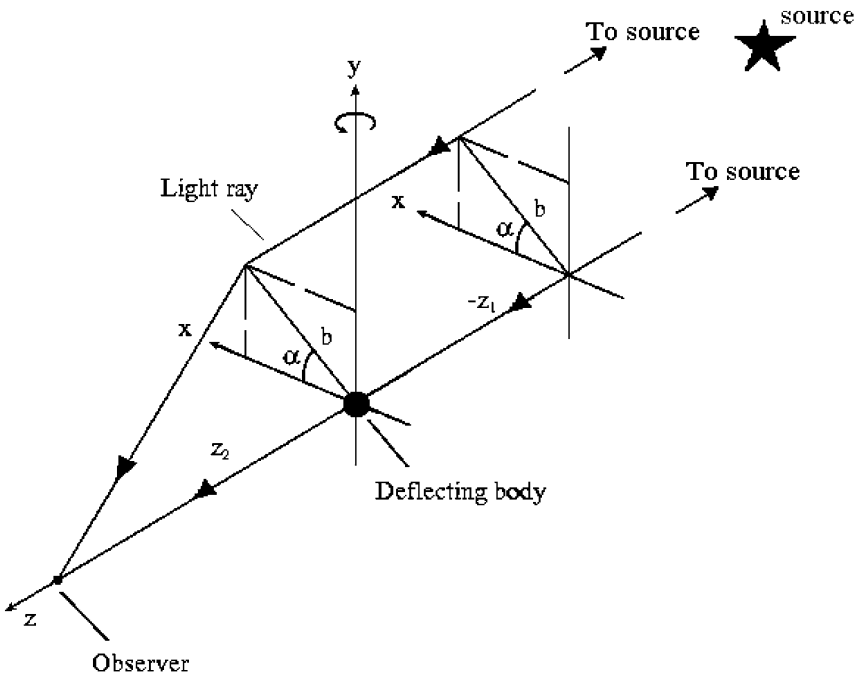

Figure 3.3. Geometry of light rays, with impact parameters b and $-b$, propagating inside a rotating shell of radius R.

3.4.4 Time delay inside a slowly rotating massive shell

In section 3.4.1 we have calculated the time delay and the deflection due to the spin of a body from the rotation of the mass inside a radius r [32]; however, there is also a spin time delay and an additional deflection due to the rotation of the external mass [51].

Inside a thin shell of mass M and radius R' rotating with slow angular velocity $\boldsymbol{\omega} = (0, 0, \omega)$, the $g_{0i} \cong h_{0i}$ components of the metric tensor are given by (3.32). Therefore, inside a rotating shell, it is not possible to synchronize clocks all around a closed path. Indeed, if we consider a clock co-rotating very slowly along a circular path with radius r ($r < R'$), when it comes back to its starting point it is advanced with respect to a clock kept there at rest (with respect to distant stars). The difference between the time read by the co-rotating clock and the clock at rest is given in [74]. Indicating with $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_2$ the unit vectors from the centre of a spherical shell to the two points on the shell where a light ray enters and leaves the sphere, respectively, we have the spin time delay due to the

rotation of the shell [74]:

$$\Delta_{GM} = -\frac{4MR_0'}{3}\boldsymbol{\omega}\cdot(\hat{\boldsymbol{r}}_1\times\hat{\boldsymbol{r}}_2). \tag{3.35}$$

A general expression for the relative spin time delay due the rotation of an external mass between two photons travelling inside the mass with (a) different impact parameters, b_1 and b_2 and (b) for any finite thickness of the external shell is given in [74]

3.4.5 Some astrophysical sources and the spin time delay due to an external rotating shell

Let us finally report the order of magnitude of the time delay corresponding to some astrophysical configurations.

For a lensing galaxy with a lens similar to that in the 'Einstein Cross' [52,53], the relative time delay of two photons travelling at a distance of $b_1 \simeq 650$ pc and $b_2 \simeq -650$ pc from the centre due to the rotation of the external mass is then [74] $\Delta t \simeq 30$ min.

As a further astrophysical example, we consider two light rays deflected by a lensing galaxy which is inside a rotating cluster or supercluster of galaxies. We then calculate the amount of time delay due to the spin of the mass rotating around the deflecting galaxy. To get an order of magnitude of the time delay, we use typical supercluster parameters [54]. By considering a galaxy at the centre of the cluster and light rays with impact parameters $b_1 \simeq 15$ kpc and $b_2 \simeq -15$ kpc (of the order of the Milky Way radius), the time delay, by assuming for simplicity a constant mass density, is [74]: $\Delta t \simeq 1$ day.

Finally, in [74] we show that if, in general, the lensing galaxy is not at the centre of the cluster, the relative spin time delay between two photons, deflected by the galaxy, that are propagating inside a rotating cluster or supercluster of galaxies may, under special conditions, be as large as several years.

3.4.6 Discussion and conclusion on spin time delay

We have derived and studied the '*spin time delay*' due to the angular momentum of a body experienced by the photons of two or more images of a source observed at a far point by gravitational lensing: this effect is due to the propagation of the photons in opposite directions with respect to the direction of the spin of the rotating body. We have analysed both the spin time delay caused by a central rotating mass and in the case of photons propagating inside a massive rotating shell with a time delay due to the rotation of the external mass.

We have also derived, in weak-field slow-motion approximation, the deflection in the path of the images due to the spin of the deflecting body. We have then compared the relative time delay of the photons due to spin with the relative time delay due to the quadrupole moment of the central body.

Finally, in order to estimate the relevance of the spin time delay in some real astrophysical configurations, we have considered some possible astrophysical cases; nevertheless, these estimates are preliminary because we need to use more accurate values for the angular momentum and the other parameters of the considered astrophysical configurations.

We can then summarize the following conclusions of our analyses:

(a) The *spin time delay* must be taken into account in the modelling of relative time delays between images observed by gravitational lensing, i.e. in addition to other time delays such as the geometrical time delay and the delay due to the quadrupole moment of the lensing body.
(b) If other smaller time delays could be modelled accurately enough and removed from the observations, we have shown that the large relative delay due to the quadrupole moment of the lensing body could be removed for some configurations of the images by using special combinations of observables. With this method, we could measure the spin time delay due to the rotation of a mass.
(c) The measurement of the spin time delay might, in principle, be a new observable for the determination of the total mass of a rotating body, i.e. of the dark matter content of such objects as galaxies, cluster and superclusters of galaxies [74].
(d) Depending on the geometry of the system, the relative spin time delay can be a quite large effect and may then be detected on Earth, in particular in systems with small angular separation and small relative time delay between the images such as B0218 + 357 [74].

3.5 Measurement of gravitomagnetism with laser-ranged satellites

In section 3.5.1 we describe a proposed measurement of the gravitomagnetism of the Earth and the Lense–Thirring effect, with a relative accuracy of the order of 1%, using the satellite LARES. LARES would also perform other basic general relativistic tests. In sections (3.5.2) and (3.5.3) we report the 1995–2004 measurements of the Lense–Thirring effect obtained by analysing the orbits of the laser-ranged satellites LAGEOS and LAGEOS II, confirming the general relativistic prediction of frame-dragging with an accuracy of about 20% and about 5% in the most recent (March 2004) analysis.

3.5.1 LARES (LAser RElativity Satellite)

The main scientific objectives of the LARES mission [23] are:

(1) To perform high-precision tests of Einstein's theory of general relativity and, in particular, the following ones:

(a) the accurate measurement of the Lense–Thirring effect due to the Earth's angular momentum and a high precision test of the Earth's gravitomagnetic field, with a relative accuracy of the order of 1%;
(b) a possible test, using the LARES perigee, of some recently proposed theories [76], based on a brane-world model, which can explain the dark-energy problem [77];
(c) improved high-precision bounds on the hypothetical very weak long-range gravitational forces, tests of the inverse square law for very weak-field gravity and an improved test of the equivalence principle corresponding to ranges of the order of few thousands km;
(d) a 10^{-6} measurement (improved by about two orders of magnitude) of the PPN (Parametrized-Post-Newtonian) parameter α_1 testing the existence of preferred frames in some alternative metric theories of gravitation;
(e) a 10^{-3} measurement of the general relativistic perigee precession of LARES and a high-precision measurement of the corresponding combination of the PPN parameters β and γ in the field of Earth. The PPN parameters β and γ test Einstein's theory of gravitation against other metric theories of gravitation; and
(f) other tests of general relativity and gravitation, such as improvements in the current limits on hypothetical spatial anisotropies of the gravitational interaction; and

(2) measurements and improved determinations in geodesy and geodynamics, in areas such as global plate tectonics, crustal deformation and variations in the Earth's rotation [22, 23].

The LARES experiment is an improved version of the LAGEOS III experiment [21]. The main differences between LARES and LAGEOS III lie in its weight and orbital eccentricity. The new LARES satellite has been designed to be about four times lighter than LAGEOS, with a total weight of about 100 kg, and to be smaller than LAGEOS, with a radius of about 16 cm. The orbital eccentricity of LARES has been proposed to be 0.04 ± 0.01, whereas the proposed orbit of LAGEOS III had an essentially zero eccentricity [23].

In [21, 22] and [23], it is shown how, by combining the measured nodal precessions of LAGEOS, $\dot{\Omega}_{\rm LAGEOS}$, and LARES, $\dot{\Omega}_{\rm LARES}$, we can get a very accurate measurement of the Lense–Thirring effect, $\dot{\Omega}^{\rm Lense–Thirring}$. The present 2004 error budget of the LARES experiment is dramatically reduced with respect to the previous estimates: present analyses show a *total statistical error in the LARES experiment of about 1%* $\dot{\Omega}^{\rm Lense–Thirring}$ *or less* over a 3 yr period. The main improvements for this substantial reduction in the total statistical error are described in [23, 55].

In addition to the high-precision measurement of the Lense–Thirring effect due to the Earth's angular momentum, the LARES experiment will provide other important general relativistic and gravitational measurements, described earlier.

In phase A of the LARES study, it is also shown how, by measuring the LARES perigee rate, we could improve our present tests of the equivalence principle by about two orders of magnitude [23]. These tests will be realized by the use of the pericentre of LARES [56]. Indeed, a very effective way of detecting a very weak Yukawa-like gravitational interaction is via a precise measurement of the pericentre precession. However, the precision of such measurement is inversely proportional to the orbital eccentricity, therefore the orbit of a new LAGEOS-type satellite with larger orbital eccentricity would be more effective in detecting a new, hypothetical, very weak gravitational force with a range of the order of two Earth radii. This would improve the present limits on such interaction by at least four orders of magnitude [57].

3.5.2 The previous 1995–2001 measurements of the Lense–Thirring effect using the node of LAGEOS and the node and perigee of LAGEOS II

3.5.2.1 Method

In section 3.3 we have seen that the whole orbital plane is dragged by the spin of the central object: this is the Lense–Thirring effect (3.5).

Let us now describe the 1995–2001 measurements of the gravitomagnetism of Earth and Lense–Thirring effect using laser-ranged satellites.

Our detection and measurement of the Lense–Thirring effect was obtained by using satellite laser-ranging data from LAGEOS (LAser GEOdynamics Satellite, NASA) and LAGEOS II (NASA and ASI, the Italian Space Agency) and the Earth gravitational field models, JGM-3 and EGM-96.

The measurement of distances has always been a fundamental issue in astronomy, engineering and science in general. So far, laser-ranging has been the most accurate technique for measuring the distances to the moon and artificial satellites [58, 59]. Short laser pulses are emitted from lasers on Earth, aimed at the target through a telescope and then reflected back by optical cube-corner retroreflectors on the moon or an artificial satellite [60], such as LAGEOS. By measuring the total round-trip travel time, one can determine the distance to a retroreflector on the moon with an accuracy of about 2 cm and to the LAGEOS satellites with an accuracy of a few millimetres.

The LAGEOS satellites are made of heavy brass and aluminum and are about 406 kg in weight. They are completely passive and covered with retroreflectors and orbit at an altitude of about 6000 km above the surface of the Earth. LAGEOS, launched in 1976 by NASA, and LAGEOS II, launched by NASA and ASI in 1992, have an essentially identical structure but they have different orbits. The semimajor axis of LAGEOS is $a \cong 12\,270$ km, the period $P \cong 3.758$ h, the eccentricity $e \cong 0.004$, and the inclination $I \cong 109.9°$. The semimajor axis of LAGEOS II is $a_{\rm II} \cong 12\,163$ km, the eccentricity $e_{\rm II} \cong 0.014$, and the inclination $I_{\rm II} \cong 52.65°$.

We analysed the laser-ranging data using the principles described in [61] and adopted the IERS conventions [62] in our modelling, except that, in the 1998 analysis, we used the static and tidal EGM-96 model [63]. Error analysis of the LAGEOS orbits indicated that the EGM-96 errors can only contribute periodic root-sum-square (rss) errors of 2–4 mm radially and, in all three directions, they do not exceed 10–17 mm. The initial positions and velocities of the LAGEOS satellites were adjusted for each 15-day batch of data, along with small variations in their reflectivities. The solar radiation pressure, Earth albedo and anisotropic thermal effects were also modelled [64–67]. In modelling the thermal effects, the orientation of the satellite spin axis was obtained from [68]. Lunar, solar and planetary perturbations were also included in the equations of motion, formulated according to Einstein's theory of general relativity with the exception of the Lense–Thirring effect which was purposely set to zero. All of the tracking-station coordinates were adjusted (accounting for tectonic motions) except for those defining the terrestrial reference frame. Adjustments were made for polar motion, and the Earth's rotation was modelled from the very long baseline interferometry-based series SPACE96 [69]. We analysed the orbits of the LAGEOS satellites using the orbital analysis and data reduction software GEODYN II [70].

The node and perigee of LAGEOS and LAGEOS II are dragged by the Earth's angular momentum. From the Lense–Thirring formula [21, 24], we get $\dot{\Omega}_I^{\text{Lense–Thirring}} \cong 31$ mas yr^{-1} and $\dot{\Omega}_{II}^{\text{Lense–Thirring}} \cong 31.5$. The argument of the pericentre (perigee in our analysis), ω, also has a Lense–Thirring drag [1]; thus, for LAGEOS we get $\dot{\omega}_I^{\text{Lense–Thirring}} \cong 32$ mas yr^{-1} and, for LAGEOS II, $\dot{\omega}_{II}^{\text{Lense–Thirring}} \cong -57$ mas yr^{-1} [24]. The nodal precessions of LAGEOS and LAGEOS II can be determined with an accuracy of the order of 1 mas yr^{-1}. Over our total observational period of about 4 yr, we obtained a RMS of the node residuals of about 4 mas for LAGEOS and about 7 mas for LAGEOS II [27]. For the perigee, the observable quantity is the product $ea\dot{\omega}$, where e is the orbital eccentricity of the satellite. Thus, the perigee precession $\dot{\omega}$ for LAGEOS is difficult to measure because its orbital eccentricity, e, is $\sim 4 \times 10^{-3}$. The orbit of LAGEOS II is more eccentric, with $e \sim 0.014$, and the Lense–Thirring drag of the perigee of LAGEOS II is almost twice as large in magnitude as that of LAGEOS. Over about 4 yr, we obtained a rms value for the residuals of the LAGEOS II perigee of about 25 mas [27], whereas the total Lense–Thirring effect on the perigee over 4 yr is $\cong -228$ mas.

To quantify and measure the gravitomagnetic effects precisely, we have introduced the parameter μ that is, by definition, one in general relativity [1] and zero in Newtonian theory.

The main error in this measurement arises from the uncertainties in the Earth's even zonal harmonics and their time variations. The unmodelled orbital effects due to lower-order harmonics are comparable to or larger than the Lense–Thirring effect. However, by analysing both the JGM-3 and EGM-96 models with their uncertainties in the even zonal harmonic coefficients and by calculating

the secular effects of these uncertainties on the orbital elements of LAGEOS and LAGEOS II, we find [24] that the main sources of error in the determination of the Lense–Thirring effect are concentrated in the first two even zonal harmonics, J_2 and J_4. We can, however, use the three observable quantities $\dot{\Omega}_{\rm I}$, $\dot{\Omega}_{\rm II}$ and $\dot{\omega}_{\rm II}$ to determine μ [24], thereby avoiding the two largest sources of error—those arising from the uncertainties in J_2 and J_4. We do this by solving the system of the three equations for $\delta\dot{\Omega}_{\rm I}$, $\delta\dot{\Omega}_{\rm II}$ and $\delta\dot{\omega}_{\rm II}$ in the three unknowns μ, J_2 and J_4, obtaining

$$\delta\dot{\Omega}^{\rm Exp}_{\rm LAGEOS\ I} + c_1\delta\dot{\Omega}^{\rm Exp}_{\rm LAGEOS\ II} + c_2\delta\dot{\omega}^{\rm Exp}_{\rm LAGEOS\ II} = \mu(31 + 31.5c_1 - 57c_2)\ \text{mas yr}^{-1} + \text{other errors} \cong \mu(60.2\ \text{mas yr}^{-1}) \tag{3.36}$$

where $c_1 = 0.295$ and $c_2 = -0.35$. Equation (3.36) for μ does not depend on J_2 and J_4 nor on their uncertainties; thus, the value of μ that we obtain is unaffected by the largest errors, which are due to δJ_2 and δJ_4, and is sensitive only to the smaller errors due to δJ_{2n}, with $2n \geq 6$.

Similarly, regarding tidal, secular and seasonal changes in the geopotential coefficients, the main effects on the nodes and perigee of LAGEOS and LAGEOS II, caused by tidal and other time variations in the Earth's gravitational field [22,26], are due to changes in J_2 and J_4. However, the tidal errors in J_2 and J_4 and the errors resulting from other unmodelled medium- and long-period variations in J_2 and J_4, including their secular and seasonal variations, are eliminated by our combination of the residuals of the nodes and perigee. In particular, most of the errors resulting from the 18.6- and 9.3-yr tides, associated with the lunar node, are eliminated in our measurement. An extensive discussion of the various error sources that can affect our result is given in [26], only a brief discussion of the error sources is given in the next section.

3.5.2.2 Results

In this section we report the 1995–2001 results of our measurements.

In figure 3.4, we show the linear combination of the residuals of the nodes of LAGEOS and LAGEOS II and perigee of LAGEOS II according to equation (3.36) to eliminate the δJ_2 and δJ_4 errors, using the Earth gravitational model JGM-3 over a 3.1-yr period and after removing ten small periodic residual signals and the small observed inclination residuals [26].

In figure 3.5, we display an improved analysis [27] (obtained with a linear combination of the residuals of the nodes of LAGEOS and LAGEOS II and perigee of LAGEOS II according to equation (3.36)) using the more recent static and tidal Earth gravitational model EGM-96. We have also refined the non-gravitational perturbations model: the total period of observations was 4 yr, longer by about 1 year than the observational period corresponding to figure 3.4. We have only removed four small periodic residual signals and the small observed

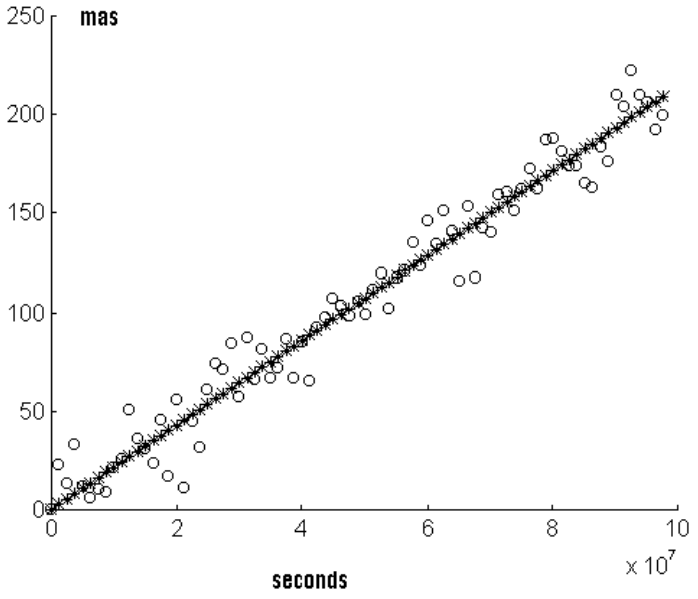

Figure 3.4. Combination of the residuals of the nodes of LAGEOS and LAGEOS II and perigee of LAGEOS II according to equation (3.36), using the Earth gravitational model JGM-3, over a 3.1 yr period. The best-fit line through these combined residuals has a slope of about $\mu_{\text{Measured}} \cong 1.1$.

inclination residuals. The removal of the periodic terms was achieved by a least-squares fit of the residuals using a secular trend and four periodic signals with periods of 1044-, 820-, 569- and 365.25-day, corresponding, respectively, to the nodal period of LAGEOS, the perigee and nodal periods of LAGEOS II, and 1 yr. The 820-day period is the period of the main odd zonal harmonics perturbations of the LAGEOS II perigee; the 1044- and 569-day periods are the periods of the main tidal orbital perturbations, with $l = 2$ and $m = 1$, which were not eliminated using equation (3.36). Some combinations of these frequencies correspond to the main non-gravitational perturbations of the LAGEOS II perigee. We note that this analysis, using EGM-96 and its accurate tidal model, is substantially independent of the removed signals, whereas the previous analysis [26], corresponding to figure 3.4, was in part sensitive to the periodic terms included in the fit. In other words, our value (figure 3.5) for the secular trend is not significantly changed by fitting additional periodic perturbations, and indeed, even the fit of the residuals with only a secular trend, with no periodic terms increases the slope by less than

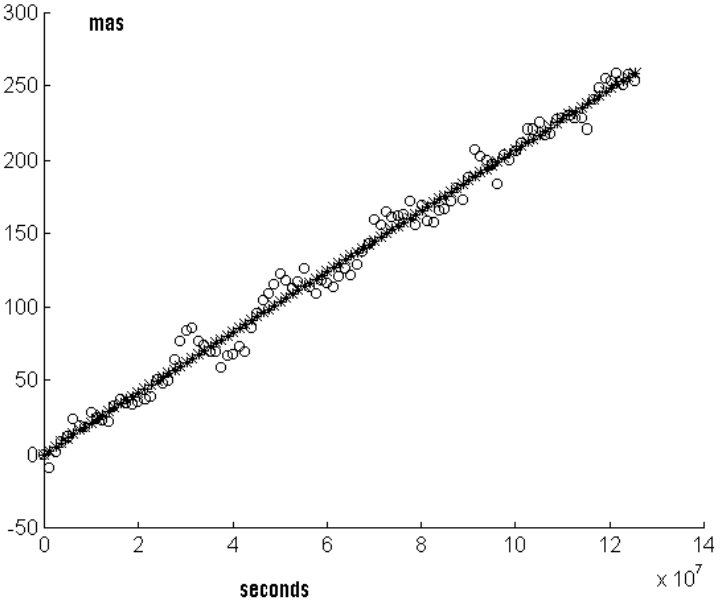

Figure 3.5. Combination of the residuals of the nodes of LAGEOS and LAGEOS II and perigee of LAGEOS II according to equation (3.36), using the Earth gravitational model EGM-96 over a 4-yr period. The best-fit line through these combined residuals has a slope of about $\mu_{\text{Measured}} \cong 1.1$.

10%. Nevertheless, in this case, the rms of the post-fit residuals increases by about four times with respect to figure 3.5.

Figure 3.6 shows the fit of the residuals obtained as in figure 3.5 but with only three periodic signals with 1044-, 820-, and 569-day periods removed.

In figures 3.4–3.6, our best-fit straight lines, through the combined residuals of nodes and perigee have, respectively, the following slopes: $\mu_1{}^{\text{Measured}} \cong 1.1$, $\mu_2{}^{\text{Measured}} \cong 1.1 \pm 0.03$, and $\mu_3{}^{\text{Measured}} \cong 1.1 \pm 0.03$, where 0.03 is the standard deviation of the fits corresponding to figures 3.5 and 3.6 using the EGM-96 gravitational model. This combined measured gravitomagnetic perturbation of the satellites' orbits corresponds in a 4-yr period to about 16 m at the LAGEOS altitude, i.e about 265 mas.

The rms of the post-fit combined residuals corresponding to figure 3.5 and 3.6 is about 9 mas. Our total systematic error is estimated to be of the order of 30–50% of μ_{GR} corresponding to figure 3.4, and of the order of 20–30% of μ_{GR} corresponding to figures 3.5 and 3.6.

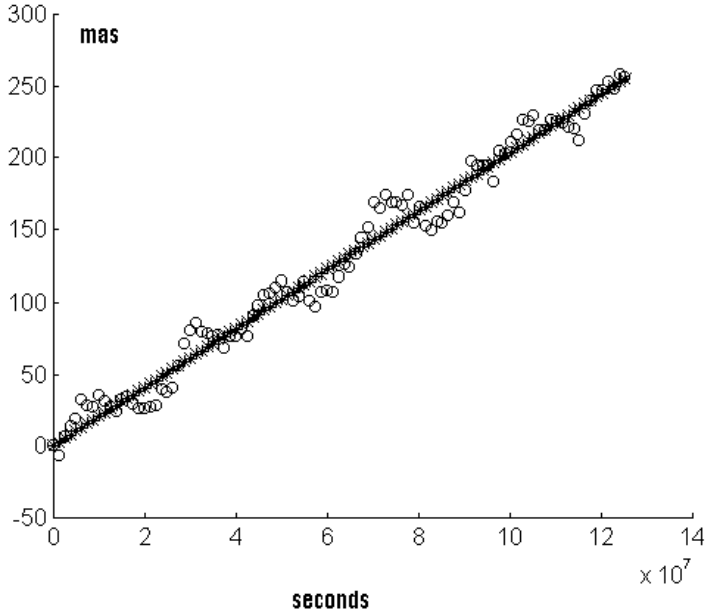

Figure 3.6. Fit of the residuals as in figure 3.5 but with removal of only three periodic signals with 1044-, 820-, and 569-day periods. The best-fit line through these combined residuals has a slope of about $\mu_{\rm Measured} \cong 1.1$.

Using the JGM-3 covariance matrix, we found the errors due to the uncertainties in the even zonal harmonics J_{2n}, with $2n \geq 6$, to be $\delta\mu^{\text{even zonals:}J_{2n}\geq J_6} \lesssim 17\%$ of $\mu_{\rm GR}$ and, using the EGM-96 covariance matrix, $\delta\mu^{\text{even zonals:}J_{2n}\geq J_6} \lesssim 13\%$ of $\mu_{\rm GR}$. The errors in the modelling of the perigee rate of LAGEOS II due to the uncertainties in the odd zonal harmonics J_{2n+1} with EGM-96 are $\delta\mu^{\text{odd zonals}} \lesssim 2\%$ of $\mu_{\rm GR}$. Using the EGM-96 tidal model, we estimated the effect of tidal perturbations and other variations in the Earth's gravitational field to be $\delta\mu^{\text{tides+other variations}} \lesssim 4\%$ of $\mu_{\rm GR}$. On the basis of analyses [26, 71] of the non-gravitational perturbations—in particular, those on the perigee of LAGEOS II—we found $\delta\mu^{\text{non-gravitational}} \lesssim 13$–$20\%$ of $\mu_{\rm GR}$, including uncertainties in the modelling of the satellites' reflectivities. The error due to uncertainties in the orbital inclinations of LAGEOS and LAGEOS II was estimated to be $\delta\mu^{\text{inclination}} \lesssim 5\%$ of $\mu_{\rm GR}$.

Taking into account all these error sources, we arrived at a total rss error $\lesssim$20–30% of $\mu_{\rm GR}$. Therefore, over an observational period of 4 yr and using

EGM-96, we determined $\mu^{\text{Measured}} = 1.1 \pm 0.3$, where 0.3 is the estimated total uncertainty due to all error sources.

Based on the 1995–2001 analyses of the orbits of the laser-ranged satellites LAGEOS and LAGEOS II, we conclude that the gravitomagnetic or Lense–Thirring effect exists and its value agrees with the prediction of Einstein's theory of general relativity.

Testing our method to measure the Lense–Thirring effect and its error budget

A basic concern in our analyses was to estimate the total error in our measurement of the Lense–Thirring effect. In order to support our measurement and the corresponding error analysis, we have performed (a) a test and (b) a preliminary blind-test simulation, explained in the following paragraphs. Finally, we describe our latest, 2001, measurement of the Lense–Thirring effect over 7.3 years of data of LAGEOS and LAGEOS II, obtained by modelling only the radiation pressure coefficient of LAGEOS II (see figure 3.8).

This 2001 measurement fully confirms and improves our previous results: the Lense–Thirring effect exists and its experimental value, $\mu \cong 1 \pm 0.3$ (± 0.3 is the estimated total systematic error), fully agrees with the general relativity prediction. It is important to note that (1) in the analysis corresponding to figure 3.8 we only modelled the radiation pressure coefficient of the satellite on LAGEOS II, i.e. the reflectivity coefficient, C_R, and no other parameters such as the accelerations along the track of the satellite as in our previous analyses corresponding to figures 3.4–3.7 (the C_{R} of LAGEOS II shows an apparent decay, in agreement with previous measurements [72]); (2) the rms of the residuals corresponding to figure 3.8 is about 10 mas, whereas the total measured signal is about 440 mas; and, finally, (3) the quality of the fit and corresponding measurement can be improved by further reducing the rms of the 15-day fits (corresponding to each point in figure 3.8) with further processing of the data using GEODYN/SOLVE, thus further reducing the rms of the final fit in figure 3.8.

(a) Testing our solution for μ using the node, perigee, mean anomaly, eccentricity and semimajor axis of LAGEOS II and node of LAGEOS. To test our measurement of the Lense–Thirring effect with the residuals of the node and perigee of LAGEOS II and the node of LAGEOS, we have also analysed the residuals of the eccentricity, mean anomaly and semimajor axis of LAGEOS II. Thus, we have produced a solution for μ by using all these orbital elements. Indeed, using the residuals of these orbital elements, we have a system of equations (for $\delta\Omega_{\text{I}}$, $\delta\Omega_{\text{II}}$, $\delta\omega_{\text{II}}$, δe_{II}, δM_{II}, and δa_{II}) in the unknowns: the Lense–Thirring effect, δJ_2 and δJ_4 errors, and a constant and a variable, once-per-revolution, along-track acceleration. Therefore, the solution of this system for μ and the corresponding fit (shown in figure 3.7) are *completely independent* of the adjusted accelerations and errors δJ_2 and δJ_4. The result of our solution with the residuals of these orbital elements and the corresponding fit for μ is shown in figure 3.7: the measured value of the Lense–Thirring effect is $\mu_{\text{Measured}} \cong 1$.

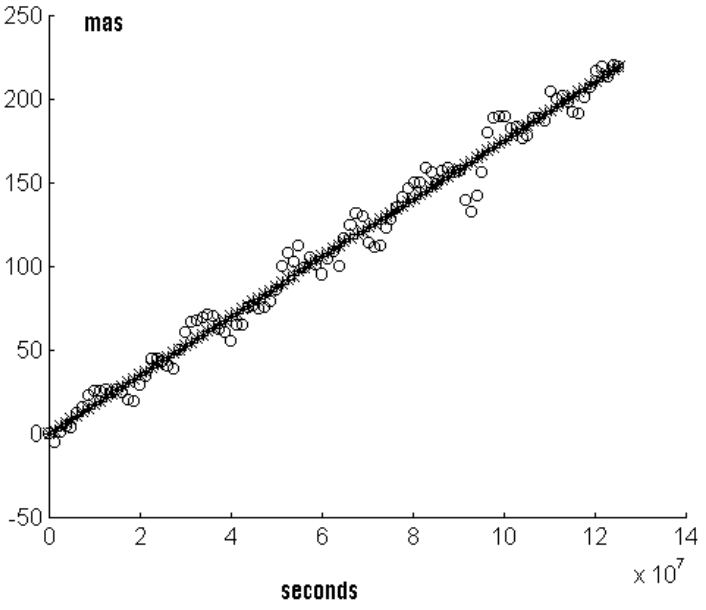

Figure 3.7. Solution of μ by using all the orbital elements of LAGEOS II. The best-fit line shown through these combined residuals has a slope of about $\mu_{\rm Measured} \cong 1$.

Thus, this further analysis and the corresponding fit test and confirm our previous result [27].

(b) Preliminary blind-test-simulation. Following a suggestion by the group at CSR at the University of Texas, Austin [72], we have performed a *preliminary* simulation and blind-test analysis. Using GEODYN, we simulated about two and a half years of data from the LAGEOS satellites with a model of perturbations in which the *a priori* value of the Lense–Thirring effect was *twice* its general relativistic value. Then, we performed our standard analysis, previously explained in this section, using the same model for the perturbations but without the Lense–Thirring effect. By our combination (3.36) of the node and perigee of LAGEOS II and the node of LAGEOS, obtained with an analysis of the *simulated* data of the orbits of these two satellites, we found a value for μ about 1.9 times the Lense–Thirring effect, i.e. about 95% of the value *a priori* set up in our simulation of their orbits and the corresponding laser-ranging data. In this analysis, we modelled the radiation pressure coefficients and along-track accelerations: the result is given in [73]. We observe that this test was just a preliminary blind-test-simulation; indeed the simulated orbits and the corresponding laser-ranging data have to be simulated by using a perturbation model which includes variations in

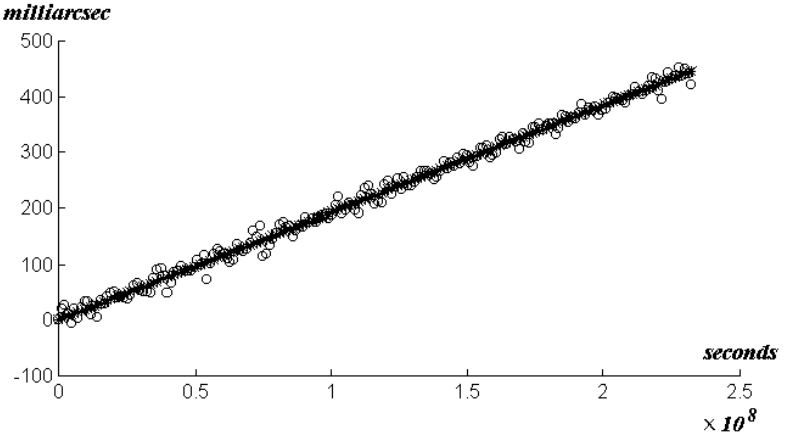

Figure 3.8. Latest 2001 measurement of the Lense–Thirring effect using LAGEOS and LAGEOS II, obtained by modelling only the radiation pressure coefficient of LAGEOS II, over nearly 8 yr of data. The best-fit line through these combined residuals has a slope of about $\mu_{\rm Measured} \cong 1 \pm 0.02$. The total estimated systematic error is about ± 0.3.

all the relevant parameters (such as J_2 and J_4) within the known uncertainties. Nevertheless, this preliminary simulation was important in that it showed the consistency of our method and of the preliminary error budget.

(c) Finally, in figure 3.8 we display our latest 2001 measurement of the Lense–Thirring effect using LAGEOS and LAGEOS II, obtained by just modelling the radiation pressure coefficient of LAGEOS II, over nearly 8 yr of data, i.e. over an observational time nearly double that of our previous analyses [73].

This recent measurement improves our previous results and fully confirms the general relativistic prediction of frame-dragging. In the analysis corresponding to figure 3.8, we have only modelled the radiation pressure coefficient on LAGEOS II and no other parameters such as the along-track accelerations. The rms of the residuals corresponding to figure 3.8 is about 10 mas

whereas the total measured signal is about 440 mas; nevertheless the quality of the fit and corresponding measurement can be improved by further reducing the rms of the 15-day fits.

3.5.3 The recent 2004 measurements of the Lense–Thirring effect using only the *nodes* of the LAGEOS satellites

3.5.3.1 Method

The accurate measurements of the Lense–Thirring effect described in this section have been obtained over a period of observation of about 10 yr by using the laser-ranging data of the satellites LAGEOS and LAGEOS II and the recently released Earth gravitational field models EIGEN-2S and GGM01S generated by the dedicated satellites CHAMP and GRACE [78–83].

The models used in this orbital analysis are described in table 3.1.

In section 3.3 we have seen how the node and perigee of a test particle are dragged by the angular momentum of a central body. *However, whereas in our previous determination of the Lense–Thirring effect, described in section 5.2, we used both the nodes of LAGEOS and LAGEOS II and the perigee of LAGEOS II, in the present analyses we have only used the nodes of LAGEOS and LAGEOS II*. Indeed, the perigee of an Earth satellite such as LAGEOS II is affected by a number of perturbations whose impact in the final error budget is not easy to assess and this was one of the two main concerns of Ries *et al* [84] (the other concern [84] was that, in the EGM-96 model, some favorable correlation of the errors of the Earth's spherical harmonics might lead to some underestimated error budget). However, this concern is absent in the present analyses. Indeed, using the previous models (JGM-3 and EGM-96), we needed three observables and thus we also needed to use the perigee of LAGEOS II. However, with the recently released solutions EIGEN-2S and GGM01S [78–83], thanks to the more accurate determination of the Earth's gravity field, it is sufficient to eliminate the uncertainty in the quadrupole moment and thus to use just two observables, i.e. the two nodes of the LAGEOS satellites. *In addition*, we have also determined the Lense–Thirring effect with EGM-96 and the use of the perigee of LAGEOS II, over about 10 years of data [73]. There was a remarkable agreement using these different techniques and the different Earth gravity models.

The nodal precessions of LAGEOS and LAGEOS II can be determined with an accuracy of the order of 1 mas yr^{-1} or less. Over our total observational period of about 10 yr, we obtained a rms of the post–fit residuals of the nodes combined with formula (3.37), of about 11 mas both with EIGEN-2S and with GGM01S.

The main error in this measurement is due to the uncertainties in the Earth's even zonal harmonics and their time variations. The un-modelled orbital effects due to the lower-order harmonics are in order of magnitude comparable to the Lense–Thirring effect (see [73]). However, by analysing the EIGEN-2S and GGM01S models and their uncertainties in the even zonal harmonics and by

Table 3.1. Models used in the orbital analysis.

Geopotential (static part)	EIGEN-2S and GGM01S
Geopotential (tides)	Ray GOT99.2
Lunisolar and planetary perturbations	JPL ephemerides DE-403
General relativistic corrections	PPN except L–T
Lense–Thirring effect	Set to zero
Direct solar radiation pressure	Cannonball model
Albedo radiation pressure	Knocke–Rubincam model
Yarkovsky–Rubincam effect	GEODYN model
Spin axis evolution of LAGEOS satellites	Farinella–Vokhroulicky–Barlier model
Station positions (ITRF)	ITRF2000
Ocean loading	Scherneck model with GOT99.2 tides
Polar motion	Estimated
Earth rotation	VLBI + GPS

calculating the secular effects of these uncertainties on the orbital elements of LAGEOS and LAGEOS II, we find that the main source of error in the determination of the Lense–Thirring effect is just due to the first even zonal harmonic, (J_2) (see later).

We can, however, use the two observable quantities $\dot{\Omega}_{\rm I}$ and $\dot{\Omega}_{\rm II}$ to determine μ [21, 22, 24], thereby avoiding the largest source of error arising from the uncertainty in J_2. We do this by solving the system of the two equations for $\delta\dot{\Omega}_{\rm I}$ and $\delta\dot{\Omega}_{\rm II}$ in the two unknowns μ and J_2, obtaining for μ:

$$\delta\dot{\Omega}^{\rm Exp}_{\rm LAGEOS\ I} + c\delta\dot{\Omega}^{\rm Exp}_{\rm LAGEOS\ II} = \mu(31 + c31.5)\ {\rm mas\ yr}^{-1} + {\rm other\ errors} \cong \mu(48.2\ {\rm mas\ yr}^{-1}) \quad (3.37)$$

where $c = 0.545$. Equation (3.37) for μ does not depend on J_2 nor on its uncertainty; thus, the value of μ that we obtain is unaffected by the largest error, due to δJ_2, and is sensitive only to the smaller uncertainties due to δJ_{2n}, with $2n \geq 4$.

Similarly, regarding tidal, secular and seasonal changes in the geopotential coefficients, the main effects on the nodes of LAGEOS and LAGEOS II caused by tidal and other time variations in Earth's gravitational field [22, 23] are due to changes in J_2; e.g. by the uncertainty in the 18.6 yr lunar tide, with the period of the Moon node, the change in J_2 due to the post-glacial rebound and by the anomalous variation in the quadrupole coefficient (see later). However, the tidal errors in J_2 and the errors resulting from other un-modelled medium- and long-period variations in J_2, including its secular and seasonal variations, are cancelled by our combination of node residuals (3.37). In particular, most of the errors resulting from the 18.6 and 9.3 yr tides, associated with the lunar node, are cancelled in our measurement. The various error sources that can affect

the measurement of the Lense–Thirring effect using the nodes of the LAGEOS satellites have been extensively treated in a large number of papers by several authors [21–24, 26, 64–68, 71]—the main error sources are treated in [73].

3.5.3.2 Results

The orbital perturbations of a satellite may be either secular and periodical. Among the secular perturbations of the node of the LAGEOS satellites, there are: the shift of the nodal line due to the even zonal harmonics of the Earth's gravitational field [41], the de Sitter effect and frame-dragging. The de Sitter effect has been measured with an accuracy of about 7×10^{-3}, its effect is only 19.2 mas on the LAGEOS node and thus its uncertainty is negligible in the Lense–Thirring measurement. However, the uncertainty in the even zonal harmonics is a crucial factor in the determination of frame–dragging since an error in one of the lower even zonal harmonics may be large enough to be indistinguishable from the Lense–Thirring nodal drag. This type of critical error is treated later. However, the periodical perturbations of the node of the LAGEOS satellites may also be a crucial factor in the determination of frame–dragging and, in particular, the uncertainty in the perturbations with a long-period compared to the period of observation may be critical. Effects with a period much shorter than the observational period are averaged out. In the present determination of the Lense–Thirring effect, we have three basic factors that make the error due to periodical effects in the measurement of frame–dragging negligible and also make this error easy to assess in the final error budget. These basic factors are:

(i) The period of the present analyses is about 10 yr and thus all the periodical perturbations of the nodes are basically averaged out apart from the 18.6 yr tide associated with the Moon node, however, the main effect of this tide is a change of the J_2 coefficient that is cancelled out using our combination of observables (3.37).

(ii) Since the original proposal of the LAGEOS III experiment [21], numerous researchers [21–24, 26, 64–68, 71] have treated the perturbations affecting the LAGEOS node in order to determine the Lense–Thirring effect and have concluded that the only critical perturbations on the nodes of the LAGEOS satellites are those due to the Earth's even zonal harmonics. However, as shown later, they contribute with an error of about 17.8 % (using the EIGEN-2S model) due to the high accuracy of the recent Earth's gravity field model EIGEN-2S (and GGM01S).

(iii) In regard to the periodical perturbations, in addition to a detailed treatment of the various perturbations affecting the LAGEOS node and their uncertainties given in [73], a simple but very meaningful test shows that the periodic perturbations cannot introduce an error larger than about 4% in our determination of the Lense–Thirring effect.

