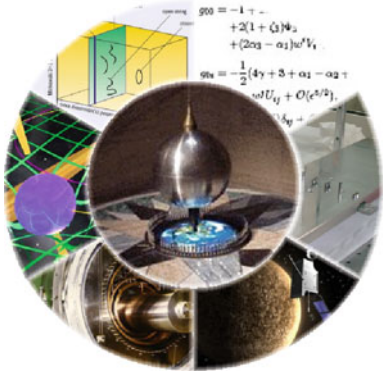
Roberto Peron · Monica Colpi
Vittorio Gorini · Ugo Moschella *Editors*

Gravity: Where Do We Stand?

Springer

Gravity: Where Do We Stand?



The logo of the SIGRAV School “Gravity: Where Do We Stand?” has been composed from a number of pictures from internet and the Group of Experimental Gravitation at IAPS-INAF. It is meant to symbolize various facets in gravitational physics research. Roberto Peron thanks Andrea Reale for his help in putting together the material.

Roberto Peron • Monica Colpi • Vittorio Gorini
Ugo Moschella
Editors

Gravity: Where Do We Stand?

Springer

Editors
Roberto Peron
Istituto di Astrofisica e Planetologia Spaziali
Istituto Nazionale di Astrofisica
Roma
Italy

Vittorio Gorini
Dipartimento di Scienza e Alta Tecnologia
Università degli Studi dell'Insubria
Como
Italy

Monica Colpi
Dipartimento di Fisica G. Occhialini
Università degli Studi di Milano-Bicocca
Milano
Italy

Ugo Moschella
Dipartimento di Scienza e Alta Tecnologia
Università degli Studi dell'Insubria
Como
Italy

ISBN 978-3-319-20223-5 ISBN 978-3-319-20224-2 (eBook)
DOI 10.1007/978-3-319-20224-2

Library of Congress Control Number: 2015949236

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

In Memoriam

Stefano Braccini (1964–2011) Stefano Braccini sadly left us in February 2011, at a time when he was working hard on the Advanced Virgo project, studying and imagining how to achieve the best possible sensitivity and secure the beginning of Gravitational Wave Astronomy.

Stefano graduated in Physics in 1992, with supervisors Carlo Bradaschia and Jean-Yves Vinet, with a thesis work on the effect of diffused light in the Virgo arms and in the central area. Subsequently he studied as Perfezionando at the Scuola Normale under the supervision of Adalberto Giazotto, until the discussion of his PhD thesis in 1996. During that time he dedicated himself to the development of the Seismic Superattenuator, a characteristic feature of the Virgo interferometer that allowed then unprecedented performance in the 10 and 100 Hz detection band. Following the thesis work, Stefano took the responsibility of completing the design of the full mirror suspension and following the construction of the nine Superattenuators, ensuring the coordination of the Rome and Pisa groups. He then worked on the commissioning of the interferometer, engaging a never-ending battle against noise. High-quality data were collected in 2009, demonstrating the fitness of Virgo for listening to the sky for years. This paved the way toward the second generation of interferometers whose development Stefano devoted many efforts to, and several of his thoughts are now integrated in the design.

Stefano took the leadership of the Pisa Virgo group in 2007, leading with enthusiasm the efforts on the detector and on the data analysis side, with special attention to rotating neutron star signals. He also dedicated much energy to the Istituto Nazionale di Fisica Nucleare (INFN), of which he was a researcher and then senior researcher. In particular, acting as referee for astroparticle experiments he was rapidly appreciated for his interest and passion for many other fields of physics research. We will remember Stefano for his achievements and his scientific and human integrity, his generosity. One way he would have liked is that we work harder than ever toward the detection of gravitational waves.

Francesco Fidecaro Università di Pisa, Pisa, Italy

Angioletta Coradini (1946–2011) Angioletta Coradini was an astrophysicist and a planetary scientist. Born in Rovereto on July 1st, 1946 she graduated in Physics from Rome University in 1970; she started her career as a planetary scientist during the pioneering times of lunar samples and from then on she devoted her whole life and tremendous energy to Planetary Science and, later, to Solar System Space science and exploration.