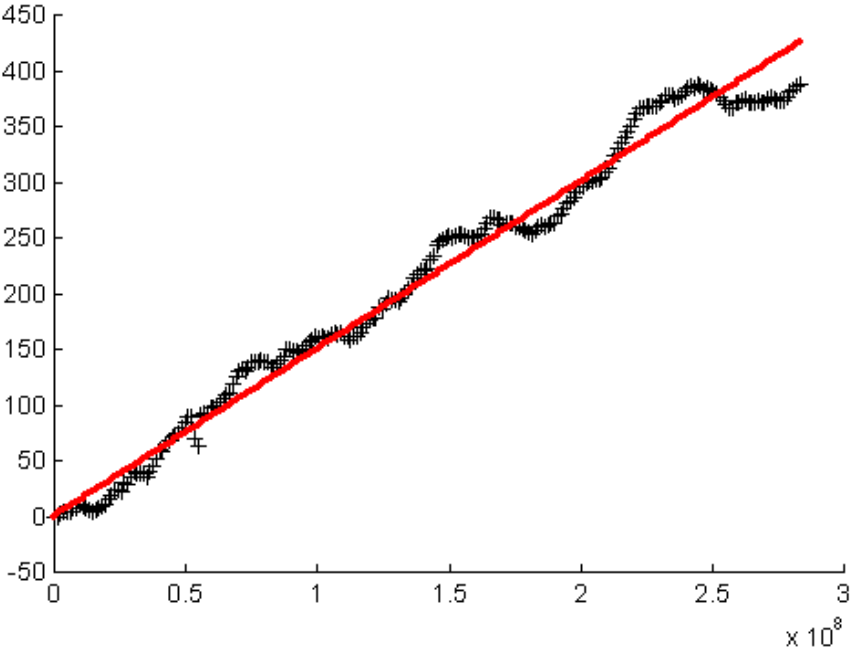

Figure 3.9. Fit of the residuals of the nodes of LAGEOS and LAGEOS II, using our combination (3.37) and the Earth model EIGEN-2S, with a secular trend only. The slope is $\mu \cong 0.99$ and the rms of the post-fit residuals is 17.5 mas.

Indeed, since the periods of the gravitational and non-gravitational orbital perturbations (but not their amplitude) are very well determined, we have also fitted for a number of periodic effects together with the secular trend. We have done a number of different fits, each with different periodical perturbations. We then compared the result in the case of the fit with a secular trend only with the various results when, together with a secular trend, we have included a different number of the main periodic perturbations. The result is that the maximum deviation of the secular trend from the case of its fit with no periodic perturbations does not exceed 4% of the Earth's frame-dragging as is clearly displayed in figures 3.9–3.12 and the corresponding captions. Of course, the rms of the fit is much smaller when we include a substantial number of periodic perturbations. Therefore, the two concerns of Ries *et al* [84], which do not impinge on the method or the value of the Lense–Thirring effect that we had obtained in our previous analyses [26–28] but rather our previous error budgets (claimed by Ries *et al* to be optimistic by a factor two or three), cannot be applied to the present analyses, as explained later and in [73]. Indeed, the first concern regarding the perturbations of the perigee of LAGEOS II is clearly absent in the present analyses since we use here only the nodes of the LAGEOS satellites and we do *not* use the

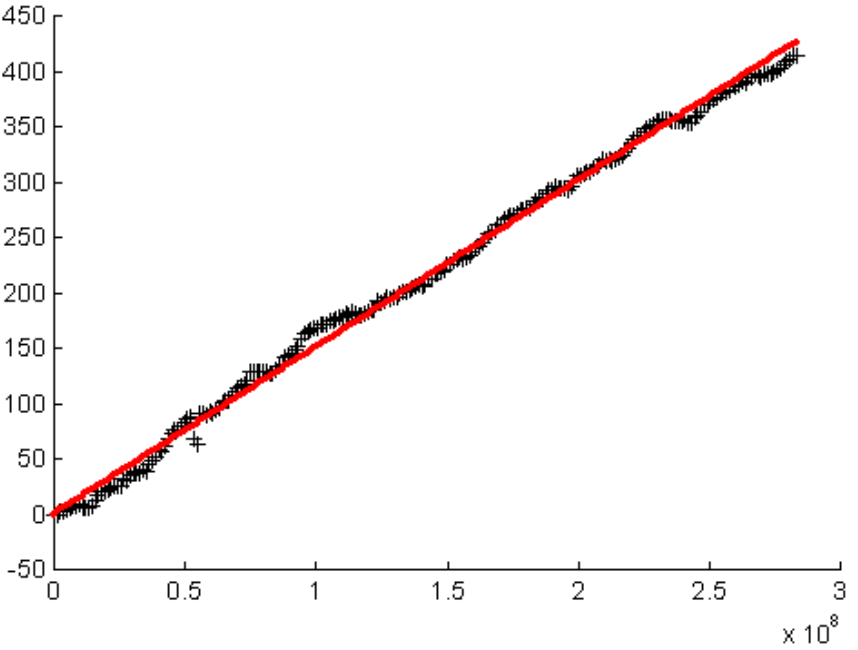

Figure 3.10. Fit of the residuals of the nodes of LAGEOS and LAGEOS II, using our combination (3.37) and the Earth model EIGEN-2S, with a secular trend plus ten periodical terms. The slope is $\mu \cong 0.97$ and the rms of the post-fit residuals is 11 mas.

perigee. Second the concern regarding the high correlation of some of the even zonal harmonics of the EGM-96 model is also substantially absent in the present analyses—indeed the Earth models we use here have a low correlation between the even zonal harmonics (see later and the related discussion in [73]).

In conclusion, in the analysis with EIGEN-2S, we have a total error budget of about 18% of the Lense–Thirring effect. Even by increasing the error due to the Earth's even zonal harmonics by 50%, we have a relative error due to the even zonals of 26.7% and a total error budget of 26.8% of the Lense–Thirring effect.

The main perturbations in our determination of the Lense–Thirring effect are described and analysed in [73].

In the present analysis, we have used EIGEN-2S and GGM01S; however, these models are preliminary in the sense that they have been obtained over relatively short periods of observations by CHAMP and GRACE. Thus, the values of the Earth's spherical harmonic coefficients may change appreciably with longer periods of observations. In particular, the uncertainties in the Earth's zonal harmonics include only tentatively systematic errors in GGM01S and are only formal errors in EIGEN-2S. However, our analysis is not sensitive to changes in the Earth's quadrupole moment—it is just affected by changes in the higher

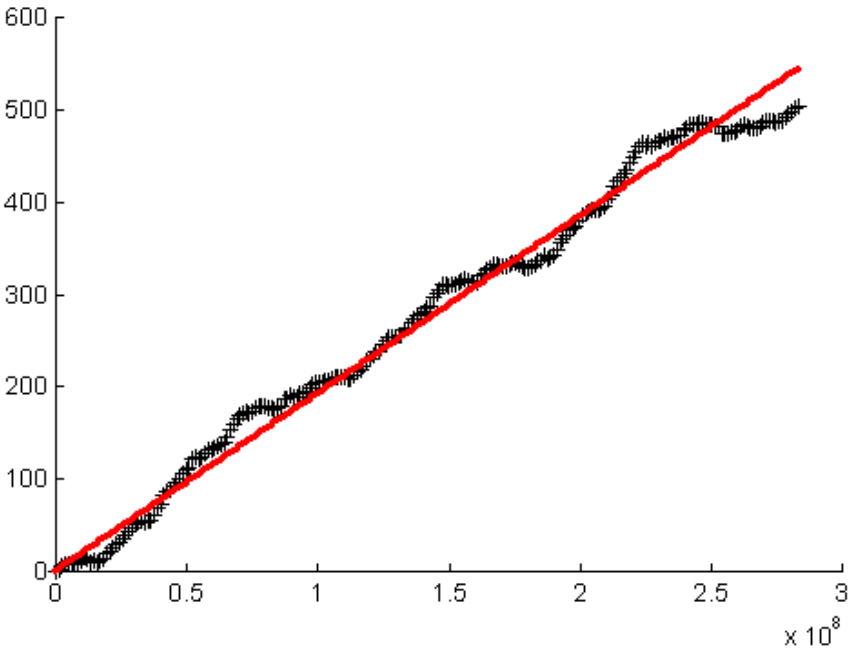

Figure 3.11. Fit of the residuals of the nodes of LAGEOS and LAGEOS II, using our combination (3.37) and the Earth model GGM01S, with a secular trend only. The slope is $\mu \cong 1.26$ and the rms of the post-fit residuals is 18 mas.

zonal harmonics. Therefore, future more accurate determinations of the even zonal coefficients and of their uncertainties might lead to different values of μ and to a different total error budget. Nevertheless, it is crucial to note that when a more accurate model using GRACE, CHAMP or GOCE becomes available, it will be straightforward to, assess *a posteriori* the total error of our present analyses very accurately. Indeed, one will just need to take the differences between the values of the even zonals of the EIGEN-2S and GGM01S models presently used with the corresponding values of the future more accurate model from GRACE, CHAMP or GOCE and consider the uncertainties in these future models. Thus, the total error in the present measurement of μ due to the uncertainties in the J_{2n} coefficients can be easily re-estimated.

In figures 3.9–3.12, we report our determination of the Lense–Thirring effect, obtained using the nodal rates of the LAGEOS and LAGEOS II satellites over a period of about 10 yr. Figure 3.9 displays the combination of the nodes according to formula (3.37), representing the measurement of the Lense–Thirring effect using the EIGEN-2S model: the slope is 0.99μ. Figure 3.10 shows our measurement of the Lense–Thirring effect using EIGEN-2S by fitting the orbital residuals with a secular trend contemporarily with ten main periodic effects: the

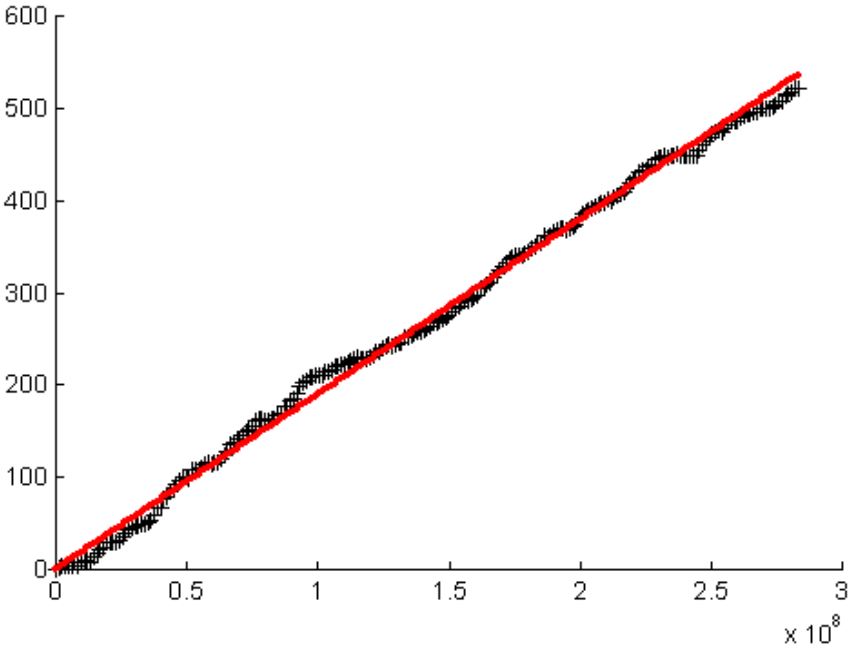

Figure 3.12. Fit of the residuals of the nodes of LAGEOS and LAGEOS II, using our combination (3.37) and the Earth model GGM01S, with a secular trend plus ten periodical terms. The slope is $\mu \cong 1.22$ and the rms of the post-fit residuals is 11 mas.

slope is 0.97μ. Figure 3.11 displays the combination of the nodes according to formula (3.37), representing the measurement of the Lense–Thirring effect using GGM01S: the slope is 1.26μ. Figure 3.12 shows our measurement of the Lense–Thirring effect using GGM01S by fitting the orbital residuals with a secular trend contemporarily with 10 periodic effects: the slope is 1.22μ.

In conclusion, by fitting our combined residuals with only a secular trend, using EIGEN-2S, we found that

$$\mu = 0.985 \pm 0.182 \tag{3.38}$$

where $\mu \equiv 1$ in general relativity. By fitting our combined residuals with a secular trend plus 10 periodic signals and, using EIGEN-2S, we found that

$$\mu = 0.965 \pm 0.182. \tag{3.39}$$

The rms of the post–fit residuals was 17.5 mas in the case of the fit of secular trend only and 11 mas in the case of the fit of a secular trend plus 10 periodic signals.

By fitting our combined residuals with a secular trend plus ten periodic signals and using GGM01S, we found that

$$\mu = 1.22 \pm 0.239.$$

The rms of the post-fit residuals was 18 mas in the case of the fit of a secular trend only and 11 mas in the case of the fit of a secular trend plus 10 periodic signals. However, since the covariance matrix of GGM01S was not available, we could not assess the total systematic error using GGM01S accurately. Nevertheless, by simply adding the absolute values of the errors due to the published uncertainties in each GGM01S even zonal coefficient, we obtained a maximum error of 23.6% in μ, due to the error in the Earth's static gravity field, and a total rss error of 23.9% of μ (since, by far the most dominating error source is due to uncertainties in the static even zonal harmonics).

By fitting our combined residuals with two six or ten periodic terms, we basically obtained the same value for the Lense–Thirring effect with a maximum variation of 4% only. Furthermore, the two determinations of the Earth's frame-dragging effect obtained with EIGEN-2S and GGM01S are practically in agreement with each other within their uncertainties ($\cong$ 18% for EIGEN-2S and $\cong$ 24% for GGM01S).

Our measured value of the Lense–Thirring effect corresponds to 97% (in the improved fit with ten frequencies and 99% in the fit with a trend only) of Einstein's theory prediction (using EIGEN-2S) and thus, since our experimental uncertainty is about 18%, it fully agrees with the general relativistic prediction.

We note that the only uncertainty in our present error budget is due to the published errors in the EIGEN-2S and GGM01S models that do not include (in EIGEN-2S) or might just underestimate the systematic errors (in GGM01S) in the determination of the even zonal harmonics. However, when more accurate models of the Earth's gravity field become available, it will be straightforward to evaluate the accuracy of the EIGEN-2S and GGM01S models we have used here (by substantially taking the difference between the corresponding coefficients). This will provide, *a posteriori*, a solid re-assessment of the error in the present determination of frame-dragging.

In addition, we also analysed the LAGEOS satellites data using the older model EGM-96 and our previous method of combining the nodes of the LAGEOS satellites with the perigee of LAGEOS II. However, the present period of analysis was about 10 yr, i.e. about 2.5 times longer than any previous period of analysis. This measurement of μ with EGM-96 was in full agreement with our previous determination of frame-dragging [73].

Finally, we note that, in addition to the accurate measurement of the Lense–Thirring effect, we have observed an anomalous increase of the Earth's quadrupole coefficient J_2 since 1998 in the orbital residuals of LAGEOS and LAGEOS II. This change we observe in J_2 is in good agreement with the J_2 variation observed by Cox and Chao [85].

In conclusion, the Lense–Thirring effect exists and its experimental value—$\mu = 0.97 \pm 0.18$—fully agrees with the prediction of general relativity [73]. Recently, in March 2004, we have obtained an improved measurement in the Earth's frame-dragging using the newest model provided by the GRACE satellites, this recent result fully agrees with general relativity with an error of $\cong$5% only [75].

References

[1] Ciufolini I and Wheeler J A 1995 *Gravitation and Inertia* (Princeton, NJ: Princeton University Press)

[2] Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco, CA: Freeman)

[3] Thirring H 1918 *Z. Phys.* **19** 33

[4] Brill D R and Cohen J M 1966 *Phys. Rev.* **143** 1011

[5] Pfister H and Braun K 1986 *Class. Quantum Grav.* **3** 335

[6] Pfister H 1995 Mach's principle *From Newton's Bucket to Quantum Gravity* ed J Barbour and H Pfister (Boston, MA: Birkhäuser) p 315

[7] Lense J and Thirring H 1918 *Phys. Z.* **19** 156
see also English translation by Mashhoon B, Hehl F W and Theiss D S 1984 *Gen. Rel. Grav.* **16** 711

[8] Friedländer B and I 1896 *Absolute und Relative Bewegung?* (Berlin: Simion)

[9] Föppl A 1904 *Sitzb. Bayer. Akad. Wiss.* **34** 5
Föppl A 1904 *Phys. Z.* **5** 416
Föppl A 1904 *Sitzb. Bayer. Akad. Wiss.* **34** 383

[10] de Sitter W 1916 *Mon. Not. R. Astron. Soc.* **76** 699

[11] Yilmaz H 1959 *Bull. Am. Phys. Soc.* **4** 65

[12] Van Patten R A and Everitt C W F 1976 *Phys. Rev. Lett.* **36** 629

[13] Pugh G E 1959 Proposal for a satellite test of the Coriolis prediction of general relativity *Weapons Systems Evaluation Group Research Memorandum No 11* (Washington, DC: The Pentagon)

[14] Schiff L I 1960 *Proc. Natl Acad. Sci.* **46** 871
Schiff L I 1960 *Phys. Rev. Lett.* **4** 215

[15] Everitt C W F 1974 The gyroscope experiment. I: General description and analysis of gyroscope performance *Experimental Gravitation* ed B Bertotti (New York: Academic) pp 331–60

[16] Braginsky V B, Polnarev A G and Thorne K S 1984 *Phys. Rev. Lett.* **53** 863

[17] Schleich W and Scully M O 1984 *Les Houches 1982, New Trends in Atomic Physics* ed G Grynberg and R Stora (Amsterdam: North-Holland) p 995

[18] Braginsky V B and Polnarev A G 1980 *Pis'ma V Zh. Eksp. Teor. Fiz. (USSR)* **31** 444 (Engl. transl. *JETP Lett.* (USA))

[19] Mashhoon B, Paik H J and Will C M 1989 *Phys. Rev.* D **39** 2825

[20] Nordtvedt K 1988 *Int. J. Theor. Phys.* **27** 1395

[21] Ciufolini I 1986 *Phys. Rev. Lett.* **56** 278

[22] Ciufolini I 1989 *Int. J. Mod. Phys.* A **4** 3083
Tapley B, Ciufolini I, Ries J C, Eanes R J and Watkins M M 1989 *NASA-ASI Study on LAGEOS III* CSR-UT publication no CSR-89-3, Austin, TX

Ciufolini I *et al* 2004 *WEBER SAT/LARES Study for INFN*
[23] Ciufolini I *et al* 1998 *LARES Phase A study for ASI*
[24] Ciufolini I 1996 *Nuovo Cimento* A **109** 1709
[25] Ciufolini I *et al* 1996 *Nuovo Cimento* A **109** 575
[26] Ciufolini I *et al* 1997 *Class. Quantum Grav.* **14** 2701
Ciufolini I *et al* 1997 *Europhys. Lett.* **39** 35
[27] Ciufolini I *et al* 1998 *Science* **279** 2100
[28] Ciufolini I 1994 *Class. Quantum Grav.* A **11** 73
[29] Ciufolini I 2004 to be published
[30] Petrov A Z 1969 *Einstein Spaces* (Oxford: Pergamon)
[31] Wheeler J A 1977 *Trans. N. Y. Acad. Sci., Series II* **38** 219
[32] Ciufolini I and Ricci F 2002 *Class. Quantum Grav.* **19** 3863
[33] Cohen J M and Mashhoon B 1993 *Phys. Lett.* A **181** 353
[34] Mashhoon, Gronwald and Theiss 1999 *Ann. Phys* **8** 135
[35] Tartaglia A 2000 *Online Preprint* http://babbage.sissa.it/abs/gr-qc/0001080
[36] Tartaglia A 2000 *Class. Quantum Grav.* **17** 783
[37] Pineault S and Roeder R C 1977 *Astron. J.* **212** 541
[38] Pineault S and Roeder R C 1977 *Astron. J.* **213** 548
[39] Rauch K P and Blandford R D 1994 *Astron. J.* **421** 46
[40] Virbhadra K S and Ellis G F R 2000 *Phys. Rev.* D **62** 084003
[41] Kaula W M 1966 *Theory of Satellite Geodesy* (Waltham: Blaisdell)
[42] Chandrasekhar S 1969 *Ellipsoidal Figures of Equilibrium* (New Haven, CT: Yale University Press)
[43] Schneider P, Ehlers J and Falco E 1992 *Gravitational Lenses* (Berlin: Springer)
[44] Hucra J *et al* 1985 *Astron. J.* **90** 691
[45] Racine R 1991 *Astron. J.* **102** 454
[46] Chae K H, Turnshek D A and Khersonsky V K 1998 *Astrophys. J.* **495** 609
[47] Will C M 1985 *Theory and Experiment in Gravitational Physics* (Cambridge: Cambridge University Press) revised edition 1993
[48] Barbour J and Pfister H (ed) 1995 *Mach's Principle. From Newton's Bucket to Quantum Gravity* (Boston, MA: Birkhäuser)
[49] Bass L and Pirani F A E 1955 *Phil. Mag.* **46** 850
[50] Weinberg S 1972 *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (New York: Wiley)
[51] Ciufolini I and Ricci F 2002 *Class. Quantum Grav.* **19** 3875
[52] Hucra J *et al* 1985 *Astron. J.* **90** 691
[53] Chae K-H, Turnshek D A and Khersonsky V K 1998 *Astrophys. J.* **495** 609
[54] Abell G O 1975 *Stars and Stellar Systems, IX Galaxies and the Universe* (Chicago, IL: University of Chicago Press) p 636
[55] Iorio L, Lucchesi D and Ciufolini I 2002 *Class. Quantum Grav.* **19** 4311
[56] Iorio L, Ciufolini I and Pavlis E 2002 *Class. Quantum Grav.* **19** 4301
[57] Nordtvedt K 1998 LARES and tests on new long range forces *LARES Phase A study for ASI* ed I Ciufolini *et al*
[58] Cohen S C and Dunn P J (ed) 1985 LAGEOS scientific results *J. Geophys. Res.* B **90** 9215
[59] Bender P and Goad C C 1979 The use of satellites for geodesy and geodynamics *Proc. 2nd Int. Symp. on the Use of Artificial Satellites for Geodesy and Geodynamics* vol II, ed G Veis and E Livieratos (National Technical University of Athens) p 145

[60] Degnan J J and Pavlis E C 1994 *GPS World* **5** 62
[61] *International Earth Rotation Service (IERS) Annual Report 1996* (Observatoire de Paris, Paris, July 1997)
[62] McCarthy D 1996 *The 1996 IERS Conventions* (Paris: Observatoire de Paris)
[63] Lemoine F G *et al* 1998 *The Development of the Joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) Geopotential Model EGM96* NASA/TP-1998-206861
[64] Rubincam D P 1988 *J. Geophys. Res.* B **93** 13803
[65] Rubincam D P 1990 *J. Geophys. Res.* B **95** 4881
[66] Rubincam D P and Mallama A 1995 *J. Geophys. Res.* B **100** 20285
[67] Martin C F and Rubincam D P 1996 *J. Geophys. Res.* B **101** 3215
[68] Farinella P, Vokrouhlicky D V Y and Barlier F 1996 *J. Geophys. Res.* B **101** 17861
[69] Gross R S 1996 *J. Geophys. Res.* B **101** 8729
[70] Pavlis D E *et al* 1997 *GEODYN II, Operations Manual* p 3
[71] Lucchesi D M 2002 *Planet Space Sci.* **50** 1067
[72] Eanes R and Ries J 1988 *Center for Space Research* (University of Texas at Austin) private communication
[73] Ciufolini I, Pavlis E and Peron R 2004 *New Astronomy* submitted
[74] Ciufolini I and Ricci F 2004 *Class. Quantum Grav.* submitted
[75] Ciufolini I, Pavlis E and Peron R 2004 *Science* submitted
[76] Dvali G 2003 Infrared modification of gravity *Nobel Symposium on Cosmology and String Theory, August 2003, Sigtuna, Sweden*
[77] Ciufolini I *et al* 2004 *LARES/WEBER-SAT Study for INFN*
[78] Pavlis N K 2003 Evaluation of some gravitational models that incorporate CHAMP data *Geophysical Research Abstracts* vol 5 (CD) Abstract EAE03-A-04523
[79] Perosanz F, Loyer S, Lemoine J M L, Biancale R, Bruinsma S and Vales N 2003 CHAMP accelerometer evaluation on two years mission *Geophysical Research Abstracts* vol 5 (CD) Abstract EAE03-A-06989
[80] Reigber Ch, Flechtner F, Koenig R, Meyer U, Neumayer K, Schmidt R, Schwintzer P and Zhu S 2002 *GRACE Orbit and Gravity Field Recovery at GFZ Potsdam—First Experiences and Perspectives, Eos. Trans. AGU, 83(47), Fall Meet. Suppl.* Abstract G12B-03
[81] Rummel R 2003 GOCE—its status and promise *Geophysical Research Abstracts* vol 5 (CD) Abstract EAE03-A-09628
[82] Tapley B D 2002 *The GRACE Mission: Status and Performance Assessment, Eos. Trans. AGU, 83(47), Fall Meet. Suppl.* Abstract G12B-01
[83] Watkins M M, Yuan D, Bertiger W, Kruizinga G, Romans L and Wu S 2002 *GRACE Gravity Field Results from JPL, Eos. Trans. AGU, 83(47), Fall Meet. Suppl.* Abstract G12B-02
[84] Ries *et al* 2004 Nonlinear gravitodynamics *The Lense–Thirring Effect* ed R Ruffini and C Sigismondi (Singapore: World Scientific) p 201
[85] Cox C M and Chao B 2002 *Science* **297** 831

Chapter 4

The special relativistic Equivalence Principle: gravity theory's foundation

Kenneth Nordtvedt
Northwest Analysis, 118 Sourdough Ridge, Bozeman, MT 59715, USA

4.1 Introduction

When Einstein formulated his grand hypothesis, the Equivalence Principle (EP) and then used that principle to make his two classic predictions—that gravity deflects light and alters clock rates—his arguments rested on only the most rudimentary feature of his special relativity theory: he essentially employed Newtonian physics. A light ray (illustrated in figure 4.1 by the finely dotted line) leaves an upwardly accelerating floor at initial angle $+\phi$ and it again meets the floor at a later time T and at horizontal distance L as determined from the two Newtonian equations

$$cT\sin\phi = \tfrac{1}{2}gT^2 \qquad \text{and} \qquad cT\cos\phi = L.$$

On reunion at time T, the light ray makes a descending angle $-\phi$ with respect to the floor: the rate per unit time for the deflection of that light ray with respect to the floor is then (in the small ϕ limit) $\mathrm{d}\phi/\mathrm{d}t \cong g/c$, or expressed as deflection rate per distance travelled, $\mathrm{d}\phi/\mathrm{d}x \cong g/c^2$. Light-ray pulses are also indicated in figure 4.1, propagating between a clock C anchored to the accelerating floor and another clock C′ anchored at height h above the floor. The *time transfer* relationship between the times the light leaves the former (t_1) and arrives at the latter (t_2) is obtained from the Newtonian equation

$$\tfrac{1}{2}gt_1^2 + c(t_2 - t_1) = h + \tfrac{1}{2}gt_2^2 \tag{4.1}$$

which, in a first approximation yields a relative rate for these times

$$\frac{\mathrm{d}t_2}{\mathrm{d}t_1} \simeq 1 + \frac{gh}{c^2}.$$

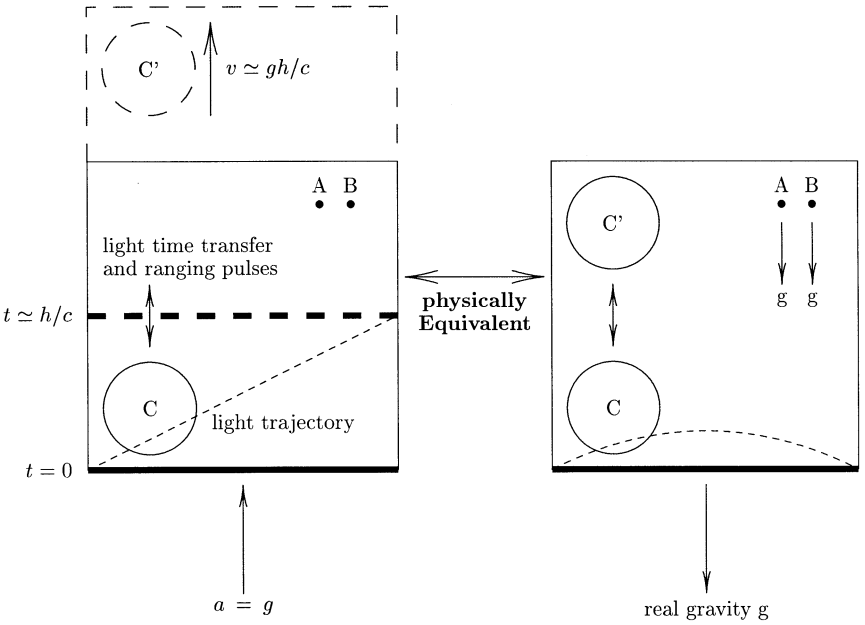

Figure 4.1. Einstein's original Equivalence Principle arguments. In the accelerating left-hand box, (1) the floor reaches bodies A and B, both at rest in inertial space, at the same time, (2) the floor's right-hand edge has accelerated upwards to meet the light ray and (3) light pulses sent out by each tick of clock C (anchored to the floor of the box) are received at a slower rate by clock C′ (anchored to the ceiling of the box) because of the latter clock's upward motion acquired during the light pulses' times of flight; and the light pulses can be reflected or transponded back to clock C. If *equivalent* phenomena are to occur in the right-hand box which is at rest in gravity, then (1) bodies A and B must fall at precisely identical rates, (2) light is deflected by gravity, (3) clock C ticks slower than clock C′ by virtue of its different location in a gravitational potential and (4) the round-trip ranging time measured by clock C is less than $2h/c$.

If round-trip ranging experiments using light had been contemplated by Einstein a century ago, he could also have predicted the local outcome of such ranging measurements by adding to equation (4.1) a relationship for the light's return trip,

$$h + \tfrac{1}{2}gt_2^2 - c(t_3 - t_2) = \tfrac{1}{2}gt_3^2 \tag{4.2}$$

which when added to the outbound time gives the round-trip's total elapsed time

$$t_3 - t_1 \cong \frac{2h}{c} - \frac{gh^2}{c^3}.$$

This EP-derived ranging time for *local* experiments is substantiated in metric theories of gravity such as general relativity and its scalar–tensor variations. The

EP predictions of light's deflection in gravity have been claimed by some to be no more than earlier predictions of mechanistic deflection of light corpuscles travelling at the finite speed c. This mechanistic viewpoint, however, would predict a speeding up of light as it approached matter, not the slowing obtained from the EP [4].

The third phenomenon illustrated in figure 4.1 consists of the generally different bodies A and B which are at rest and located side by side in inertial space. The upwardly accelerating floor then meets both of these bodies simultaneously: indeed it was Einstein's contemplation of this identity of free fall which led him to his principle.

Requiring these observational results to occur also in gravity by virtue of the EP, the interpretations must now be that the local gravitational acceleration g (1) deflects a transversely propagating light ray, (2) changes clock frequencies f with altitude h and (3) increases the speed of light (as measured by a ground clock) by the previously derived rates,

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = \frac{1}{f}\frac{\mathrm{d}f}{\mathrm{d}h} = \frac{1}{c}\frac{\mathrm{d}c}{\mathrm{d}h} = \frac{g}{c^2}$$

and (4) different bodies A and B *fall* in gravity at precisely identical rates. Special relativity played almost no role in arriving at these conclusions.

But the EP can predict a number of additional novel phenomena. By fully utilizing special relativity when exploring implications of the EP, converting it into the Special Relativistic Equivalence Principle (SREP), further effects can be predicted which include (1) *geodetic* precession of a body's inertial orientation as it free-falls non-vertically in gravity, (2) a relativistic ($1/c^2$ order) contribution to the precession of the major axes of gravitational orbits (such as Mercury's) and (3) a *gravitomagnetic* precession of a body's inertial orientation by virtue of a moving source of gravity, as well as a general gravitational interaction between mutually moving masses and between moving mass and light.

The derivation of these new consequences of *equivalence* follows the spirit of the original EP arguments. Novel phenomena are first derived as they occur in gravity-free, accelerated laboratories. To analyse body and light ray trajectories, clock rates and behaviour of other experimental devices, we set up a *master* inertial frame with its observer and clock at rest and, from that perspective, the calculations of clock, body and light behaviours can be performed. In this gravity-free inertial frame, light rays travel along straight lines at a unique speed c, free bodies move at constant velocities and arbitrarily moving clocks 'tick' at the special relativistic *proper* rate

$$\mathrm{d}\tau = \mathrm{d}t\sqrt{1 - v(t)^2/c^2} \tag{4.3}$$

expressed in terms of the rate $\mathrm{d}t$ of a clock at rest in the master inertial frame. A 'ground' floor of clocks are synchronously given equal and constant (properly measured by accompanying accelerometers) upward accelerations. To

keep the interpretations of various measurable phenomena as straightforward and free of controversy as possible, the experimental observables are confined to measurements made on this ground floor of accelerating clocks (later, of course, an equivalent array of clocks is deployed on the actual *ground* in a local gravitational field). Special relativity's Lorentz transformation, used to relate event coordinates as measured in two inertial frames which move at constant velocity relative to each other, is needed: in the case of a transformation to a frame moving at speed v in the $y\text{-}direction$, for example, new coordinates are related to original ones by

$$\begin{aligned} \mathrm{d}t' &= \gamma(\mathrm{d}t - v\,\mathrm{d}y/c^2) \\ \mathrm{d}y' &= \gamma(\mathrm{d}y - v\,\mathrm{d}t) \\ \mathrm{d}x' &= \mathrm{d}x \\ \mathrm{d}z' &= \mathrm{d}z \end{aligned}$$

with

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}.$$

The types of *gedanken* experiments analysed in this gravity-free situation of a ground floor of upwardly accelerating clocks are shown in the bottom picture of figure 4.4. Both bodies and light rays which are given free trajectories on initially leaving the accelerating ground floor are considered. The bodies may carry clocks and have extension (orientation). At future times, there will be reunions of the body (clock) trajectories and light trajectories with that of the upwardly accelerating ground floor. Various measurable quantities are then recorded at these reunion events: such measurements include the elapsed proper times of various clocks, the body orientations, horizontal locations of reunions, etc. *The SREP then requires identical results for the same measurable quantities in gravity, as shown in the top picture of figure 4.4. In order to achieve this identity of results, unique gravity-induced modifications to the speed of light function, to the body equation of motion and to the clock rate function are determined and rotations of an inertial rod with respect to the ground during free fall motion are required.*

Consider a rod which travels at constant velocity and without rotation through gravity-free inertial space. ('Non-rotation' of the rod can be established, for instance, by attached accelerometers which record no centrifugal forces.) As shown in figure 4.2, the trajectory of this rod is twice crossed by that of an upwardly accelerating ground floor of the non-inertial laboratory. In the instantaneous rest frames of those two crossing events, the orientations of the rod with respect to the ground are determined and found to differ. When the SREP is then invoked and a rod free-falling in gravity (and free of absolute rotation) is considered, this same change in orientation will be required but that rotation must now be interpreted as a precession of the rod's inertial orientation by virtue of its motion through the local gravity—*geodetic precession* or, by virtue of the motion of the source of gravity, *gravitomagnetic precession*.

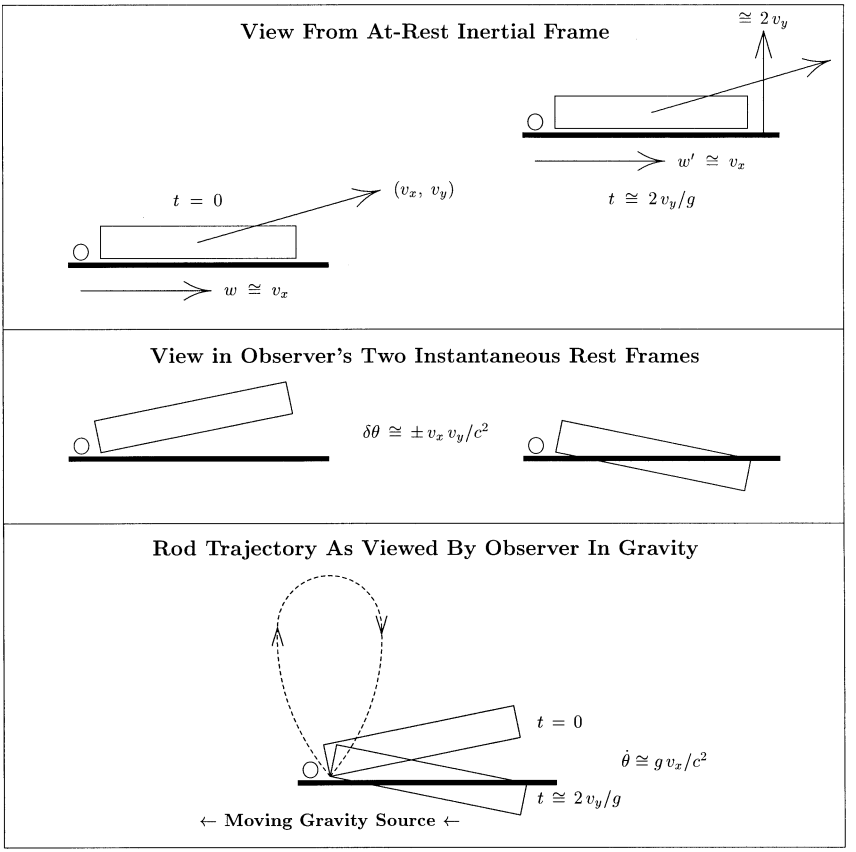

Figure 4.2. Gravitomagnetic precession from the Equivalence Principle. The top scene shows an upwardly accelerating floor and a non-rotating rod moving freely through gravity-free space. Floor and rod meet twice and an observer moves at constant proper velocity along the floor to be present at both events. The middle scene shows the meetings in the two instantaneous rest frames of the observer. The relativity of simultaneity in the Lorentz transformation for time results in different rotations of the rod in the two events. The EP calls for the same observable outcomes in gravity: this predicts the gravitomagnetic rotation relative to the ground of an *inertially non-rotating* rod due to a moving source of the gravity. The bottom scene also shows that the rod's gravitational free-fall trajectory is not vertical as viewed from the observer. This specifies the local gravitomagnetic contribution to the gravitational equation of motion.

4.2 Gravitomagnetic precession due to moving gravity source

As viewed from a master inertial frame (top panel of figure 4.2), at time $t = 0$ a horizontal rod leaves a floor with horizontal velocity component v_x and vertical

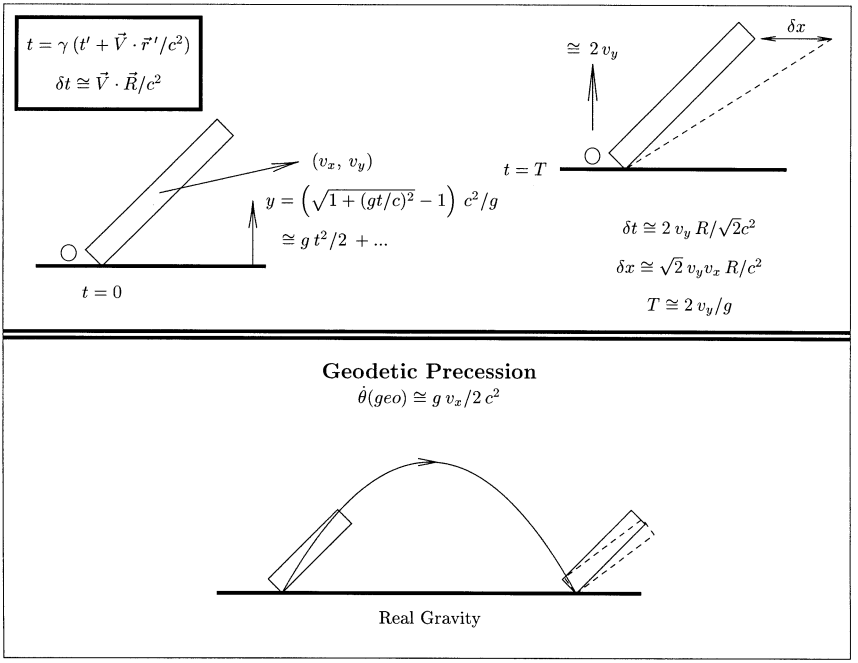

Figure 4.3. Geodetic precession from the Equivalence Principle. The top scenes shown from perspective of a master inertial frame show an intrinsically non-rotating rod both when it leaves and when it again meets an upwardly accelerating floor. After Lorentz transformations to the instantaneous rest frames of the observer (fixed on the floor) indicated by symbols ◯, the orientations of the rod relative to the floor are shown by the dotted rod. The key *time* Lorentz transformation responsible for the rod reorientation is shown in the upper left-hand corner of the figure. Assuming the SREP, the figure's bottom scene shows the same rotation of the rod in gravity but which now must be interpreted as *geodetic* rotation of the inertial frame which moves through gravity with the rod.

velocity v_y [3]. An observer travels along the floor at constant horizontal *proper* speed w selected so as to arrive at the future reuniting event of rod and floor. The floor accelerates upward as $y \cong gt^2/2$. In the $t' = 0$ rest frame of this observer, the *time* Lorentz transformation indicates different times as measured in the master inertial frame for the two ends of the rod

$$t = \gamma(t' + \boldsymbol{V}\cdot\boldsymbol{r}'/c^2) \qquad \text{with } \gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{4.4}$$

with the right-hand side of the rod having the later time t value. With the rod initially moving up from the floor, the middle panel of figure 4.2 indicates the rod's initial orientation as seen in the observer's rest frame. The difference

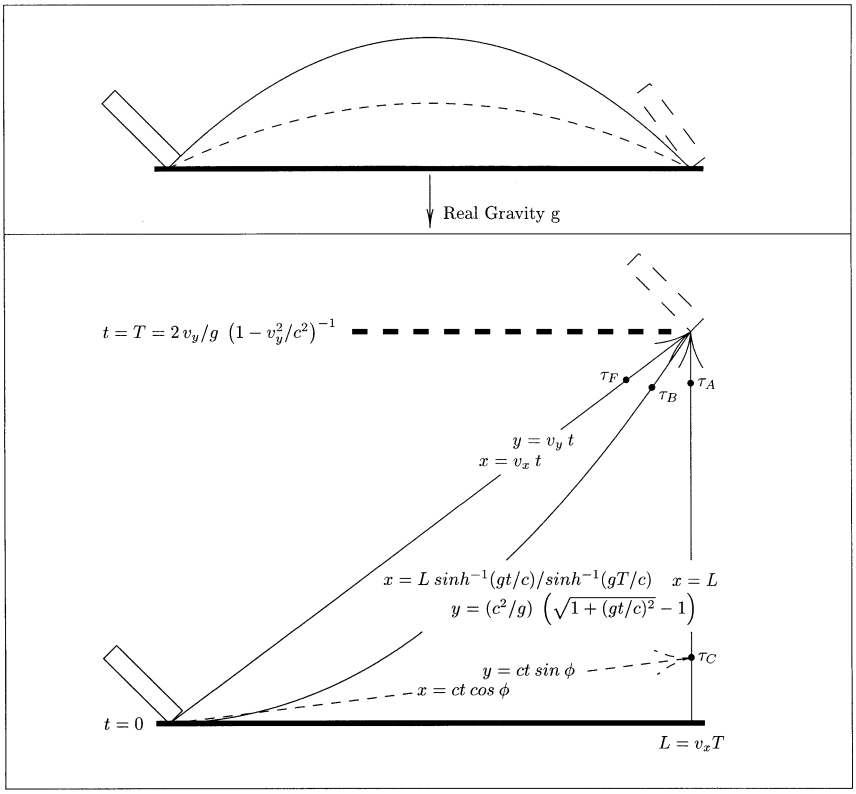

Figure 4.4. Clock/rod and light ray leave and rejoin the ground: two views. The bottom view of events is as seen by a master inertial observer at rest in gravity-free space. At time $t = 0$, a light ray (broken line) is launched at angle ϕ and a non-accelerating, non-rotating rod with clock is launched at angle $\tan^{-1}(v_y/v_x)$ (full line), and then ray and rod meet an upwardly accelerating 'ground' clock at the latter's times τ_C and τ_A, respectively. Another 'ground' clock moves at constant proper speed w to the right to also meet the rod/clock at the reunion event. The non-accelerating (free falling) clock records the time τ_F for its reunion event, and the right-moving 'ground' clock records the time $\tau_B = \tau_A\sqrt{1 - w^2/c^2}$. The trajectories of the three clocks and the light ray as recorded in the master inertial frame are indicated. The top view shows the same physical events occuring in gravity. The SREP requires all observables, such as the clock readings at the reunions etc, to have identical values in the two situations.

between rod orientation angles seen in the two frames is readily evaluated to be $v_x\,v_y/c^2$ in leading relativistic order.

The upwardly accelerating floor meets the rod again at time $T \cong 2v_y/g$ by which time the floor is travelling upward at speed of about $2v_y$. In the

instantaneous rest frame of the observer at this second meeting of floor with rod, the Lorentz transformation given in equation (4.4) can again be used to find that the master inertial frame time for the right-hand end of the rod is later than that for the rod's left-hand end. But since the rod is now travelling down relative to the floor at speed of about v_y, the relationship between rod orientations as seen in the master inertial frame and instantaneous rest frame of the observer is now reversed as also shown in the middle panel of figure 4.2: the latter orientation is now turned down from the horizontal orientation by angle which is again $v_x v_y/c^2$. Dividing by the total elapsed time T between the two events, one obtains a precession rate for the rod relative to the floor

$$\dot{\Theta}_{\mathrm{LT}} \cong \frac{g\, v_x}{c^2} \tag{4.5}$$

labelled 'LT' in recognition of the pioneering work of Lense and Thirring concerning this precession in general relativity theory [6]. As seen from a frame of reference at rest with the observer, a rod is launched (almost) vertically into gravitational free-fall. Upon return to the ground, the rod has rotated while nevertheless not experiencing internal centrifugal accelerations. An observer moving at constant velocity along an upwardly accelerating floor detects his/her own motion: there is a preferred frame on this floor established by special relativity. In gravity, however, the only available explanation for this rotation is the observer's presence in a gravitational field and the leftward horizontal motion of the gravitational source relative to the observer's frame. *In proximity to a moving source of gravity, the local inertial frame must rotate!* The slight non-verticality of the free-fall trajectory which is another consequence of gravitomagnetism is discussed in section 4.4.

4.3 Geodetic precession due to motion through gravity

The top panel of figure 4.3 illustrates the geodetic precession case. Two observers are fixed on the floor: one is located where a rod is launched upward from the floor and another is located where the rod again meets the upwardly accelerating floor. It is convenient to orient the rod at 45 degrees with respect to the floor—at this orientation, the two different Lorentz contractions of the rod seen in the instantaneous rest frames of the two observers produce identical angular change in the rod with the floor and the discussion is simplified. The instantaneous rest frame of the observer at the $t = 0$ event coincides with the master inertial frame, so the solid rod indicates the orientation measured in that observer's instantaneous rest frame. But the second observer is moving upward at speed of about $2v_y$ when the second meeting of rod and floor occurs. Therefore, the Lorentz transformation of times given in equation (4.4) must again be used to understand this latter event. At some time in the second observer's instantaneous rest frame for the meeting, the *time* Lorentz transformation measures a time difference for the rod's two ends as seen in the master inertial frame of $\delta t \cong 2v_y R/\sqrt{2}c^2$ with R being the length

of the rod. In this time interval, the right-hand end of the rod moves distance $\delta x \cong \sqrt{2} v_x v_y R/c^2$ further to the right, thereby decreasing the angle between the rod and the floor in amount $v_x v_y/c^2$. Dividing by the total time T between these events then yields the precession rate relative to the floor,

$$\dot{\Theta}_{\text{geodetic}} \cong \frac{1}{2}\frac{g v_x}{c^2}. \tag{4.6}$$

Since the observers in this case are at rest with respect to the source of gravity, this precession of the inertial rod must be explained as being due to the motion of that rod transversely through the gravitational field, i.e. *geodetic* precession.

4.4 General consideration of the observables

A rod with clock moves at constant velocity and without rotation through the master inertial frame as shown in figure 4.4. At $t = 0$, its lower end '1' leaves the floor (ground) which is upwardly accelerating. Expressed in the master inertial frame which, for convenience, is selected to coincide with the instantaneous rest frame of the floor at $t = 0$, the trajectories of the rod's two ends are

$$x_1(t) = v_x t \qquad \text{and} \qquad x_2(t) = x_1(t) + X \tag{4.7}$$

$$y_1(t) = v_y t \qquad \text{and} \qquad y_2(t) = y_1(t) + Y \tag{4.8}$$

with X, Y, v_x and v_y all positive [2]. The 'fixed ground' clocks have no horizontal motion and the common vertical motion

$$y(t) = \frac{c^2}{g}\left(\sqrt{1 + (gt/c)^2} - 1\right) \tag{4.9}$$

which manifests constant acceleration g as measured by accelerometers accompanying these clocks. The y-motion given in equation (4.9) catches up with $y_1(t)$ from equation (4.8) at master inertial frame time

$$T = \frac{2v_y}{g}\frac{1}{1 - v_y^2/c^2} \tag{4.10}$$

which event occurs at horizontal location

$$L = v_x T \tag{4.11}$$

with the floor moving upward at speed

$$V = \frac{2v_y}{1 + v_y^2/c^2} \tag{4.12}$$

as measured in the master inertial frame. The rod's vertical velocity relative to the floor, as measured in the rest frame of the floor at the reunion event, is obtained using the special relativistic transformation rule for velocities:

$$v'_y = \frac{v_y - V}{1 - v_y V/c^2} = -v_y$$

an unsurprising result. At this reunion event, the horizontal velocity of the rod as measured in the instantaneous rest frame of the ground is found to be equal to its original horizontal velocity, so the trajectory's locally measured arrival angle is the negative of the original locally measured departure angle.

In the instantaneous rest frame of the floor at reunion with the rod end '1', the master inertial frame event coordinates are

$$t'_1 = \gamma T(1 - v_y V/c^2) \qquad x'_1 = v_x T \qquad y'_1 = \gamma(v_y - V)T$$

with

$$\gamma = \frac{1}{\sqrt{1 - V^2/c^2}}.$$

In this frame, and at the moment its end '1' meets the floor, we also want to know where the rod's other end '2' is. From the time transformation of special relativity, we have

$$t'_1 = \gamma(t_2 - V(Y + v_y t_2)/c^2)$$

which gives

$$t_2 = T + Y\frac{V}{c^2 - v_y V}.$$

The location of rod end '2' at that moment is then

$$x'_2 = X + v_x\left(T + Y\frac{V}{c^2 - v_y V}\right)$$

$$y'_2 = \gamma\left(Y + (v_y - V)\left(T + Y\frac{V}{c^2 - v_y V}\right)\right).$$

The orientations of the rod at the two crossings of the floor can now be compared. Constructing the tangents of the angles the rod makes with the floor in the two instances, measured in each case in the floor's instantaneous rest frame,

$$\tan\phi = (y_2 - y_1)/(x_2 - x_1) = \frac{Y}{X}$$

$$\tan\phi' = (y'_2 - y'_1)/(x'_2 - x'_1) = \frac{Y}{X}\gamma\frac{c^2 - V^2}{c^2 + v_x V Y/X - v_y V}$$

the difference between these angles represents a change in the rod's orientation relative to the floor in a clockwise sense. In the limit of small vertical velocities of the rod, this rotation angle is

$$\delta\phi \simeq \frac{v_x v_y}{c^2}(1 - \cos 2\phi).$$

The $\cos 2\phi$ term of this expression is simply due to the change in the Lorentz contraction of the rod as its velocity components (with respect to the floor) have changed from (v_x, v_y) to $(v_x, -v_y)$. The remaining constant term of the expression is equivalent to a secular precession rate

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{1}{2}\frac{g v_x}{c^2} \tag{4.13}$$

which confirms the conclusion in section 4.3. The SREP requires this precession also to occur for an inertial rod which is on a free-fall trajectory in gravity.

How dramatic it would have been in the era 1907–11 when Einstein had still no theory of gravity but only his Equivalence Principle, if he had publically predicted not only that inertial frames are local, not global, and undergo free-fall acceleration in gravity but also that if these frames are moving non-radially in that gravity, they must rotate with respect to more distant inertial frames! It remained until just after Einstein's publication of his complete theory of general relativity for Willem deSitter in 1916 to discover by calculation the full *geodetic* precession contribution to the Moon's perigee rotation rate with respect to distant inertial space, one-third of which has here been shown to follow from the SREP [12].

Additional observables can be established by considering a number of clocks, some in free motion, some fixed in position on the upwardly accelerating ground floor and others moving at *constant proper speed* along the upwardly accelerating ground floor. Each of these clocks undergoes an interval of elapsed proper time which depends on its specific motion in the master inertial frame

$$\mathrm{d}\tau_{\mathrm{i}} = \sqrt{1 - v_{\mathrm{i}}(t)^2/c^2}\,\mathrm{d}t \tag{4.14}$$

with $\mathrm{d}t$ being the elapsed proper time increment of a clock at rest in the master inertial frame. Using the previously derived *master* time of reunion of the rod end '1' (also carrying a clock) with the ground, given by equation (4.10), this free clock on the rod records this reunion event at an elapsed proper time since launch

$$\begin{aligned}\tau(v_y, v_x)_{\mathrm{F}} &= T\sqrt{1 - (v_x^2 + v_y^2)/c^2} \\ &= \frac{2v_y}{g}\sqrt{1 - (v_x^2 + v_y^2)/c^2}\,\frac{1}{1 - v_y^2/c^2}.\end{aligned} \tag{4.15}$$

The trajectory of the fixed ground clock's trajectory is given by equation (4.9): integrating the proper time expression given by equation (4.14) then gives that clock's elapsed proper time between the launch event and the reunion event with the free-falling clock

$$\begin{aligned}\tau_{\mathrm{A}} &= \int_0^T \sqrt{1 - (\mathrm{d}y/\mathrm{d}t)^2/c^2}\,\mathrm{d}t = \frac{c}{g}\sinh^{-1}(gT/c) \\ &= \frac{c}{g}\sinh^{-1}\left(\frac{2v_y}{c}\frac{1}{1 - v_y^2/c^2}\right)\end{aligned} \tag{4.16}$$

which is independent of v_x, unlike the case for the elapsed proper time of the free (freely falling) clock.

A third type of clock permits an interesting variation on this experiment in which the same ground clock records both the launch from and reunion with the ground of the freely falling clock. This is achieved by giving that ground clock an initial velocity w to the right such that it arrives at horizontal location L simultaneously with the freely falling clock. Because of the upward acceleration of the ground, its horizontal velocity does not remain constant as seen in the master inertial frame: it moves according to

$$\mathrm{d}x/\mathrm{d}t = w\sqrt{1-(\mathrm{d}y/\mathrm{d}t)^2/c^2}$$

which, however, fulfils the requirement that no horizontal force acts on the clock

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}x/\mathrm{d}t}{\sqrt{1-v^2/c^2}} = 0$$

and that equal intervals of x are travelled per unit of proper time recorded on the horizontally moving clock. Since we want the simultaneous arrival of the *free-falling* clock and the clock moving along the ground, this requires

$$\int_0^T (\mathrm{d}x/\mathrm{d}t)\,\mathrm{d}t = w\int\sqrt{1-(\mathrm{d}y/\mathrm{d}t)^2/c^2}\,\mathrm{d}t = L$$

requiring an initial horizontal speed w which is greater than that of the freely falling clock

$$w = v_x\frac{gT/c}{\sinh^{-1}(gT/c)} = v_x\left(1+\frac{1}{6}\frac{g^2T^2}{c^2}+\cdots\right)$$

with T given in equation (4.10). The proper time of the reunion event as recorded by this moving clock is $\tau(w)_\mathrm{B} = \tau_\mathrm{A}\sqrt{1-w^2/c^2}$. Since w is in excess of v_x, in the frame of reference travelling to the right with this moving clock B, the freely falling clock is not launched vertically: it must instead be launched to the left of vertical (see top view in figure 4.5) at angle $\Theta \cong -2v_xv_y/3c^2$ (for non-relativistic speeds v_x and v_y) and, more generally, at an angle

$$\tan\Theta = \frac{v_x-w}{1-wv_x/c^2}\frac{\sqrt{1-w^2/c^2}}{v_y}.$$

These elapsed proper times, $\tau(v_y, v_x)_\mathrm{F}$, τ_A and $\tau(w)_\mathrm{B}$, and the horizontal location L of the reunion event from equation (4.11) are observables which must all be reproduced in the equivalent gravity environment if the SREP is to be fulfilled.

Some of these observables are relevant to the case in which the inertially moving rod is replaced by a light ray. Its trajectory in the master inertial frame is

$$x = ct\cos\phi \qquad y = ct\sin\phi.$$

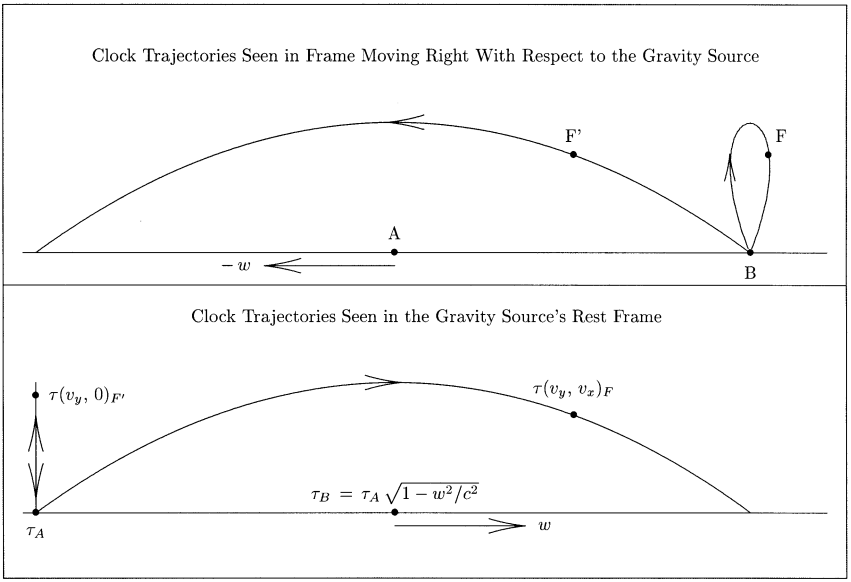

Figure 4.5. Clock trajectories seen in two frames of reference. Freely falling and ground clocks, marked F, F′, A, B, in gravity are shown from two frames of reference—the lower viewpoint is at rest with respect to the gravity source and the upper viewpoint moves to the right at speed w (the source of gravity moves to left at speed w). In the frame in which gravity's source is at rest, clock F′ is launched into vertical free fall and ground clock A waits at rest for the reunion. In the same frame, clock F is on a free-falling trajectory which moves to the right, and clock B moves on the ground at constant velocity to the right to meet the return of F to the ground. Proper times at the various reunions of these clocks and other related observables are calculated in the gravity-free but accelerating *ground floor* situations; and fulfilment of the SREP requires that those results must all be reproduced in each of these two illustrated situations in gravity (the four clock times are shown in the lower view). This specifies modifications of the gravitational equations of motion when these equations are stated in the frame moving with respect to the source of the gravity, including *gravitomagnetic* terms which are proportional to the velocity of the gravity's source.

The time recorded by a clock at rest in the master inertial frame for the reunion of the light ray with the ground is then

$$T' = \frac{2c}{g}\frac{2\sin\phi}{(\cos\phi)^2}$$

which then determines the elapsed proper time for the ground clock at reunion with the light ray

$$\tau_{\mathrm{C}} = \frac{c}{g} \sinh^{-1}\left(\frac{2\sin\phi}{\cos^2\phi}\right). \tag{4.17}$$

This proper time is simply obtainable from equation (4.16) by taking the limit of a free body which travels at the speed of light. The second observable is the light-ray launch angle which results in a horizontal location for the reunion of light ray and the ground equal to that for the rod:

$$\tan\phi = \frac{gL}{2c^2}.$$

There is, of course, no elapsed proper time for a 'clock on a light ray'.

4.4.1 Moving gravity source

Trajectories of these various clocks and the light ray in the gravity environment are shown in figure 4.5. The lower picture gives the scene in the rest frame of the gravity source and the *at rest* ground clocks. The upper picture gives the scene in the frame of the clock which moves to the right so as to record the reunion of the freely falling clock launched to the right. When the unusual motion of the freely falling clock in the upper scene is also required to occur in gravity, additional *gravitomagnetic-like* acceleration terms to the freely falling clock's equation of motion are required which are in proportion to the motion of the gravity source in that frame of observation.

Another consequence of performing measurements in the frame moving with the ground clocks at speed w is a change in the measured local value of gravitational acceleration. Since this moving clock will experience an elapsed proper time smaller than that of the ground clocks at rest,

$$\tau(w)_{\mathrm{B}} = \tau_{\mathrm{A}}\sqrt{1 - w^2/c^2} \tag{4.18}$$

and the vertical speed of the launched freely falling body is enhanced as measured in this frame (time dilation),

$$v'_y = \frac{1}{\sqrt{1 - w^2/c^2}} v_y$$

observers accompanying the horizontally moving ground clocks record a local gravitational acceleration of

$$g(w) = \frac{1}{1 - w^2/c^2} g \cong (1 + w^2/c^2)g.$$

The SREP's enforcement of this will require further modifications of the gravitational equation of motion when expressed in frames in which the source of gravity is moving.

A clock launched vertically in the frame which is not horizontally moving relative to gravity's source can also be viewed from the rest frame of the ground clocks which travel to the right. The elapsed proper time for the freely falling clock in that case is obtained from equation (4.15) with $v_x = 0$

$$\tau(v_y, 0)_{\mathrm{F}'} = \frac{2v_y}{g}\frac{1}{\sqrt{1 - v_y^2/c^2}}.$$

In the frame of reference moving horizontally along the ground at speed w, this situation is seen as a clock launched into gravitational free fall and moving to the left, along with another clock also moving to the left along the ground such that it meets the freely falling clock at its reunion with the ground. A further *gravitomagnetic* acceleration term will be required to obtain equivalent observational outcomes when this situation is considered in gravity.

4.5 Requirements for equivalent predictions in gravity

All the phenomena and situations considered in the preceding sections must be considered again in an environment of real gravitational acceleration g as measured on the ground. The outcomes for all the observables previously obtained by kinematical calculations in gravity-free space must be reproduced under identical arrangements in the gravity environment if the SREP is valid. To achieve this, $1/c^2$ order gravitational corrections to the equations of motion for freely falling bodies, to the expression for the *proper* tick rates of clocks in gravity and to the speed of light function are required [3]. Expressing each of these three equations in terms of a proper time variable τ which represents the elapsed time of clocks at rest, the modified rate for clocks in general motion and at general altitude above the ground is assumed to be

$$\mathrm{d}\tau(\boldsymbol{r}, \boldsymbol{v}) = \mathrm{d}\tau\left(1 - \frac{1}{2}\frac{v^2}{c^2} + a_1\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right). \tag{4.19}$$

The equation of motion for bodies freely falling in the gravity is assumed to be

$$\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}\tau^2} = \boldsymbol{g}\left(1 + a_2\frac{v^2}{c^2} + a_3\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2} + a_5\frac{(\hat{g}\cdot\boldsymbol{v})^2}{c^2}\right) + a_4\frac{\boldsymbol{g}\cdot\boldsymbol{v}\boldsymbol{v}}{c^2} \tag{4.20}$$

with $\boldsymbol{v} = \mathrm{d}\boldsymbol{r}/\mathrm{d}\tau$. And the light speed function in gravity is assumed to be

$$c(\boldsymbol{r}) = \frac{|\mathrm{d}\boldsymbol{r}|}{\mathrm{d}\tau} = c\left(1 + a_6\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right). \tag{4.21}$$

Values for the numerical coefficients in these three equations, $a_1 \ldots a_6$, are sought so that the observables previously obtained kinematically in gravity-free inertial

space are reproduced in the corresponding situation in gravity. A unique solution will result.

A freely falling clock is launched with the same initial velocity used previously—(v_x, v_y). The horizontal equation of motion from equation (4.20) is first considered:

$$\frac{\mathrm{d}^2x}{\mathrm{d}\tau^2} = -a_4\frac{g}{c^2}\frac{\mathrm{d}y}{\mathrm{d}\tau}\frac{\mathrm{d}x}{\mathrm{d}\tau}.$$

Since the right-hand side of this equation is already proportional to $1/c^2$, the Newtonian trajectory for the freely falling body

$$x(\tau) = v_x\tau \qquad y(\tau) = v_y\tau - \tfrac{1}{2}g\tau^2$$

can be employed in its evaluation. Integrating this horizontal equation of motion, demanding that the distance given in equation (4.11) is reached at the proper time given by equation (4.15):

$$\frac{2v_xv_y}{g}\frac{1}{1-v_y^2/c^2} = v_x\tau_G + \int_0^{\tau_G}\mathrm{d}\tau\int_0^{\tau}\frac{\mathrm{d}^2x}{\mathrm{d}\tau^2}\,\mathrm{d}\tau \qquad \text{to order } 1/c^2$$

requires

$$a_4 = -2.$$

The vertical equation of motion from equation (4.20) is now considered:

$$\frac{\mathrm{d}^2y}{\mathrm{d}\tau^2} = -g\left(1 + a_2\frac{(\mathrm{d}y/\mathrm{d}\tau)^2 + (\mathrm{d}x/\mathrm{d}\tau)^2}{c^2} - a_3\frac{gy}{c^2} + (a_5-2)\frac{(\mathrm{d}y/\mathrm{d}\tau)^2}{c^2}\right) \tag{4.22}$$

in which the result for a_4 has been incorporated. Since the proper time for the reunion of clock and ground as recorded by the ground clock is given by equation (4.16) and is independent of the horizontal speed of the body. This result can only emerge when solving equation (4.22) if

$$a_2 = 0.$$

The remaining dimensionless coefficients in equation (4.20) are fixed by using the Newtonian motion on the right-hand side, integrating from the initial vertical position 0 and speed v_y, and requiring both the return of the freely falling clock to the ground and the reversal of the vertical velocity to $-v_y$ to occur at time τ_A. This yields

$$a_3 = -1 \qquad a_5 = 0.$$

The proper time of the reunion with the ground as recorded by the freely falling clock is obtained by integrating the clock rate expression given in equation (4.19). Demanding that the result be equal to the kinematically derived amount given in equation (4.15) yields the value of the 'red-shift' coefficient in equation (4.19)

$$a_1 = -1.$$

It should be noted that this derivation of the gravitational 'red-shift' of clock rates did not employ light-ray propagation between differently located clocks. Combining these results for the clock rate expression and the equation of motion expression, their coefficients now determined, the locally measured acceleration rate for a body instantaneously at rest is found to be dependent on altitude

$$g(y)_{\text{local}} = g\left(1 - \frac{gy}{c^2}\right).$$

In the limit of small initial elevation angles, light rays move in gravity along the curves

$$y(x) = \frac{g}{2c^2}x(x_\gamma - x).$$

The proper elapsed time of ground clocks for the reunion of the light ray with the ground has already been determined and is given in equation (4.17). Demanding this same elapsed proper time in gravity, the light-ray speed function is assumed and integration over the light trajectory is performed to obtain the total elapsed time. Corrections to the light trajectory of order $1/c^2$ need not be considered as they will generate only $1/c^4$ order corrections to the result. Therefore,

$$\tau_{\text{C}} = \int_{(0,0)}^{(x_\gamma,0)} \frac{\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}}{c(\boldsymbol{r})}.$$

This is fulfilled for the coefficient value

$$a_6 = -1$$

which appears in the light speed function, equation (4.21). Combining this result with the clock rate expression given by equation (4.19), the locally measured speed of light is found to be independent of altitude in gravity.

In conclusion: *In a frame of reference at rest with respect to a source of gravity which locally (at the ground) produces a gravitational acceleration g and speed of light c as measured by clocks at rest on the ground, then the equivalence of all local phenomena to that which occurs in an accelerated but force-free environment requires the following* $1/c^2$ *order modifications to the local equations of motion for bodies, clocks and light* [4]

$$\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}\tau^2} = \boldsymbol{g}\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) - 2\frac{\boldsymbol{g}\cdot\boldsymbol{v}\boldsymbol{v}}{c^2} \tag{4.23}$$

$$\mathrm{d}\tau(\boldsymbol{r},\boldsymbol{v}) = \mathrm{d}\tau\left(1 - \frac{1}{2}\frac{v^2}{c^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) \tag{4.24}$$

$$c(\boldsymbol{r}) = c\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right). \tag{4.25}$$

4.5.1 Geometrical interpretation

This body equation of motion is obtainable from the particle Lagrangian

$$L = \frac{1}{2}v^2\left(1 + \frac{1}{4}\frac{v^2}{c^2}\right) + \boldsymbol{g}\cdot\boldsymbol{r}\left(1 + \frac{1}{2}\frac{v^2}{c^2}\right)$$

which, to the exhibited $1/c^2$ order, is equivalent to a *geometrical* least action principle

$$\delta A = 0 = \delta\int\sqrt{g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu} \qquad \mu, \nu = t, x, y, z \tag{4.26}$$

with the dominant *time–time* component of the metric tensor being altered from the Minkowski metric value $\eta_{tt} = 1$, while the spatial values remain unchanged; $\eta_{xx} = \eta_{yy} = \eta_{zz} = -1/c^2$

$$g_{tt} \;=\; (1 - \boldsymbol{g}\cdot\boldsymbol{r}/c^2)^2.$$

The light speed function given by equation (4.25) then follows from the *null-geodesic* assumption

$$g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu = 0 \qquad \text{for light} \tag{4.27}$$

and the clock rate equation (4.24) is the Lagrangian invariant

$$\mathrm{d}\tau = \sqrt{g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu}.$$

4.5.2 Moving gravity source

An equation of motion for freely falling bodies which is valid in more general frames in which the source of the gravity moves would be informative. In this situation, additional acceleration terms must be considered which are functions of the gravity source's velocity $\boldsymbol{v}_\mathrm{s}$. By considering the previous phenomena from a reference frame which moves at constant velocity along the ground, three such terms can be determined:

$$\delta\left(\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}\tau^2}\right) = \frac{1}{c^2}(a_7\boldsymbol{g}\cdot\boldsymbol{v}\boldsymbol{v}_\mathrm{s} + a_8\boldsymbol{v}\cdot\boldsymbol{v}_\mathrm{s}\boldsymbol{g} + a_9 v_\mathrm{s}^2\boldsymbol{g}). \tag{4.28}$$

Because the source velocity $\boldsymbol{v}_\mathrm{s}$ is orthogonal to the local gravity direction in these situations, a number of other possible acceleration terms proportional to $\boldsymbol{g}\cdot\boldsymbol{v}_\mathrm{s}$ are not brought into play and so remain undetermined by these SREP arguments.

In the case of the clock originally launched up and to the right, with a ground clock following along the ground so as to arrive at the reunion of clock with ground, we recall that in the frame which follows the ground clock, the freely falling clock was launched not vertically upward but at an angle to the vertical of

$\Theta \cong -2v_x v_y/3c^2$ (to leading order in $1/c^2$). As illustrated in figure 4.5, it then moved on a closed trajectory which finished at its starting point on the ground. Demanding this outcome from the x-component of our body's equation of motion with the source of gravity moving to the left, then requires

$$-\frac{v_x v_y}{c^2}\frac{2v_y}{g} + a_7 \int_0^{2v_y/g} \mathrm{d}t \int_0^{2v_y/g} \frac{wg}{c^2}(v_y - gt)\,\mathrm{d}t \cong 0$$

with $w \cong v_x$. This requires $a_7 = 2$. And, as previously indicated, the vertical acceleration in this frame is not g, it is $g(w) = g(1 + w^2/c^2)$ which requires $a_9 = 1$.

If the case of the clock vertically launched in the original frame is now considered in the frame moving to the right at speed w, the vertical speed with which it was launched is $v_y(1 - w^2/2c^2)$ to lowest order in $1/c^2$, while the total proper time for the ground clock travelling to the left to meet the freely falling clock at reunion with the ground is as given in equation (4.18). Since this moving clock's proper time runs at a rate slower than that of the ground clocks at rest in this frame by the factor

$$\frac{\mathrm{d}\tau}{\mathrm{d}\tau(w)} = \sqrt{1 - w^2/c^2} \cong 1 - \frac{1}{2}\frac{w^2}{c^2}$$

the vertical acceleration of the freely falling clock must be $g(1 - w^2/c^2)$ to lowest order in w^2. This fixes the final coefficient in equation (4.28) to be $a_8 = -2$.

The entire equation of motion, equation (4.28) plus the contributions from equation (4.23), is then

$$\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}\tau^2} = \boldsymbol{g}\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2} + \frac{v_s^2}{c^2}\right) - \frac{2}{c^2}\boldsymbol{g}\cdot\boldsymbol{v}\boldsymbol{v} + \frac{2}{c^2}\boldsymbol{v}\times(\boldsymbol{v}_s\times\boldsymbol{g}). \tag{4.29}$$

A moving gravity source also changes the speed of light function. A Lorentz transformation to the frame travelling to the right at speed w relates the launch angle of the light ray which will be seen in this frame to the original launch angle

$$\tan\phi' = \frac{\sin\phi\sqrt{1 - w^2/c^2}}{\cos\phi - w/c}$$

or for small angles

$$\phi' \cong \phi(1 + w/c).$$

But the maximum height above the ground which the light ray reaches is unchanged by this transformation and is given approximately by

$$h \cong \frac{1}{2}\phi^2\left(\frac{\mathrm{d}c}{c\mathrm{d}y}\right)^{-1}.$$

The gravitomagnetically modified light speed function

$$c(\boldsymbol{r}) = c\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}(1 - 2\hat{c}\cdot\boldsymbol{v}_\mathrm{s}/c)\right) \tag{4.30}$$

is required to achieve this equivalent result: $\hat{c}$ is the unit vector in the direction of light propagation and again $\boldsymbol{v}_\mathrm{s}$ is the velocity of the source of gravity.

These SREP results for a moving source are in agreement with what one obtains by applying a Lorentz transformation to the metric field previously found in the gravity source's rest frame. From the transformation rule for a second rank tensor,

$$g'_{\mu\nu} = \sum_\alpha \sum_\beta \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu}\, g_{\alpha\beta}$$

and the lowest-order expression of the Lorentz transformation,

$$\begin{aligned} \boldsymbol{r} &\cong \boldsymbol{r}' - \boldsymbol{v}_\mathrm{s}t' \\ t &\cong t' - \boldsymbol{v}_\mathrm{s}\cdot\boldsymbol{r}'/c^2 \end{aligned}$$

a spatial vector of (mixed time–space) metric components is obtained:

$$g'_{i0} = g'_{0i} \cong 2\boldsymbol{g}\cdot\boldsymbol{r}(\boldsymbol{v}_\mathrm{s})_i/c^4 \qquad \text{components } i = x, y, z$$

which, when inserted into the action principle given by equation (4.26), generates the new Lagrangian term

$$\delta L = -2\boldsymbol{g}\cdot\boldsymbol{r}\boldsymbol{v}_\mathrm{s}\cdot\boldsymbol{v}/c^2. \tag{4.31}$$

4.6 Periastron precession

Just about any modification from an inverse square central acceleration law causes the major axis of Keplerian orbits to precess in inertial space. This holds, in particular, for the modifications to the equation of motion which result from the SREP as given by equation (4.23). Consider a body which is close to being in a circular orbit around a central body. Small perturbations are considered from the mean circular motion so that the time evolution of the *eccentric* deviations from circularity can be derived and compared to the mean orbital motion. Starting with the radial and tangential equations of motion

$$\begin{aligned} \frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} &= \boldsymbol{g}(\boldsymbol{r}, \boldsymbol{v})\cdot\hat{r} + \omega^2 r \\ \frac{\mathrm{d}}{\mathrm{d}\tau}(r^2\omega) &= r\boldsymbol{g}(\boldsymbol{r}, \boldsymbol{v})\cdot\hat{t} \end{aligned}$$

small perturbations are considered about a circular orbit, $r \to r + x(\tau), \omega \to \omega + \delta\omega(\tau)$. The needed acceleration components from equation (4.23) are

$$\begin{aligned} \boldsymbol{g}\cdot\hat{r} &= -g(1 + gx/c^2) \\ \boldsymbol{g}\cdot\hat{t} &= \frac{2gv}{c^2}\frac{\mathrm{d}x}{\mathrm{d}\tau} \end{aligned}$$

with v being the horizontal velocity of the mean circular orbit. The linearized equation for radial perturbation $x(\tau)$ then becomes

$$\frac{\mathrm{d}^2x}{\mathrm{d}\tau^2} + \left(3\omega^2 + \frac{\mathrm{d}g}{\mathrm{d}r} - \frac{3g^2}{c^2}\right)x = 0$$

with the radial tidal gradient of the solar system's total acceleration field $\mathrm{d}g/\mathrm{d}r$ added, and the relationship $g = v\omega$ being used to simplify the radial and tangential $1/c^2$ order perturbation terms. The resulting radial perturbation is simple *eccentric* harmonic motion with arbitrary amplitude and phase determined by initial conditions. But this eccentric motion's frequency ω_o is specific, and relative to the orbital frequency, it is shifted by the SREP modifications of the dynamics to be slightly less than what it will be due solely to the tidal gradient $\mathrm{d}g/\mathrm{d}r$. This increased frequency difference between orbital and eccentric motions appears in space as an addition to the total precession rate of the orbit's major axis in the positive sense of the orbital motion (prograde precession), and of amount

$$\delta(\omega - \omega_\mathrm{o}) \cong \frac{3}{2}g^2/(\omega c^2) \cong \frac{3}{2}\frac{v^2}{c^2}\omega.$$

Prior to Einstein's development of his special relativity theory in 1905 and the formulation of his EP beginning in 1907, a century of astronomical observations had already discovered about a 43 arcsec/century precession rate for Mercury's orbit in excess of what could be understood from consideration of the Newtonian perturbations by the other known planets in the solar system. Half this anomalous precession is here accounted for from the SREP:

$$\left(\frac{3}{2}\frac{v^2}{c^2}\omega\right)_\mathrm{Mercury} \cong 22.5 \text{ arcsec/century.} \tag{4.32}$$

4.6.1 A historical speculation

As early as December 1907, Einstein mentioned in a letter to a friend that, 'I am now occupied with a relativistic treatment of the law of gravity, with which I hope to explain the anomalous secular change in the perihelion of Mercury.' And he added in a footnote, 'Up to now the thing doesn't appear to want to succeed' [5]. Had Einstein arrived at the SREP's prediction, equation (4.32), about this time? By then he certainly was in a position to extend his EP to a full SREP. Perhaps he had done so but chose not to publish the consequences of a full special relativistic generalization of his principle because this perihelion prediction was only half the *known* anomaly in Mercury's orbital motion? Yes, his prediction of light deflection from the EP was also only half that which would eventually emerge from his complete gravity theory of 1915/16 but in 1907 neither the full theory's prediction for light deflection nor its experimental measurement during the eclipse of 1919 were available to create a conflict.

However, continued work toward a complete relativistic theory of gravity may have been spurred on by such an anomalous early EP-inspired estimate which produced contributions to Mercury's perihelion precession rate with magnitude being a simple fraction of the observed anomaly of 43 arcsec/century. From several letters from Einstein to colleagues written around the end of 1915, Einstein mentioned that one of the things which had kept him searching right up until the end for a better metric tensor theory of gravity was that his 'old theory' only explained half Mercury's anomalous perihelion precession. And then when he recalculated this effect in November 1915 using the new vacuum field equations of his final metric tensor theory of general relativity and did obtain the full anomaly 'without any special hypothesis', he mentioned in another letter that this produced one of the strongest emotional experiences of his career: 'for a few days I was beside myself with joyous excitement'. It appears clear that the Mercury orbit anomaly played a continuous and key role in Einstein's search for a new theory of gravity. Many narratives of this scientific revolution seem to have minimized this part of the story and the focus on the later confirmation of the theory with the measurement of the deflection of light during the 1919 eclipse further overshadowed the perihelion precession phenomenon.

4.7 Summary

Incorporating the special relativity theory more fully into Einstein's principle of equivalence between the phenomena in accelerated frames of reference and that in local gravitational fields has led to the prediction of a number of additional effects in post-Newtonian gravity. These include *geodetic precession* of local inertial frames which follow non-radial, free-falling trajectories through gravity, precession of Mercury's perihelion and *gravitomagnetic* forces between matter proportional to the velocities of both source matter and acted-upon matter, as well as *gravitomagnetic* precession. And the original predictions of Einstein's EP are, of course, also predicted—universal reduction of clock rates and both deflection and slowing of light in gravity.

The SREP predictions do not generally account for the entire physical effects which are now routinely measured by experiments. Within the general class of locally Lorentz-invariant, complete metric theories of gravity—all of which fulfil the SREP—a variety of calculated post-Newtonian gravitational effects are now listed and expressed in terms of two dimensionless parameters, γ and $\beta^* = 2\beta - 1$, which identify and quantify the post-Newtonian features of the metric theories which go beyond the local physics specified by the Equivalence Principles.

$$d\tau = dt\left(1 - \frac{1}{2}\frac{v^2}{c^2} - \mathbf{1}\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right)$$

$$c(\boldsymbol{r}) = c\left(1 - (\mathbf{1}+\gamma)\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right)$$

$$\boldsymbol{\Omega}_{\rm geo} = (\mathbf{1/2} + \gamma)\frac{\boldsymbol{g} \times \boldsymbol{v}}{c^2}$$

$$\Omega_{\rm Merc} = (\mathbf{3/2} + 2\gamma - \beta^*/2)\frac{v^2}{c^2}\omega$$

$$\boldsymbol{a}_{\rm grav-mag} = (\mathbf{2} + 2\gamma)\frac{\boldsymbol{v} \times (\boldsymbol{g} \times \boldsymbol{v}_{\rm s})}{c^2}$$

$$\frac{M_{\rm G}}{M_{\rm I}} = 1 + (\mathbf{1} + \gamma - 2\beta^*)\frac{1}{2Mc^2}\int \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|}\,{\rm d}^3x\,{\rm d}^3y.$$

SREP contributions are shown in bold numbers. These parameterized post-Newtonian (PPN) expressions for different (albeit theoretically connected) gravitational effects have been known for decades [8–10]; indeed, it was my awareness of the contributions to these several phenomena which were independent of the specifics of the particular metric theory that motivated this investigation. It is the SREP which dictates these universal contributions to post-Newtonian gravity.

Acknowledgment

This work was performed with support by National Aeronautics and Space Administration contract NASW-00011 and grant NAG8-1811.

Appendix

Beginning with an underlying metric field theory of gravity which is locally Lorentz-invariant, a $1/c^2$ order, N-body Lagrangian can generally be derived. The part of this Lagrangian which is independent of the specifics of the metric theory and which manifests both local Lorentz invariance and the EP is

$$\begin{aligned} L_{\rm SREP} = &\sum_i \left(\frac{1}{2}m_i v_i^2 + \frac{1}{8c^2}m_i v_i^4\right) \\ &+ \frac{G}{2}\sum_{i,j}\frac{m_i m_j}{r_{ij}}\left(1 - \frac{1}{2c^2}(\boldsymbol{v}_i \cdot \boldsymbol{v}_j + \boldsymbol{v}_i \cdot \hat{\boldsymbol{r}}_{ij}\hat{\boldsymbol{r}}_{ij} \cdot \boldsymbol{v}_j)\right) \\ &+ \frac{G}{4c^2}\sum_{i,j}\frac{m_i m_j}{r_{ij}}(\boldsymbol{v}_i - \boldsymbol{v}_j)^2 \end{aligned}$$

with the first line by itself being Lorentz-invariant to $1/c^2$ order but the additional Lorentz-invariant term on the second line being also needed in order to fulfil the EP. Focusing on one of the N-bodies in the presence of $N - 1$ other quasi-static sources of gravity seen by the selected body, one can expand this Lagrangian about a chosen origin, rescale the time variable into the proper time variable at this origin and then reproduce the SREP-derived equation of motion given

by equation (4.23). Giving the source bodies motions $\boldsymbol{v}_s$, the SREP-derived gravitomagnetic equation of motion corrections found in equation (4.29) can also be obtained.

But there is additional $1/c^2$ order gravitational physics beyond the SREP. It results from two Lagrangian terms

$$L_\gamma = \gamma \frac{G}{2c^2} \sum_{i,j} \frac{m_i m_j}{r_{ij}} (\boldsymbol{v}_i - \boldsymbol{v}_j)^2$$

$$L_{\beta^*} = -\beta^* \frac{G^2}{2c^2} \sum_{i,j,k} \frac{m_i m_j m_k}{r_{ij} r_{ik}}$$

with the indices i, j, k each being summed over the N bodies [11]. The two new coupling strength parameters have special values in general relativity, $\gamma_{GR} = 1$ and $\beta^*_{GR} = 1$ ($\beta = (1+\beta^*)/2$ is the more traditional PPN, *Eddington* coefficient) [7] but they individually have different values in scalar–tensor metric theories, for example. In addition to contributing to additional gravitomagnetic interaction, the Lagrangian term L_γ produces a global non-Euclidean geometry for the arena of physical events and objects. But locally this deviation from the Euclidean nature of space can be *delayed*. At a chosen locality $\boldsymbol{r}_0$, a sequence of spatial coordinate transformations involving first a rescaling of the spatial coordinates

$$\boldsymbol{x}' = (1 + \gamma U(\boldsymbol{r}_0)/c^2)\boldsymbol{x}$$

with $U(\boldsymbol{r}_0)$ being the Newtonian potential at $\boldsymbol{r}_0$ of the gravitational sources, and $\boldsymbol{x} = \boldsymbol{r} - \boldsymbol{r}_0$, and then the nonlinear *warping* of the coordinates

$$\boldsymbol{x}' = \boldsymbol{\rho} + \frac{1}{2c^2}\gamma \boldsymbol{g} \rho^2 - \frac{1}{c^2}\gamma \boldsymbol{g} \cdot \boldsymbol{\rho}\boldsymbol{\rho}$$

the locality only experiences the onset of non-Euclidean spatial effects at the quadratic order in laboratory size. The nonlinear Lagrangian term $L_{\beta*}$ produces three-body gravitational interactions and it also produces modifications to the gravitational potential between two bodies whose strength is proportional to the square of one mass or the other and depends on the inverse square of body separation. Neither of these two Lagrangian terms can be inferred by SREP arguments: a full field theory of gravity is required for their specification.

References

[1] Because the rod is moving essentially straight up or straight down in the instantaneous rest frames of the observer at each end of the rod's free flight, a horizontal rod experiences no reorientations due to Lorentz contractions; this simplifies the analysis of the gravitomagnetic precession.

[2] $\sqrt{X^2 + Y^2}$ is not the *proper* length of the rod; the coordinates X, Y record the Lorentz-contracted position of the rod's end '2' at the instant that end '1' has

coordinates 0, 0, and as seen from a frame in which the rod moves with velocity components (v_x, v_y).

[3] Several derived results for observables calculated in the accelerating, gravity free laboratory have been done exactly within special relativity (to all orders in v/c and gt/c). They could therefore be ingredients for SREP predictions in ultra-relativistic contexts. In this paper, however, these expressions are approximated to only the necessary order for being equated to the corresponding expressions for the same observables calculated to $1/c^2$ order in gravity.

[4] Some have argued that the deflection of light prediction of the EP just reproduces an old Newtonian calculation for a 'particle' travelling initially at speed of light through gravity. This is a hasty conclusion. Simple Newtonian acceleration would imply that the 'particle' speeds up during its passage by a gravitating body, acquiring speed $v^2 \simeq c^2 + 2Gm/r$ when at distance r from body of mass m and therefore *reducing* its transit time-of-flight through gravity. But the SREP predicts that light's speed, measured by clock at some fixed position, slows down as $v^2 \simeq c^2 - 2Gm/r$. Globally, general relativity effectively doubles the light speed reduction near gravitating bodies as a result of the *curvature of space* produced in the complete theory. This 'Shapiro time delay' (see 'Fourth test of general relativity' *Phys. Rev. Lett.* **13** 789), is now routinely measured in solar system radar ranging experiments.

[5] I thank Professor John Stachel for bringing this correspondence to my attention; it is found in *The Collected Papers of Albert Einstein, Vol 5, The Swiss Years: Correspondence 1902-1914* ed Klein M J *et al* 1993 (Princeton, NJ: Princeton University Press) p 82. See also, Pais A 1982 *'Subtle is the Lord': The Science and the Life of Albert Einstein* (New York: Oxford University Press) p 182.