The range of her scientific interests and activities in Planetary Sciences was broad and spanned from her early study in the statistical classification of lunar samples to theoretical problems in the field of Solar System formation and evolution; such as modelling of processes of grain accretion for planetesimals growth in gaseous disks, formation of giant planets and their satellites, and models of the thermal evolution of solid bodies in the Solar System. She published more than 200 scientific papers, many of them reference works in their respective fields.

At the end of the 1980s, her personal Space Era started, and she directed her interest and energy to the field of Space Science and instrument design: a fleet of imaging spectrometers built under her responsibility are evidence of her dedication and still are flying the empty spaces of the Solar System on board spacecraft to Mars, to Venus, to the comets, to the icy realm of Jupiter and beyond.

Angioletta formed a generation of planetary scientists, inspired and led her younger colleagues with her energy and optimism, and promoted the development of a Planetary Science community in Italy from scratch with her enthusiasm and her charisma. She was a generous and warm person, always ready to forget many animated discussions with the breaking of her infectious smile.

As an enthusiastic and a passionate scientist, Angioletta was deeply curious of science and of the world in general, and did not stop at the questions and puzzles posed by her field, but was always ready to be diverted and interested by something for her completely new and unknown, a new challenge to respond to, a new book to study.

This is why we want to remember her in this book: the SIGRAV School at Villa Olmo delved in scientific topics far from her field, but she sponsored and promoted it with all her energy and passion. To Angioletta we devote this effort and this book.

Priscilla Cerroni Istituto di Astrofisica e Planetologia Spaziali (IAPS-INAF), Roma, Italy

Simeone Peron (1935–2012) Simeone Peron, here with his son Roberto

Preface

The SIGRAV (Italian Society of Relativity and Gravitation) Graduate School in Contemporary Relativity and Gravitational Physics has become a traditional meeting point in Italy for research in the field of general relativity and gravitational physics. The 2009 (VIIIth) edition of this School—Gravity: Where Do We Stand?—has been envisaged as an exploration path of gravitation, seen as a pervasive phenomenon acting at all scales, from the microscopic (the realm of a still-to-be-discovered satisfactory quantum theory of gravitation) to laboratory and Solar System experiments, up to astrophysical phenomena and cosmological scenarios. Each step of this path turns out to be full of insight, full of ideas, full of discoveries. This is not by chance for at least three reasons. First, as stated above, gravitation seems to play a major role in a whole range of important phenomena, from the build-up of spacetime itself (an oxymoron, we may say) to the formation and evolution of astrophysical structures, up to the very origin and evolution of the universe. Second, we have had, for almost a century, an amazing theory (general relativity) whose structure, and especially consequences, we are still exploring. Third, today we have at our disposal an ever-increasing amount of experimental data, from various techniques, and we can confidently speak about *precision tests*.

The School was attended by more than 50 students from all over the world and the lecturers, all outstanding scientists, were able to cover a substantial part of current research areas, from foundations to frontiers. The days of the School—also thanks to the beautiful venue of Villa Olmo, on Lake Como—have been nice and relaxed. As is customary for the school, after the end of this edition we started working on the preparation of the related book. This took a considerable amount of time, also due to new chapters which we have added to meet the suggestions by the book referees (whom we thank for their advice). All contributions were fully updated to the time the book went to print and we are finally able to offer the text to the attention of the reader.

This is also the place to remember two persons who are not among us anymore. Stefano Braccini has been a passionate and competent lecturer, and as such contributed to the success of the School. The news of his untimely passing filled us with

dismay. Angioletta Coradini, at that time Director of Istituto di Fisica dello Spazio Interplanetario (IFSI-INAF), was a passionate supporter of the School. All of us miss such a great person, as well as a first-class scientist. On paper ... The reader can find a remembrance of them by Francesco Fidecaro and Priscilla Cerroni.

The school was made possible thanks to the support of SIGRAV, the University of Insubria, the Department of Physics and Mathematics of the Insubria University, the National Institute of Nuclear Physics (INFN) and Galileian Plus S.r.l. (for which we warmly thank Massimiliano Chersich). We are grateful to the secretarial conference staff of the Center Alessandro Volta, in particular to Chiara Stefanetti and Francesca Gamba for their assistance, precious help and kindness.