[6] Lense J and Thirring H 1918 *Phys. Z.* **19** 156

[7] Eddington A 1920 *Space, Time and Gravitation* (First Harper Torchbook edition 1959) p 126 and appendix note 10

[8] Nordtvedt K 1968 *Phys. Rev.* **169** 1017

[9] Nordtvedt K 1972 *Science* **178** 1157

[10] Will C M 1993 *Theory and Experiment in Gravitational Physics* revised edn (New York: Cambridge University Press)

[11] Nordtvedt K 1985 *Astrophys. J.* **297** 390

[12] de Sitter W 1916 *Mon. Not. R. Astron. Soc.* **77** 155

Chapter 5

Lunar laser ranging: a comprehensive probe of post-Newtonian gravity

Kenneth Nordtvedt
Northwest Analysis, 118 Sourdough Ridge, Bozeman, MT 59715, USA

5.1 Introduction

The precise fit of the lunar laser ranging (LLR) data to theory yields a number of the most exacting tests of Einstein's field theory of gravity, general relativity, because almost any alternative theory of gravity predicts a number of changes (from that produced by general relativity) in the lunar orbit which would be readily detected in the LLR data. Some of the most interesting and fundamental of such theory-dependent effects and which are particularly well measured by LLR include (1) a difference in the free-fall rate of the Earth and Moon toward the Sun due to gravity theory's nonlinear structure acting on the gravitational binding energy within the Earth, (2) a time variation in Newton's gravitational coupling parameter, $G \rightarrow G(t)$, related to the expansion rate of the universe and (3) precession of the local inertial frame (relative to distant inertial frames) because of the Earth–Moon system's motion through the Sun's gravity.

Measurements of the round-trip travel times of laser pulses between Earth stations and sites on the lunar surface have been made on a frequent basis ever since the Apollo 11 astronauts placed the first passive laser reflector on the Moon in 1969. Today about 15 000 such range measurements are archived and available for use by analysis groups wishing to fit the data to theoretical models for the general relativistic gravitational dynamics of the relevant bodies, the speed of light function in the solar system, tidal distortions of Earth and Moon, atmospheric corrections to light propagation, etc. An individual range measurement today has a precision of about a centimetre (one-way) but a new generation of observing program plans to improve this range measurement precision down to a millimetre.

Because of the large number of range measurements, some of the key length parameters which describe the lunar orbit are already estimated with precision of a few millimetres, and key lunar motion frequencies to fractional precisions of a few parts in 10^{12}.

Because both the Earth's mass and that of the Moon are sufficiently large, the orbits of these bodies can be modelled as single orbital 'arcs' extending over three decades through time. The complete model used to fit the many range measurements contains in excess of a hundred parameters, P_m, which are optimally adjusted from their nominal model values $P_m^{(o)}$ by amounts $\delta P_m = P_m - P_m^{(o)}$ determined in a weighted least-squares fit type procedure:

$$\text{Minimize} \qquad \sum_{i,j=1}^{N} W_{ij} \sum_{m,n=1}^{M} [f(m)_i \delta P_m - r_i][f(n)_j \delta P_n - r_j]$$

with the N range measurements being identified by the labels i and j and the M model parameters being identified by the labels m and n. W_{ij} are the weightings given to each measurement (pair) and are usually taken to be diagonal in ij and inversely proportional to the square of inferred measurement errors; the *residuals* r_i are the differences between observed and calculated range values, $r_i = R_{\text{obs}}(t_i) - R_{\text{calc}}(t_i)$; and the remaining functions $f(m)_i$ are the parameter *partials* which give the sensitivity of the modelled (calculated) range to change in each model parameter value

$$f(m)_i = \frac{\partial R_{\text{calc}}(t_i)}{\partial P_m}$$

evaluated at the time t_i of the ith range measurement.

Among the very many model parameters, the information needed for testing relativistic gravity theory is concentrated in only a handful of orbital features. The needed orbital parameters are connected with four key oscillatory contributions to the lunar motion, the *eccentric, evective, and variational* motions and the *parallactic inequality*, which are illustrated in figure 5.1. The eccentric motion produces an oscillatory range contribution proportional to $\cos(A)$, A being the anomalistic (eccentric) phase and is a natural and undriven perturbation of circular motion. The *variation* is driven by the Sun's leading order quadrupolar tidal field and produces a range contribution proportional to $\cos(2D)$, D being the synodic phase from the *new moon*. The *parallactic inequality* is driven by the Sun's next-order octupolar tidal field and its range perturbation $\cos(D)$ has a monthly period. The *evection* is a hybrid range perturbation proportional to the eccentric motion as modified by the *variation* and having a time dependence $\cos(2D - A)$. The eccentric and evective motions, which alter the times of eclipses, were discovered by the ancients: the variation and parallactic inequalities, which do not alter the times of eclipses, were only found during and after the era of Newton.

The amplitude of the parallactic inequality, L_{PI}, is unusually sensitive to any difference in the Sun's acceleration rate of the Earth and Moon [2]. The

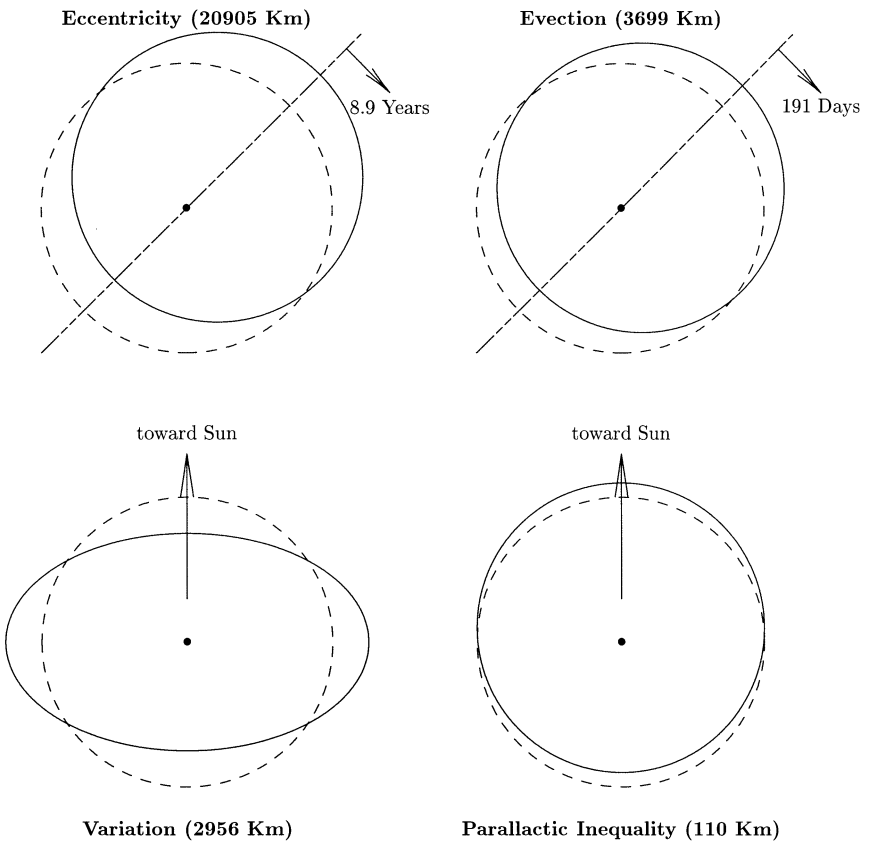

Figure 5.1. Lunar orbit's four main perturbations. Four lunar orbit perturbations from a nominal circular orbit (dotted) are shown. They produce oscillatory Earth–Moon range terms: the eccentric oscillation $\sim \cos(A)$, the variation oscillation $\sim \cos(2D)$, the parallactic inequality oscillation $\sim \cos(D)$, and the evective oscillation $\sim \cos(2D - A)$, with respective amplitudes indicated. Key tests of general relativity are achieved from precise measurements of amplitudes or phase rates of these perturbations. Measurement of the amplitude of the parallactic inequality determines whether Earth and Moon fall toward the Sun at same rate. Measurements of the synodic phase D and anomalistic (eccentric) phase A rates and rate of change of these rates determine the deSitter precession of the lunar orbit and time rate of change of Newton's G.

frequency of the eccentric motion, the *anomalistic* frequency $\dot{A}$, when compared to other lunar frequencies determines the precession rate of the Moon's perigee. This rapid precession, which completely rotates the orbit's major axis in about 8.9 yr, is primarily driven by the Sun's tidal acceleration but there is a leading order relativistic contribution to this precession rate interpreted as an actual

rotation of the local inertial frame, the *de Sitter precession*, resulting from motion through the Sun's field of gravity. From measurement of the time rates of change of the Moon's *anomalistic* and *synodic* frequencies, $\ddot{A}$ and $\ddot{D}$, a rather clean measurement can be made of a time rate of change of Newton's coupling parameter G. The Earth–Moon range model can be expressed in terms of these primary contributions

$$r(t) = L_0 - L_{\rm ecc}\cos(A) - L_{\rm evc}\cos(2D - A) - L_{\rm var}\cos(2D) - L_{\rm PI}\cos(D) + \cdots$$

with phases advancing as $A = A_o + \dot{A}(t - t_o) + \ddot{A}(t - t_o)^2/2 + \cdots$ and similarly for the synodic phase D. The LLR measurement of $L_{\rm PI}$, $\dot{A}$, $\dot{D}$, $\ddot{A}$, and $\ddot{D}$ forms the foundation for the gravity theory tests.

5.2 Dynamical equations for bodies, light and clocks

LLR comprehensively tests the $1/c^2$ order, gravitational N-body equations of motion which analysis groups integrate to produce orbits for the Earth, Moon and other relevant solar system bodies. The Sun–Earth–Moon system dynamics is symbolically illustrated in figure 5.2, with the rest of the solar system bodies sufficiently considered at the Newtonian level of detail. The Earth moves with velocity $\boldsymbol{V}$ and acceleration $\boldsymbol{A}$ with respect to the Sun, while the Moon is moving at velocity $\boldsymbol{V} + \boldsymbol{u}$ and acceleration $\boldsymbol{A} + \boldsymbol{a}$. (If *preferred frame* effects were to be considered for cases when gravity is not locally Lorentz-invariant, the Sun's cosmic velocity $\boldsymbol{W}$ also becomes involved [4].) There are a variety of post-Newtonian forces acting on the Earth and Moon through the Sun, each other and on themselves (self-forces) which are dependent on these general motions. Included in these are *nonlinear* gravitational forces for which each mass element of the Earth and Moon experiences forces due to the interactive effect of the Sun's gravity with the other mass elements of the same body or of the other neighbouring body. The accelerations of individual mass elements of Earth also induce accelerations on the other mass elements of Earth and similarly with the Moon. Acceleration of the Earth induces an acceleration of the Moon. Altogether, these $1/c^2$ order accelerations produce a rich assortment of modifications of the Earth–Moon range which LLR can measure.

The N-body equation of motion in metric gravity has been formulated in the literature for the completely general case [16]. Not observing any violations of local Lorentz invariance or breakdown of conservation laws in solar system gravity, I here give special consideration to the fully conservative, locally Lorentz-invariant, Lagrangian-based gravitational equation of motion (plus the cosmological variation of Newton's G). For N bodies in general motion and configuration and valid for a broad class of plausible metric theories of gravity, scalar–tensor theories in particular, the order $1/c^2$ equations of motion for these

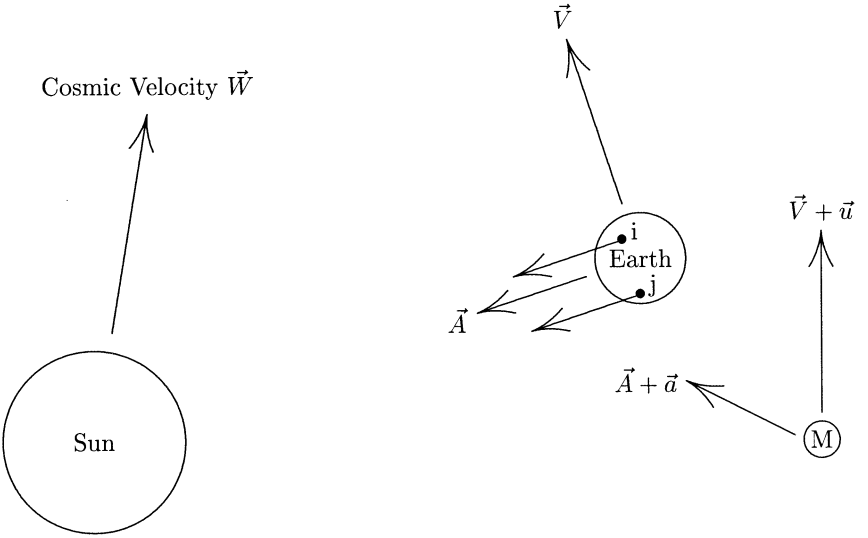

Figure 5.2. Velocities and accelerations of Sun, Earth and Moon. When formulating the Earth–Moon dynamics in the solar system barycentric frame, there are post-Newtonian force terms acting between Sun, Earth and Moon which depend on either the velocity or acceleration vectors of both the Earth and Moon. Body *self-accelerations* also result from the inductive inertial forces acting between the mutually accelerating mass elements (i, j) within each of these bodies. The intrinsic nonlinearity of gravity also produces net external forces on these bodies proportional not only to the presence of other bodies, but also to their internal gravitational binding energies. The motional, accelerative and nonlinear contributions to the three-body system's dynamics, taken collectively, make LLR a comprehensive probe of the post-Newtonian dynamics of metric gravity in the general case. If the dynamics is not locally Lorentz invariant, then the velocity $\boldsymbol{W}$ of the solar system through the cosmos leads to novel forces and resulting observable effects in LLR proportional to $\boldsymbol{W}$ (or its square) but such effects have not been seen.

N bodies take the form

$$(A) \qquad \boldsymbol{a}_i = \left(1 + \frac{\dot{G}}{G}(t - t_o)\right)\left(\frac{M(G)}{M(I)}\right)_i \boldsymbol{g}_i$$

$$(B) \qquad -\beta^* \sum_{j\neq i}\left(\sum_{k\neq i}\frac{\mu_k}{r_{ik}} + \sum_{k\neq j}\frac{\mu_k}{r_{jk}}\right)\boldsymbol{g}_{ij}$$

$$(C) \qquad +(2\gamma+2)\sum_{j\neq i}\boldsymbol{v}_i \times (\boldsymbol{v}_j \times \boldsymbol{g}_{ij})$$

$$(D) \qquad +\tfrac{1}{2}\sum_{j\neq i}[(2\gamma+1)v_i^2 + (2\gamma+2)v_j^2 - 3(\boldsymbol{v}_j \cdot \hat{r}_{ij})^2]\boldsymbol{g}_{ij}$$

$$
\begin{aligned}
& - (4\gamma + 2)[\boldsymbol{g}_{ij} \cdot \boldsymbol{v}_j(\boldsymbol{v}_j - \boldsymbol{v}_i) + \boldsymbol{g}_{ij} \cdot \boldsymbol{v}_i \boldsymbol{v}_i] \\
(E) \qquad & + \tfrac{1}{2} \sum_{j \neq i} \frac{\mu_j}{r_{ij}} [(4\gamma + 3)\boldsymbol{a}_j + \boldsymbol{a}_j \cdot \hat{r}_{ij} \hat{r}_{ij}] \\
(F) \qquad & - \tfrac{1}{2} v_i^2 \boldsymbol{a}_i - \boldsymbol{a}_i \cdot \boldsymbol{v}_i \boldsymbol{v}_i - (2\gamma + 1) \sum_{j \neq i} \frac{\mu_j}{r_{ij}} \boldsymbol{a}_i
\end{aligned} \tag{5.1}
$$

with $\boldsymbol{v}_i = \mathrm{d}\boldsymbol{r}_i/\mathrm{d}t$, $\boldsymbol{a}_i = \mathrm{d}\boldsymbol{v}_i/\mathrm{d}t$, $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ and $i, j, k = 1, \ldots, N$. The speed of light factor $1/c^2$ has been set equal to one in lines B to F to simplify presentation. The body gravitational mass strengths $\mu_i = GM(G)_i$ are indicated along with the Newtonian acceleration vectors

$$
\boldsymbol{g}_{ij} = \frac{\mu_j}{r_{ij}^3} \boldsymbol{r}_{ji} \qquad \text{and} \qquad \boldsymbol{g}_i = \sum_{j \neq i} \boldsymbol{g}_{ij}
$$

γ and β (with $\beta^* = 2\beta - 1$) are two *Eddington* parameters which quantify deviations in metric gravity theory from Einstein's pure tensor theory in which both these parameters equal one. Several lines of this total equation of motion warrant individual descriptions and brief discussions.

(1) Line A. If the metric theory *Eddington* parameters γ and β differ from their general relativistic values $\gamma_{\mathrm{GR}} = \beta_{\mathrm{GR}} = 1$, application of the equation of motion relativistic corrections from lines B through F to a body's internal gravity finds that the gravitational to inertial mass ratio of a celestial body depends on its gravitational self-energy content [1]:

$$
\frac{M(G)}{M(I)} = 1 - (4\beta - 3 - \gamma) \frac{G}{2Mc^2} \int \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|} \, \mathrm{d}^3x \, \mathrm{d}^3y + O(1/c^4). \tag{5.2}
$$

Another way to view this ratio is in terms of a spatially varying gravitational coupling parameter G

$$
G(\boldsymbol{r}, t) \cong G_\infty [1 - (4\beta - 3 - \gamma) U(\boldsymbol{r}, t)/c^2]
$$

in which a body with a significant part of its mass–energy coming from its gravitational binding energy experiences the additional acceleration

$$
\delta \boldsymbol{a}_i \cong - \frac{\partial \ln M_i}{\partial G} c^2 \nabla G
$$

with the leading gravitational energy contribution to body mass being the Newtonian contribution

$$
\frac{\partial M}{\partial \ln G} = - \frac{G}{2c^2} \int \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \, \mathrm{d}^3r \, \mathrm{d}^3r'.
$$

When cosmological equations from a metric theory are considered, Newton's coupling parameter G will also generally be found to vary in time in

proportion to the Hubble expansion rate of the universe

$$\frac{\dot{G}}{G} \sim (4\beta - 3 - \gamma)H. \tag{5.3}$$

The presently most precise way to measure any deviation of β from its general relativistic value is through measurement of the $M(G)/M(I)$ ratio of Earth using LLR data.

(2) Line B. Gravity couples to itself, thereby producing nonlinear gravitational forces among and between bodies.

(3) Line C. Just as pairs of moving charges generate magnetic forces between themselves in proportion to the velocities of both charges, pairs of moving masses generate *gravitomagnetic* forces between themselves. This force acts between the mutually moving Earth and Moon and contributes to the necessary Lorentz contraction of the lunar orbit as viewed from the solar system barycenter.

(4) Line D. Masses in motion both produce and couple to gravitational fields differently than masses at rest. The package of velocity-dependent acceleration terms in this line plus line C lead to the local Lorentz invariance of gravity. Any further modifications of this package (beyond the γ-dependence) will lead to additional terms in the equation of motion with one or two powers of body velocities being replaced by the velocity $\boldsymbol{W}$ of the solar system relative to the universe *preferred frame*. A variety of preferred frame effects which would then result have been empirically ruled out in LLR and other solar system observations [5].

(5) Line E. Accelerating masses generate inductive gravitational forces on other proximite masses.

(6) Line F. The inertia of a mass is altered by its motion and by its proximity to other masses. The combination of terms from this line plus line E are necessary in order that a body's gravitational self-energy contributes to its total inertial mass in accord with special relativity's prescription $M = E/c^2$. This modification of inertia is part of the $M(G)/M(I)$ calculation for a celestial body.

LLR measures the round-trip time of the propagation of light between two separate body trajectories, and this measurement is made by a specific clock moving on a particular trajectory. So in the solar system barycentric and spatially isotropic coordinates employed to express the body equations of motion given by equation (5.1), there are also requirements for the post-Newtonian modifications to the light coordinate speed function and to the clock rates, these respectively being

$$c(\boldsymbol{r}, t) \cong c_\infty[1 - (1 + \gamma)U(\boldsymbol{r}, t)/c^2] \tag{5.4}$$

and

$$\mathrm{d}\tau \cong \mathrm{d}t[1 - v^2/2c^2 - U(\boldsymbol{r}, t)/c^2] \tag{5.5}$$

in which $U(\boldsymbol{r})$ is the total Newtonian gravity potential function due to solar system bodies

$$U(\boldsymbol{r}, t) = \sum_j \int \frac{G\rho(\boldsymbol{r}'(t))_j}{|\boldsymbol{r} - \boldsymbol{r}'(t)|} \, \mathrm{d}^3 r'. \tag{5.6}$$

Because the Earth moves in the solar system barycentric frame and it rotates at rate $\boldsymbol{v}$, there must be two corrections applied to an Earth surface location $\boldsymbol{a}$: first there is the Lorentz contraction of the extended body

$$\delta \boldsymbol{a} \cong -\boldsymbol{a} \cdot \boldsymbol{V}\boldsymbol{V}/2c^2$$

and because of special relativity's non-absolute nature of time simultaneity there is a further displacement of the rotating Earth surface locations

$$\delta \boldsymbol{a} \cong \boldsymbol{V} \cdot \boldsymbol{a}(\boldsymbol{v} \times \boldsymbol{a})/c^2.$$

These light and clock equations and special relativistic body distortion effects play only supportive (but necessary) roles in fitting LLR data: the main science emerges from the body equations of motion as given by equation (5.1).

5.3 LLR's key science-related range signals

Associated with each feature of gravitational theory which is tested by LLR, there are specific range signals in the LLR data whose measurements yield the information about theory. Several of these signals are here described.

5.3.1 Violation of the universality of free-fall

Because celestial bodies have gravitational self-energies (internal gravitational binding energies), they will generally possess gravitational to inertial mass ratios which differ from each other as indicated in line A of equation (5.1) and given by equation (5.2). But there are other ways in which bodies may accelerate at different rates toward other bodies. Within the paradigm that forces between objects are carried by a field, an additional long-range interaction in physical law generates a force between bodies i and j which will typically have the static limit form

$$\boldsymbol{f}_i = K_i \nabla_i \frac{K_j}{r_{ij}} \mathrm{e}^{-\mu r_{ij}}. \tag{5.7}$$

The bodies' coupling strengths K_i and K_j, except in special cases such as metric scalar–tensor gravity in which $K_i \sim M_i$, will be attributes of the bodies which are different than total mass–energy (*non-metric* coupling); and the dependence on distance of this force will be either inverse square if the field is massless or Yukawa-like if the underlying field transmitting this force between bodies has mass. Such a new force will produce a difference in the Sun's acceleration of the Earth and Moon, because the latter two bodies are of different compositions—the

Earth has a substantial iron core while the Moon is composed of low-Z mantle-like materials. The fractional difference in acceleration rates of Earth and Moon amounts to

$$|\delta \boldsymbol{a}_{EM}/\boldsymbol{g}_{\mathrm{S}}| = \frac{K_{\mathrm{S}}}{GM_S}\left(\frac{K_{\mathrm{M}}}{M_{\mathrm{M}}} - \frac{K_{\mathrm{E}}}{M_{\mathrm{E}}}\right)(1+\mu R)\mathrm{e}^{-\mu R}$$

and it will supplement any difference in the accelerations resulting from the possible anomalies in the bodies' gravitational to inertial mass ratios due to gravitational self-energies. LLR has become a sufficiently precise tool for measuring $|\delta \boldsymbol{a}_{\mathrm{EM}}|$ and it now competes favorably with ground-based laboratory measurements looking for the composition-dependence of free-fall rates. LLR is also the premier probe for measuring a body's $M(G)/M(I)$ ratio as given by equation (5.2).

If the Earth and Moon fall toward the Sun at different rates due to either of the mechanisms discussed here, then the lunar orbit is polarized along the solar direction. Detailed calculation of this polarization reveals an interesting interactive feedback mechanism which acts between this $\cos(D)$ polarization and the $\cos(2D)$ Newtonian solar tide perturbation of the lunar orbit, the *variation*). The result is an amplification of the synodic perturbation

$$\begin{aligned}\delta r(t)_{\mathrm{ME}} &= \frac{3}{2}\frac{\Omega}{\omega}RF(\Omega/\omega)\delta_{\mathrm{ME}}\cos D \\ &\cong 2.9\times 10^{12}\delta_{\mathrm{ME}}\cos D \text{ cm}\end{aligned} \tag{5.8}$$

with $\delta_{\mathrm{EM}} = |(\boldsymbol{a}_{\mathrm{E}} - \boldsymbol{a}_{\mathrm{M}})/\boldsymbol{g}_{\mathrm{S}}|$, R is distance to the Sun, Ω and ω are the sidereal frequencies of solar and lunar motion and D is the lunar phase measured from new moon. The feedback amplification factor for the lunar orbit is already $F(\Omega/\omega) \cong 1.75$: it grows further with larger orbits with an interesting resonance divergence for an orbit about twice the size as that of the Moon [14, 15]. Computer integration of the complete equation (5.1) for the Sun–Earth–Moon system dynamics confirms these analytically estimated polarization sensitivities.

The most recent fits of the LLR data find no anomalies in the $\cos(D)$ amplitude to a precision of 4 mm, so from equation (5.8) δ_{ME} is constrained to be less than 1.3×10^{-13}. Neglecting any possible composition dependence and, using equation (5.2) with an estimate for the fractional gravitational self-energy of the Earth being 4.5×10^{-10}, the following constraint on a combination of the two *Eddington* parameters is

$$|4\beta - 3 - \gamma| \leq 4\times 10^{-4}. \tag{5.9}$$

If metric gravity is a combination of scalar and tensor interactions, the small size of this constraint is an approximate measure of the scalar interaction strength compared to the dominant tensor interaction. One scenario which could explain today's weakness of the scalar interaction is illustrated in figure 5.3. Scalar–tensor metric gravity involves one coupling function $V(\phi)$: the slope of this function

gives the strength of the scalar interaction and, in combination with the function's curvature, also determines gravity's $1/c^2$ order nonlinearity. Near an extremum of $V(\phi)$, the Eddington parameters are given by simple properties of the coupling function:

$$1 - \gamma \cong \frac{1}{2}\left(\frac{\mathrm{d}\ln V(\phi)}{\mathrm{d}\phi}\right)^2 \tag{5.10}$$

$$\beta - 1 \cong \frac{1-\gamma}{8}\frac{\mathrm{d}^2 \ln V(\phi)}{\mathrm{d}\phi^2}. \tag{5.11}$$

As the universe expands, the dynamical equations for the background scalar field will drive the scalar to a minimum of the coupling function, if it exists, and where γ and β take their general relativistic values. Scalar gravity turns itself off naturally if an 'attractor' exists in its coupling function $V(\phi)$. But that process, being dynamical, should not be entirely complete today, and the small remnant of the scalar interaction may still be detectable by sufficiently precise testing of relativistic gravity using LLR and other experiments [8, 9].

The LLR result can also place limits on the spatial gradient of the *fine structure constant*, $\alpha = e^2/\hbar c$, in the proximity of the Sun. If α is a function of a scalar field whose source includes ordinary matter, a spatial gradient of α near bodies should exist and composition-dependent accelerations of other objects toward this body should occur:

$$\delta \boldsymbol{a}_i = -\frac{\partial \ln M_i}{\partial \ln \alpha} c^2 \frac{\boldsymbol{\nabla}\alpha}{\alpha}.$$

The dominant electromagnetic contribution to the mass–energy of different elements is due to the electrostatic energy among the Z nuclear protons. This energy fractionally varies by an order of magnitude (from a few parts in 10^4 to a few parts in 10^3) as one proceeds through the periodic table from low-Z to high-Z elements. For the Earth with its iron core and the Moon composed almost entirely of mantle-like materials, one can conclude from the LLR constraint on δ_{ME} that any gradient of α due to and toward the Sun is quite small compared to the Sun's gravitational field $\boldsymbol{g}_{\mathrm{S}}$:

$$\frac{c^2|\boldsymbol{\nabla}\ln\alpha|}{|\boldsymbol{g}_{\mathrm{S}}|} \leq 4 \times 10^{-10}.$$

This should be compared with the best constraints on the time variation of α, which, in units of the Hubble expansion rate, are substantially weaker:

$$H\frac{\dot{\alpha}}{\alpha} \leq 10^{-5}.$$

This suggests that, unless there are unusual sources for the scalar field which controls the value of α, e.g. sources which are present in an average cosmological context but which do not concentrate in ordinary matter or other special situations, then today's LLR constraint on the spatial gradient of α is the significant present measure of the constancy of α.

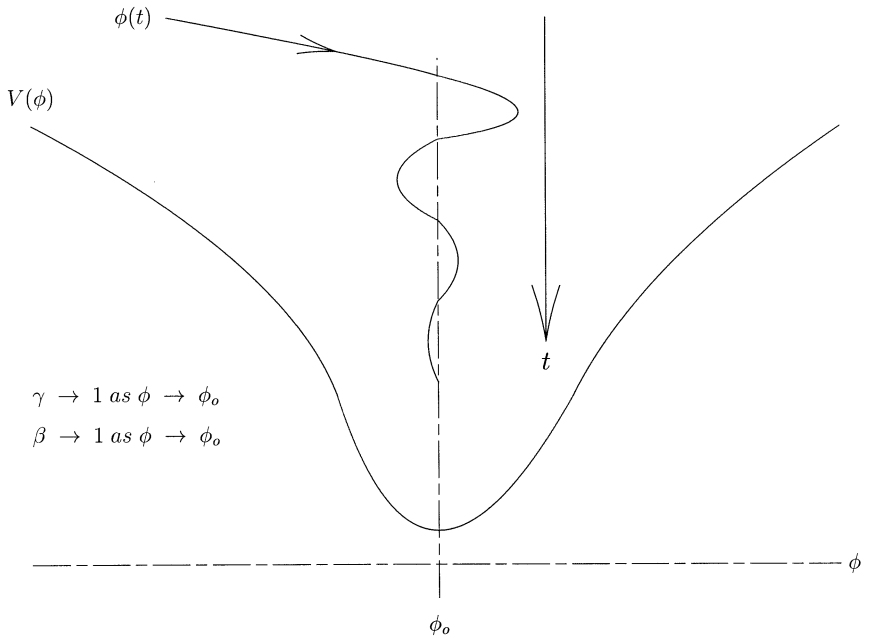

Figure 5.3. Typical cosmological dynamics of a background scalar field is shown if that field's coupling function $V(\phi)$ has an attracting point ϕ_o. The strength of the scalar interaction's coupling to matter, proportional to the derivative of the coupling function, weakens as the attracting point is approached: so in a scalar–tensor metric theory, for example, the Eddington parameters γ and β both approach the pure tensor gravity values of one.

5.3.2 Geodetic precession of the local inertial frame

Because the Earth and Moon travel at different velocities through the Sun's gravitational field, terms from lines D and F of equation (5.1) are present which accelerate the Moon relative to Earth. A particularly interesting part of the relative acceleration is proportional to both $\boldsymbol{V}$ and $\boldsymbol{u}$ and the Sun's acceleration with, as shown in figure 5.2, $\boldsymbol{V}$ being the velocity of the Earth relative to the Sun, and $\boldsymbol{u}$ the velocity of the Moon relative to Earth. These terms form deSitter's Coriollis-like acceleration

$$\delta \boldsymbol{a}_{\mathrm{M}} = 2\boldsymbol{\Omega}_{\mathrm{dS}} \times \boldsymbol{u} \tag{5.12}$$

with

$$\boldsymbol{\Omega}_{\mathrm{dS}} = \frac{2\gamma + 1}{2} \frac{GM_{\mathrm{s}}}{c^2 R^3} \boldsymbol{R} \times \boldsymbol{V} \tag{5.13}$$

and its geometrical interpretation is the local precession of the inertial frame at rate $\boldsymbol{\Omega}_{\mathrm{dS}}$ which amounts to about 19.2 mas yr^{-1}. The effect of this perturbing

acceleration on the orbit is primarily an additional rate of perigee precession with respect to distant inertial space. This is measured by comparing the Moon's anomalistic frequency $\dot{A}$ (rate of eccentric motion) with its synodic frequency $\dot{D}$ (rate of monthly phase) and with the latter converted into lunar sidereal frequency ω (orbital rate) by adding to $\dot{D}$ the annual rate Ω which is provided by results from other solar system experiments. Sidereal minus anomalistic frequency of lunar motion includes deSitter's precessional rate as a supplement to the Newtonian tidal contributions to perigee precession. These lunar frequencies are measured from range signal perturbations whose size grows linearly in time. The Moon's range from Earth includes several dominant oscillatory contributions:

$$\delta r_{\mathrm{ME}} = L_{\mathrm{ecc}} \cos(A) + L_{\mathrm{var}} \cos(2D) + L_{\mathrm{evc}} \cos(2D - A) + \cdots$$

with L_{ecc} being the amplitude of eccentric motion, L_{var} the amplitude of solar tidal perturbation called the *variation* and L_{evc} being the amplitude of the hybrid *evection* perturbation due to both the solar tidal force and the eccentric motion of the Moon. The least-squares fit of the LLR data, which yields the best estimates for the two key lunar frequencies, will then involve the parameter 'partials':

$$\begin{aligned} \frac{\partial \delta r_{\mathrm{ME}}}{\partial \dot{A}} &= -t(L_{\mathrm{ecc}} \sin(A) - L_{\mathrm{evc}} \sin(2D - A)) \\ \frac{\partial \delta r_{\mathrm{ME}}}{\partial \dot{D}} &= -2t(L_{\mathrm{var}} \sin(2D) + L_{\mathrm{evc}} \sin(2D - A)) \end{aligned}$$

The precision of the measurement of the deSitter precession grows particularly with the total time of the LLR experiment, not only because of the growing quantity and quality of the accumulated range measurements but also because of the linear growth in signal sensitivity. The most recent fit of the LLR data confirms the presence of the geodetic precession with precision of 0.07 mas yr^{-1} [10].

5.3.3 Time evolution of gravity's coupling strength *G*

The evolution of Newton's coupling parameter G over time results in proportional evolutions for both the radial size and frequencies of the lunar motion. Slightly different orbital changes occur when a torque (indicated by $\dot{L}$) acts on the orbit:

$$\begin{aligned} \frac{\dot{r}}{r} &= -\frac{\dot{G}}{G} + 2\frac{\dot{L}}{L} \\ \frac{\dot{\omega}_{\mathrm{n}}}{\omega_{\mathrm{n}}} &= 2\frac{\dot{G}}{G} - 3\frac{\dot{L}}{L}. \end{aligned}$$

During the earlier years of the LLR experiment, the mean orbital radius signal

$$\delta r(t)_{\mathrm{ME}} = \left(2\frac{\dot{L}}{L} - \frac{\dot{G}}{G}\right) r(t - t_0)$$

was used to measure $\dot{G}$. However, this involved estimating and subtracting a contribution to $\dot{r}$ which results from the orbital torque exerted on the Moon by the ocean tidal bulges on Earth which, because of friction, lag in angle from the direction toward the Moon. The inclination and 18.6 yr precession of the lunar orbit's plane result in a modulation of the tidal contribution to $\dot{r}$ which helps to separate the two perturbations after sufficient years of data have been accumulated. But the data set produced by LLR has, in recent years, become sufficiently extended in time so that the range signals associated with frequency shifts, which grow quadratically in time, are becoming dominant in the fit for $\dot{G}$. Recall that the two lunar phases can be expanded in terms of the initial phase, rate and *acceleration*:

$$D(t) = D + \dot{D}(t - t_o) + \tfrac{1}{2}\ddot{D}(t - t_o)^2 + \cdots \tag{5.14}$$

$$A(t) = A + \dot{A}(t - t_o) + \tfrac{1}{2}\ddot{A}(t - t_o)^2 + \cdots. \tag{5.15}$$

The synodic frequency is, by definition, equal to the difference in the lunar sidereal rate and the solar sidereal rate around the Earth,

$$\dot{D} = \omega - \Omega$$

while the Moon's *anomalistic* rate is derivable from the underlying equation of motion and can be expressed in the form

$$\dot{A} = \omega - \frac{3}{4}\frac{\Omega^2}{\omega} - \frac{225}{32}\frac{\Omega^3}{\omega^2} - \cdots - (\gamma + 1/2)\frac{GM}{c^2 R}\Omega + \cdots$$

which consists of the classical Newtonian expression plus relativistic modifications, with the dominant geodetic precession contribution shown. From these two expressions, the solar sidereal rate and its acceleration can then be expressed as follows

$$\Omega = \dot{A} - \dot{D} + \frac{3}{4}\frac{(\dot{A} - \dot{D})^2}{\dot{A}} + \cdots$$

$$\dot{\Omega} = \ddot{A} - \ddot{D} + \cdots.$$

While the lunar phases A and D suffer accelerations due to any tidal torques acting between Earth and Moon, the solar rate Ω is not affected by the tidal torques. Acceleration of this solar rate is, therefore, a rather pure measure of a time variation in G. Noting from equations (5.14) and (5.15) that the partials for $\ddot{A}$ and $\ddot{D}$ will grow in amplitude quadratic in time

$$\frac{\partial R_{\rm calc}}{\partial \ddot{A}} = \frac{1}{2}t^2(L_{\rm ecc}\cos(A) - L_{\rm evc}\cos(2D - A)) \tag{5.16}$$

$$\frac{\partial R_{\rm calc}}{\partial \ddot{D}} = t^2(L_{\rm var}\cos(2D) + L_{\rm evc}\cos(2D - A)) \tag{5.17}$$

it follows that the formal error in measuring $\dot{G}$ decreases as the inverse square of the time span T of LLR observations. For a uniform time distribution of observations, one obtains with

$$\frac{\dot{G}}{G} = \frac{1}{2}\frac{\dot{\Omega}}{\Omega}$$
$$\left(\frac{\delta\dot{G}}{G}\right)_{\mathrm{RMS}} = \sqrt{\frac{360}{N}}\,\frac{1}{\Omega T^2}\,\frac{\sigma}{\sqrt{4L_{\mathrm{var}}^2 + 3L_{\mathrm{evc}}^2}}$$

with σ being the rms size of individual range measurement errors and N the total number of measurements spread over the time T. A recent fit of almost 30 yr of LLR data yields the following excellent measurement constraint [10]:

$$\frac{\dot{G}}{G} \cong (0 \pm 1.1) \times 10^{-12}\ \mathrm{yr}^{-1}. \tag{5.18}$$

This amounts to about 1/60 of the observed Hubble expansion rate of the universe. With the precision of this measurement now growing quadratically in time, LLR should continue indefinitely to be at the cutting edge in the measuring of $\dot{G}$.

5.4 An additional Yukawa interaction?

When the supplementary interaction given by equation (5.7) is of a Yukawa nature, $\mu \neq 0$, it contributes to the precession of the periastron for a near-circular orbit of radius r by an amount

$$\frac{\delta(\omega - \omega_{\mathrm{o}})}{\omega} = \frac{1}{2}\frac{K_i K_j}{G M_i M_j}(\mu r)^2 \exp(-\mu r)$$

with ω and ω_{o} being the orbit's sidereal and eccentric frequencies, respectively. This perturbation of the precession rate also occurs if the Yukawa force is metric, $K_i \sim M_i$, or non-metric. With the Moon's perigee precession rate measured to a precision of 0.07 mas yr^{-1} and showing no anomaly, then for Yukawa ranges in the vicinity of that for the maximum sensitivity of lunar perturbation, $\mu r = 2$, the strength of the Yukawa force is decisively constrained

$$\frac{|K_{\mathrm{E}} K_{\mathrm{M}}|}{G M_{\mathrm{E}} M_{\mathrm{M}}} \leq 5 \times 10^{-12}\left(\frac{4}{(\mu r)^2}\exp(\mu r - 2)\right).$$

5.5 Gravitomagnetism

Line C of the complete N-body gravitational equation of motion given by equation (5.1) indicates a post-Newtonian gravitational force proportional to the velocities of both bodies in the interaction and, in analogy with electromagnetic

theory, it has been called the *gravitomagnetic* interaction. From line C of equation (5.1), this acceleration is

$$\delta \boldsymbol{a}_i = (2+2\gamma)\sum_{j\neq i}\frac{Gm_j}{c^2 r_{ij}^3}(\boldsymbol{r}_{ij}\boldsymbol{v}_i\cdot\boldsymbol{v}_j - \boldsymbol{r}_{ij}\cdot\boldsymbol{v}_i\boldsymbol{v}_j).$$

It often has been claimed that the presence of gravitomagnetism within the total gravitational interaction has not been experimentally confirmed and measured. Indeed, different experiments have been developed explicitly to observe the effects of this historically interesting prediction of general relativity. But this gravitomagnetic acceleration already plays a large role in producing the final shape of the lunar orbit, albeit in conjunction with the rest of the total equation of motion: the precision fit of the LLR data indicates that gravitomagnetism's presence and specific strength in the equation of motion can hardly be in doubt. Because both the Earth and Moon are moving in the solar system barycentric frame—the frame in which the dynamical equations are formulated and then integrated into orbits—a gravitomagnetic interaction exists between these two bodies, the Earth having velocity $\boldsymbol{V}(t)$ and the Moon's being $\boldsymbol{V}(t)+\boldsymbol{u}(t)$, as seen in figure 5.2. As a result of these mutual motions, perturbations to the Earth–Moon range from the gravitomagnetic acceleration are proportional to both V^2 and Vu:

$$\begin{aligned}\delta r(t) &\cong \frac{Gm_e}{r^2}\left(-\frac{4}{3\omega^2}\frac{V^2}{c^2}\cos(2D)+\frac{2}{\omega\Omega}\frac{Vu}{c^2}F(\Omega/\omega)\cos(D)\right)\\ &\cong -530\cos(2D)+525\cos(D)\ \text{cm}. \qquad (5.19)\end{aligned}$$

As previously discussed, the amplitudes of the lunar motion at both these periods (monthly and semi-monthly) are determined to better than half a centimetre precision in the total orbital fit to the LLR data. It would be impossible to understand this fit of the LLR data without the participation of the gravitomagnetic interaction in the underlying model and with a strength very close to that provided by general relativity, $\gamma = 1$. As in electromagnetic theory, the velocity-dependent force terms in lines C and D of equation (5.1) can be changed individually by formulating the dynamics in different frames of reference but the very ability to reformulate the equations of motion in different frames without introducing new frame-dependent terms depends on the local Lorentz invariance (LLI) of gravity. It is the entire package of velocity-dependent, post-Newtonian terms which includes the gravitomagnetic terms, lines C plus D of equation (5.1) that produces the LLI: the *Eddington* parameter γ represents the only freedom in the structure of this LLI package. Our confidence in the exhibited structure of this total collection of velocity-dependent terms is established in proportion to the precision with which the various *preferred frame*, LLI-violating effects in the solar system proportional to W^2, WV and Wu have been found to be absent [5]. LLR has been one of the main contributors in establishing gravity's LLI through null measurements of several W-dependent effects [4, 6, 7, 18].

5.6 Inductive inertial forces

Inductive forces are shown on line E of equation (5.1): in such forces, the acceleration of one mass element induces an acceleration of another proximite mass element (e.g. i and j in figure 5.2). From line E of equation (5.1), we have

$$\delta \boldsymbol{a}_i = \sum_{j \neq i} \frac{G m_j}{2c^2 r_{ij}} ((4\gamma + 3)\boldsymbol{a}_j + \boldsymbol{a}_j \cdot \hat{r}_{ij} \hat{r}_{ij}). \tag{5.20}$$

These accelerations play a key part in altering the inertial masses of the Earth and Moon because of their internal gravitational binding energies: either the absence or an anomalous strength of these inductive forces would translate directly into differences between the acceleration rates of these whole bodies toward the Sun. A polarization of the Moon's orbit in the solar direction, as previously discussed, would result. The forces, equation (5.20), acting between the mass elements of Earth, for example, by themselves would lead to an anomalous polarization of the lunar orbit of very large magnitude:

$$\delta r(t) \cong 130 \cos(D) \text{ m}. \tag{5.21}$$

Only when these inductive forces are combined with the other post-Newtonian inertial forces shown on line F of equation (5.1) does the total inertial self-force of a body become

$$\begin{aligned} \delta \boldsymbol{f} = & -\frac{1}{c^2}\left(\frac{1}{2}\sum_i m_i v_i^2 - \frac{G}{2}\sum_{i,j}\frac{m_i m_j}{r_i j}\right)\boldsymbol{a} \\ & -\frac{1}{c^2}\left[\sum_i m_i \boldsymbol{v}_i \boldsymbol{v}_i - \frac{G}{2}\sum_{i,j}\frac{m_i m_j}{r_{ij}^3}\boldsymbol{r}_{ij}\boldsymbol{r}_{ij}\right]\cdot \boldsymbol{a}. \end{aligned}$$

The first line of this total self-force is now the expected inertial force due to the internal kinetic energy and gravitational binding energy within the body. The second line represents contributions to the body's internal *virial* which, when totalled over all internal force fields, vanishes for a body in internal equilibrium and experiencing negligible external tidal-like forces. These self-forces of a body are an integral part of the determination of the total gravitational to inertial mass ratio of bodies discussed previously, and in general relativity, they are cancelled by the equal contributions of internal energies to a body's gravitational mass. They were explicitly discussed here in order to show the large size of such inductive force contributions which must necessarily be taken into account in the fit of theory to the LLR data.

Acknowledgment

This work has been supported by the National Aeronautics and Space Administration through contract NASW-97008 and NASW-98006.

References

[1] Nordtvedt K 1968 *Phys. Rev.* **169** 1017
[2] Nordtvedt K 1968 *Phys. Rev.* **170** 1186
[3] deSitter W 1916 *Mon. Not. R. Astron. Soc.* **77** 155
[4] Nordtvedt K 1973 *Phys. Rev.* D **7** 2347
[5] Nordtvedt K 1987 *Astrophys. J.* **320** 871
[6] Nordtvedt K and Will C 1972 *Astrophys. J.* **177** 775
[7] Nordtvedt K 1996 *Class. Quantum Grav.* **13** 1309
[8] Damour T and Nordtvedt K 1993 *Phys. Rev. Lett.* **70** 2217
[9] Damour T and Nordtvedt K 1993 *Phys. Rev.* D **48** 3436
[10] Williams J G, Boggs D H, Dickey J O and Folkner W M 2001 Lunar laser tests of gravitational physics *Proc. Marcel Grossmann Meeting IX, 2–8 July 2000, Rome, Italy* ed R Ruffini (Singapore: World Scientific)
[11] Shapiro I I *et al* 1988 *Phys. Rev. Lett.* **61** 2643
[12] Bertotti B, Ciufolini I and Bender P L 1987 *Phys. Rev. Lett.* **58** 1062
[13] Williams J G, Newhall X X and Dickey J O 1996 *Phys. Rev.* D **53** 6730
[14] Nordtvedt K 1995 *Icarus* **114** 51
[15] Damour T and Vokrouhlicky D 1996 *Phys. Rev.* D **53** 4177
[16] Will C and Nordtvedt K 1972 *Astrophys. J.* **177** 757
[17] Nordtvedt K 1996 *Class. Quant. Grav.* **13** 1317
[18] Müller J, Nordtvedt K and Vokrouhlicky D 1996 *Phys. Rev.* D **54** 5927
[19] Müller J and Nordtvedt K 1998 *Phys. Rev.* D **58** 2001

Chapter 6

The early Universe and the cosmic microwave background

Amedeo Balbi
Dipartimento di Fisica, Università di Roma 'Tor Vergata'
and INFN, Sezione di Roma II
Via della Ricerca Scientifica
00133, Roma, Italy

6.1 Introduction

In the past few years, cosmology has experienced enormous progress. Our understanding of the physics of the early Universe, its evolution and current large-scale structure, now lies on firmer grounds than in the past. This was due, on one side, to breakthroughs in theoretical research and, on the other side, to the impressive advancements made in observational techniques, which has allowed a large quantity of high-quality data to be collected. A fundamental role in entering what has been dubbed 'the era of precision cosmology' has been played by the study of the cosmic microwave background (CMB). In the first part of this chapter, I briefly outline the basics of the standard cosmological model and some key elements of the early Universe. Next, I give a pedagogical exposition of the physics of CMB anisotropy, showing its importance as a cosmological probe. Finally, I highlight the progress made in CMB investigation in the last decade, from the results of the COBE satellite, that opened a new era in the investigation of the cosmos to the recent WMAP results, ending up with some future prospects from the forthcoming Planck mission.

6.2 The standard cosmological model

Cosmology has a standard model, which provides a well-established framework in which to understand the global properties of the physical Universe. There is

a strong interplay between fundamental physics and cosmology, since the early Universe is a natural laboratory for high-energy physics. What follows gives a very sketchy picture of the main elements of the standard cosmological model. I refer the interested reader to general books on the subject, such as those by Kolb and Turner [19] and Peebles [30], for a thorough exposition.

6.2.1 The big bang model

The big bang model (or, more precisely, the Friedmann–Robertson–Walker (FRW) model) provides a very successful description of the physical Universe from very early times ($t \sim 10^{-2}$ s) to the present. It can easily explain some key features of the observed Universe, such as

- the expansion law,
- the abundance of light elements,
- the existence of the cosmic microwave background (CMB) and
- the age of the oldest objects observed.

Furthermore, it provides a framework where the gravitational instability scenario that explains the growth of cosmic structures can be easily accommodated.

The Universe appears to be homogeneous and isotropic on scales comparable to its present observable volume. The geometry of such a Universe is described by the Robertson–Walker metric[1]:

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu = \mathrm{d}t^2 - R^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right] \quad (6.1)$$

where k is a parameter assuming the values $1, 0, -1$ for positive, null or negative space curvature, respectively. The scale factor $R(t)$ describes the expansion of the Universe. This is often parametrized in terms of a scale factor normalized to unity at present: $a(t) \equiv R(t)/R_0$. The time coordinate t is the proper time. The coordinates r, θ, ϕ are called *comoving*: they label the position of observers at rest in the expanding frame.

The *proper distance* in the Robertson–Walker metric is defined as

$$d \equiv R(t)\int_0^r \frac{\mathrm{d}r'}{\sqrt{1-kr'^2}}. \quad (6.2)$$

The quantity $H \equiv \dot{R}/R = \dot{a}/a$ is the *Hubble parameter* describing the expansion rate of the Universe. The *Hubble time*, $t \equiv H^{-1}$, gives the characteristic time scale of the expansion. In $c = 1$ units, H^{-1} also identifies a characteristic length scale, the *Hubble radius*, giving the approximate size of the visible Universe. The present value of the Hubble parameter, H_0, is called the *Hubble constant*. This is usually parameterized in terms of the adimensional quantity h as $H_0 =$

[1] Units are chosen so that $c = 1$.

100 h km s^{-1} Mpc^{-1}. The Hubble expansion law is obtained by deriving the proper distance with respect to time:

$$v \equiv \dot{d} = \frac{\dot{R}}{R} R \int_0^r \frac{\mathrm{d}r'}{\sqrt{1 - kr'^2}} \tag{6.3}$$

that is

$$v = Hd. \tag{6.4}$$

The dynamics of the Universe, i.e. the time evolution of the scale factor, is governed by the Friedmann equation, which can be derived by the Einstein equation using as the stress-energy tensor that of an ideal fluid with time-dependent energy density $\rho(t)$ and pressure $p(t)$:

$$\left(\frac{\dot{R}}{R}\right)^2 + \frac{k}{R^2} = \frac{8\pi G}{3}\rho. \tag{6.5}$$

Imposing stress–energy tensor conservation results in the equation:

$$\mathrm{d}[R^3(\rho + p)] = R^3 \,\mathrm{d}p. \tag{6.6}$$

Assuming a generic equation of state $p = w\rho$ with w independent of time, the latter gives

$$\rho \propto R^{-3(1+w)}. \tag{6.7}$$

For example, for radiation $p = \rho/3$ and $\rho \propto R^{-4}$; for matter $p = 0$ and $\rho \propto R^{-3}$; and for a cosmological constant (vacuum energy) $p = -\rho$ and $\rho \propto$ constant. The evolution of the scale factor when the Universe is dominated by one of these components is found by solving the Friedmann equation: $R \propto t^{1/2}$ for a radiation-dominated Universe, $R \propto t^{2/3}$ for a matter-dominated Universe and $R \propto \exp(Ht)$ (with $H =$ constant) for a vacuum-energy-dominated Universe.

Deriving the Friedmann equation with respect to time, one gets

$$\frac{\ddot{R}}{R} = -\frac{4\pi G}{3}(\rho + 3p). \tag{6.8}$$

The Universe is expanding at present, so that $\dot{R} > 0$ today. If, in the past, the right-hand side was always negative (or $\rho + 3p > 0$), then there must have been some finite time when $R = 0$. This time is usually set as $t = 0$ and is called the big bang.

The *causal horizon* (or *particle horizon*), $r_{\mathrm{H}}(t)$, is the distance covered at time t by a light signal emitted at time $t = 0$: all the points which, at time t, are further away than $r_{\mathrm{H}}(t)$ have not had enough time to get in causal contact. The horizon length is calculated by imposing $\mathrm{d}s = 0$ (light-like interval) in the Robertson-Walker metric:

$$\int_0^t \frac{\mathrm{d}t'}{R(t')} = \int_0^{r_H} \frac{\mathrm{d}r}{\sqrt{1 - kr^2}} \tag{6.9}$$

so that the proper length of the horizon is

$$d_{\mathrm{H}}(t) = R(t) \int_0^t \frac{\mathrm{d}t'}{R(t')}. \tag{6.10}$$

The photon wavelength is affected as any other length by the expansion of the Universe. The relative variation of the observed wavelength λ_{o} with respect to the emitted wavelength λ_{e} due to the expansion is the *redshift* z:

$$z \equiv \frac{\lambda_{\mathrm{o}} - \lambda_{\mathrm{e}}}{\lambda_{\mathrm{e}}}. \tag{6.11}$$

The redshift is related to the scale factor by

$$1 + z = \frac{R_0}{R} = \frac{1}{a}. \tag{6.12}$$

By defining the *critical density*

$$\rho_{\mathrm{c}} = \frac{3H^2}{8\pi G} \tag{6.13}$$

and the *density parameter*

$$\Omega = \frac{\rho}{\rho_{\mathrm{c}}} \tag{6.14}$$

the Friedmann equation can be rewritten as

$$k = H^2 R^2 (\Omega - 1) \tag{6.15}$$

relating the space curvature to the quantity of matter in the Universe. Note that this equation applies at any time and that Ω and ρ_{c} vary as the Universe expands.

6.2.2 Inflation

Despite its success, the big bang model has a number of shortcomings. Rather than being real inconsistencies of the theory, these problems are essentially related to questions about the initial state of the Universe that cannot be answered by the FRW model itself.

The first problem of the big bang model is usually referred to as the *horizon problem*. One manifestation of this problem is given by the fact that we receive the thermal background radiation left over from the very early, very hot stages of the Universe with nearly the same temperature from any point of the sky. However, regions of the sky separated by angles larger than about 1° were outside the causal horizon when those photons were emitted, preventing any physical process from creating the observed uniformity. A closely related problem arises when trying to explain the existence of density perturbations on scales larger than the horizon at early times. Primeval inhomogeneities on scales of cosmological interest cannot

have been produced by causal, microphysical processes taking place in the early Universe.

Another problem of the big bang model is known as the *flatness problem*. If we use equation (6.15) to trace back in time the evolution of the density parameter Ω, we find that, as $R \to 0$, Ω has to be closer and closer to 1. More quantitatively, it turns out that in order to have Ω of order unity today, it had to be fixed with enormous precision at early times. For example, $\Omega(1\ \mathrm{s}) = 1 \pm \mathcal{O}(10^{-16})$. Were Ω not set this close to unity, the Universe would either have quickly recollapsed or it would have expanded so rapidly as to reach a temperature of 3K in a tiny fraction of a second. Equation (6.15) can also be used to relate the radius of curvature of the Universe, defined as $R_{\mathrm{curv}} \equiv R(t)|k|^{-1/2}$, to the Hubble radius H^{-1}: $H^{-1}/R_{\mathrm{curv}} = |\Omega - 1|^{-1/2}$. So the flatness problem can be restated by saying that the radius of curvature of the Universe had to be much bigger than the Hubble radius at early times: $R_{\mathrm{curv}}(1\ \mathrm{s}) \gtrsim 10^8 H^{-1}$.

Finally, fundamental physics theories predict the presence of a variety of stable, massive particles, with very small annihilation cross sections, created in the very early Universe. There is no way in the standard big bang model of preventing these *unwanted relics* from becoming the dominant component in the present Universe and contributing to the total energy density in such a way that $\Omega \gg 1$.

Starting from the pioneering work done by Starobinsky [43] and by Guth [14], it became clear that a class of models, grouped under the generic term *inflation*, can provide a mechanism for solving the problems of the big bang model. The basic idea behind inflation is that, at some very early time, the comoving Hubble radius decreased in time:

$$\mathrm{d}(H^{-1}/R)/\mathrm{d}t < 0. \tag{6.16}$$

This is the opposite of what happens in the standard big bang model. It is easy to check that this condition is satisfied as long as $\ddot{R} > 0$, which means that the Universe had to undergo a phase of accelerated expansion.

Condition (6.16) immediately solves the flatness problem, as can be seen by plugging it into equation (6.15): now, as the Universe inflates, Ω is pushed closer and closer to 1, no matter what the initial value is. Stated differently, during inflation the radius of curvature grew much bigger than the Hubble radius, so that the observable portion of the Universe appears to be flat. Inflation also easily solves the horizon problem. During inflation, small, causally connected regions of the Universe rapidly grew much faster than the causal horizon. Regions of the Universe that appear to be causally disconnected at late times were actually in causal contact before inflation began. The accelerated expansion of the Universe during inflation also diluted any unwanted relic, whose density rapidly became negligible with respect to the total energy density.

Inflation is not an alternative to the big bang model. It is an additional ingredient, a mechanism added to the model at very early times to explain its

evolution at later times. Actually, the inflationary phase lasts for a very short time, after which the Universe evolves according to the standard big bang model.

There is no universally accepted and tested model for inflation. There are a number of viable candidates, all of which are based in one way or another on the dynamics of a weakly coupled, homogeneous scalar field ϕ [1,21]. In its simplest form, the equation of motion of such a field is

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0 \tag{6.17}$$

and its energy density and pressure are given by

$$\rho_\phi = \tfrac{1}{2}\dot{\phi}^2 + V(\phi) \tag{6.18}$$

$$p_\phi = \tfrac{1}{2}\dot{\phi}^2 - V(\phi) \tag{6.19}$$

The expansion of the Universe contributes a friction term in the equation of motion through H. The exact shape of the potential V depends on the specific model of inflation. Note from equation (6.8) that the condition for inflation $\ddot{R} > 0$ requires that $\rho_\phi + 3p_\phi < 0$. This is satisfied as long as $\dot{\phi}^2 < V(\phi)$, i.e. if the field potential energy overcomes its kinetic energy. This implies that, during inflation, the field must be moving very slowly down the potential hill. In fact, a common solution to the field equation of motion is based on the so-called *slow-roll* approximation, which assumes that the field acceleration $\ddot{\phi}$ is negligible, so that

$$3H\dot{\phi} \simeq -V'(\phi). \tag{6.20}$$

The conditions for the slow-roll assumption to hold are given by

$$\epsilon \equiv \frac{1}{16\pi G}\left(\frac{V'}{V}\right)^2 \ll 1 \qquad |\eta| \equiv \frac{1}{8\pi G}\left|\frac{V'}{V}\right| \ll 1 \tag{6.21}$$

where ϵ and η are called the *slow-roll parameters*. Clearly, given an arbitrary V, the existence of a slow-roll regime ensures that inflation can take place. The potential remains roughly constant during slow roll and $\dot{\phi}^2 \ll V$, so that the Friedmann equation is simply:

$$H^2 \simeq \frac{8\pi G}{3}V(\phi) \tag{6.22}$$

since, as R grows, the term k/R^2 rapidly decays and can be neglected. This Friedmann equation has an exponential solution for R. The logarithmic amount of expansion between time t_1 and t_2, called the *number of e-foldings* N, is then given by

$$N \equiv \ln\left(\frac{R(t_2)}{R(t_1)}\right) = \int_{t_1}^{t_2} \mathrm{d}t\, H = -8\pi G \int_{\phi_1}^{\phi_2} \mathrm{d}\phi\, \frac{V(\phi)}{V'(\phi)}. \tag{6.23}$$

About 70 e-foldings (i.e. an expansion by a factor $\sim 10^{30}$) are enough to solve the problems of the big bang model: in realistic models of inflation, this is obtained in about 10^{-35} s. Inflation comes to an end when the field reaches the minimum of the potential and it starts rapid, damped oscillations, dissipating its energy through particle creation (a process called *reheating*). From now on, the evolution of the Universe can be described by the standard big bang model.

One important feature of inflation is that it provides a mechanism for generating super-horizon primordial density perturbations in the early Universe. Broadly speaking, the mechanism goes as follows: consider a generic quantum fluctuation $\delta\phi(\boldsymbol{x}, t)$ in the scalar field ϕ. The Fourier expansion coefficients of this fluctuation are $\delta\phi_k$. During inflation, the wavelength of each Fourier component will rapidly grow much bigger than the causal horizon. When this happens, the corresponding fluctuation will 'freeze', since no causal mechanism will be able to influence its evolution. At later times, long after inflation ends, each wavelength will re-enter the horizon and the associated component of the fluctuation will be seen as a density perturbation. Note that there is no way of producing such a mechanism in classical cosmology: in the standard big bang model, a certain comoving scale becomes smaller than the causal horizon at some given time and remains inside the horizon ever after. In a similar way, inflation also produces a stochastic background of gravitational waves. Gravitational waves correspond to tensor perturbations in the spacetime metric, while density perturbations are scalar. Density perturbations produced during inflation are *adiabatic* or *isentropic*: they are genuine curvature perturbations in the spacetime metric and leave the ratio of matter and radiation (or of any other two species) constant at any point in space. Furthermore, they are *Gaussian distributed* (or very close to Gaussian). The power spectrum of density perturbations produced by inflation in the slow-roll approximation is quite simple:

$$P_{\rm s}(k) = A_{\rm s}k^{n_{\rm s}} \qquad P_{\rm t}(k) = A_{\rm t}k^{n_{\rm t}} \tag{6.24}$$

for scalar and tensor density perturbations respectively, with

$$n_{\rm s} = 1 - 4\epsilon + 2\eta \qquad n_{\rm t} = 2\epsilon. \tag{6.25}$$

Of course, since in the slow-roll regime η and ϵ must both be very small, inflationary models usually predict a scalar spectral index very close to 1, a property termed *scale invariance*. Similarly, the power spectrum of tensor perturbations should be roughly constant, since $n_{\rm t} \simeq 0$. The ratio of the amplitude of tensor and scalar perturbations must satisfy the so-called consistency relation $r \equiv A_{\rm t}/A_{\rm s} = 13.6\epsilon$. Measuring the power spectrum of density perturbations is then a powerful tool for testing the inflationary parameters.

Summing up, the inflation mechanism proved quite powerful as a refinement of the standard big bang model and is now considered as an important ingredient of the standard cosmological model. Independently of the details of the specific model, the inflationary scenario makes a number of testable predictions:

- the Universe must be very close to flat;
- primordial density perturbations in the Universe are Gaussian distributed, adiabatic and have a power-law power spectrum; and
- a stochastic background of gravitational waves should be present in the Universe.

Furthermore, constraining the slow-roll parameters by measuring the exact shape of the power spectrum of primordial perturbations can rule out specific models of inflation.

6.2.3 The cosmic budget

The evolution of the Universe in the big bang model is essentially determined by its content. The total density parameter in a multi-component Universe is the sum of the density parameters of the single components:

$$\Omega = \sum_i \Omega_i. \tag{6.26}$$

Assuming that each component has an equation of state of the sort $p = w\rho$, with w independent of time, the Friedmann equation can be written as

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2\left[\sum_i \Omega_i a^{-3(1+w_i)} + (1-\Omega)a^{-2}\right] \tag{6.27}$$

where the density parameters are evaluated at present time. One of the main tasks of observational cosmology is to obtain accurate estimates of the parameters in the right-hand side of the Friedmann equation: the Hubble constant and the contributions to Ω from the various components in the Universe. Let us analyse each of these components in turn.

6.2.3.1 Radiation

The radiation component of the Universe (relativistic particles) has equation of state $p_{\mathrm{R}} = \rho_{\mathrm{R}}/3$. When the Universe is radiation dominated, the scale factor evolves as $a \propto t^{1/2}$. According to the standard cosmological model, today the radiation in the Universe is made of the cosmic microwave background photons and three species of relic massless neutrinos. The present radiation density can be expressed in terms of the photon temperature T, as

$$\rho_{\mathrm{R}} = \frac{\pi^2}{30} g_\star T^4 \tag{6.28}$$

where $g_\star$ counts the total number of effectively massless degrees of freedom. This can be computed, giving $g_\star = 3.36$, while the cosmic microwave background average temperature is accurately measured to be $T = 2.725 \pm 0.001$ K. Thus, today the radiation gives a totally negligible contribution to the critical density: $\Omega_{\mathrm{R}} = 4.31 \times 10^{-5} h^{-2}$.

6.2.3.2 Matter

The equation of state of matter or non-relativistic particles is $p_{\rm M} = 0$, so that, during matter domination, the scale factor evolves as $a \propto t^{2/3}$. The most familiar contribution to matter in the Universe comes from baryons (or nucleons). The abundance of light elements produced in the early Universe is strongly dependent on the baryon-to-photon ratio, which is directly related to the present baryon density. Measurement of primordial abundances of D, ^{3}He, ^{4}He, ^{7}Li are a strong probe of the baryon density and indicate that baryons contribute to roughly 5% of the critical density. If $\Omega \sim 1$, as predicted by inflation and now accurately confirmed by cosmological observations, most of the Universe is not made of the same stuff of which we are made!

There is strong observational evidence that a large contribution (about 30%) to the critical density comes from so-called *dark matter*. Theoretically, the most plausible candidate for dark matter is some heavy, weakly-interacting massive particle, left from the very early stages of the evolution of the Universe. The standard picture for the production of such a relic is as follows. The candidate particle is assumed to be initially in thermal equilibrium with the primordial plasma, so that its abundance decreases as $\exp(-M_{\rm X}/T)$ where $M_{\rm X}$ is the particle mass and T is the photon temperature. When the interaction rate of the particle, Γ, becomes smaller than the expansion rate of the Universe, H, the particle decouples from the thermal plasma and its abundance becomes constant (a moment known as *freeze-out*). Then, a cosmologically relevant relic abundance can be achieved provided the particle has a large enough mass and a small enough interaction rate. There are many candidates for dark matter (for example, supersymmetric partners): unfortunately, since it interacts so weakly, direct detection of dark matter proves challenging. Some light on the nature of dark matter can be shed by accurate measurements of its present density by cosmological observations.

6.2.3.3 Dark energy

In its most general form, Einstein equation includes a so-called cosmological term Λ in addition to the familiar stress–energy tensor:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu} + \Lambda g_{\mu\nu}. \tag{6.29}$$

Adding a cosmological constant term is completely equivalent to introducing a new contribution to the stress-energy tensor from a component with

$$\rho_{\rm V} = \Lambda/8\pi G \qquad p_{\rm V} = -\Lambda/8\pi G. \tag{6.30}$$

It can be shown that this is exactly the kind of contribution resulting from zero-point fluctuations of quantum fields or *vacuum energy*. The equation of state of vacuum energy is $p_{\rm V} = -\rho_{\rm V}$ and the Universe expands exponentially when it is vacuum dominated: $a \propto \exp[(\Lambda/3)^{1/2} t]$.

The introduction of this seemingly harmless contribution to the energy density of the Universe has unfortunately disturbing implications. First of all, any estimate of plausible values for the vacuum energy density from fundamental physics exceeds the critical density ρ_c by at least 40 orders of magnitude, while observational cosmology sets the total energy density of the Universe at roughly the critical value, $\Omega \sim 1$. One might hope that some mechanism is leading to an exact cancellation of the contributions to the vacuum energy, so that it is exactly $\rho_V = 0$: however, such a mechanism is currently unknown. The situation is even more puzzling, since recent observations of distant type Ia supernovae [33, 37] have shown that we live in a Universe that has just entered a vacuum-dominated epoch, starting a phase of accelerated expansion. This means that the cosmological constant term is still very small compared to theoretical estimates but it is large enough ($\rho_V/\rho_c \sim 0.7$) to be cosmologically relevant in the present Universe. There seems to be a serious fine-tuning problem: if Λ is non-zero, then why is it so small? Furthermore, given the observed value of Λ, the vacuum energy was never important in the past evolution of the Universe but it is starting to be the dominant contribution at present time: we then seem to live in a very special moment in the Universe, an annoying coincidence indeed!

The vacuum-energy problem may, in fact, be the biggest mystery of modern physics [38]. A possible way to alleviate it, and one that has interesting and testable implications for cosmology, is to consider a generalization of the cosmological constant term, that has been termed *dark energy*. As shown when discussing inflation, a scalar field ϕ with effective potential $V(\phi)$ has an equation of state with $w = (\dot{\phi}^2/2 - V)/(\dot{\phi}^2/2 + V)$. Any value of w such that $1 + 3w < 0$ results in an accelerated expansion, so it is dynamically equivalent to a cosmological constant. The interesting feature of these models is that they admit *tracking solutions*, in which the dark energy can reach the present value starting from a very different set of initial conditions. This mitigates the fine tuning and coincidence problems but, of course, leaves open the questions about the nature of the field ϕ. Cosmological constraints to w can be able to discriminate among dark energy models by saying something about the scalar field potential V. An excellent review on dark energy from the point of view of both cosmology and fundamental physics is [27].

6.3 The cosmic microwave background

The cosmic microwave background (CMB) is one of the primary tools for investigating the physics of the early Universe and constraining the parameters of the standard cosmological model. It provides a picture of the Universe when it was only a few hundred thousand years old and its physics can be described by simple thermodynamics and linear perturbation theory.

6.3.1 The primordial plasma and the CMB

The CMB was serendipitously discovered in 1964 by radio astronomers Arno Penzias and Robert Wilson, as an excess noise in the radio antenna they were testing at the Bell Labs in Homdel, New Jersey [32]. The CMB intensity observed by Penzias and Wilson was highly isotropic (i.e. it was independent of the direction of observation in the sky) and resulted in being consistent with the emission expected from a black body at a temperature of about 3 K.

The existence of a thermal background radiation has a natural explanation in the standard hot big bang model [9]. According to this model, the temperature in the early Universe is so high that neutral atoms cannot exist. Frequent Thomson scattering in the primordial plasma maintains photons and free electrons in thermodynamical equilibrium. Such a system is characterized by a black body energy spectrum and is completely described by thermodynamical quantities, like the black body temperature T. The plasma optical depth (i.e. the mean number of collisions experienced by a photon from a certain time to the present) is given by

$$\tau(\eta) = -\int_{\eta_0}^{\eta} \mathrm{d}\eta'\, c\sigma_{\mathrm{T}} n_{\mathrm{e}}(\eta') a \tag{6.31}$$

where η_0 is the present conformal time (with $\mathrm{d}\eta = c\,\mathrm{d}t/a$), σ_{T} is the Thomson cross section and n_{e} is the number density of free electrons. The mean free path of photons, $\dot{\tau}^{-1} = 1/\sigma_{\mathrm{T}} n_{\mathrm{e}} a$, is very close to zero in the primordial plasma. The fraction of free electrons X_{e} at any given time is approximately governed by the Saha equilibrium equation:

$$\frac{X_{\mathrm{e}}^2}{1 - X_{\mathrm{e}}} = \frac{(2m_{\mathrm{e}}kT)^{3/2}}{n_{\mathrm{H}}} \mathrm{e}^{-B/kT} \tag{6.32}$$

where $B = 13.6$ eV is the binding energy of hydrogen. The formation of neutral hydrogen atom, a process known as *recombination*, can take place as the Universe cools down at about $T_\star \simeq 3000$ K: when this happens, X_{e} rapidly drops to zero and photons can travel essentially unimpeded. The transition from $\dot{\tau}^{-1} \simeq 0$ to $\dot{\tau}^{-1} \to \infty$ is quite rapid and happens at $a_\star \simeq 10^{-3}$ or $t_\star \sim 10^5$ yr after the big bang. The CMB is made of the photons we receive from this epoch, cooled down by the Universe expansion so that $T_0 = T_\star a_\star \simeq 3$ K. Its black body energy spectrum is a result of the matter–radiation thermodynamical equilibrium existing at early times.

6.3.2 The anisotropy of the CMB

The standard picture for understanding structure formation in the Universe is based on gravitational instability: the observed large-scale structure formed by gravitational amplification of small density perturbations generated in the early Universe. In such a scenario, the presence of anisotropies in the temperature distribution of the CMB is unavoidable, as density fluctuations leave an imprint

in the CMB at the time of photon–matter decoupling. The first calculations of the expected anisotropy of the CMB were done by Sachs and Wolfe [39], who predicted the level of anisotropy induced by fluctuations in the gravitational potential and by Silk [41], who computed the amplitude of density fluctuations at recombination needed to produce galaxies. It was immediately clear that, despite the high level of isotropy observed by Penzias and Wilson, fluctuations in the CMB temperature had to exist in order to explain the level of inhomogeneity observed in the present Universe.

The largest anisotropy observed in the CMB is not intrinsic but originates from the fact that our reference frame (the Solar System) moves at speed $v_\odot$ relatively to the CMB photons. This gives rise to a dipole anisotropy due to the Doppler effect:

$$\Delta T/T \approx (v_\odot/c)\cos(\theta) \tag{6.33}$$

where θ is the angle between the direction of motion and the line of sight[2].

Intrinsic CMB fluctuations generated at or before recombination are called *primary anisotropies*. The main contributions to primary anisotropies are as follows.

- Intrinsic fluctuations: $\Delta T/T \propto \delta$, where δ is the matter density perturbation at recombination: if the perturbations leave unchanged the entropy of the radiation per baryon (*adiabatic fluctuations*), then $\Delta T/T = 1/4\delta_\gamma = 1/3\delta$.
- Velocity-induced fluctuations: $\Delta T/T \propto v/c$, where v is the peculiar velocity of the matter at decoupling: photons get extra energy when scattered by matter in motion.
- Gravitational potential fluctuations: $\Delta T/T \propto \delta\phi/c^2$, where $\delta\phi$ is the fluctuation in the gravitational potential at decoupling: photons leaving the perturbed region have to 'climb' out of the potential well, experiencing gravitational redshift (Sachs–Wolfe effect [39]).

Interactions experienced by the background photons between recombination and the present may give rise to sub-dominant effects, the so-called *secondary anisotropies*. Possible sources of secondary anisotropies include gravitational lensing [6], reionization of the intergalactic medium, inverse Compton scattering by free electrons in hot intracluster gas (thermal Sunyaev–Zel'dovich effect [44]), variation of the gravitational potential after decoupling (integrated Sachs–Wolfe effect [39]), passage through nonlinear structures (Rees–Sciama effect [36]), etc.

6.3.3 The statistics of the CMB

In inflationary models, primordial density perturbations follow a Gaussian statistics: the probability of having a density contrast $\delta(\boldsymbol{x}) \equiv \delta\rho(\boldsymbol{x})/\bar{\rho}$, at some point of space $\boldsymbol{x}$ and at some initial time t_i is proportional to $\exp(-\delta^2/\sigma^2)$. Small

[2] Incidentally, from the observed dipole amplitude, which is $\sim$ 3 mK [4], we measure $v_\odot \sim$ 600 km s^{-1}

deviations from Gaussianity are predicted in non-standard inflationary models and in non-inflationary scenarios: however, the observed upper limits to these deviations are very small and, therefore, I will restrict the discussion to the case of Gaussian density perturbations.

The evolution of density perturbations is more easily understood by expanding the density contrast in plane waves:

$$\delta(\boldsymbol{x}, t) = \frac{V}{(2\pi)^3} \int_{\mathrm{V}} \mathrm{d}^3 k \, \delta_k(t) \exp(-\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}) \tag{6.34}$$

where $V = L^3$ is the fundamental volume and periodic boundary conditions have been imposed. Every component of the expansion, $\delta_k(t)$, describes the evolution of a perturbation of given characteristic scale $\lambda \equiv 2\pi/k$. Note that the Fourier components of the density field only depend on the module of the wavenumber, k, because of the isotropy of the Universe. An appealing consequence of working in Fourier space is that Gaussian initial conditions imply that each coefficient δ_k is a Gaussian random variable with zero average, $\langle \delta_k \rangle = 0$, and random phases, i.e. the modes corresponding to different wave numbers are uncorrelated:

$$\langle \delta_k \delta_k' \rangle = 0 \qquad \text{if } k \neq k' \tag{6.35}$$

and their evolution can be followed independently. The symbol $\langle \cdot \rangle$ denotes the average on the statistical ensemble, i.e. on every possible realization of the statistical field. The statistical properties of Gaussian random density perturbations are completely described by the power spectrum $P(k) \equiv \langle |\delta(k)|^2 \rangle$. A power-law power spectrum, $P(k) = Ak^n$, as predicted by inflation, is usually assumed.

The evolution of perturbations in linear regime leaves their statistical properties unchanged. So the temperature fluctuation of the CMB on the sky, $\delta T/T$, is a two-dimensional random Gaussian field: such a field is completely described in terms of the two-point correlation function:

$$C(\alpha) = \left\langle \frac{\delta T}{T}(\hat{\gamma}_1) \frac{\delta T}{T}(\hat{\gamma}_2) \right\rangle \tag{6.36}$$

where α is the angle between the directions of observation $\hat{\gamma}_1$ and $\hat{\gamma}_2$.

The anisotropy as a function of the direction of observation can be expanded in spherical harmonics:

$$\frac{\delta T}{T}(\hat{\gamma}) = \sum_{lm} a_{lm} Y_{lm}(\hat{\gamma}). \tag{6.37}$$

The assumption of Gaussianity implies that each coefficient a_{lm} is a Gaussian random variable, with zero mean,

$$\langle a_{lm} \rangle = 0 \tag{6.38}$$

and covariances

$$\langle a_{lm} a^{\star}_{l'm'} \rangle = C_l \delta_{ll'} \delta_{mm'}. \tag{6.39}$$

The coefficients $C_l \equiv \langle |a_{lm}|^2 \rangle$ represent the angular power spectrum of the CMB anisotropy in multipoles space. The C_ls are independent on the azimuthal index m as a consequence of the isotropy of the Universe. The angular correlation function becomes

$$C(\alpha) = \sum_l \frac{2l+1}{4\pi} C_l P_l(\cos\alpha) \tag{6.40}$$

where P_l are the Legendre polynomials.

Then, for Gaussian initial conditions, the angular power spectrum C_l carries all the statistical information on the angular temperature anisotropy of the CMB.

The variance of the temperature fluctuations is given by:

$$\left\langle \left| \frac{\delta T}{T} \right|^2 \right\rangle = \sum_l \frac{2l+1}{4\pi} C_l. \tag{6.41}$$

Each C_l is associated with an angular scales θ given by

$$\theta \approx \frac{180^\circ}{l}. \tag{6.42}$$

The causal horizon at decoupling subtends an angular scale on the sky which is approximately given by

$$\theta_{\rm H} \approx 1^\circ \Omega^{1/2} \tag{6.43}$$

so the power spectrum for $l \lesssim 200$ is left unchanged by physical processes occurring prior to the decoupling[3] and is fixed only by initial conditions while, for $l \gtrsim 200$, it will show a strong dependence on cosmological parameters.

The angular power spectrum of the CMB can be computed in a given cosmological model by following the evolution of each mode of the density perturbations. Each a_{lm} will be a superposition of contributions from different modes, $a_{lm}(k)$, and the linearity of the evolution implies that each $a_{lm}(k)$ is proportional to the initial density perturbation δ_k. Figure 6.2 shows an example of a theoretical power spectrum calculated for a fiducial cosmological model. Note the absence of features at low multipoles and the presence of a series of peaks in the higher l region. I will say more on the dependence of these peaks on cosmological parameters later on.

A final note on the statistics of the C_ls. They represent the variances of independent Gaussian random variables: $C_l = \langle |a_{lm}|^2 \rangle$. The ensemble average should be computed on any possible realization of the random field but, in real,

[3] Of course, post-recombination effects can alter these regions of the spectrum.

life we only have one Universe to observe. It can be shown that a maximum-likelihood estimator of C_l can be computed as

$$\tilde{C}_l = \frac{1}{2l+1} \sum_{m=-l}^{l} |a_{lm}|^2. \tag{6.44}$$

The estimator $\tilde{C}_l$ is unbiased, i.e. its ensemble average is equal to C_l. Furthermore, it is distributed as a chi-square with $2l + 1$ degrees of freedom, so that its variance is

$$\mathrm{Var}(\tilde{C}_l) = \frac{2}{2l+1} C_l. \tag{6.45}$$

This unavoidable uncertainty, deriving from the fact that we only have $2l + 1$ samples to estimate a given C_l, is called *cosmic variance* and represents the fundamental theoretical limit to the accuracy of any measurement of the angular power spectrum.[4]

6.3.4 Computing the anisotropy

As I will show in more detail later, the observed level of anisotropy of the CMB is very small, about 10^{-5} with respect to the average. This means that the Universe was extremely smooth at recombination and that linear perturbation theory can be applied in order to calculate the expected CMB temperature anisotropy.

The full general-relativistic treatment of the evolution of linear perturbations in an expanding Universe was first developed by Lifshitz in 1946 [22]. In this approach, small metric perturbations $h_{\mu\nu}$ are added to a flat, homogeneous and isotropic spacetime metric $\eta_{\mu\nu}$ and the corresponding perturbed Einstein equations are solved.

The fact that each density perturbation mode evolves independently greatly simplifies the calculation. However, calculating the expected CMB anisotropy for a given cosmological model is not a trivial task at all. The evolution of perturbations has to be followed from very early times to the present, using a complicated set of coupled differential equations [23]. In a standard adiabatic inflationary scenario, the quantities to be computed include the matter density contrast δ and velocity $\beta \equiv v/c$, the metric perturbation $h_{\mu\nu}$ and the radiation distribution function perturbation f_γ and density contrast δ_γ.

Peebles and Yu [31] first derived the equation that governs the evolution of linear fluctuations in the photon brightness. After some manipulations, this can be rewritten in Fourier space as an equation for the k-component of the CMB temperature fluctuation observed in the sky direction $\hat{\gamma}$ at conformal time η, indicated by $\Delta = \Delta(\hat{\gamma}, k, \eta)$:

$$\dot{\Delta} + \mathrm{i}kc\mu\Delta + H = -\dot{\tau}(\delta_\gamma/4 + \mu\beta - \Delta). \tag{6.46}$$

[4] A larger uncertainty arises when observing limited regions of the sky. In this case the variance is increased because of the fact that even fewer modes are available due to the finite size of the observed patch. This is called *sample variance*.

Here, the dot indicates the derivative with respect to conformal time. The reference frame was chosen so that the z-axis is aligned with the $\boldsymbol{k}$ vector for each Fourier mode and $\boldsymbol{k} \cdot \hat{\gamma} \equiv k\mu$. All of the metric perturbations terms are collected in the quantity $H \equiv (1 - 3\mu^2)\partial h_{33}/\partial t - (1 - \mu^2)\partial h/\partial t$.

Over the years, many groups have developed numerical codes to solve the coupled set of equations for Δ and for the other relevant physical quantities in an exact way, for any particular cosmological model. The standard approach was to make use of the Legendre expansion

$$\Delta(\mu, k, \eta) = \sum_l (-\mathrm{i})^l (2l + 1)\Delta_l(k, \eta) P_l(\mu) \tag{6.47}$$

resulting in a hierarchy of differential equations for the Δ_l at arbitrary l. The CMB angular power spectrum can then be computed using the formula:

$$C_l = \frac{2}{\pi} \int \mathrm{d}k^2 \, P(k) |\Delta_l(k, \eta_0)|^2. \tag{6.48}$$

In 1996, Seljak and Zaldarriaga [40] developed a fast algorithm based on the line of sight solution:

$$\Delta(\eta_0)\mathrm{e}^{\mathrm{i}kc\mu\eta_0} = \int_0^{\eta_0} \mathrm{d}\eta \, [\dot{\tau}(\delta_\gamma/4 + \mu\beta) + H]\mathrm{e}^{-\tau}\mathrm{e}^{\mathrm{i}kc\mu\eta}. \tag{6.49}$$

Using the identity $\mathrm{e}^{\mathrm{i}kc\mu\eta} = \sum_l (-i)^l (2l + 1) j_l[kc(\eta)] P_l(\mu)$, where j_l are the Bessel functions, allows the l dependence to be isolated in a purely geometrical model-independent factor that can be precomputed for all the models. This opened up the possibility of computing a large number of theoretical predictions for the CMB anisotropy in different cosmological models in a reasonable time.

Simplified analytical treatments can be used to obtain a physical intuition on the different processes governing the generation of anisotropies in the CMB, and to understand how the peaks in the power spectrum depend on cosmological parameters [18]. One useful approximation is to assume that recombination happens instantaneously, and that matter and radiation are perfectly coupled ($\dot{\tau}^{-1} = 0$) before recombination. This simplifies the previous line of sight integral, since now the dominant contribution to the anisotropy originates from an infinitely thin shell at $\eta = \eta_\star$, and is given by

$$\Delta_l(\eta_0) \approx \Delta_0(\eta_\star) j_l[kc(\eta_\star - \eta_0)] \tag{6.50}$$

where I have made use of the identity $\Delta_0 = \delta_\gamma/4$. Since the Bessel function $j_l(x)$ is sharply peaked at $l \simeq x$, the main contribution to each C_l will come from k modes such that $kc(\eta_\star - \eta_0) \equiv kD_\star \simeq l$, where $D_\star$ is the distance to the shell at $\eta_\star$ where matter and radiation decouple. It then only remains to compute the source term $\Delta_0(\eta_\star)$.

The equation for Δ_0 in this tight-coupling approximation is

$$\ddot{\Delta}_0 + \frac{\dot{a}}{a}\frac{\dot{\mathcal{R}}}{1+\mathcal{R}}\dot{\Delta}_0 + k^2 c_s^2 \Delta_0 = -\frac{k^2}{3}\phi \tag{6.51}$$

where $\mathcal{R} \equiv 3\rho_B/4\rho_\gamma$ and $c_s = c/[3(1+\mathcal{R})]^{1/2}$ is the sound velocity. The quantity

$$\phi \equiv \frac{1}{2k^2}\left(\ddot{h} + \frac{\dot{a}}{a}\frac{\mathcal{R}}{1+\mathcal{R}}\dot{h}\right) \tag{6.52}$$

is the Newtonian gravitational potential.

So the evolution of perturbations prior to recombination in the tight-coupling regime is essentially governed by a forced harmonic oscillator equation. This has to be completed with the standard generalization of the continuity equation, the Euler equation and the Poisson equation, in order to describe the velocity, pressure, density and gravitational potential of the fluid.