Roberto Peron
Monica Colpi
Vittorio Gorini
Ugo Moschella

Contents

Contributors

Garry W. Angus Astrophysics, Cosmology and Gravity Centre, University of Cape Town, Rondebosch, South Africa

Ignatios Antoniadis Department of Physics, CERN-Theory Division, Geneva, Switzerland

CPHT (UMR CNRS 7644) Ecole Polytechnique, Palaiseau, France

Stefano Braccini INFN-Sezione di Pisa, Pisa, Italy

Marta Burgay INAF-Osservatorio Astronomico di Cagliari, Selargius, Cagliari, Italy

Simone Dell'Agnello INFN-LNF, Frascati, Rome

Antonaldo Diaferio Dipartimento di Fisica, Università degli Studi di Torino, and Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Torino, Torino, Italy

Francesco Fidecaro "Enrico Fermi" Physics Dept., University of Pisa and INFN, Pisa, Italy

Valerio A. Iafolla Istituto di Astrofisica e Planetologia Spaziali (IAPS), Istituto Nazionale di Astrofisica (INAF), Roma, Italy

Stefano Liberati SISSA and INFN, Trieste, Italy

Luca Lusanna Polo Scientifico, Sezione INFN di Firenze, Sesto Fiorentino, Florence, Italy

Manuele Martini INFN-LNF, Frascati, Rome

David Mattingly University of New Hampshire, Durham, USA

Roberto Peron Istituto di Astrofisica e Planetologia Spaziali (IAPS), Istituto Nazionale di Astrofisica (INAF), Roma, Italy

Andrea Possenti INAF-Osservatorio Astronomico di Cagliari, Selargius, Cagliari, Italy

Vyacheslav G. Turyshev Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Clifford M. Will University of Florida, Gainesville, USA

Introduction

Roberto Peron

> Bastò questa fuggevole impressione a farmi vedere la situazione in modo nuovo: se era vero che lo spazio con qualcosa dentro è diverso dallo spazio vuoto perché la materia vi provoca una curvatura o tensione che obbliga tutte le linee in esso contenute a tendersi o curvarsi allora la linea che ognuno di noi seguiva era una retta nel solo modo in cui una retta può essere retta cioé deformandosi di quanto la limpida armonia del vuoto generale è deformata dall'ingombro della materia, ossia attorcigliandosi tutto in giro a questo gnocco o porro o escrescenza che è l'universo nel mezzo dello spazio.
>
> ...
>
> Questa era però ancora un'immagine schematica, come se avessimo a che fare con un solido dalle pareti lisce, una compenetrazione di poliedri, un aggregato di cristalli; in realtà lo spazio in cui ci muovevamo era tutto merlato e traforato, con guglie e pinnacoli che si irradiavano da ogni parte, con cupole e balaustre e peristili, con bifore e trifore e rosoni, e noi mentre ci sembrava di piombar giù dritto in realtà scorrevamo sul bordo di modanature e fregi invisibili, come formiche che per attraversare una città seguono percorsi tracciati non sul selciato delle vie ma lungo le pareti e i soffitti e le cornici e i lampadari.
>
> Italo Calvino, Le Cosmicomiche, La forma dello spazio

Why another book on gravitation? Well, why continue speaking and reasoning on gravitation? Of course, gravitation is amongst the four known fundamental interactions of nature. Of course, gravitation shapes in many ways the world around us, from planets to stars to the dynamics of the universe as a whole. In fact, these are good reasons to continue to speak and investigate on this phenomenon (and worth more than a book). I think, however, there is a more subtle hint we can follow. It relates to the superb idea — started with the work of Gauss and Riemann on differential geometry and achieved by Einstein in his general theory of relativity — that gravitation is not exactly like the other fundamental interactions, but instead the result of a characteristic of spacetime itself. This finding is not granted and, in fact, related empirical evidence (think about the Equivalence Principle) has been under heavy empirical scrutiny for many decades. Yet, the experimental corpus that has been accumulated in favour of this view is impressive; and we are left with a tool

R. Peron (✉)
Istituto di Astrofisica e Planetologia Spaziali (IAPS), Istituto Nazionale di Astrofisica (INAF), Via Fosso del Cavaliere 100, Roma, Italy
e-mail: roberto.peron@iaps.inaf.it

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_1

which provides us with something very deep on those (remember Kant) concepts which shape the world and the way we see it.