We can make some additional simplifying assumption. First of all, we can treat the gravitational potential ϕ as independent of η, which is true in a matter-dominated Universe. The other simplification is to assume that any characteristic time scale is small compared to the Universe expansion time scale, so that we can neglect $\dot{a}/a$. With these hypotheses, the equation becomes

$$\ddot{\Delta}_0 + k^2 c_s^2 \Delta_0 = -k^2\phi/3 \tag{6.53}$$

or

$$(1+\mathcal{R})\ddot{\Delta}_0 + \tfrac{1}{3}k^2c^2\Delta_0 = -\tfrac{1}{3}k^2(1+\mathcal{R})\phi. \tag{6.54}$$

The solution to this equation is

$$\Delta_0(\eta) = A_1\cos(kc_s\eta) + A_2\sin(kc_s\eta) - (1+\mathcal{R})\frac{\phi}{c^2}. \tag{6.55}$$

If we fix the initial conditions as $\dot{\Delta}_0(0) = 0$ and $\Delta_0(0) = -2\phi/3c^2$ (the latter comes from Poisson and Euler equations), we obtain

$$\Delta_0 = (\tfrac{1}{3}+\mathcal{R})\frac{\phi}{c^2}\cos(kc_s\eta) - (1+\mathcal{R})\frac{\phi}{c^2}. \tag{6.56}$$

The physical interpretation of this solution is very simple. The baryon-to-photon ratio $\mathcal{R}$ acts like an effective mass in the harmonic oscillator equation through the factor $(1+\mathcal{R})$. The baryons tend to collapse due to self-gravitation. The restoring force is provided by the radiation pressure $k^2c^2/3$. This sets up acoustic oscillations in the baryon–photon fluid (the sound velocity c_s quantifies the resistance of the fluid to compression): the higher $\mathcal{R}$ is, the larger the amplitude of the oscillations is. The driving force term due to gravitation, constant in our approximation, simply displaces the zero point of the oscillations. Increasing $\mathcal{R}$ enhances this displacement and gives more amplitude to compressions over rarefactions, because of the increased inertia of the fluid.

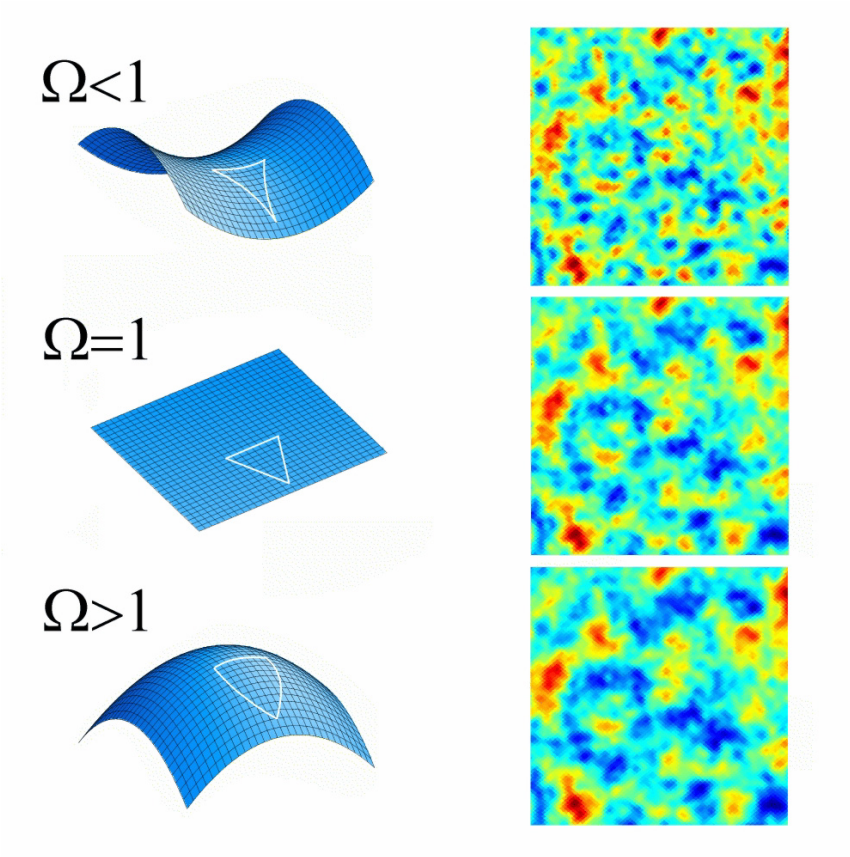

Figure 6.1. The effect of the geometry of the Universe on the CMB temperature anisotropy pattern. Because of light-ray geodesic deviation, the same linear scale is seen under different angles depending on the curvature of the metric (see figures on the left). This effect modifies the apparent typical dimension of the temperature fluctuations (as shown on the simulated CMB maps on the right).

If we freeze the oscillations at the time of decoupling, $\eta_\star$, each mode will be caught in a different stage of oscillation. The total power will have the largest contributions $|\Delta_0(\eta_\star)|^2$ from modes having $kc_s\eta_\star \equiv kS_\star = m\pi$, where I have introduced the sound horizon at recombination, $S_\star$. This results in a harmonic series of peaks in the angular power spectrum at locations $l_m = ml_1$, where the location of the first peak is $l_1 = \pi D_\star/S_\star$. Odd peaks are due to compression of the fluid, even peaks to rarefaction: so the odd peaks will generally be higher than the even peaks because of $\mathcal{R}$. Increasing the baryon content will enhance this effect. The fact that the k modes corresponding to peaks in the power spectrum are

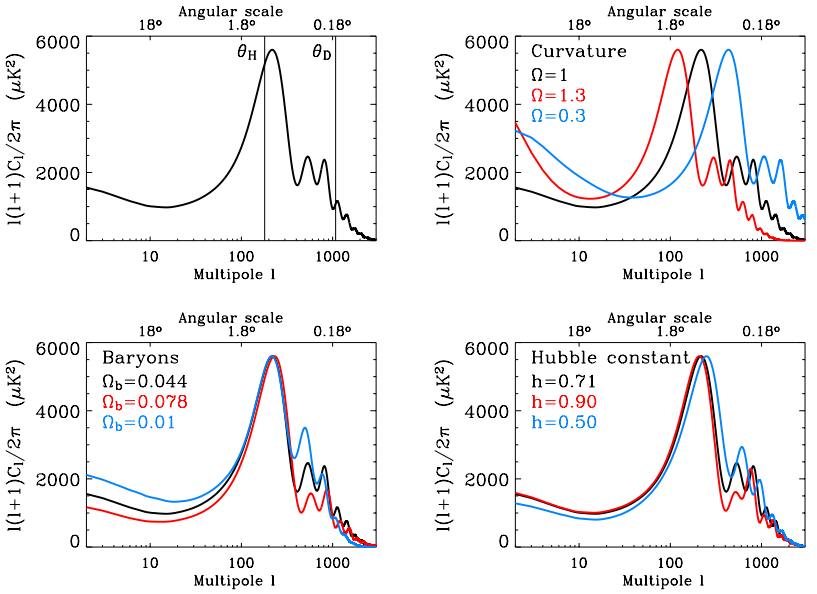

Figure 6.2. The angular power spectrum of the CMB temperature anisotropy. In the upper left-hand panel, the theoretical power spectrum corresponding to the inflationary model which best fits current observations. The vertical lines correspond to the approximate angular scale of the causal horizon θ_H and to the damping scale θ_D. In the remaining plots, the effect of varying some cosmological parameters with respect to the fiducial model: clockwise, the curvature of the Universe, related to the total density parameter Ω, the Hubble constant $H_0 = 100h$ km s^{-1} Mpc^{-1} and the density parameter of baryons Ω_b.

fixed by the size of the sound horizon at recombination $S_\star \equiv c_s\eta_\star$ explains why the peaks were called 'acoustic'. While in a flat Universe, the relation between k modes and multipole l is, as we saw, fixed by the distance $D_\star$, in curved space this is not so simple. For example, if the Universe is open, then the geodesics are curved with respect to the flat case, so that the same physical scale on the last scattering surface is seen today under a smaller angle. The position of the peaks then moves to higher ls. This projection effect is illustrated in figure 6.1.

To complete this rapid overview of the physics of CMB anisotropy, I should mention an effect arising from the failure of the simple tight-coupling approximation used in the previous discussion. Approaching recombination, the mean free path of the photon's interaction with matter increases gradually and the hypothesis of perfect coupling breaks down. A photon can escape from an over-dense region through random walk: this results in the fact that perturbations in the photon fluid on scales smaller than a certain characteristic length (related to the mean free path of the photons) are erased. This shows up as a damping (known

as *Silk damping* [41]) of the peaks in the power spectrum at high ls. The damping length is given by $\lambda_D^2 \sim k_D^{-2} \sim c^2 \int d\eta\, \dot{\tau}^{-1}$ and, for flat models, corresponds to $l \sim 1000$ in the power spectrum. The damping factor is approximately $e^{-(k/k_D)^2}$.

Finally, variations in the gravitational potential (that was held constant in the simplified picture discussed earlier) boost the amplitude of the peaks through the forcing term ϕ. Since the potential changes during the radiation-dominated epoch, decreasing the Hubble constant h while keeping fixed the other parameters will increase this effect by delaying the matter-radiation equality.

Many additional effects exist that can only be followed through a more detailed treatment. The results of a full numerical solution of the exact equations governing the evolution of the perturbations is shown in figure 6.2, showing the angular power spectrum as a function of different cosmological parameters. Despite the many simplifying assumptions made in the previous discussion, the qualitative behaviour described here is clearly visible in the results of these accurate calculations.

6.4 Past, present and future of CMB observation

As I showed in the previous sections, there is a lot to learn from observations of the CMB anisotropy. The last decade has been a period of intense experimental activity in this field, which has resulted in a number of impressive achievements and in a huge progress in our understanding of physical cosmology.

6.4.1 The COBE satellite

NASA launched the COBE (COsmic Background Explorer) satellite in 1989, with the purpose of performing full sky observations of the CMB from space (figure 6.3). The COBE results were first announced in 1992, causing a revolution in observational and theoretical cosmology.

The FIRAS (Far Infra-Red Absolute Spectrometer) instrument aboard COBE measured the energy spectrum of the CMB with stunning precision. The black body nature of the CMB was proved conclusively, signing a huge success of the big bang model (figure 6.4) [11]. The CMB temperature measured by FIRAS is 2.725 ± 0.001 K [24].

The DMR (Differential Microwave Radiometer) instrument detected, for the first time, tiny temperature fluctuations in the CMB, of the order of 10^{-5}, at angular scales of about 7° [42]. The importance of this result cannot be overemphasized. The CMB anisotropy measured by COBE was interpreted as being cosmological in origin, reflecting inhomogeneities in the distribution of matter in the universe at the time of decoupling. This is crucial for understanding the initial conditions that seeded the formation of the large-scale structure we observe in the present Universe. The microwave sky observed by COBE is shown in the upper map of figure 6.6 [3]. This map does not include the dipole anisotropy

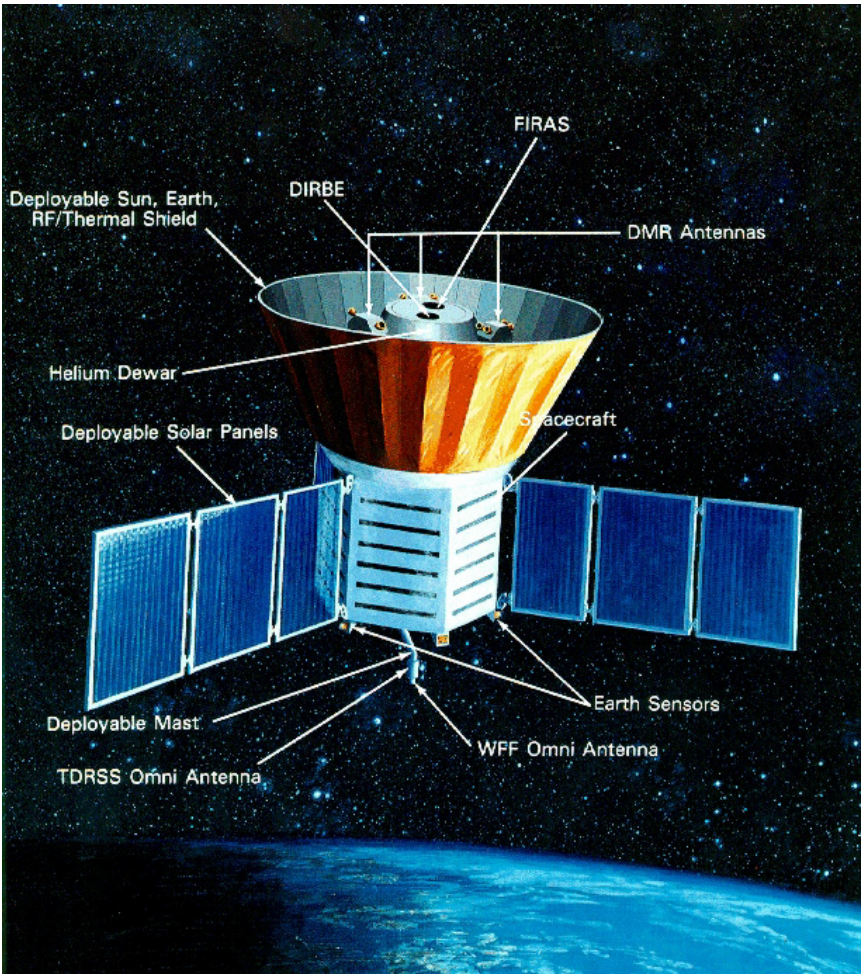

Figure 6.3. NASA's COBE satellite.

(corresponding to temperature differences in directions making an angle of 180°) of order 10^{-3} of the average temperature, which is due to our motion with respect to the background photons.

6.4.2 The hunt for the peaks

In the decade following the release of the COBE results, the experimental efforts focused on measuring the CMB anisotropy at intermediate and small angular scales. These scales, encoding most information on the early universe and on cosmological parameters, were not accessible to COBE because of its low angular

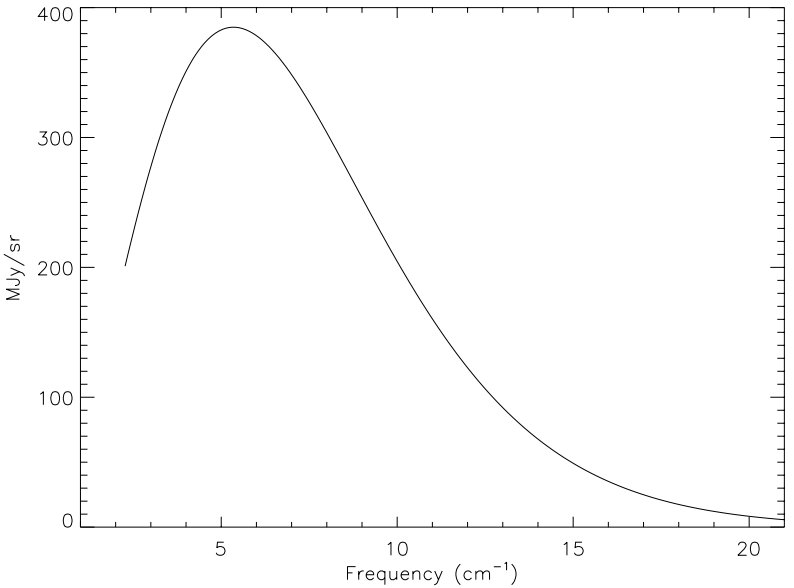

Figure 6.4. The energy spectrum of the CMB, as measured by the FIRAS instrument on board COBE, superimposed to the prediction for a black body at temperature 2.725 K. The measured data (whose error bars are not visible at this scale) are indistinguishable from the theoretical curve. (From [11].)

resolution. Several experiments, conducted from 1992 to 1998, either from the ground or from balloon-borne payloads, explored the CMB angular power spectrum in the region between few arcminutes to about one degree. Although each single experiment could only probe a narrow band in l-space, the combined measurements seemed to indicate a rise in the power spectrum at $l \sim 200$.

Thanks to the progress in detector technology, between 1998 and 2000 the experiments TOCO [25], BOOMERanG [7] and MAXIMA [16] were able independently, for the first time, to resolve clearly the first acoustic peak in the angular power spectrum. BOOMERanG and MAXIMA also produced the first high-resolution (about 10 arcmin) maps of the CMB, although on small patches of the sky. The detection of the first peak served to support the inflationary scenario and allowed the total energy density of the universe to be measured with unprecedented accuracy. This turned out to be very close to the critical value, $\Omega \simeq 1$, corresponding to a flat universe [2, 7].

Later, in 2001, the DASI [15], BOOMERanG [8] and VSA [13] experiments detected hints of a second acoustic peak in the CMB power spectrum, further strengthening the case for the adiabatic nature of primordial fluctuations. Then, in

Table 6.1. WMAP instrumental features (from [4]).

Channel	K	Ka	Q	V	W
Central frequency (GHz)	22.8	33	40.7	60.8	93.5
Bandwidth (GHz)	5.5	7	8.3	14	20.5
Angular resolution (FWHM)	0.82°	0.62°	0.49°	0.33°	0.21°
Sensitivity (μK per 0.3° pixel)	35	35	35	35	35
Number of channels	4	4	8	8	16

2002, the Archeops [5] experiments secured the measurement of the first acoustic peak and the CBI [28] and ACBAR [20] experiments explored the spectrum at smaller angular scales, measuring the expected damping of primary anisotropy.

6.4.3 The WMAP satellite

The WMAP (Wilkinson Microwave Anisotropy Probe) satellite[5], launched by NASA aboard a Delta rocket on 30 June 2001, represents the state-of-the-art of CMB experiments. In many ways, WMAP is a follow-up to COBE. It was designed to make full-sky map of CMB anisotropy by looking at temperature differences in the sky, using differential radiometers in five frequency bands. WMAP scans large regions of the sky in relatively short times, with a strong cross-linking among observations performed at different times: this is very useful to control systematic effects and correlated instrumental noise. WMAP operates from the L2 Lagrangian point, completing a full sky coverage in a six-month period. WMAP detector technology is based on HEMT (High Electron Mobility Transistor) radiometers, passively cooled at about 90 K. WMAP's main instrumental features are summarized in table 6.1

Results of the first year of observations by WMAP (August 2001–02), corresponding to two full-sky surveys, were announced at the beginning of 2003 (see [4] and companion papers cited therein). Data collected later are currently being analysed. Figure 6.6 shows the CMB map produced by WMAP at 94 GHz. The pattern of anisotropy is clearly consistent with that observed by COBE after four years of observation. WMAP has 30 times better resolution than COBE. When the WMAP map is degraded at COBE resolution, the difference map is below the instrumental noise level.

Figure 6.7 shows the CMB temperature anisotropy power spectrum measured by WMAP. This is the best currently available measurement of the power spectrum and is cosmic variance limited up to $l \simeq 350$. WMAP results provide an extraordinary confirmation of the theoretical predictions. The presence of at least two acoustic peaks in the power spectrum is evident. The

[5] http://map.gsfc.nasa.gov

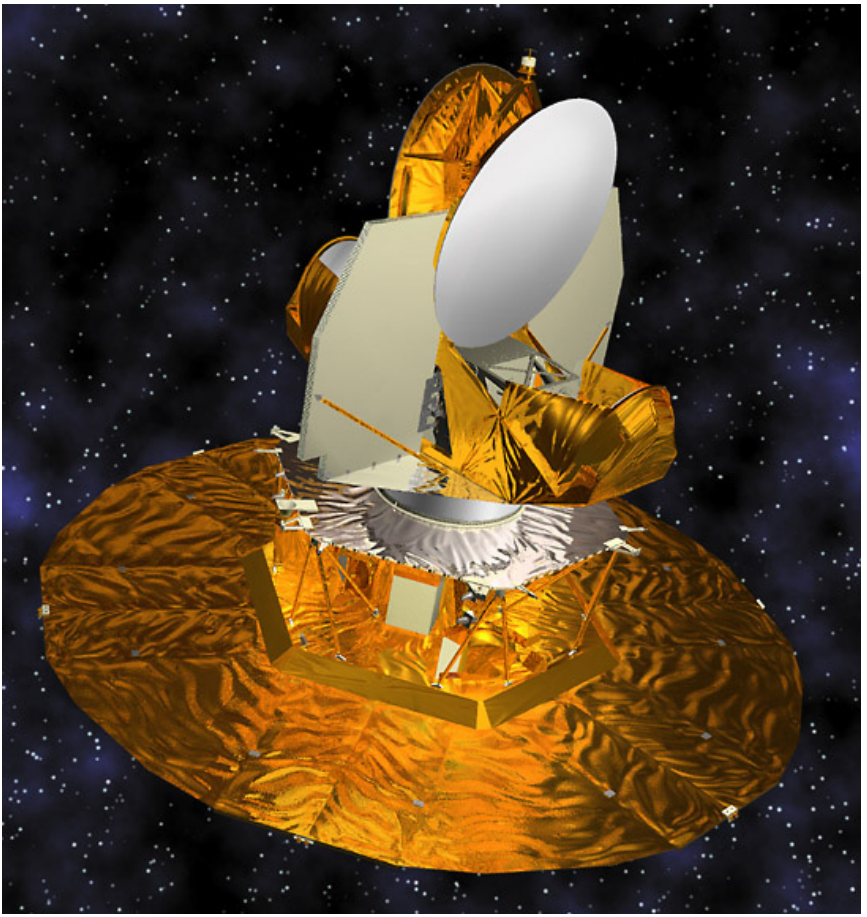

Figure 6.5. NASA's WMAP satellite (courtesy of NASA/WMAP science team).

cosmological interpretation of these results lends further support to the standard cosmological model based on big bang plus inflation. A flat universe, with adiabatic, Gaussian, scale-invariant primordial density fluctuations is perfectly consistent with the WMAP data. The values of the main cosmological parameters, estimated using the WMAP data, are summarized in table 6.2. These values are generally more precise than those obtained with other kinds of observations, and are consistent with them. For example, the baryon density at recombination measured by WMAP is in agreement with big bang nucleosynthesis predictions and measurements of the primordial abundance of light elements [10, 26, 34]; the Hubble constant value agrees with the measurement by the Hubble Space Telescope [12]; the age of the Universe is consistent with the value from stellar observables [17, 35]; and finally, the dark matter content of the Universe is in

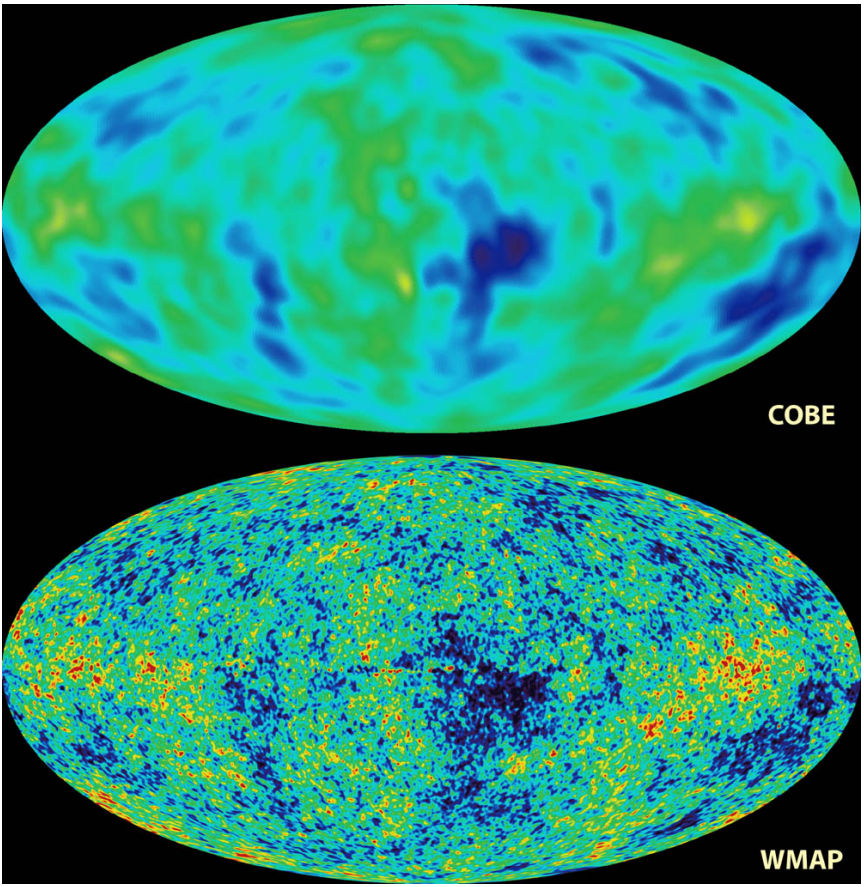

Figure 6.6. Maps of CMB temperature anisotropy. The sky, in galactic coordinates, is represented in Mollweide projection. Bright spots are hotter the average, dark spots colder. Temperature scale is about ± 100 μK. In the upper panel, the map produced by COBE/DMR after four years of observation [3]. In the lower panel, the map produced by WMAP after one year of observation [4]. WMAP has 30 times better resolution than COBE. The contributions from galactic emission and the dipole anisotropy have been subtracted from the maps. (Courtesy NASA/WMAP science team.)

agreement with the one derived by the large-scale matter distribution [45]. The low value of the matter density, combined with the fact that $\Omega \simeq 1$, confirms that most of the energy density in the Universe is provided by dark energy, as recently indicated by high-redshift type Ia supernovae observations [33, 37]. The outstanding concordance among completely different kinds of observations testifies to the level of maturity reached by cosmology in recent times.

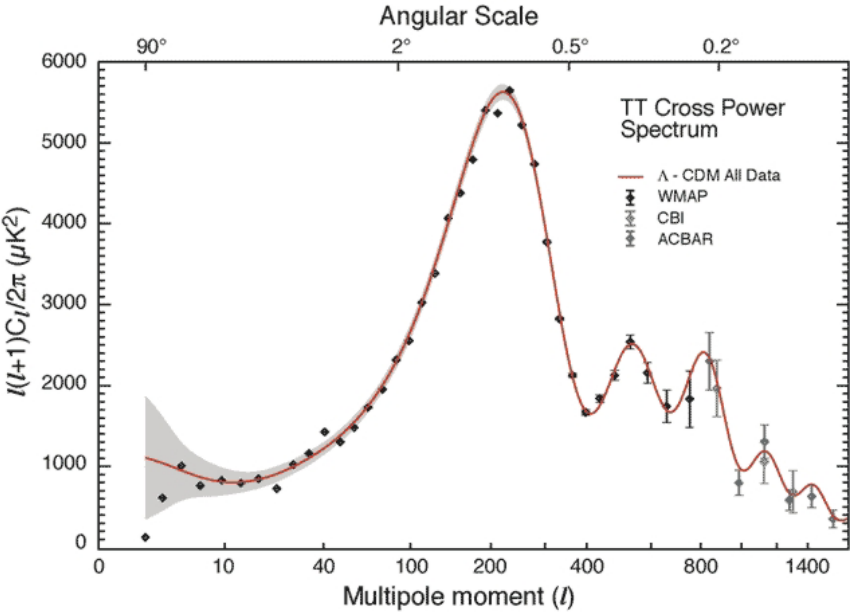

Figure 6.7. The angular power spectrum of CMB temperature anisotropy. The dots are the measurements from the WMAP, ACBAR and CBI experiments. The continuous curve is the theoretical model which best fits the data. The grey region represents the cosmic variance uncertainty for this theoretical model. (From [4].)

Table 6.2. Some cosmological parameters estimated by WMAP (from [4]).

Parameter	Symbol	Value
Total density	Ω	1.02 ± 0.02
Baryon density	$\Omega_{\rm b}$	0.044 ± 0.004
Dark matter density	$\Omega_{\rm m}$	0.27 ± 0.04
Dark energy density	Ω_Λ	0.73 ± 0.04
Equation of state of dark energy	w	< -0.78 (95% C.L.)
Hubble constant (km s^{-1} Mpc^{-1})	H_0	71^{+4}_{-3}
Age of the Universe (Gy)	t_0	13.7 ± 0.2
Optical depth of the Universe	τ	0.17 ± 0.04
Spectral index of primordial density perturbations	$n_{\rm s}$	0.93 ± 0.03

WMAP mission has been approved for four years of operation in L2. In the next few years, further data and analysis will provide more and more detailed cosmological information.

6.4.4 The Planck Surveyor

Despite its extraordinary achievements, the WMAP mission does not represent the end of the story. Much remains to be told about the CMB temperature anisotropy. On one hand, WMAP angular resolution does not allow the damping tail of the CMB power spectrum to be investigated: although the first two acoustic peaks in the spectrum are now accurately resolved, higher ls are affected by large uncertainties. Other experiments, especially interferometers, are starting to unveil the small angular scale details of the anisotropy pattern but much work needs to be done. On the other hand, WMAP maps are still affected by a non-negligible instrumental noise, which strongly reduces the possibility of direct pixel space analyses.

ESA's Planck Surveyor[6], planned for launch in 2007, will represent the third-generation CMB space mission (figure 6.8). The main product of the Planck mission will be full-sky maps in nine frequency bands between 30 and 900 GHz. Planck frequency coverage will be the widest ever for a single microwave experiment. This is crucial for separating the various components that constitute the observed signal and will allow the investigation of a large variety of poorly known astrophysical processes, both galactic and extragalactic. Planck will carry on board two different instruments: the HFI (High Frequency Instrument), based on bolometric detectors, and the LFI (Low Frequency Instrument), which uses HEMT radiometers. Exploiting this redundancy and comparison among measurements will be extremely important for the detection and removal of systematics. The main experimental features of Planck are summarized in table 6.3.

Planck's instrumental sensitivity will be several times better than WMAP's. The design of Planck's detectors and optics (a 1.5 m primary mirror and a off-axis secondary, coupled to an array of corrugated horns in the focal plane) will allow the best possible resolution to be obtained at each frequency, making it possible to resolve details of a few arcmin in the sky.

The accuracy of the CMB angular power spectrum measurement by Planck will be limited by cosmic variance and by unavoidable foreground contamination, over the entire range of angular scales relevant to the primary CMB anisotropy, i.e. from $l = 2$ up to $l \sim 1000$, well below the damping scale. This will allow the vast amount of cosmological information encoded in the CMB to be extracted. Planck will be able to measure the cosmological parameters to unprecedented accuracy, minimizing the need of external input from other observations.

The full-sky maps produced by Planck will have a signal-to-noise ratio much larger than 1: this means that Planck's maps will be real pictures of the Universe at recombination. This will allow the accurate investigation of the physical processes which affect the CMB statistics beyond the angular power spectrum, such as small deviations from Gaussianity of the primordial fluctuations, predicted in some theoretical scenarios.

[6] http://astro.estec.esa.nl/Planck

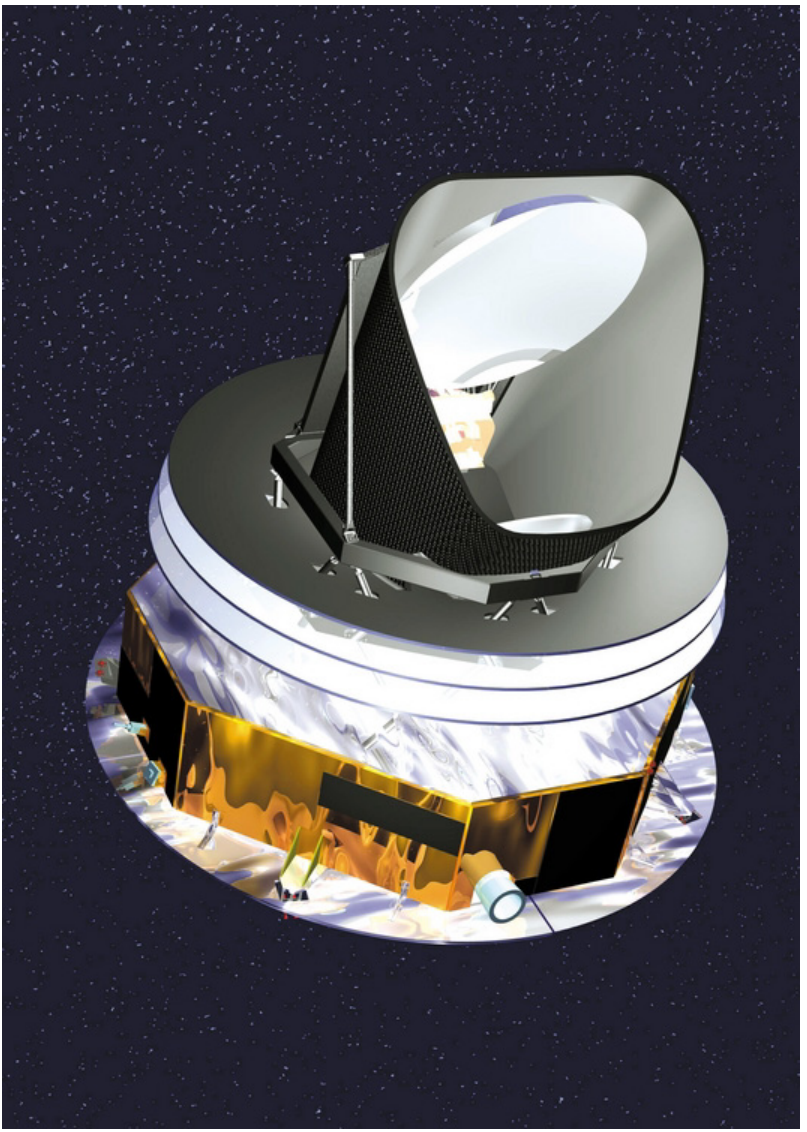

Figure 6.8. ESA's Planck Satellite.

6.5 Conclusions

Cosmology has developed into a fully mature science. The parameters of the big bang model are now known with great accuracy and the constraints are expected to get tighter in the future. Inflation has not been falsified and its main predictions

Table 6.3. Estimated performance of Planck instruments.

Instrument	LFI			HFI					
Detector technology	HEMT			Bolometers					
Detector temperature	20 K			0.1 K					
Cooling requirements	Active			Active					
Central frequency (GHz)	30	44	70	100	143	217	353	545	857
Bandwidth (GHz)	6	8.8	14	33	47	72	116	180	283
Angular resolution (FWHM, arcminutes)	33	24	14	9.2	7.1	5	5	5	5
Sensitivity (temperature) (μK per pixel)	2.0	2.7	4.7	2.0	2.2	4.8	14.7	147	6700
Sensitivity (polarization) (μK per pixel)	2.8	3.9	6.7	—	4.2	9.8	29.8	—	—
Number of unpolarized detectors	0	0	0	4	4	4	4	4	4
Number of polarized detectors	4	6	12	0	8	8	8	0	0

are strikingly consistent with observations. The results obtained using completely different cosmological probes are in remarkable agreement among themselves, as well as with theoretical predictions. Nonetheless, many fundamental questions are still open [29]. The pace of experimental and theoretical progress, however, does not seem to be close to a halt.

References

[1] Albrecht A and Steinhardt P J 1982 *Phys. Rev. Lett.* **48** 1220
[2] Balbi A *et al* 2000 *Astrophys. J.* **545** L1
[3] Bennett C L *et al* 1996 *Astrophys. J.* **464** L1
[4] Bennett C L *et al* 2003 *Astrophys. J. Suppl.* **148** 1
[5] Benoît A *et al* 2003 *Astron. Astrophys.* **399** L19
[6] Blanchard A and Schneider J 1987 *Astron. Astrophys.* **184** 1
[7] de Bernardis P *et al* 2000 *Nature* **404** 955
[8] de Bernardis P *et al* 2002 *Astrophys. J.* **564** 559
[9] Dicke R H, Peebles P J E, Roll P G and Wilkinson D T 1965 *Astrophys. J.* **142** 414
[10] D'Odorico S *et al* 2001 *Astron. Astrophys.* **368** L21
[11] Fixsen D J, Cheng E S, Gales J M, Mather J C, Shafer R A and Wright E L 1996 *Astrophys. J.* **473** 576
[12] Freedman W L *et al* 2001 *Astrophys. J.* **553** 47
[13] Grainge K *et al* 2003 *Mon. Not. R. Astron. Soc.* **341** L23
[14] Guth A H 1981 *Phys. Rev.* D **23** 347
[15] Halverson N W *et al* 2002 *Astrophys. J.* **568** 38
[16] Hanany S *et al* 2000 *Astrophys. J.* **545** L5
[17] Hansen B M S *et al* 2002 *Astrophys. J.* **574** L155
[18] Hu W, Sugiyama N and Silk J 1997 *Nature* **386** 37
[19] Kolb E W and Turner M S 1990 *The Early Universe* (Reading, MA: Addison-Wesley)
[20] Kuo C-L *et al* 2002 *Astrophys. J.* astro-ph/0212289
[21] Linde A D 1982 *Phys. Lett.* B **108** 389
[22] Lifshitz E M 1946 *J. Phys. (Moscow)* **10** 116
[23] Ma C P and Bertschinger E 1995 *Astrophys. J.* **455** 7
[24] Mather J C, Fixsen D J, Shafer R A, Mosier C and Wilkinson D T 1999 *Astrophys. J.* **512** 511
[25] Miller A D *et al* 1999 *Astrophys. J.* **524** L1
[26] O'Meara J M *et al* 2001 *Astrophys. J.* **552** 718
[27] Padmanabhan T 2003 *Phys. Rep.* **380** 235
[28] Pearson T J *et al* 2003 *Astrophys. J.* **591** 556
[29] Peebles P J E 2003 *Preprint* astro-ph/0311435
[30] Peebles P J E 1993 *Principles of Physical Cosmology* (Princeton, NJ: Princeton University Press)
[31] Peebles P J E and Yu J T 1970 *Astrophys. J.* **162** 815
[32] Penzias A A and Wilson R W 1965 *Astrophys. J.* **142** 419
[33] Perlmutter S *et al* 1999 *Astrophys. J.* **517** 565
[34] Pettini M and Bowen D V 2001 *Astrophys. J.* **560** 41
[35] Reid I N 1997 AJ **114** 161
[36] Rees M and Sciama D 1968 *Nature* **519** 611

[37] Riess A G *et al* 2001 *Astrophys. J.* **560** 49
[38] Rugh S E and Zinkernagel H 2002 *Studies in History and Philosophy of Modern Physics* **33** 663
[39] Sachs R K and Wolfe A M 1967 *Astrophys. J.* **147** 73
[40] Seljak U and Zaldarriaga M 1996 *Astrophys. J.* **469** 437
[41] Silk J 1968 *Astrophys. J.* **151** 459
[42] Smoot G F *et al* 1992 *Astrophys. J.* **396** L1
[43] Starobinsky A A 1979 *JETP Lett* **30** 682
[44] Sunyaev R A and Zel'dovich Ya B 1972 *Comments Ap. Space Sci.* **4** 173
[45] Verde L *et al* 2002 *Mon. Not. R. Astron. Soc.* **335** 432

Chapter 7

Strings, gravity and particle physics

Augusto Sagnotti[1] *and Alexander Sevrin*[2]
[1] *Dipartimento di Fisica*
Università di Roma 'Tor Vergata', INFN—Sezione di Roma 'Tor Vergata', Via della Ricerca Scientifica, 100133 Roma, Italy
[2] *Theoretische Natuurkunde, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium*

7.1 Introduction

One of the main achievements of physics is certainly the reduction of all forces in Nature, no matter how diverse they might appear at first sight, to four fundamental types: gravitational, electromagnetic, weak and strong. The last three, in particular, are nicely described by the standard model, a Yang–Mills gauge theory where the gauge group $SU(2)_L \times U(1)_Y \times SU(3)_{QCD}$ is spontaneously broken to $U(1)_{em} \times SU(3)_{QCD}$. A gauge theory is a generalization of Maxwell's theory of electromagnetism whose matrix-valued potentials satisfy nonlinear field equations even in the absence of matter and the corresponding gauge bosons are the quanta associated with their wave modes. For instance, the W and Z bosons, quanta of the corresponding W_μ and Z_μ gauge fields, are charged under one or more of the previous gauge groups and are, thus, mutually interacting, an important feature well reflected by their nonlinear field equations. The other key ingredient of the standard model, the spontaneous breaking of $SU(2)_L \times U(1)_Y$ to $U(1)_{em}$, is a sort of Meissner effect for the whole of space time that is held responsible for screening the weak force down to very short distances. It relies on a universal low-energy description of the phenomenon in terms of scalar modes and, therefore, the search for the residual Higgs boson (or, better, Brout–Englert–Higgs or BEH boson) is perhaps the key effort in experimental particle physics today. Whereas the resulting dynamics is very complicated, the standard model is *renormalizable* and this feature allows reliable and consistent perturbative analyses of a number of quantities of direct interest for particle physics. These

have now been tested by very precise experiments, as we have heard in several Moriond talks and, therefore, leaving aside the BEH boson that is yet to be discovered, a main problem today is ironically the very good agreement between the current experiments and the standard model, with the consequent lack of clear signals for new physics in this domain.

Despite the many successes of this framework, a number of aesthetic and conceptual issues have long puzzled the theoretical physics community: in many respects, the standard model does not have a compelling structure, while gravity cannot be incorporated in a satisfactory fashion. In fact, gravity differs in crucial respects from the other fundamental forces, since it is very weak and plays no role in atomic and nuclear physics: for instance, the Newtonian attraction in a hydrogen atom is lower than the corresponding Coulomb force by an astonishing factor—42 orders of magnitude. Moreover, the huge ratio between Fermi's constant G_{F} and Newton's constant G_{N} that determines the strength of the weak and gravitational interactions at low energies, $G_{\mathrm{F}}/G_{\mathrm{N}} \sim 10^{35}\hbar^2c^{-2}$, poses by itself a big puzzle, usually called the *hierarchy problem*: it is unnatural to have such a large number in a fundamental theory and, in addition, virtual quantum effects in the vacuum mixing the different interactions would generally make such a choice very unstable. *Supersymmetry*, an elegant symmetry between boson and fermion modes introduced in this context by J Wess and B Zumino in the early 1970s, can alleviate the problem by stabilizing the hierarchy but does not eliminate the need for such unnatural constants. It also predicts the existence of Fermi and Bose particles degenerate in mass and, therefore, it cannot be an exact feature of our low-energy world, while attaining a fully satisfactory picture of supersymmetry breaking is a major challenge in present attempts.

In sharp contrast with the other three fundamental forces, Newtonian gravity is purely attractive so that, despite its weakness in the microscopic realm, it dominates the large-scale dynamics of our universe. General relativity encodes these infrared properties in a very elegant way and, taken at face value as a quantum theory, it would associate with the gravitational interaction an additional fundamental carrier, the graviton, that would be on the same footing as the photon, the gluons and the intermediate W and Z bosons responsible for the weak interaction. The graviton would be a massless spin-2 particle, and the common tenet is that its classical Hertzian waves have escaped direct detection for a few decades only due to their feeble interactions with matter. Unlike standard model interactions, however, general relativity is not renormalizable, essentially because the gravitational interaction between point-like carriers that, as we shall see in more detail at the end of section 7.2, is measured by the effective coupling

$$\alpha_{\mathrm{N}}(E) \sim G_{\mathrm{N}}E^2/\hbar c^5 \tag{7.1}$$

grows rapidly with energy, becoming strong at the Planck scale $E_{\mathrm{Pl}} \approx 10^{19}$ GeV, defined so that $\alpha_{\mathrm{N}}(E_{\mathrm{Pl}}) \approx 1$. This scale, widely beyond our means of investigation if not of imagination itself, is, in principle, explored by virtual quantum processes and, as a result, unpleasant divergences arise in the

quantization of general relativity that, in modern terms, seems to provide, at most, an effective description of gravity at energies well below the Planck scale. This is the *ultraviolet problem* of Einstein gravity and this state of affairs is not foreign. Rather, it is somewhat reminiscent of the way in which the Fermi theory describes the weak interactions well below the mass scale of the intermediate bosons, $E_{\rm W} \approx 100$ GeV, where the effective Fermi coupling $\alpha_{\rm F}(E) \sim G_{\rm F}E^2/\hbar^3c^3$ becomes of order one. It is important to keep in mind that this analogy, partial as it may be, lies at the heart of the proposed link between string theory and the fundamental interactions.

String theory provides a rich framework for connecting gravity to the other forces and, indeed, it does so in a way that has the flavour of the modifications introduced by the standard model in the Fermi interaction: at the Planck scale new states appear, in this case actually an infinity of them, that result in an effective weakening of the gravitational force. This solves the ultraviolet problem for four-dimensional gravity but the resulting picture, still far from complete, raises a number of puzzling questions that still lack a proper answer and which are, thus, actively investigated by many groups. One long-appreciated surprise, of crucial importance for the ensuing discussion, is that string theory, in its more popular, or more tractable, supersymmetric version, requires that our spacetime include *six additional dimensions*. Despite the clear aesthetic appeal of this framework, however, let us stress that, in dealing with matters that could be so far beyond the currently accessible scales, it is fair and wise to avoid untimely conclusions, keeping also an eye on other possibilities. These include a possible thinning of the spacetime degrees of freedom around the Planck scale, that would solve the ultraviolet problem of gravity in a radically different fashion. For the Fermi theory, this solution to its ultraviolet problem would assert the impossibility of processes entailing energies or momenta beyond the weak scale. While this is clearly not the case for weak interactions, we have no fair way to exclude that something of this sort could actually take place at the Planck scale, on which we have currently no experimental clues. This can be regarded as one of the key points of the canonical approach to quantum gravity, long pursued by a smaller community of experts in general relativity.

With this proviso, we can return to string theory, the main theme of our discussion. Ideally, one should demand from it two characteristics: some sort of uniqueness, in order to make such a radical departure from the standard model, a four-dimensional field theory of point particles, more compelling and some definite path for connecting it to the low-energy world. The first goal has been achieved, to a large extent, in the last decade after the five *supersymmetric* string models, usually called type IIA, type IIB, heterotic SO(32) (or, for brevity, HO), heterotic $\mathrm{E}_8 \times \mathrm{E}_8$ (or, for brevity, HE) and type I, have been argued to be equivalent as a result of surprising generalizations of the *electric magnetic duality* of classical electrodynamics. Some of these string dualities are nicely suggested by perturbative string theory and, in fact, can also connect other non-supersymmetric ten-dimensional models to the five superstrings, while others

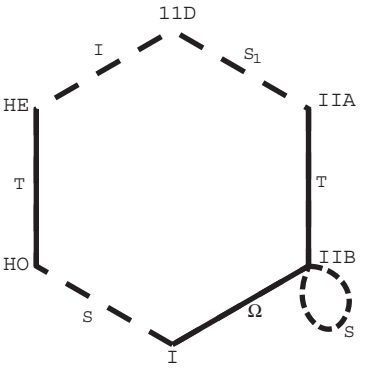

Figure 7.1. The five ten-dimensional superstring theories are dual to one another. The bold full lines denote perturbative dualities, while the broken ones indicate non-perturbative ones. At strong coupling, both type IIA and heterotic $E_8 \times E_8$ strings develop an additional large dimension, a circle (S_1) and an interval (I_1) respectively. Therefore, they are both described by an 11-dimensional theory but this bears no direct relation to strings.

rest on the unique features of ten-dimensional supergravity. Supergravity is an elegant extension of general relativity, discovered in the mid-1970s by S Ferrara, D Z Freedman and P van Nieuwenhuizen, that describes the effective low-energy dynamics of the light superstring modes, where additional local supersymmetries require corresponding gauge fields, the *gravitini*, and bring about, in general, other matter fields. In ten dimensions, supergravity is *fully determined* by the type of supersymmetry involved, (1,0), (1,1) or (2,0), where the numbers count the (left and right) Majorana–Weyl ten-dimensional supercharges and, in the first case, by the additional choice of a Yang–Mills gauge group and this rigid structure allows one to make very strong statements [1]. The end result is summarized in the duality hexagon of figure 7.1, where the bold links rest on perturbative string arguments, while the broken ones reflect non-perturbative features implied by ten-dimensional supergravity. The resulting picture, provisionally termed 'M-Theory', has nonetheless a puzzling feature: it links the ten-dimensional superstrings to the eleven-dimensional Cremmer–Julia–Scherk (CJS) supergravity, that can be shown to bear no direct relation to strings!

An additional vexing problem is that the reduction from ten dimensions to our four-dimensional spacetime entails a deep lack of predictivity for the low-energy parameters that depend on the *size and shape* of the extra dimensions. This fact reflects the absence of a global minimum principle for gravity, similar to

[1] This counting is often a source of confusion: in four dimensions a Weyl spinor has two complex, or four real, components, while in ten dimensions the corresponding minimal Majorana–Weyl spinor has 16 real components, four times as many. Thus, the minimal (1,0) ten-dimensional supersymmetry is as rich as $\mathcal{N} = 4$ in four dimensions, while a similar link holds between the (1,1) and (2,0) cases and $\mathcal{N} = 8$ in four dimensions.

those that determine the ground states of a magnet in a weak external field below its Curie temperature or the spontaneous breaking of the electroweak symmetry in the standard model and represents a stumbling block in all current approaches that aim at deriving our low-energy parameters from string theory. It has long been hoped that a better understanding of string dynamics would help in bypassing this difficulty but, to date, no concrete progress has been made on this crucial issue. Thus, ironically, by what we currently understand, string theory appears to provide a unique answer to the problem of including gravity in the standard model but the four-dimensional remnants of this uniqueness are at least classes of theories. Supersymmetry again has a crucial effect on this problem, since it basically *stabilizes* the internal geometry, much along the lines of what we have seen for the hierarchy between the electroweak and Planck scales but, as a result, the sizes and shapes (moduli) of the extra dimensions are apparently *arbitrary*. This is the *moduli problem* of supersymmetric vacua, a problem indeed, since the resulting low-energy parameters generally depend on the moduli. However, the breaking of supersymmetry, a necessary ingredient to recover the standard model at low energies if we are to describe Fermi and Bose fields of different masses, tends to destabilize the background spacetime. The end result is that, to date, although we know a number of scenarios to break supersymmetry within string theory, that we shall briefly review in section 7.5, we have little or no control on the resulting spacetimes once quantum fluctuations are taken into account.

The following sections are devoted to some key issues raised by the extension from the standard model to string theory, in an attempt to bring some of the main themes of current research to the attention of the interested reader, while using as starting points basic notions of electrodynamics, gravitation and quantum mechanics.

In a similar spirit, we are also happy to contribute to the Proceedings of the 2002 SIGRAV School, where part of this material was presented in a lecture given by the first author.

7.2 From particles to fields

The basic tenet from which our discussion may well begin is that *all matter is apparently made of elementary particles*, while our main theme will be to illustrate why this may not be the end of the story. Particles exchange mutual forces, and the Coulomb force between a pair of static point-like charges q_1 and q_2,

$$|\boldsymbol{F}_{\mathrm{C}}| \sim \frac{|q_1 q_2|}{r^2} \tag{7.2}$$

with an intensity proportional to their product and inversely proportional to the square of their mutual distance, displays a remarkable similarity to the Newton force between a pair of static point-like masses m_1 and m_2,

$$|\boldsymbol{F}_N| \sim \frac{|G_{\mathrm{N}} m_1 m_2|}{r^2}. \tag{7.3}$$

Actually, it has long been found more convenient to think of these basic forces in two steps: some 'background' charge or mass distribution affects the surrounding space creating a *field* that, in its turn, can affect other 'probe' charges or masses, sufficiently small not to perturb the background significantly. In the first case, the classical dynamics is encoded in the Maxwell equations that relate the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$ to electric charges and currents and, as a result, both fields satisfy in vacuum wave equations of the type

$$\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} - \nabla^2\phi = 0. \tag{7.4}$$

These entail retardation effects due to the finite speed c with which electromagnetic waves propagate and as first recognized by Lorentz and Einstein, provide the route to special relativity.

With gravity, the situation is more complicated, since the resulting field equations are highly nonlinear. According to Einstein's general relativity, the gravitational field is a distortion of the spacetime geometry that replaces the Minkowski metric $\eta_{\mu\nu}$ with a generic metric tensor $g_{\mu\nu}$, used to compute the distance between two nearby points as

$$\mathrm{d}s^2 = g_{\mu\nu}(x)\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{7.5}$$

Material bodies follow *universally* curved trajectories that reflect the distorted geometries, while the metric $g_{\mu\nu}$ satisfies a set of nonlinear wave-like equations where the energy–momentum of matter appears as a source. In fact, the nonlinear nature of the resulting dynamics reflects the fact that the gravitational field carries energy and is, therefore, bound to act as its own source. These observations extend a familiar fact: in the local uniform gravitational field $\boldsymbol{g}$ near the earth ground, Newtonian bodies fall according to

$$m_{\mathrm{i}}\boldsymbol{a} = m_{\mathrm{g}}\boldsymbol{g} \tag{7.6}$$

and the equality of the inertial and gravitational masses m_{i} and m_{g} makes this motion *universal*. The resulting 'Equivalence Principle' is well reflected in the distorted spacetime geometry that has inevitably a universal effect on test bodies. The modification in (7.5) cannot be the whole story, however, since a mere change of coordinates can do this to some extent, a simple example being provided by the transition to spherical coordinates in three-dimensional Euclidean space, that turns the standard Euclidean metric

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \tag{7.7}$$

into

$$\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2. \tag{7.8}$$

This simple example reflects a basic ambiguity met when describing the gravitational field via a metric tensor, introduced by the freedom available in the

choice of a coordinate system. Strange as it may seem, this is but another, if more complicated, instance of the ambiguity met when describing the Maxwell equations in terms of the potentials $\boldsymbol{A}$ and Φ, defined via

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \qquad \boldsymbol{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial \boldsymbol{A}}{\partial t} \tag{7.9}$$

a familiar fact of classical electrodynamics. This ambiguity, in the form of *gauge transformations* of parameter Λ,

$$\boldsymbol{A} \to \boldsymbol{A} + \nabla\Lambda \qquad \Phi \to \Phi - \frac{1}{c}\frac{\partial \Lambda}{\partial t} \tag{7.10}$$

does not affect measurable quantities like $\boldsymbol{E}$ and $\boldsymbol{B}$. A suitable combination of derivatives of $g_{\mu\nu}$, known as the Christoffel connection $\Gamma^{\mu}_{\nu\rho}$, is the proper gravitational analogue of the electrodynamic potentials $\boldsymbol{A}$ and Φ. Note the crucial difference: in gravity the potentials are derivatives of the metric field, a fact that has very important consequences, since it essentially determines equation (7.1). In a similar fashion, the gravitational counterparts of the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields can be built from the Riemann curvature tensor $R^{\mu}{}_{\nu\rho\sigma}$, essentially a curl of the Christoffel connection $\Gamma^{\mu}_{\nu\rho}$, that thus contains second derivatives of $g_{\mu\nu}$. Summarizing, gravity manifests itself as a curvature of the spacetime geometry that falling bodies are bound to experience in their motion.

Note that equation (7.10) can also be cast in the equivalent form:

$$\boldsymbol{A} \to \boldsymbol{A} + \frac{\hbar c}{\mathrm{i}q}\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar c}\nabla\mathrm{e}^{\mathrm{i}q\Lambda/\hbar c} \qquad \Phi \to \Phi - \frac{\hbar}{iq}\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar c}\frac{\partial}{\partial t}\mathrm{e}^{\mathrm{i}q\Lambda/\hbar c} \tag{7.11}$$

a rewriting that has a profound meaning, since it is telling us that, in electrodynamics, the effective gauge parameter is a pure phase,

$$\beta = \mathrm{e}^{\mathrm{i}q\Lambda/\hbar c}. \tag{7.12}$$

Quantum mechanics makes this interpretation quite compelling, as can be seen by the following simple reasoning. In classical mechanics, the effect of electric and magnetic fields on a particle of charge q is described by the Lorentz force law,

$$\boldsymbol{F} = q\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{B}\right) \tag{7.13}$$

while quantum mechanics makes use of the Hamiltonian H or the Lagrangian L, from which this force can be obtained by differentiation. Thus, H and L are naturally bound to involve the potentials and so does the non-relativistic Schrödinger equation

$$-\frac{\hbar^2}{2m}\left(\nabla - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}\right)^2\psi + q\Phi\psi = \mathrm{i}\hbar\frac{\partial\psi}{\partial t} \tag{7.14}$$

that maintains its form after a gauge transformation only provided the wavefunction ψ transforms as

$$\psi \to \mathrm{e}^{(\mathrm{i}q/\hbar c)\Lambda}\psi \tag{7.15}$$

under the electromagnetic gauge transformation (7.11), thus leaving the probability density $|\psi|^2$ unaffected. Note that the electromagnetic fields can also be recovered from commutators of the *covariant derivatives* in (7.14): for instance

$$\left[\partial_i - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}_i, \partial_j - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}_j\right] = -\frac{\mathrm{i}q}{\hbar c}(\partial_i \boldsymbol{A}_j - \partial_j \boldsymbol{A}_i) = -\frac{\mathrm{i}q}{\hbar c}\varepsilon_{ijk}\boldsymbol{B}_k. \tag{7.16}$$

If special relativity is combined with quantum mechanics, one is inevitably led to a multi-particle description: quantum energy fluctuations $\Delta E \sim mc^2$ can generally turn a particle of mass m into another and, therefore, one cannot forego the need for *a theory of all particles of a given type*. Remarkably, the field concept is naturally tailored to describe particles, for instance *all* the identical photons in nature and it does so in a relatively simple fashion, via the theory of the harmonic oscillator. A wave equation in fact emerges, from the continuum limit of coupled harmonic oscillators, a basic fact nicely reflected by the corresponding normal modes, as can be seen letting

$$\Phi(\boldsymbol{x}, t) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} f(t) \tag{7.17}$$

in equation (7.4). Quantum mechanics associates with the resulting harmonic oscillators

$$\frac{\mathrm{d}^2 f}{\mathrm{d}t^2} + c^2\boldsymbol{k}^2 f = 0 \tag{7.18}$$

equally spaced spectra of excitations, that represent *identical* particles, each characterized by a momentum $\boldsymbol{p} = \hbar\boldsymbol{k}$, the *photons* in the present example. The allowed energies are

$$E_n(\boldsymbol{k}) = \hbar c|\boldsymbol{k}|(n + \tfrac{1}{2}) \qquad (n = 0, 1, \ldots) \tag{7.19}$$

and the equally spaced spectra allow an identification of the nth excited state with a collection of n photons. Note the emergence of the zero-point energy $\frac{1}{2}\hbar c|\boldsymbol{k}|$, a reflection of the uncertainty principle to which we shall return in the following. Let us add that a similar reasoning for fermions would differ in two respects. First, the Pauli principle would only allow $n = 0, 1$ for each $\boldsymbol{k}$, while, for the general case of massive fermions with momentum $\boldsymbol{p} = \hbar\boldsymbol{k}$, the allowed energies would be, in general,

$$E = \sqrt{c^2\boldsymbol{p}^2 + m^2c^4}(n - \tfrac{1}{2}) \qquad (n = 0, 1). \tag{7.20}$$

Note the *negative* zero-point energy which should be compared with the *positive* zero-point energy for bosons. Incidentally, equal numbers of boson and fermion

types degenerate in mass would result in an exactly vanishing zero-point energy, a situation realized in models with *supersymmetry*.

This brings us naturally to a brief discussion of the *cosmological constant problem*, a wide mismatch between macroscopic and microscopic estimates of the vacuum energy density in our universe. Note that, in the presence of gravity, an additive contribution to the vacuum energy has sizeable effects: energy, just like mass, gives rise to gravitation and, as a result, a vacuum energy appears to endow the universe with a corresponding average curvature. Macroscopically, one has a time scale $t_{\rm H} \sim 1/H \sim 10^{17}$ s, where the Hubble constant H characterizes the expansion rate of our universe, and a simple dimensional argument associates to it an energy density $\rho_{\rm M} \sim H^2c^2/G_{\rm N}$. One can attempt a theoretical estimate of this quantity, following Ya B Zel'dovic, taking into account the zero-point energies of the quantum fields that describe the types of particles present in nature. A quantum field, however, even allowing no modes with wavelengths below the Planck length $\ell_{\rm Pl} = \hbar c/E_{\rm Pl} \approx 10^{-33}$ cm, the Compton wavelength associated with the Planck scale where, as we have seen, gravity becomes strong, would naturally contribute via its zero-point fluctuations a Planck energy per Planck volume, or $\rho_{\rm m} \sim E_{\rm Pl}^4/(\hbar c)^3$. Using equation (7.1) to relate $G_{\rm N}$ to $E_{\rm Pl}$, the ratio between the theoretical estimate of the vacuum energy density and its actual macroscopic value is then

$$\frac{\rho_{\rm M}}{\rho_{\rm m}} \sim \left(\frac{\hbar H}{E_{\rm Pl}}\right)^2 \approx 10^{-120}. \tag{7.21}$$

This is perhaps the most embarrassing failure of contemporary physics and, to many theorists, it has the flavour of the black body problem, where a similar mismatch led eventually to the formulation of quantum mechanics. In a supersymmetric world, the complete microscopic estimate would give a vanishing result since, as we have seen, fermions and bosons give opposite contributions to the vacuum energy. Still, with supersymmetry broken at a scale $E_{\rm s}$ in order to allow for realistic mass differences $\delta M \sim E_{\rm s}/c^2$ between bosons and fermions, one would essentially recover the previous estimate but for the replacement of $E_{\rm Pl}$ with the supersymmetry breaking scale $E_{\rm s}$, so that, say, with $E_{\rm s} \sim 1$ TeV, the ratio in (7.21) would become about 10^{-88}, with an improvement of about 30 orders of magnitude. These naïve considerations should suffice to motivate the current interest in the search for realistic supersymmetric extensions of the standard model with the lowest scale of supersymmetry breaking compatible with current experiments where, if we also account for the contribution of gravity that here we ignored for the sake of simplicity, more sophisticated cancellations can allow the bound to be reduced much further. We should stress, however, that no widely accepted proposal exists today, with or without supersymmetry or strings, to resolve this clash between theoretical physics and the observed large-scale structure of our universe.

We have thus reviewed how all *identical* particles of a given type can be associated with the normal modes of a single field. While these are determined by

the *linear* terms in the field equations, the corresponding nonlinear terms mediate transformations of one particle species into others. This 'micro-chemistry', the object of particle physics experiments, is regulated by conservation laws and, in fact, the basic reaction mechanisms in the standard model are induced by proper generalizations of the electromagnetic 'minimal substitution' $\nabla \to \nabla - (\mathrm{i}q/\hbar c)A$. The basic idea, as formulated by Yang and Mills in 1954, leads to the nonlinear generalization of electrodynamics that forms the conceptual basis of the standard model and can be motivated in the following simple terms. As we have seen, the electromagnetic gauge transformation

$$U = \mathrm{e}^{\mathrm{i}q\Lambda/\hbar c} \tag{7.22}$$

is determined by a pure phase that can be regarded as a one-by-one unitary matrix, as needed, say, to describe the effect of a rotation around the z-axis of a three-dimensional Euclidean space on the complex coordinate $x + \mathrm{i}y$. Thus, one might well reconsider the whole issue of gauge invariance for an arbitrary rotation or, more generally, for $n \times n$ unitary matrices U. What would happen then? First, the electrodynamic potentials would become matrices themselves, while a gauge transformation would act on them as

$$\nabla - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A} \to U\left(\nabla - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}\right)U^{\dagger}. \tag{7.23}$$

Moreover, the analogues of the electric and magnetic fields would become *nonlinear* matrix-valued functions of the potentials, as can be seen repeating the derivation in (7.16) for a matrix potential $\boldsymbol{A}_\mu$, for which

$$\begin{aligned}\left[\partial_\mu - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}_\mu, \partial_\nu - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}_\nu\right] &= -\frac{\mathrm{i}q}{\hbar c}\left(\partial_\mu \boldsymbol{A}_\nu - \partial_\nu \boldsymbol{A}_\mu - \frac{\mathrm{i}q}{\hbar c}[A_\mu, A_\nu]\right) \\ &= -\frac{\mathrm{i}q}{\hbar c}F_{\mu\nu}.\end{aligned} \tag{7.24}$$

Note that the matrix $(\boldsymbol{A}_\mu)_i^j$ and the Christoffel symbol $(\Gamma_\mu)_\nu^\rho$ are actually very similar objects, apart from the fact the latter is not an independent field but a combination of derivatives of $g_{\mu\nu}$.

The resulting Yang–Mills equations

$$\left[\partial_\mu - \frac{\mathrm{i}q}{\hbar c}\boldsymbol{A}_\mu, F^{\mu\nu}\right] = \frac{4\pi}{c}\boldsymbol{J}^\nu \tag{7.25}$$

to be compared with the more familiar Maxwell equations of classical electrodynamics, indeed contain nonlinear (quadratic and cubic) terms that determine the low-energy mutual interactions of gauge bosons. For instance, the familiar Gauss law of electrodynamics becomes

$$\nabla \cdot \boldsymbol{E} - \frac{\mathrm{i}q}{\hbar c}(\boldsymbol{A}\cdot\boldsymbol{E} - \boldsymbol{E}\cdot\boldsymbol{A}) = 4\pi\rho \tag{7.26}$$

that cannot be written in terms of $\boldsymbol{E}$ alone. Note also that the Yang–Mills analogues of $\boldsymbol{E}$ and $\boldsymbol{B}$ are not gauge invariant. Rather, under a gauge transformation,

$$F_{\mu\nu} \to e^{(\mathrm{i}q/\hbar c)\Lambda} F_{\mu\nu} \mathrm{e}^{-(\mathrm{i}q/\hbar c)\Lambda} \tag{7.27}$$

so that the actual observables are more complicated in these non-Abelian theories. An example is, for instance, $\mathrm{tr}(F_{\mu\nu}F^{\mu\nu})$, while a more sophisticated, non-local one, is the Wilson loop

$$\mathrm{tr}P \exp\left(\frac{\mathrm{i}q}{\hbar c}\oint_{\gamma} A_\mu \,\mathrm{d}x^\mu\right) \tag{7.28}$$

where P denotes path ordering, a prescription to order the powers of A_μ according to their origin along the path γ. This non-Abelian generalization of the Aharonov–Bohm phase is of key importance in the problem of quark confinement.

The standard model indeed includes fermionic matter in the form of quark and lepton fields, whose quanta describe three families of (anti)particles but only the leptons are seen in isolation, so that the non-Abelian SU(3) color force is held responsible for the permanent confinement of quarks into neutral composites, the hadrons. The basic interactions of quarks and leptons with the gauge bosons are simple to characterize: as we anticipated, they are determined by minimal substitutions of the type $\nabla \to \nabla - (\mathrm{i}q/\hbar c)\boldsymbol{A}$ but some of them violate parity or, in more technical language, are *chiral*. This fact introduces important constraints due to the possible occurrence of *anomalies* quantum violations of classical conservation laws. To give an idea of the difficulties involved, it suffices to consider the Maxwell equations in the presence of a current,

$$\partial_\mu F^{\mu\nu} = J^\nu. \tag{7.29}$$

Consistency requires that the current be conserved, i.e. that $\partial_\mu J^\mu = 0$ but, in the presence of parity violations, quantum effects can also violate current conservation, making (7.29) inconsistent. Remarkably, the fermion content of the standard model passes this important test, since all potential anomalies cancel among leptons and quarks.

Another basic feature of the standard model is related to the spontaneous breaking of the electroweak symmetry, responsible for screening the weak force down to very short distances or, equivalently, for the masses of the $\mathrm{W}^\pm$ and Z bosons. This is achieved by the BEH mechanism, whereby the whole of spacetime hosts a quartet of scalar fields responsible for the screening. Making a vector massive costs a scalar field which provides the longitudinal polarization of the corresponding waves, so that three scalars are eaten up to build the W^+, W^- and Z bosons, while a fourth massive scalar is left over: this is the Higgs or, more properly, the BEH particle, whose discovery would be a landmark event in particle physics.

After almost three decades, we are still unable to study the phenomenon of *quark confinement* in fully satisfactory terms but we have a host of numerical

Figure 7.2. The first diagram shows a typical contribution to the self-energy of the electron. The virtual particle/anti-particle pair behaves as a small electric dipole, thereby screening the electron charge. Turning to quarks and the strong interaction, in quantum chromodynamics the diagrams of the first kind (where now the wavy line denotes a gluon, rather than a photon) are accompanied by additional ones of the second kind, since the gluons are themselves charged. A direct calculation shows that the anti-screening effect wins, leading to *asymptotic freedom*.

evidence and simple semi-quantitative arguments to justify our expectations. Thus, in quantum electrodynamics (QED) (see figure 7.2), the uncertainty principle fills the actual vacuum with virtual electron–positron pairs, vacuum fluctuations that result in a partial *screening* of a test charge. This, of course, can also radiate and absorb virtual photons that, however, cannot affect the picture since they are uncharged. However, the Yang–Mills vacuum (see figure 7.2) is dramatically affected by the radiation of virtual gauge bosons that are charged and tend to *anti-screen* a test charge. The end result of the two competing effects depends on the relative weight of the two contributions and the colour force in quantum chromodynamics (QCD) is actually dominated by anti-screening. This has an impressive consequence, known as *asymptotic freedom*: quark interactions become feeble at high energies or short distances, as reflected in the experiments on deep inelastic scattering. A naïve reverse extrapolation would then appear to justify intense interactions in the infrared, compatibly with the evident impossibility of finding quarks outside hadronic compounds but no simple quantitative proof of quark confinement has been attained to date along these lines. On the contrary, even if the weak interactions are also described by a Yang–Mills theory, no subtle infrared physics is expected for them, compatible with the fact that leptons are commonly seen in isolation: at scales beyond the Compton wavelength of the intermediate bosons, $\ell_{\mathrm{W}} \sim 10^{-16}$ cm, the resulting forces are, in fact, screened by the BEH mechanism!

While more can be said about the standard model, we shall content ourselves with these cursory remarks, with an additional comment on the nature of the spontaneous breaking. This ascribes the apparent asymmetry between, say, the short-range weak interactions and the long-range electromagnetic interactions to an asymmetry of the vacuum, much in the same way as the magnetization of a bar can be related to a proper hysteresis. As a result, although hidden, the symmetry is still present and manifests itself in full power in high-energy virtual processes, making the theory *renormalizable* like QED, a crucial result recognized by the Nobel prize awarded to G 't Hooft and M Veltman in 1999. A by-product of the

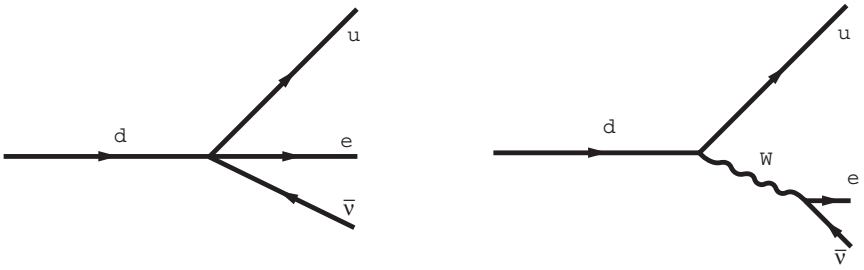

Figure 7.3. At energies significantly below 100 GeV, beta-decay is well described by a four-fermion interaction. However, at higher energies, the interaction is resolved or 'smeared out' by the exchange of a W boson.

BEH mechanism is a simple relation between the Fermi constant and, say, the W mass M_{W}

$$G_{\mathrm{F}} \sim \frac{\hbar^3 \alpha}{c M_{\mathrm{W}}^2} \tag{7.30}$$

with α a dimensionless number of the order of the QED fine-structure constant. This again reflects the fact that the weak forces are completely screened beyond the Compton wavelength of their carriers, $\lambda_{\mathrm{W}} \sim \hbar/cM_{\mathrm{W}}$, but an equivalent, rather suggestive way of stating this result is to note that the growth of the effective fine-structure function

$$\alpha_{\mathrm{F}}(E) \sim \frac{G_{\mathrm{F}} E^2}{\hbar^3 c^3} \tag{7.31}$$

actually *stops* at the electroweak scale $E_{\mathrm{W}} \approx \sqrt{\hbar^3 c^3}/\sqrt{G_{\mathrm{F}}}$ to leave room for an essentially constant coupling. This transition results from the emergence of new degrees of freedom that effectively smear out the local four-Fermi interaction into QED-like exchange diagrams, as in figure 7.3.

Given these considerations, it is tempting and natural to try to repeat the argument for gravity, constructing the corresponding dimensionless coupling,

$$\alpha_{\mathrm{N}}(E) \sim \frac{G_{\mathrm{N}} E^2}{\hbar c^5}. \tag{7.32}$$

The relevant scale is now the Planck scale $E_{\mathrm{Pl}} \approx 10^{19}$ GeV but the problem is substantially subtler, since now energy itself is to be spread and this is where strings come into play. According to figures 7.4 and 7.5, a simple, if rather crude, argument to this effect is that if a pair of point masses experiencing a hard gravitational collision are replaced with strings of length ℓ_{s}, asymptotically only a fraction of their energies is effective in the interaction, so that $\alpha_{\mathrm{N}}(E)$ actually saturates to a finite limiting value, $G_{\mathrm{N}}\hbar/\ell_{\mathrm{s}}^2 c^3$. This simple observation can be taken as the key motivation for strings in this context and, indeed, a detailed

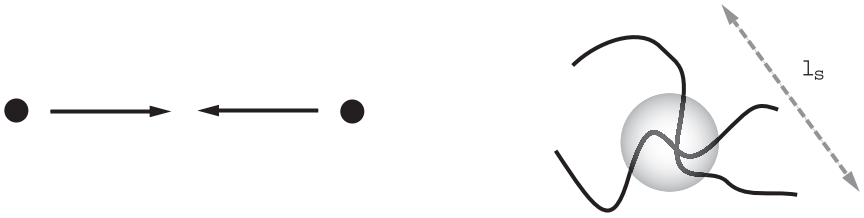

Figure 7.4. A hard gravitational collision can be softened if point particles are replaced by strings of length ℓ_s. The energy is then spread over their length, while at high energies the effective gravitational coupling $\alpha_N(E) = G_N E^2/\hbar c^5$ is replaced by $\alpha_N(E) \times ((\hbar c/E)/\ell_s)^2$, according to the fraction of the energy effective in the collision.

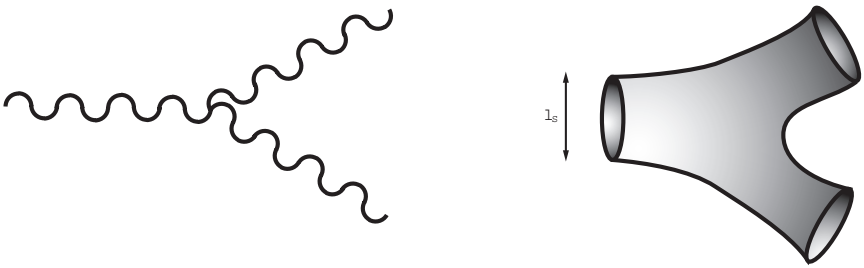

Figure 7.5. Replacing point particles with strings, a graviton three-point interaction (left) is 'smeared' to such an extent that, in the resulting three-string vertex (right), even a localized interaction point can no longer be found.

analysis shows that the ultraviolet problem of gravity is absent in string theory. A subtler issue is to characterize what values of ℓ_s one should actually use, although naïvely the previous argument would tend to identify ℓ_s with the Planck length $\ell_{Pl} \approx 10^{-33}$ cm.

7.3 From fields to strings

This brings us naturally to strings that clearly come in two varieties, *open* or *closed*. It is probably quite familiar that an ordinary vibrating string has an infinity of harmonics that depend on the boundary conditions at its ends but whose frequencies are essentially multiples of a fundamental tone. In a similar fashion, a single relativistic string has an infinity of tones, naïvely related to an infinity of masses according to

$$m^2 \sim N\omega^2 \qquad (N \geq 1) \tag{7.33}$$

with N an integer and, thus it apparently describes an infinity of massive particle species. There is a remarkable surprise, however: the dynamics of strings requires

a higher-dimensional Minkowski space and typically turns the previous relation into

$$m^2 \sim (N-1)\omega^2 \tag{7.34}$$

so that string spectra actually include massless modes, as needed to describe *long-range forces*. A more detailed analysis would reveal that open strings include massless vectors, while closed strings include massless spin-2 fields. Therefore, not only is one softening the gravitational interactions by spreading mass or energy but one is also recovering, without further ado, gauge bosons and gravitons from the string modes.

A closer look would reveal that strings can also describe spacetime fermions, with the chiral interactions needed in the standard model. Their consistency, however, rests on a new mechanism, discovered by M B Green and J H Schwarz, that supplements the ordinary anomaly cancellations at work in the standard model with the contributions of new types of particles. In their simplest manifestation, these have to do with a *two-form* field, a peculiar generalization of the electrodynamic potential A_μ bearing an antisymmetric pair of indices, so that $B_{\mu\nu} = -B_{\nu\mu}$. The corresponding field strength, obtained as in electrodynamics from its curl, is, in this case, the *three-form* field $H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu}$. Two-form fields have a very important property: their basic electric sources are *strings*, just like the basic electric sources in the Maxwell theory are *particles*. Thus, in retrospect, a $B_{\mu\nu}$ field is a clearcut signature of an underlying string extension. A field of this type is always present in the low-energy spectra of string models but is absent in the CJS supergravity that, for this reason, as stressed in the Introduction, bears no direct relation to strings.

We have already mentioned that there are apparently several types of string models, all defined in spacetimes with a number of extra dimensions. At present, the only direct way to describe their interactions is via a perturbative expansion. Truly enough, this is essentially the case for the standard model as well but for strings we still somehow lack a way to go systematically beyond perturbation theory. There is a framework, known as string field theory, vigorously pursued over the years by a small fraction of the community and, most notably, by A Sen and B Zwiebach, that is starting to produce interesting information on the string vacuum state but it is still a bit too early to give a fair assessment of its real potential in this respect. Indeed, even the very concept of a string could well turn out to be provisional, a convenient artifice to describe in one shot an infinity of higher-spin fields, much in the spirit of how a generating function in mathematics allows one to describe conveniently in one shot an infinity of functions and, in fact, string theory appears, in some respect, as a BEH-like phase of a theory with higher spins [2]. This is another fascinating, difficult and deeply related subject, pursued over the years mostly in Russia and mainly by E Fradkin and M Vasiliev.

[2] The standard model contains particles of spin-1 (the gauge bosons), $\frac{1}{2}$ (the quarks and leptons) and 0 (the BEH particle) and possibly of spin-2 (the graviton), while the massive string excitations have arbitrarily high spins.

String theory allows two types of perturbative expansions. The first is regulated by a dimensionless parameter, g_s, that takes the place in this context of the fine-structure constants present in the standard model, while the second is a low-energy expansion, regulated by the ratio between typical energies and a string scale $M_s \sim \hbar/c\ell_s$ related to the 'string size' ℓ_s. In the following, we shall use the two symbols ℓ_s and $\alpha' = \ell_s^2$ interchangeably to characterize the string size. A key result of the 1970s, due mainly to J Scherk, J H Schwarz and T Yoneya, is that in the low-energy limit the string interactions embody both the usual gauge interactions of the standard model and the gravitational interactions of general relativity. Thus, to reiterate, string theory embodies, by necessity, long-range electrodynamic and gravitational quanta, with low-energy interactions consistent with the Maxwell (or Yang–Mills) and Einstein equations.

The extra dimensions require that a spacetime version of symmetry breaking be at work to recover our four-dimensional world. The resulting framework draws from the original work of Kaluza and Klein, and has developed into the elegant and rich framework of Calabi–Yau compactifications but some of its key features can be illustrated by a simple example. To this end, let us consider a massless scalar field ϕ that satisfies, in five dimensions, the wave equation

$$\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} - \nabla^2\phi - \frac{\partial^2\phi}{\partial y^2} = 0 \tag{7.35}$$

where the fifth coordinate has been denoted by y. Now suppose that y lies on a circle of radius R, so that $y \sim y + 2\pi R$ or, equivalently, impose periodic boundary conditions in the y-direction. One can then expand ϕ in terms of a complete set of eigenfunctions of the circle Laplace operator, plane waves with quantized momenta, writing

$$\phi(x, y) = \sum_{n\in Z} \phi_n(x) \mathrm{e}^{\mathrm{i}ny/R}. \tag{7.36}$$

Plugging this expansion into the Klein–Gordon equation shows that, from the four-dimensional viewpoint, the mode coefficients $\phi_n(x)$ describe independent fields with masses n/R, satisfying

$$\frac{1}{c^2}\frac{\partial^2\phi_n}{\partial t^2} - \nabla^2\phi_n + \frac{n^2}{R^2}\phi = 0. \tag{7.37}$$

At low energies, where the massive modes are frozen, the extra dimension is thus effectively screened and inaccessible, since only quanta of the zero-mode field can be created. Simple as it is, this example suffices to show that the spectrum of massive modes reflects the features of the internal space, in that it depends on the radius R. By a slight complication, for instance playing with anti-periodic modes, one could easily see how even the numbers and types of low-lying modes present generally reflect the features of the internal space. This is perhaps the greatest flaw in our current understanding: the four-dimensional manifestations of a given

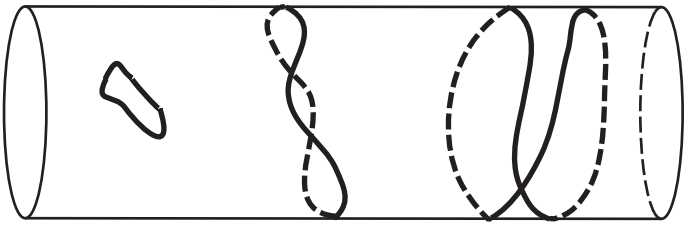

Figure 7.6. Contrary to point particles, closed strings can wind in a non-trivial way around a compact direction. Here we show, from left to right, three examples of closed strings with winding numbers 0, 1 and 2, respectively.

string and, in particular, the properties of its light particles are manifold, since they depend on the size and shape of the extra dimensions. Let us stress that, while in the electroweak breaking, we dispose of a clear minimum principle that drives the choice of a vacuum, no general principle of this type is available in the presence of gravity. Therefore, despite many efforts over the years, we have, at present, no clearcut way to make a dynamical choice between the available possibilities and, as a result, we are still not in a position to give clearcut string predictions for low-energy parameters. Nonetheless, these possibilities include, rather surprisingly, four-dimensional worlds with gauge and matter configurations along the lines of the standard model, although inheriting chiral interactions from higher dimensions would naïvely appear quite difficult. We may thus be driven to keep an eye on a different and less attractive possibility, as with the old, ill-posed problem, of deriving, from first principles, the sizes of the Keplerian orbits. As we now understand, these result from accidental initial conditions and a similar situation for the four-dimensional string vacuum, while clearly rather disturbing, cannot be fairly dismissed at the present time.

Still, in moving to string theory as the proper framework to extend the standard model, it would be reassuring to foresee some sort of uniqueness in the resulting picture, at least in higher dimensions. Remarkably this was achieved, to a large extent, by the mid 1990s and we have now good reasons to believe that all ten-dimensional superstring models are somehow equivalent to one another. The basic equivalences between the four superstring models of oriented closed strings, IIA, IIB, heterotic SO(32) and heterotic $E_8 \times E_8$, and the type I model of unoriented closed and open strings, usually called *string dualities*, are summarized in figure 7.1. The bold links, labelled by T and Ω, can be explicitly established in string perturbation theory, while the additional broken links rely on non-perturbative arguments that rest on the unique features of the low-energy ten-dimensional supergravity. We can now comment a bit on the labels, beginning with the T duality.

When a particle lives in a circle, the de Broglie wave $e^{ipx/\hbar}$ can be properly periodic only if the momentum is quantized in units of the inverse radius R, i.e. if

$p = n\hbar/R$. We have already met the field counterpart of this property, when we described how a massless five-dimensional field would manifest itself to a four-dimensional observer as an infinite tower of massive fields. A closed string can also be endowed with a centre-of-mass momentum, quantized for the same reason in units of $1/R$ and, thus, a single string spectrum would appear, by necessity, to a lower-dimensional observer as a tower of string spectra. However, a closed string can also wrap around the circle an arbitrary number of times, so that, in fact, a closed string coordinate admits expansions of the type

$$X(\sigma,\tau) = x + (2\alpha')\frac{m}{R}\tau + 2nR\sigma + \frac{\mathrm{i}}{2}\sqrt{2\alpha'}\sum_{k\neq 0}\left(\frac{\alpha_k}{k}\mathrm{e}^{-\mathrm{i}2k(\tau-\sigma)} + \frac{\tilde{\alpha}_k}{k}\mathrm{e}^{-\mathrm{i}2k(\tau+\sigma)}\right) \tag{7.38}$$

where τ in this context, replaces, the 'proper time' of particle dynamics while σ labels the points of the string. Note that the third term implies that $X(\pi,\tau) = X(0,\tau) + 2\pi nR$, as pertains to a closed string winding n times around a circle. The spectrum of the string as seen from the uncompactified dimensions will have the form

$$M^2 = \frac{\hbar^2}{c^2}\left(\frac{m^2}{R^2} + \frac{n^2R^2}{\alpha'^2}\right) + \cdots \tag{7.39}$$

where the dots stand for contributions due to the higher frequencies of the string (see, e.g. equation (7.33)). While the first term in (7.39) is familiar from ordinary quantum mechanics, the second, which, as we have seen, reflects the possibility of non-trivial windings, is new and intrinsically 'stringy'. Notice that equation (7.39) displays a remarkable symmetry: one cannot distinguish somehow between a string propagating on a circle of radius R and another propagating on a circle with the 'dual' radius α'/R ! We have actually simplified matters to some extent, since, in general, T-duality affects the fermion spectra of closed strings. Upon circle compactification, it thus maps the two heterotic models and the two type II models into one another, providing two of the bold duality links in figure 7.1.

The other bold link, labelled by Ω, reflects an additional peculiarity, the simultaneous presence of two sets of modes in a closed string (the 'right-moving' α and 'left-moving' $\tilde{\alpha}$ modes in equation (7.38)). If a symmetry is present between them, as is the case only for the type IIB model, one can use it to combine states but string consistency conditions require, in general, that new sectors emerge. As a result, combining, in this fashion, states of closed strings, one is generally led to introduce open strings as well. This construction, now commonly called an *orientifold*, was introduced long ago by one of the present authors and was then widely pursued over the years at the University of Rome 'Tor Vergata'. It links the type IIB and type I models in the diagram also, offering new perspectives on the issue of string compactification.

The additional broken links in figure 7.1 are harder to describe in simple terms but can be characterized as analogues, in this context, of the electric–

magnetic duality of Maxwell's electrodynamics. It is indeed well known that, in the absence of sources, the electric–magnetic duality transformations $\boldsymbol{E} \to \boldsymbol{B}$ and $\boldsymbol{B} \to -\boldsymbol{E}$ are a symmetry of the Maxwell equations:

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= 0 \qquad \nabla \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} \\ \nabla \cdot \boldsymbol{B} &= 0 \qquad \nabla \times \boldsymbol{B} = \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} \end{aligned} \tag{7.40}$$

but it is perhaps less appreciated that the symmetry can be extended to the general case, at the expense of turning electric charges and currents into their magnetic counterparts. Whereas these are apparently not present in nature, Yang–Mills theories generically, but not necessarily, predict the existence of heavy magnetic poles, of masses $M \sim M_{\mathrm{W}}/\alpha_{\mathrm{e}}$, with α_{e} a typical (electric) fine-structure constant. Thus, we might well have failed to see existing magnetic poles, due to their high masses, about 100 times larger than those of the W and Z bosons! In fact, QED could also be formulated in terms of magnetic carriers but, for Quantum Mechanics, that adds an important datum: the resulting magnetic fine-structure constant α_{m} would be enormous, essentially the inverse of the usual electric one. More precisely, magnetic and electric couplings are not independent but are related by Dirac quantization conditions, so that

$$\alpha_{\mathrm{e}} \sim \frac{1}{\alpha_{\mathrm{m}}} \tag{7.41}$$

and, therefore, it is the smallness of α_{e} that favours the usual electric description, where the actual interacting electrons and photons are only mildly different from the corresponding free quanta, on which our intuition about elementary particles rests.