This book comes from an attempt of putting together in a consistent way important pieces of knowledge in the current development of research in what can be called "gravitation physics", both from the theoretical and experimental points of view. The focus of this path, however, has been shifted to issues which can be termed as "foundational" or that can be related, in any case, to fundamental questions. This was in fact was the initial idea behind the organization of the 2009 Villa Olmo SIGRAV School "Gravity: Where do we stand?", and that, with some adjustments, is presented here.

The chapter by Clifford Will, *Gravity: Newtonian, Post-Newtonian and General Relativistic*, opens the stage, providing a very thorough introduction to gravitational theory from the theoretical point of view, stressing along the way the manifold points of contact with empirical evidence given by experiments, both in the Solar System and in the astrophysical context. Going from Newtonian gravity to general relativity and from various fields of application (and approximations) of Einsteinian theory to real-world situations, a wide view of our knowledge on gravitational machinery is offered. Of particular interest is the discussion of the motion of extended fluid bodies (Newtonian hydrodynamics) and the derivation of post-Minkowskian and post-Newtonian theories by iteration of the so-called *relaxed Einstein field equation*. It is also striking the possibility of applying these approximation schemes to situations where apparently these should not work properly, in relatively strong-field systems such as the compact binaries. The overall sensation is of walking in an extremely rich garden: at every step, at every angle, new species appear, while the general pattern of the garden itself is present everywhere .

A series of chapters follows, in which gravitational phenomena are explored in the environment most directly available to us for exploration: the Earth and (in selected cases) the Solar System. This is the realm to the so-called *weak-field* condition, in which the gravitational source masses (and the related gravitational effects) are so small that a basic, linear approximation of the exact theory suffices in taking into account the observable gravitational phenomenology. This regime is nearest to our everyday experience. Yet we could have surprises. An example could be the presence of the so-called *gravitomagnetic* phenomena, analogous to the most familiar ones in electromagnetism. Another is the need, in very precise metrological applications, to take into account at least the most direct relativistic effects, such as time dilation.

Valerio Iafolla, in his chapter *The Newtonian Gravity and Some of Its Classical Tests*, offers us a tour exploring the phenomenology of gravitational phenomena around us, in (as is customary to say) a laboratory environment. This field of research has a long history, starting with Galileo's hints and experiments and proceeding along the centuries up to present times. Indeed, the very basic questions (present in the title of this chapter) are as valid now as they were centuries ago: Why does gravitation seem so different with respect to the other known interactions of nature? What precisely is the magnitude of the gravitational interactions? Is gravitation really a long-range interaction? These questions are after all very simple, yet

pertain to aspects so fundamental that each generation of scientists has something to say (and to do) about them. A common feature of the experiments designed to answer to these questions is that they deal with the measurement of very small accelerations in a rather noisy environment. A technique is described which makes use of advanced devices based on the concept of an "accelerometer": a test mass attached to a frame, and whose motion is precisely sensed. This instrument, in a variety of configurations, is applied to build experiments to give an answer to the above-mentioned questions. Different types of acceleration-sensitive devices are at the core of experiments designed to test the Equivalence Principle in its weak formulation, to measure the Newtonian gravitational constant, to verify the behaviour of gravitational attraction at laboratory scales. The range of applications of these instruments and techniques extends to geophysical uses and to interplanetary missions.

Extending the scale of gravitational investigations to the size of planetary bodies, the general relativistic phenomenology starts appearing. This is not by chance: thinking in terms of distances, a rough measure of the relativistic signal is given by the Schwarzschild radius GM/c^2, with M the mass of the main body producing spacetime curvature. A back-of-the-envelope calculation shows that bodies of mass comparable at least to the terrrestrial planets are necessary to produce a signal detectable with current metrology. For the Earth, this number amounts to a few centimeters. Nowadays, thanks to laser tracking, it is possible to follow the motion of objects around Earth with this level of precision. The chapter by Roberto Peron, *Fundamental Physics with the LAGEOS Satellites*, describes the experiments made possible by this technique and by the availability of satellites which are the closest man-made realizations to the ideal concept of a *test mass*. In these experiments, the data at our disposal are so precise that the biggest effort is to develop a modelization of comparable quality. Indeed, the motion of these objects is extremely sensitive to changes in the Earth-system environment (which is the main reason for their development). Putting together these high-quality data and state-of-the-art dynamical models and orbit determination software allows filtering out this "noise" and revealing a perhaps unsuspected relativistic world, with several effects in action. Amongst them, the already cited gravitomagnetic phenomena, and a smaller realization of one of the basic tests of general relativity, that is the pericentre precession (with the Earth taking the place of the Sun).

Nature gave us a wonderful instance where to verify the machinery of gravitation: the Earth–Moon system. Indeed, the Sun–Earth–Moon triad constitutes a paradigmatic example of a three-body system, which shows considerable complexity even at the Newtonian level (and in fact it has been an important hint to centuries of developments in mechanics, celestial mechanics and theoretical physics). Since the Apollo 11 mission, retroreflector arrays are present on the Moon surface. These arrays are used as reference points for laser ranging and are routinely tracked by stations on Earth. In fact, this technique provides an uninterrupted time series of measurements lasting more than forty years. Manuele Martini and Simone Dell'Agnello, in the chapter *Probing Gravity with Next Generation Lunar Laser Ranging*, describe the current activity in this field of research, which along the years saw many important results in fundamental physics and geophysics. An

important goal is the development of a new generation of retroreflectors, capable of complying with the advanced capabilities of ground stations such as APOLLO. In particular, a facility to characterize the retroreflectors on the thermal and optical sides is described.

The chapter by Vyacheslav Turyshev, *Space-based Tests of Relativistic Gravitation*, shifts our attention to the wide arena of Solar System dynamics. This is the place where evidence has been first found for general relativistic dynamics; think about the well-known secular precession of Mercury's orbit around the Sun and the deflection of starlight passing near the Sun (both involving the mass of our star curving spacetime in a sensible way). With the advent of the space age, we started to send spacecraft around in the Solar System. This means exploring gravitation too: each spacecraft can be considered to be with a very good approximation a test mass, and its motion — accurately tracked — contains information on the gravitational dynamics. In practice, almost every deep-space mission has implied "piggyback" experiments using radiometric tracking of spacecraft. Most important results have been obtained in the past decades, and others are expected in the near future. Advanced technologies, such as laser tracking over interplanetary distances, promise to extend this task even further, which we may call "Solar System gravitational surveying and mapping".

Evidence for the relativistic dynamics of the gravitational field is provided not only at the "local" scale of the Solar System but also in a "more astrophysical" context. Indeed, we expect to find traces of this dynamics at all accessible scales, given the importance of gravitation in the formation and evolution of astrophysical structures till the grand arena of cosmology. Two of these traces are particularly significant: the quest for gravitational waves and the study of compact objects.

Gravitational waves are currently a universally accepted consequence of general relativity theory, and a manifestation of gravitation being a field phenomenon, that is something extended in space(time). Indirect evidence for them came from the PSR1913+16 system, whose orbital evolution is compatible with the loss of energy due to gravitational wave emission. After Joseph Weber's pioneering attempts at measuring gravitational waves in the 1960s, this search has never stopped, with more and more sensitive devices. Stefano Braccini and Francesco Fidecaro, in the chapter *The Detection of Gravitational Waves*, carefully describe this search, starting from the mathematical description of their generation and propagation. The statistical characterization of possible sources is extremely important, since this knowledge is useful to constrain event detection probabilities. The big challenge is of course the smallness of the signal to be detected; for this reason, in recent years the focus has shifted more and more from resonant bars to interferometric detectors, whose features (especially their wide sensitivity band) seem promising in view of actual detections. The story of gravitational waves search is the story of fight against noise. When this fight will be won (soon, we hope), as is commonly said, a new window will be open for our observation of the universe.

Compact objects are a wide class of astrophysical entities. Among them, radio pulsars (neutron stars emitting collimated beams of radio waves) are particularly interesting for gravitational physics. Some of them constitute, in effect, very stable

clocks and the study of the emitted signal can be used to constrain several of their dynamic features. In their chapter, *The Role of Binary Pulsars in Testing Gravity Theories*, Andrea Possenti and Marta Burgay concentrate on a particular class of radio pulsars, the so-called relativistic binary pulsars, and show how a careful analysis of the *times of arrival* of the radio pulses reveals several properties of these systems. The analyses can be carried out in the usual post-Newtonian framework or, when the two objects of the system are close enough, in the post-Keplerian one. The impressive constraints obtained from these analyses once more highlight the "special" role of general relativity among possible theories of gravitation.