In string theory, the 'electric' coupling is actually determined by the vacuum expectation value (vev) of a ubiquitous massless scalar field, the *dilaton* (closely related to the Brans–Dicke scalar, a natural extension of general relativity), according to

$$g_{\mathrm{s}} = \mathrm{e}^{\langle\varphi\rangle}. \tag{7.42}$$

At this time, we have no direct insight into g_{s} that, in general, could be spacetime dependent. It is thus interesting to play with these S dualities, that indeed fill the missing gaps in figure 7.1. A surprise is that both the type IIA and the $\mathrm{E}_8 \times \mathrm{E}_8$ heterotic models develop, at strong coupling, an additional dimension, invisible in perturbation theory but macroscopic if g_{s} is large enough. The emergence of the additional dimension brings into the game the CJS supergravity, the unique supergravity model in 11 dimensions that, however, cannot be directly related to strings: as we have stressed, it does not contain a $B_{\mu\nu}$ field, although it does contain a three-index field, $A_{\mu\nu\rho}$, related to corresponding higher-dimensional solitonic objects, that we shall briefly return to in the next section, the M2- and M5-branes. This is the puzzling end of the story alluded to in the

introduction: duality transformations of string models, that supposedly describe the microscopic degrees of freedom of our world, link them to a supergravity model with no underlying string. This is indeed, in some respect, like ending up with pions with no clue on the underlying 'quarks'! This beautiful picture was contributed in the last decade by many authors, including M Duff, A Font, P Horava, C M Hull, L Ibanez, D Lust, F Quevedo, A Sen, P K Townsend and, most notably, by E Witten.

Let us conclude this section by stressing that a duality is a complete equivalence between the spectra of two apparently distinct theories. We have met one example of this phenomenon earlier, when we discussed the case of T duality: winding modes find a proper counterpart in momentum modes and *vice versa*. Now, in relating the heterotic SO(32) model, say, to the type I string, their two sets of modes have no way to match directly. For instance, a typical open-string coordinate

$$X = x + (2\alpha')\frac{m}{R}\tau + 2nR\sigma + \mathrm{i}\sqrt{2\alpha'}\sum_{k\neq 0}\frac{\alpha_k}{k}\mathrm{e}^{-\mathrm{i}k\tau}\cos(k\sigma) \tag{7.43}$$

is vastly different from the closed-string expansion met previously since, for one matter, it involves a single set of modes. How can a correspondence of this type hold? We have already stumbled on the basic principle, when we said that typically Yang–Mills theories also describe magnetic poles. These magnetic poles are examples of *solitons*, stable localized blobs of energy that provide apparently inequivalent descriptions of wave quanta, to which we now turn.

7.4 From strings to branes

A number of field theories admit solitonic solutions, blobs of energy whose shape is stabilized by nonlinear couplings. A simple example is provided by the 'kink' that interpolates between the two minima of the potential shown in figure 7.7. It can be regarded as a model for a wall separating a pair of Curie–Weiss domains in a ferromagnet. Its stability can be argued by noting that any attempt to deform it, say, to the constant vacuum $\phi = a$ would cost, in one dimension, an energy of the order of $LV(0)$, where L is the size of the region where the field theory lives and $V(0)$ is the height of the potential barrier. For a macroscopic size L, this becomes an infinite separation and the solution is thus stable. Moreover, its energy density, essentially concentrated in the transition region, results in a *finite* total energy, $E = m^3/12\lambda$, where m denotes the mass of the elementary scalar field, defined as expanding around one of the minima of the potential, $\phi = \pm a$. This energy defines the mass of the soliton and, as anticipated, blows up in the limit of small coupling λ. The 't Hooft–Polyakov monopole works, in three dimensions, along similar lines: any attempt to destroy it would cost an infinite energy. As is usually said, these objects are *topologically stable* and, in fact, their stability can be ascribed to the conservation of a suitable (topological) charge that, for the

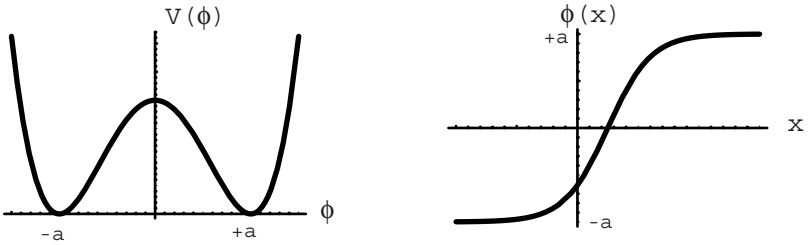

Figure 7.7. A scalar field theory in $1+1$ dimensions (denoting the scalar field as $\phi(t, x)$), with a potential $V(\phi) = \lambda(\phi^2 - a^2)^2$ (shown on the left) admits a static, finite-energy solution that interpolates between the two vacua $\phi = \pm a$ (shown on the right), known as a 'kink'. Its energy or mass is $m^3/12\lambda$, where m is the mass of the elementary scalar field. In the perturbative regime, i.e. for small coupling λ, the kink is, therefore, very heavy.

monopole, is simply its magnetic charge. A further feature of solitons is that their energy is proportional to an inverse power of a coupling constant, as we have seen for the kink. This is simple to understand in general terms: the nonlinear nature of the field equations is essential for the stability of solitons that, therefore, should disappear in the limit of small coupling!

A localized distribution of energy and/or charge is indeed a modern counterpart of our classical idea of a particle. It is probably familiar that an electron has long been modelled in classical electrodynamics, in an admittedly *ad hoc* fashion, as a spherical shell with a total charge e and a finite radius a, associating the resulting electrostatic energy

$$E \sim \frac{e^2}{a} \tag{7.44}$$

with the electron mass. In a similar fashion, the localized energy distribution of a soliton is naturally identified with a particle, just like an energy distribution localized along a line is naturally identified with an infinite string, while its higher dimensional analogues define generalized *branes*. Thus, for instance, the 'kink' describes a particle in $1+1$ dimensions, a string in $1+2$ dimensions, where the energy distribution is independent of a spatial coordinate, and a domain wall or *two-brane* in $1+3$ dimensions, where the energy distribution is independent of two spatial coordinates. These are, therefore, new types of 'quanta', somehow missed by our prescription of reading particle spectra from free wave equations. Amusingly enough, one can argue that the two descriptions of particles are only superficially different, while the whole picture is well fitted with quantum mechanics. The basic observation is that these energy blobs have typically a spatial extension

$$\Delta \sim \frac{\hbar}{Mc} \tag{7.45}$$

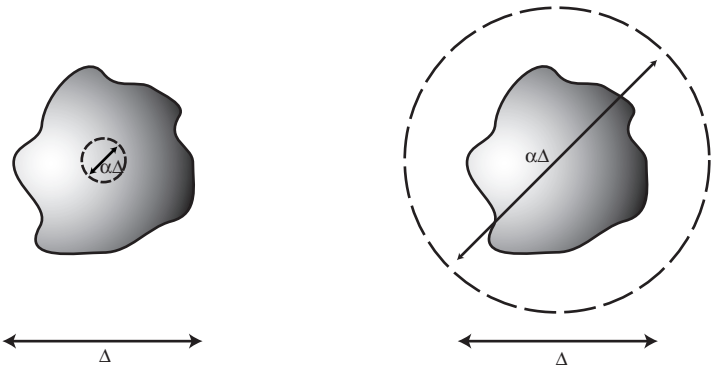

Figure 7.8. An 'artistic' impression of a soliton of size Δ. The broken circle depicts its Compton wavelength $\lambda_C \sim \alpha\Delta$, with α a dimensionless fine-structure constant. The first figure refers to the weak-coupling regime, where α is very small and, as a result, the Compton wavelength λ_C is much smaller than the soliton size Δ. In this regime, the soliton can be regarded as a classical object. The second figure refers to the strong-coupling regime, where α is very large and, as a result, the Compton wavelength λ_C is much larger than the soliton size Δ. In this regime, the soliton can be regarded as an ordinary light quantum without any inner substructure.

where M denotes a typical mass scale associated with a BEH-like phenomenon, since they basically arise from regions where a transition between vacua takes place, typically of the order of the Compton wavelength (7.45). In addition, the energy stored in these regions, that determines the mass of the soliton, is

$$M_{\rm sol} \sim \frac{M}{\alpha} \tag{7.46}$$

with α a typical fine-structure constant. At weak coupling (small α) we have quantitative means to explore the phenomenon further, but $M_{\rm sol} \gg M$, so that the Compton wavelength of the soliton is well within its size. In other words, in the perturbative region the soliton is a *classical* object. However, in the strong-coupling limit (large α), the soliton becomes light while its Compton wavelength spreads well beyond its size, so that its inner structure becomes immaterial: we are then back to something very similar in all respects to an ordinary quantum. Solitons are generally interacting objects. For instance, magnetic poles typically experience the magnetic dual of the usual Coulomb force. This is reflected in the fact that, being solutions of nonlinear equations, they cannot be superposed. In special cases, their mutual forces might cancel and then, quite surprisingly, the corresponding nonlinear field equations allow a superposition of different solutions. This is a typical state of affairs in supersymmetric theories, realized when special inequalities, called 'BPS bounds', are saturated.

The T and S dualities discussed in the previous section can also be seen as maps between ordinary quanta and solitons. The former simpler case involves the interchange of momentum excitations with winding modes that, as we have stressed, describe topologically inequivalent closed-string configurations on a circle, while the latter rests on similar operations involving solitons in spacetime. These, as in the two examples we have sketched in this section, can be spotted from the field equations for the low-energy string modes but some of their features can be discussed in simple, general terms. To this end, let us begin by rewriting the $(1 + 3)$-dimensional Maxwell equations (7.40) in covariant notation, while extending them to the form

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} J_{\rm e}^\nu \tag{7.47}$$

$$\varepsilon^{\mu\nu\rho\sigma}\partial_\nu F_{\rho\sigma} = \frac{4\pi}{c} J_m^\nu \tag{7.48}$$

where, in addition to the more familiar electric sources $J_{\rm e}$, we have also introduced magnetic sources J_m, that affect the Faraday–Neumann–Lenz induction law and the magnetic Gauss law. Note how a current $J_{\rm e}^\mu$ is naturally borne by particles, with $J_{\rm e}^\mu \sim q u^\mu$ in terms of their charge and four-velocity. In $D > 4$, however, the ε tensor carries D indices and, consequently, J_m carries, in general, $D - 3$ indices, while its sources are extended objects defined via $D - 4$ Lorentz indices. Thus, a magnetic pole is a particle in four dimensions as a result of a mere accident. In six dimensions, for instance, the magnetic equations would become

$$\varepsilon^{\mu\nu\rho\sigma\tau\lambda}\partial_\sigma F_{\tau\lambda} = \frac{4\pi}{c} \tilde{J}_m^{\mu\nu\rho} \tag{7.49}$$

so that, by the previous reasoning, a magnetic pole would bear a pair of indices, as pertains to a surface. In other words, it would be a two-brane. The argument can be repeated for a general class of tensor gauge fields, $B_{\mu_1\ldots\mu_{p+1}}$, in D dimensions: their electric sources are p-branes, while the corresponding magnetic sources are $(D - 4 - p)$-branes. These tensor gauge fields are typically part of low-energy string spectra, while the corresponding 'electric' and 'magnetic' poles show up as solutions of the complete low-energy equations for the string modes. As we have seen, they define new types of 'quanta' that are to be taken into account: in fact, 'branes' of this type are the missing states alluded to at the end of the previous section!

As stressed by J Polchinski, a peculiarity of string theory makes some of the 'branes' lighter than others in the small-coupling limit and, at the same time, simpler to study. The first feature is due to a string modification of equation (7.46), that, for these 'D-branes', happens to depend on $\sqrt{\alpha} \sim g_{\rm s}$, rather than on α, as is usually the case for ordinary solitons. The second feature is related to the possibility of *defining* string theory in the presence of D-branes via a simple change of boundary conditions at the string ends. In other words, D-branes absorb and radiate strings. In analogy with ordinary particles, D-branes can be

characterized by a *tension* (mass per unit volume) and a *charge* that defines their coupling to suitable tensor gauge fields. While their dynamics is prohibitively complicated, in the small coupling limit, they are just rigid walls and so one is effectively studying some sort of Casimir effect induced by their presence. The idea is hardly new: for instance, the familiar Lamb shift of QED is essentially a Casimir effect induced by the atom. What is new and surprising in this case, however, is that the perturbation theory around D-branes can be studied in one shot for the whole string spectrum. In other words, for an important class of phenomena that can be associated to D-branes, a *macroscopic* analysis of the corresponding field configurations can be surprisingly accompanied by a *microscopic* analysis of their string fluctuations. This is what makes D-branes far simpler than other string solitons, for instance the M5-brane, on which we have very little control at this time. The mixing of left and right closed-string modes met in the discussion of orientifolds in the context of string dualities can also be given a spacetime interpretation along these lines: it is effected by apparently non-dynamical 'ends of the world', usually called O-planes. There is also an interesting possibility, well realized in perturbative open-string constructions: while branes, being physical objects, are bound to have a positive tension, one can allow different types of O-planes, with both negative and positive tension. While the former are typical ingredients of supersymmetric vacua, the latter can induce interesting mechanisms of supersymmetry breaking that we shall mention briefly in the next section.

7.5 Some applications

The presence of branes in string theory provides new perspectives on a number of issues of crucial conceptual and practical import. In this section, we comment briefly on some of them, beginning with the amusing possibility that our universe is associated with a collection of branes and then moving on to brief discussions of black hole entropy and color-flux strings.

7.5.1 Particle physics on branes?

One is now confronted with a fully novel situation: as these 'branes' are extended objects, one is naturally led to investigate the physics of their interior or, in more pictorial terms, the physics as seen by an observer living on them. To this end, it is necessary to study their small oscillations that define the light fields or, from what we have said in the previous sections, the light species of particles seen by the observer. These will definitely include the scalars that describe small displacements of the 'branes' from their equilibrium positions and possibly additional light fermionic modes. A surprising feature of D-branes is that their low-energy spectra also include *gauge fields*. Both scalars and gauge fields arise from the fact that open strings end on D-branes (seen from the brane, their intersections are point-like) and are, in fact, associated with

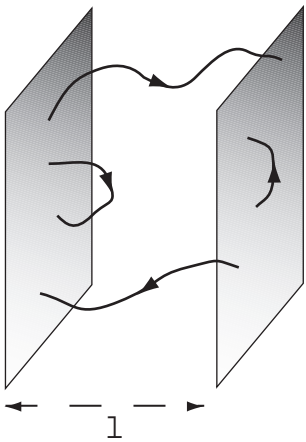

Figure 7.9. A system consisting of two parallel D-branes, to which oriented open strings can attach in four different ways. The masses of the gauge fields associated to the four types of open strings are proportional to the shortest distances between the branes they connect. When the mutual distance l between the branes is not zero, one is thus describing two massless gauge fields, with a corresponding unbroken U(1) $\times$ U(1) symmetry, and two additional massive W-like fields. However, when $l \to 0$ the two W-like fields become massless as well, while the gauge symmetry enhances to U(2). In this geometric setting for the BEH mechanism, the Higgs scalar describes the fluctuations of the branes relative to one another, while its vacuum value defines their relative distance l.

string fluctuations transversal or longitudinal to the branes, respectively. In addition, when several branes coincide non-Abelian gauge symmetries arise, as summarized in figure 7.9. In equivalent terms, the mutual displacement of branes provides a geometric perspective on the BEH mechanism. Moreover, the low-energy dynamics of gauge fields on a Dp-brane is precisely of the Yang–Mills type but, at higher energies, interesting stringy corrections come into play. While a proper characterization of the general case is still an open problem, in the Abelian case of a single D-brane and in the limit of slowly varying electric and magnetic fields, string theory recovers a beautiful action proposed in the 1930s by Born and Infeld to solve the singularity problem of a classical point-like electric charge, as originally shown by E Fradkin and A Tseytlin. Let us explain this point briefly. Whereas in the usual Maxwell formulation the resulting Coulomb field

$$\boldsymbol{E} = \frac{q}{r^2}\hat{\boldsymbol{r}} \tag{7.50}$$

where $\hat{\boldsymbol{r}}$ denotes the unit radial vector, leads to an infinite energy, in string theory the Maxwell action for the static case is modified and takes the form

$$-\frac{1}{2}\boldsymbol{E}^2 \to \frac{1}{2\pi\alpha'}\sqrt{1-2\pi\alpha'\boldsymbol{E}^2} \tag{7.51}$$

so that equation (7.50) is turned into

$$\boldsymbol{E} = \frac{q}{\sqrt{r^4 + (2\pi\alpha')^2}}\hat{\boldsymbol{r}}. \tag{7.52}$$

As a result, the electric field strength saturates to $q/2\pi\alpha'$, much in the same way as the speed of a relativistic particle in a uniform field saturates to the speed of light c, an analogy first stressed in this context by C Bachas. Thus, once more string theory appears to regulate divergences, as we have already seen in connection with the ultraviolet problem of gravity.

Summarizing, the world volume of a collection of Dp-branes is, by construction, a $(p+1)$-dimensional space that contains, in principle, the correct types of light fields to describe the particles of the standard model. This observation has changed our whole perspective on the Kaluza–Klein scenario, is at the heart of current attempts to model our universe as a collection of intersecting D-branes and brings about a novelty that we would like to comment upon briefly. The issue at stake is, again, the apparently unnatural hierarchy between the electroweak and Planck scales, on which this scenario offers a new geometric perspective, since in a 'brane world' gauge and matter interactions are confined to the branes, while gravity spreads in the whole ambient space. One can thus provide a different explanation for the weakness of gravity: most of its Faraday lines are spread throughout the internal space and are, thus, simply 'lost' for a brane observer. This is the essence of a proposal made by I Antoniadis, N Arkani-Ahmed, S Dimopoulos and G Dvali, that has stimulated a lot of activity in the community in recent years. For instance, with n extra circles of radius R one would find that a $(4+n)$-dimensional Newton constant G_{4+n} for bulk gravity induces, for two point-like masses on the brane, an effective Newton constant $1/G_4 \sim R^n/G_{4+n}$. This result can be obtained by adding the contributions of the extra circles or, more simply, purely on dimensional grounds. Playing with the size R, one can start with $G_{4+n} \sim (1/\mathrm{TeV})^{2+n}$ and end up with the conventional $G_4 \sim 1/(10^{19}\ \mathrm{GeV})^2$, if $R \sim 10^{32/n} \times 10^{-4}$ fm, so that if $n \geq 2$, the resulting scenario is not obviously excluded. The phenomenon would manifest itself as a striking change in the power law for the Newton force (7.3) which, for $r < R$, would behave like $1/r^{2+n}$, a dramatic effect indeed, currently investigated by a number of experimental groups at scales somewhat below the millimetre. In a similar fashion, one can also conceive scenarios where the string size ℓ_s is also far beyond the Planck length but a closer inspection shows that, in all cases, the original hierarchy problem has been somehow rephrased in geometrical, although possibly milder, terms: all directions parallel to the world brane should be far below the millimetre, at least $\mathcal{O}(10^{-16}\ \mathrm{cm})$, if no new phenomena are to be present in the well-explored gauge interactions of the standard model at accessible energies, so that a new hierarchy emerges between longitudinal and transverse directions. The literature also contains interesting extensions of this scenario with infinitely extended curved internal dimensions, where gravity can, nonetheless,

be localized on branes but this simpler case should suffice to give a flavour of the potential role of branes in this context.

It is also possible to complicate this picture slightly to allow for the breaking of supersymmetry. To date, we have only one way to introduce supersymmetry breaking into closed strings working at the level of the full string theory, as opposed to its low-energy modes: Bose and Fermi fields can be given different harmonic expansions in extra dimensions. For instance, referring to the case in section 7.3, if along an additional circle Bose fields are periodic while Fermi fields are antiperiodic, the former inherit the masses k/R, while the latter are lifted to $(k + 1/2)/R$, with supersymmetry broken at a scale $\Delta M \sim 1/R$. This is the Scherk–Schwarz mechanism, first fully realized in models of oriented closed strings by S Ferrara, K Kounnas, M Porrati and F Zwirner, following a previous analysis of R Rohm. Branes and their open strings, however, allow new possibilities, known in the literature, respectively, as 'brane supersymmetry' and 'brane supersymmetry breaking', that we would like to briefly comment upon. Of course, the mere presence of branes, extended objects of various dimensions, breaks some spacetime symmetries and, in fact, one can show that a single brane breaks at least half of the supersymmetries of the vacuum but more can be done by suitable combinations of them. Thus, the first mechanism follows from the freedom to use, in the previous construction, directions parallel or transverse to the 'brane world' to separate Fermi and Bose momenta. While momenta along parallel directions reproduce the previous setting, orthogonal ones, in principle, *cannot* separate brane modes. However, a closer inspection reveals that this is only true for the low-lying excitations, while the massive ones, affected by the breaking, feed it via radiative corrections to the low-lying modes, giving rise to a gravitational analogue of the 'seesaw' mechanism, with $\Delta M \sim \sqrt{G_{\rm N}}/R^2$. Finally, the second mechanism can induce supersymmetry breaking in our world radiatively from other non-supersymmetric branes, with the interesting possibility of attaining a low vacuum energy in the observable world.

By and large, however, one is again led to a puzzling end: a sort of 'brane chemistry' allows one to concoct an observable world out of these ingredients, much in the spirit that associates chemical compounds with the basic elements of the Periodic Table and, eventually, to electrons and nuclei. However, the problem alluded to in the previous sections is still with us: we presently have no plausible way of selecting a preferred configuration to connect string theory to our low-energy world, although one can well construct striking realizations of the standard model on intersecting branes, as first shown by the string groups at the Universidad Autonoma de Madrid and at the Humboldt University in Berlin.

7.5.2 Can strings explain black hole thermodynamics?

As we have stressed, D-branes can be given a *macroscopic* description as solutions of the nonlinear field equations for the light string modes and, at the same time, a *microscopic* description as emitters and absorbers of open strings.

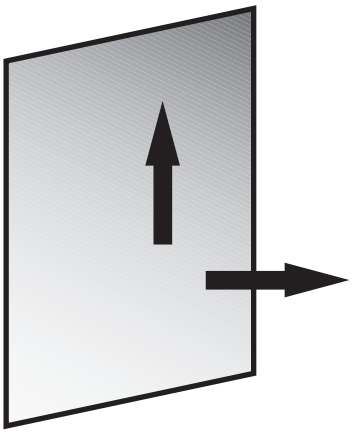

Figure 7.10. Brane supersymmetry and the gravitational seesaw. If the momenta of Bose and Fermi brane fields are separated along a direction orthogonal to the D-brane, only the massive brane excitations feel the effect, that is then transmitted via radiative corrections to the low-lying excitations, relevant for standard model physics.

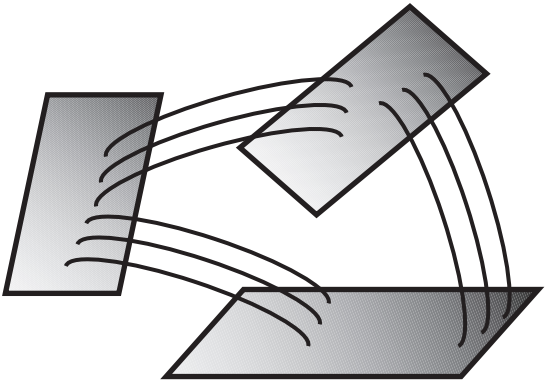

Figure 7.11. The world as a collection of interacting branes. Amusingly, we have come all this way from our matter seen as a collection of point particles, recovering at large scales something quite reminiscent of our starting point.

If for a black hole, both descriptions were available, one would be naturally led to regard the open string degrees of freedom as its own excitations. This appears to provide a new perspective on a well-known result of S Hawking, that associates a blackbody spectrum of radiation at a characteristic temperature T_{H} with the formation of a black hole. Since, as originally stressed by J Bekenstein, the resulting conditions for the mass variation of the hole have the flavour of thermodynamics, D-branes offer the possibility of associating with this

thermodynamics a corresponding statistical mechanics, the relevant microstates being their own excitations.

In the presence of a static isotropic source of mass M at the origin, the Minkowski line element is deformed to

$$\mathrm{d}s^2 = \left(1 - \frac{2G_\mathrm{N}M}{rc^2}\right) c^2\,\mathrm{d}t^2 - \left(1 - \frac{2G_\mathrm{N}M}{rc^2}\right)^{-1} \mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2. \quad (7.53)$$

This expression holds *outside* the source, while the special value of the radial coordinate $r_\mathrm{h} = 2G_\mathrm{N}Mc^{-2}$ corresponds to the *event horizon*, that can be characterized as the minimum sphere centred at the origin that is accessible to a far-away observer. For most objects, r_h lies deep inside the source itself (e.g. for the sun $r_\mathrm{h} \approx 3$ km, to be compared with the solar radius $R \approx 10^6$ km), where equation (7.53) is no longer valid, but one can conceive a source whose radius is inferior to r_h and this is called a *black hole*: according to classical general relativity, any object coming from outside and crossing the horizon is trapped inside it forever. Over the past decade, astrophysical observations have given strong, if indirect, clues that black holes are ubiquitous in our universe.

As anticipated, however, Hawking found that black holes are not really black if quantum mechanics is properly taken into account. Rather, quantizing a field theory in a background containing a black hole, he showed that, to an external observer, the hole appears to radiate as a black body with temperature

$$T_\mathrm{H} = \frac{c^3\hbar}{8\pi k_\mathrm{B} G_\mathrm{N} M} \quad (7.54)$$

where k_B denotes Boltzmann's constant. This amazing phenomenon that can be made plausible by noting that a virtual particle–antiparticle pair popping up in the neighbourhood of the horizon can have such a dynamics that one of the two crosses the horizon, while the other, forced by energy conservation to materialize as a real particle, will do so by absorbing and carrying away part of the gravitational energy of the black hole. In analogy with the second law of thermodynamics, given the temperature T_H one can associate to a black hole an *entropy*

$$\frac{1}{k_\mathrm{B}} S_H = \frac{4\pi G_\mathrm{N} M^2}{c\hbar} = \frac{1}{4} A_\mathrm{H} \ell_\mathrm{Pl}^{-2} \quad (7.55)$$

where A_H is the area of the horizon and ℓ_Pl is the Planck length $\ell_\mathrm{Pl} = \sqrt{G_\mathrm{N}\hbar/c^3} \approx 10^{-33}$ cm, that we have repeatedly met in the previous sections. This expression, known as the Bekenstein–Hawking formula, reflects a universal behaviour: the entropy of any black hole is one-quarter of the area of its horizon in Planck units.

Several questions, that have long puzzled many experts arise:

- As anything crossing the horizon disappears leaving only thermal radiation behind, the S-matrix of a system containing a black hole no longer seems to

be unitary, thus violating a basic tenet of quantum mechanics. This is known as the *information paradox*.

- Entropy is normally a measure of the degeneracy of microstates Σ in some underlying microscopic description of a physical system, determined by Boltzmann's formula,
$$S = k_{\mathrm{B}} \log \Sigma. \tag{7.56}$$
Since the entropy (7.55) of a black hole is naturally a huge number, how can one exhibit such a wealth of microstates?
- Equation (7.54) clearly shows that the more mass is radiated away from the black hole, the hotter this becomes. What then is the endpoint of black hole evaporation?

Within string theory, there is a class of black holes where these problems can be conveniently addressed, the so-called extremal black holes, that correspond to BPS objects in this context. The simplest available example is provided by a source that also carries an electric charge Q. The coupled Maxwell–Einstein equations would give, in this case, the standard Coulomb potential for the electric field, together with the modified line element ($G_{\mathrm{N}}M^2 > Q^2$)

$$\begin{aligned} \mathrm{d}s^2 = &\left(1 - \frac{2G_{\mathrm{N}}M}{rc^2} + \frac{G_{\mathrm{N}}Q^2}{r^2c^4}\right)\mathrm{d}t^2 \\ &- \left(1 - \frac{2G_{\mathrm{N}}M}{r\,c^2} + \frac{G_{\mathrm{N}}Q^2}{r^2c^4}\right)^{-1}\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2 \end{aligned} \tag{7.57}$$

that generalizes equation (7.53). Note that the additional terms in (7.57) have a nice intuitive meaning: $Q^2/2r$ is the electrostatic energy introduced by the charge in the region beyond r and this contribution gives rise to a repulsive gravitational effect. The event horizon, defined again as the smallest sphere surrounding the hole that is accessible to a far-away observer, would now be

$$r_{\mathrm{H}} = c^{-2}\left(G_{\mathrm{N}}M + \sqrt{(G_{\mathrm{N}}M)^2 - G_{\mathrm{N}}Q^2}\right). \tag{7.58}$$

A source with a radius smaller than r_{H} would be a Reissner–Nordstrom black hole, with temperature and entropy given by

$$\begin{aligned} T_{\mathrm{H}} &= \frac{c^3\hbar\sqrt{(G_{\mathrm{N}}M)^2 - G_{\mathrm{N}}Q^2}}{2\pi k_{\mathrm{B}}\left(G_{\mathrm{N}}M + \sqrt{(G_{\mathrm{N}}M)^2 - G_{\mathrm{N}}Q^2}\right)^2}, \\ \frac{S_{\mathrm{H}}}{k_{\mathrm{B}}} &= \frac{\pi}{c\hbar G_{\mathrm{N}}}\left(G_{\mathrm{N}}M + \sqrt{(G_{\mathrm{N}}M)^2 - G_{\mathrm{N}}Q^2}\right)^2 = \frac{1}{4}A_{\mathrm{H}}l_p^{-2}. \end{aligned} \tag{7.59}$$

For a given value of Q, if $M \to Q/\sqrt{G_{\mathrm{N}}}$, the temperature vanishes, so that the black hole behaves somehow in this limiting (BPS) case as if it were an

elementary particle. Such a black hole is called extremal: its mass is tuned so that the tendency to gravitational collapse is precisely balanced by the electrostatic repulsion. This limiting case entails a manifestation of the phenomenon alluded to in section 7.4: although the Maxwell–Einstein equations are highly nonlinear, one can actually superpose these extremal solutions.

Extremal black holes of this type can be described in string theory in relatively simple terms. One of the simplest configurations involves the type IIB string theory compactified on a five-dimensional torus, together with a D5-brane and a D1-brane wrapped n_5 and n_1 times respectively around the torus. This BPS configuration is characterized by two topological numbers, n_1 and n_5, but one needs a slight complication of it since, being the only BPS state with these charges, it leads to a vanishing entropy, consistently with equation (7.59). However, suitable excitations, involving open strings ending on the D-branes and wrapping in various ways around the torus, are also BPS and can be characterized by a single additional quantum number, n_e. Many open string configurations now correspond to a given value of n_e and counting them one can obtain a *microscopic* estimate of the entropy. One can then turn to IIB supergravity on the 5-torus, constructing a BPS solution of its field equations that involves the *three* charges mentioned earlier, to calculate its event horizon, its temperature and, finally, to obtain the corresponding *macroscopic* estimate for the entropy. The exact agreement between the two estimates is then striking. Since this original example was discussed by A Strominger and C Vafa, many other black hole configurations have been studied, while the analysis has been successfully extended to nearly extremal ones. These results, however, rely heavily on supersymmetry and serious difficulties are met in attempts to extend them to non-supersymmetric black holes.

The analysis of nearly extremal black holes also appears to provide a clue to the information paradox. Studying a configuration slightly away from extremality, it was indeed found that Hawking radiation can be associated with the *annihilation* of pairs of open strings, each ending on a D-brane, that give rise to open strings remaining on the brane and to closed strings leaving it. The resulting radiation turns out to be exactly thermal, while temperature and radiation rate are in perfect agreement with a Hawking-like calculation. Almost by construction, this process is unitary and so the information that seemed lost appears to be left in the D-branes.

7.5.3 AdS/CFT: strings for QCD mesons or is the universe a hologram?

In the previous section, we saw that the entropy of a black hole is proportional to the area of its horizon. This is remarkable, since one can argue that black holes maximize the entropy. Indeed, assume for a moment that one had managed to construct a physical system in a given volume V with a mass $M - \delta M$ slightly inferior to that of a black hole whose horizon spans the surface surrounding V but with an entropy $S + \delta S$ slightly larger than that of the black hole. Throwing in a bit of matter would then create a black hole while simultaneously lowering

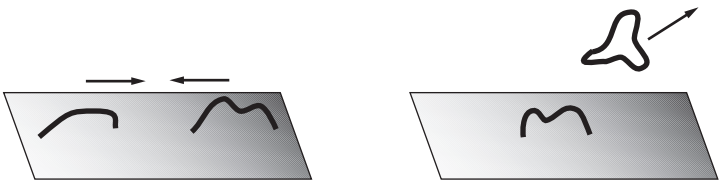

Figure 7.12. The D-brane picture of Hawking radiation. A pair of open strings collide, giving rise to a closed string that leaves the brane. As a result, Hawking radiation reaches the bulk via the emission of closed strings.

the entropy, thereby violating the fundamental law of thermodynamics. This observation led 't Hooft to propose the *Holographic Principle*: in a complete theory of quantum gravity, it should be possible to describe the physics of a certain region of spacetime in terms of degrees of freedom living on the surface surrounding it, while the information stored should be limited to roughly one bit per Planck area unit.

Over the past few years, concrete realizations of the Holographic Principle have been constructed, most dramatically in the context of the so-called AdS/CFT correspondence. In its simplest form, this arises if the type IIB string theory is defined in a ten-dimensional spacetime with the topology of a five-dimensional sphere (S^5) multiplied by a five-dimensional anti-de Sitter space (AdS_5), a non-compact manifold whose boundary can be identified with four-dimensional Minkowski space. This geometry describes the region around the horizon for a stack of n D3-branes, that in the large-n limit actually invades the whole of space time. On the one hand, there are, therefore, D-branes, that as we have seen host a Yang–Mills theory, while, on the other, there is a corresponding string background and J Maldacena conjectured that the resulting string theory (which includes gravity) in the bulk of AdS_5 is exactly *equivalent* (dual) to an $\mathcal{N} = 4$ U(n) super Yang–Mills theory in its border, the four-dimensional Minkowski space. This remarkable correspondence actually reflects a number of unusual equivalences between string amplitudes: for instance, as shown in figure 7.13, a one-loop diagram for open strings, obtained by widening an ordinary field theory loop into an annulus, can alternatively be regarded as a tree-level diagram for closed strings. In other words, the distinction between closed and open strings and, thus, between gravity and gauge fields is somewhat blurred in string theory. The conjecture was particularly well tested in the regime where the size of the strings is very small compared to the radii of AdS_5 and S^5 and where the string coupling constant is also small, so that the string theory is well described by classical supergravity. In the dual picture, this corresponds to the U(n) Yang–Mills theory in the limit where both n and the 't Hooft coupling $g^2_{\rm YM}n$ are large, i.e. in its deep quantum mechanical regime. Still, some quantities protected by supersymmetry match admirably in the two descriptions, confirming this surprising correspondence between theories defined in different spacetime dimensions. Tests at intermediate

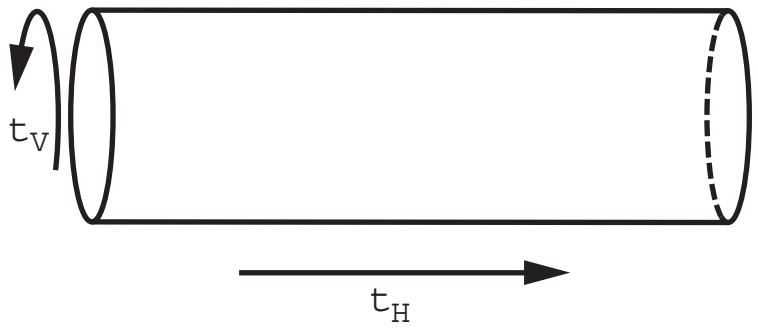

Figure 7.13. A surprising equivalence in string theory. An annulus diagram can be regarded as a loop diagram for open strings (with vertical time t_V) or, equivalently, as a closed string tree diagram (with horizontal time t_H).

regimes are much harder and are still largely lacking but no contradictions have emerged so far. In this way, gravity would ideally provide a tool for studying quark confinement but with a new ingredient: the colour flux tubes penetrate an additional dimension of spacetime.

We have thus come somehow full circle. The string idea originated from attempts made in the 1960s to model the strong interaction amongst mesons via narrow flux tubes, that culminated in the well known work of G Veneziano. With the advent of QCD, this picture was abandoned, since the flux tubes were regarded as a manifestation of QCD itself, while strings were proposed, as we have seen, as a tool to attain a finite quantum gravity. However, many people kept looking for a string-like description of the colour flux-tubes and, with the advent of the AdS/CFT correspondence, this was indeed realized to some extent, albeit once more in a supersymmetric setting that is free of many of the intricacies of QCD. Again, difficulties in the various types are met when one tries to proceed away from supersymmetry to come closer to our real confining low-energy world.

Acknowledgments

The work of the first author was supported in part by INFN, by the European Commission RTN programmes HPRN-CT-2000-00122 and HPRN-CT-2000-00148, by the INTAS contract 99-1-590, by the MURST-COFIN contract 2001-025492 and by the NATO contract PST.CLG.978785. The work of the second author was supported in part by 'FWO-Vlaanderen' through project G.0034.02, by the Federal Office for Scientific, Technical and Cultural Affairs through the Interuniversity Attraction Pole P5/27 and by the European Commission RTN programme HPRN-CT-2000-00131, in which he is associated with the University of Leuven.

We are grateful to C Angelantonj, J Lemonne, J Troost, G Stefanucci and F Zwirner for useful suggestions and comments on the manuscript.

The spirit of this review suggests that we refrain from giving detailed references to the original literature, contenting ourselves with a number of books and reviews that introduce the various topics addressed in this paper, and where the interested reader can find further details.

References

[1] Some recent books on the standard model:
Mandl F and Shaw G 1984 *Quantum Field Theory* (Chichester, UK: Wiley)
Peskin M E and Schroeder D V 1995 *An Introduction To Quantum Field Theory* (Reading, MA: Addison-Wesley)
Weinberg S 1995 *The Quantum Theory of Fields* (Cambridge: Cambridge University Press)

[2] Two books on general relativity:
Weinberg W 1972 *Gravitation and Cosmology* (New York: Wiley)
Wald R M 1984 *General Relativity* (Chicago, IL: Chicago University Press)

[3] A review on the ultraviolet problem of quantum gravity:
Alvarez E 1989 *Quantum Gravity: A Pedagogical Introduction To Some Recent Results Rev. Mod. Phys.* **61** 561

[4] Two recent reviews of the BEH mechanism, with historical remarks:
Brout R *A Brief Course in Spontaneous Symmetry Breaking. I: The Paleolitic Age* arXiv:hep-th/0203096
Englert F *A Brief Course in Spontaneous Symmetry Breaking. II: Modern Times: The BEH Mechanism* arXiv:hep-th/0203097

[5] Some technical books on string theory:
Green M B, Schwarz J H and Witten E 1987 *Superstring Theory* 2 vols (Cambridge: Cambridge University Press)
Polchinski J 1998 *String Theory* 2 vols (Cambridge: Cambridge University Press)
Kiritsis E *Introduction to Superstring Theory* arXiv:hep-th/9709062
Lüst D and Theisen S 1989 Lectures On string theory *Lecture Notes Phys.* **346** 1

[6] Some important old reviews on String theory:
Jacob M (ed) 1974 *Dual Theory* (Amsterdam: North-Holland)
Scherk J 1975 An introduction to the theory of dual models and strings *Rev. Mod. Phys.* **47** 123

[7] A popular book on string theory:
Greene B R 1999 *The Elegant Universe* (New York: Norton)

[8] A recent review on canonical quantum gravity:
Rovelli C 2000 The century of the incomplete revolution: Searching for general relativistic quantum field theory *J. Math. Phys.* **41** 3776, arXiv:hep-th/9910131

[9] Some reviews on the cosmological constant problem:
Weinberg S 1989 The cosmological constant problem *Rev. Mod. Phys.* **61** 1
Witten E *The Cosmological Constant from the Viewpoint of String Theory* arXiv:hep-ph/0002297

Carroll S M 2001 The cosmological constant *Living Rev. Rel.* **4** 1, arXiv:astro-ph/0004075

[10] Two reviews on open strings and their applications:
Dudas E 2000 Theory and phenomenology of type I strings and M-theory *Class. Quantum Grav.* **17** R41, arXiv:hep-ph/0006190
Angelantonj C and Sagnotti A *Open Strings* arXiv:hep-th/0204089

[11] Some books and reviews on field theory solitons:
Rajaraman R 1982 *Solitons and Instantons* (Amsterdam: North-Holland)
Coleman S 1985 *Aspects Of Symmetry* (Cambridge: Cambridge University Press)
't Hooft G *Monopoles, Instantons and Confinement* arXiv:hep-th/0010225

[12] A review on string solitons:
Duff M J, Khuri R R and Lu J X 1995 *Phys. Rep.* **259** 213, arXiv:hep-th/9412184

[13] Some reviews on D-branes and their applications:
Polchinski J, Chaudhuri S and Johnson C V *Notes on D-Branes* arXiv:hep-th/9602052
Bachas C P *Lectures on D-branes* arXiv:hep-th/9806199
Johnson C V *D-brane primer* arXiv:hep-th/0007170

[14] A less technical review on D-branes:
Sevrin A 2000 From strings to branes: a primer *Les Arcs 2000, Electroweak Interactions and Unified Theories*

[15] Some reviews on duality relations:
Olive D I 1996 Exact electromagnetic duality *Nucl. Phys. Proc. Suppl.* A **45** 88
Olive D I 1996 *Nucl. Phys. Proc. Suppl.* **46** 1, arXiv:hep-th/9508089
Duff M J 1996 M theory (the theory formerly known as strings) *Int. J. Mod. Phys.* A **11** 5623, arXiv:hep-th/9608117
Sen A 1997 Unification of string dualities *Nucl. Phys. Proc. Suppl.* **58** 5, arXiv:hep-th/9609176
Townsend P K *Four Lectures on M-theory* arXiv:hep-th/9612121

[16] Some reviews on the AdS/CFT correspondence:
Witten E *New Perspectives in the Quest for Unification* arXiv:hep-ph/9812208
Aharony O, Gubser S S, Maldacena J M, Ooguri H and Oz Y 2000 Large N field theories, string theory and gravity *Phys. Rep.* **323** 183, arXiv:hep-th/9905111
Bianchi M 2001 (Non-)perturbative tests of the AdS/CFT correspondence *Nucl. Phys. Proc. Suppl.* **102** 56, arXiv:hep-th/0103112

[17] Some reviews on black holes in string theory:
Maldacena J M *Black Holes in String Theory*, arXiv:hep-th/9607235
Callan C 1997 Black holes in string theories: some surprising new developments *Les Arcs 1997, Electroweak Interactions and Unified Theories* 185
David J R, Mandal G and Wadia S R *Microscopic Formulation of Black Holes in String Theory* arXiv:hep-th/0203048

[18] A recent review on the holographic principle:
Bousso R *The Holographic Principle* arXiv:hep-th/0203101

[19] Some reviews on heterotic-string phenomenology:
Dienes K R 1997 String theory and the path to unification: A review of recent developments *Phys. Rep.* **287** 447, arXiv:hep-th/9602045
Quevedo F *Lectures on Superstring Phenomenology*, arXiv:hep-th/9603074

[20] Some reviews on large extra dimensions and open strings:

Lykken J D 1997 String model building in the age of D-branes *Nucl. Phys. Proc. Suppl.* A **52** 271, arXiv:hep-th/9607144

Ibanez L E *New Perspectives in String Phenomenology from Dualities* arXiv:hep-ph/9804236

Ibanez L E 2000 The second string (phenomenology) revolution *Class. Quantum Grav.* **17** 1117, arXiv:hep-ph/9911499

Antoniadis I *String and D-brane Physics at Low Energy* arXiv:hep-th/0102202

Antoniadis I and Benakli K 2000 Large dimensions and string physics in future colliders *Int. J. Mod. Phys.* A **15** 4237, arXiv:hep-ph/0007226

Bachas C P 2000 Scales of string theory *Class. Quantum Grav.* **17** 951, arXiv:hep-th/0001093

Ibanez L E *Standard Model Engineering with Intersecting Branes* arXiv:hep-ph/0109082

Blumenhagen R, Kors B, Lust D and Ott T *Intersecting Brane Worlds on Tori and Orbifolds* arXiv:hep-th/0112015