The conspicuous amount of evidence gathered in the last decades may induce one to think the scene is very clear and no big issues are on the horizon. In fact, it is quite the opposite. The challenges come from many fronts. First of all, there are many "darks": dark matter, dark energy, with all the related issues. As analogous cases in the past, this could imply at least two alternatives: there is something missing in the counts, or we do not understand correctly the basic laws. Given the possible amount of "missing matter", it is legitimate to ask ourselves whether we fully understand general relativity and its consequences. A basic question could be: what is the behaviour of gravitation at very large distances? Surprises can appear: indeed, as of today, there is a lack of *precise* experimental knowledge. There is then the question on the behaviour of gravitation at extreme conditions and, more in general, its desirable reconciling with quantum principles. We do not currently have at our disposal a quantum theory of gravitation at the same time self-consistent and capable of providing definite answers to precise experimental setups.

Given the mathematical complexity of general relativity equations, the link between formal solutions and the results of experiments is in general quite tricky, once left the trivial cases. In other words, the development of a consistent relativistic metrology is something achieved only in part and needs constant improvements to keep the pace of experimental results. Luca Lusanna, in his chapter *Non-inertial Frames in Special and General Relativity*, faces this delicate issue. A theory of global non-inertial frames in special and general relativity is introduced and used to treat asymptotically Minkowskian spacetimes. The very interesting possibility is suggested that dark matter and dark energy could be effects related to the choice of the coordinate system.

The problems one has to face in interpreting astrophysical data are discussed from a peculiar point of view by Antonaldo Diaferio and Garry W. Angus, in their chapter, *The Acceleration Scale, Modified Newtonian Dynamics, and Sterile Neutrinos*. The galactic dynamics phenomenology seems to call for an acceleration scale below which a modification of Newtonian dynamics would be necessary: this is the objective of the MOdified Newtonian Dynamics (MOND). This theory, with a minimum of basic hyphotheses, seems to neatly explain the features of galactic rotation curves as well as other important features. The basic principles justifying this scenario are possibly still to be unveiled, but this is another indication that progress is certainly welcome in the way we understand gravitation.

Another aspect currently under theoretical and experimental scrutiny is represented by the low-energy consequences of gravitation at the quantum level. In this

respect, several theories predict some type of violation of Lorentz symmetry. Stefano Liberati and David Mattingly, in their chapter *Lorentz Breaking Effective Field Theory Models for Matter and Gravity: Theory and Observational Constraints*, thoroughly explore this issue in the context of the Standard Model Extension. By applying techniques of effective field theory, they review several constraints provided by current observations.

In the background, there is of course the quest for a quantum theory of gravitation in its full shape, which would be a fundamental step in the unification of all known interactions in nature. Several attempts have been made over the years, with many challenges to overcome, not least being able to answer to definite experimental questions. Among them, one of the most crossed paths is string theory. Ignations Antoniadis, in his chapter *Possible Low Energy Manifestations of Strings and Gravity*, provides a discussion of the basic motivations of such a theory and briefly introduces its perturbative version. The definitely innovative view of spacetime it offers opens the way to observational attempts, both in particle colliders and in experiments dedicated to the search for new short-range forces.

Part I
Foundations and Solar System Tests

Gravity: Newtonian, Post-Newtonian, and General Relativistic

Clifford M. Will

Abstract We present a pedagogical introduction to gravitational theory, with the main focus on weak gravitational fields. We begin with a thorough survey of Newtonian gravitational theory. After a brief introduction to general relativity, we develop the post-Minkowskian formulation of the field equations, which is ideally suited to studying weak-field gravity. We then discuss applications of this formulation, including post-Newtonian theory, the parametrized post-Newtonian framework, and gravitational radiation.

1 Introduction

General relativity is a theory of "gravity as geometry." It provides a remarkable new conception of the nature of gravity, totally at odds with the idea of gravity as conceived by Isaac Newton. It makes predictions of remarkable new and exotic phenomena: black holes, space-time singularities, the expanding universe.

But, as we approach the centenary of the 1915 publication of the full theory, we have learned that there is a wealth of important physical and astrophysical phenomena in which general relativity is important, but for which the full, exact theory is not needed. These phenomena range from the dynamics of the solar system, where many important experimental tests of general relativity have been carried out, to binary pulsar systems, to the dynamics of stars in galactic cores containing supermassive black holes, to the inspiral of binary systems of neutron stars or black holes, which will be important sources of detectable gravitational waves. These are situations where the gravitational potentials are small (in units of c^2) and speeds are low (in units of c). These cases lend themselves to methods that forego any attempt to find exact solutions of Einstein's equations, but instead attempt to develop systematic approximations to solutions of the equations.

Among these methods are the "post-Minkowskian" and "post-Newtonian" approximations that will be the main topic of this chapter. In post-Minkowskian theory, the strength of the gravitational field is measured by the gravitational constant G, and the Einstein field equations are formally expanded in powers of G.

C. M. Will (✉)
University of Florida, Gainesville, USA
e-mail: cmw@physics.ufl.edu

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_2

At zeroth post-Minkowskian order there is no field, and one deals with Minkowski space-time. At first post-Minkowskian order, the gravitational field appears as a correction of order G to the Minkowski metric, and the (linearized) field equations are integrated to obtain this correction. The correction is refined by terms of order G^2 in the second post-Minkowskian approximation, and the process is continued until the desired degree of accuracy is achieved. When augmented by an assumption of slow motions, the method is known as post-Newtonian theory. A slow-motion assumption is natural, as the virial theorem requires that kinetic energies be comparable to gravitational potentials, so that weak fields, characterized by (potentials)$/c^2 \ll 1$, go hand in hand with slow motions, $(v/c)^2 \ll 1$.

These methods have proven in recent years to be very powerful, and unreasonably effective in describing important relativistic phenomena. Because the methods can be carried out to higher orders in the approximation, they can handle situations where the fields are not so weak or the motions not so low. For example, high-order calculations in post-Newtonian theory of the late inspiral of a compact binary system have proven to agree remarkably well with results from numerical relativity, in regimes where the two techniques overlap.

Accordingly, this chapter is designed as a pedagogical introduction to these methods. Because they yield Newtonian gravity as the first-order approximation, we begin with a fairly lengthy exposition of Newtonian theory (Sect. 2), one that goes beyond the simplest undergraduate level of "inverse square law" –"elliptical orbit" concepts. After a brief introduction to general relativity (Sect. 3), we describe the post-Minkowskian approach (Sect. 4). This is an exact reformulation of Einstein's equations that nicely lends itself to a systematic sequence of approximations. Including slow motions yields post-Newtonian theory, along with the parametrized post-Newtonian (PPN) formalism, used to discuss alternative theories of gravity and their experimental tests (Sect. 5). Finally, we discuss the use of these methods to analyze gravitational radiation (Sect. 6).

Because this chapter is meant to be pedagogical, we shall not provide extensive references to the primary literature, but will instead give a bibliography and suggestions for further reading at the end. This chapter is based in part on a forthcoming book, *Gravity: Newtonian, post-Newtonian, Relativistic*, coauthored by Eric Poisson. I am grateful to my coauthor and to Cambridge University Press for permission to use portions of the book draft in this chapter.

2 Newtonian Gravity

The gravitational theory of Newton is an extremely good representation of gravity for a host of situations of practical and astronomical interest. It accurately describes the structure of the Earth and the tides raised on it by the Moon and Sun. It gives a detailed account of the orbital motion of the Moon around the Earth, and of the planets around the Sun. Of course, it is now well-established that Newtonian theory is not an exact description of the laws of gravitation. But apart from specialized

situations requiring very high precision, such as the orbit of Mercury, timing in the Global Positioning System, or very long baseline radio interferometry (VLBI), Newtonian gravity rules the solar system.

For the overwhelming majority of stars in the universe, Newtonian gravity also rules. The structure and evolution of the Sun and other "main sequence" stars can be completely and accurately treated using Newtonian gravity. Only for extremely compact stellar objects, such as neutron stars and, of course, black holes, is general relativity important. Newtonian gravity is also perfectly capable of handling the structure and evolution of galaxies and clusters of galaxies, and can even be used in the context of an expanding universe, as long as the relevant distance scales are shorter than the Hubble scale.

Generally speaking, the criterion that we use to decide whether to employ Newtonian gravity or general relativity is the magnitude of the quantity

$$\varepsilon \sim \frac{GM}{c^2 r} \sim \frac{v^2}{c^2}\,, \tag{1}$$

where G is the Newtonian gravitational constant, c is the speed of light, and where M, r, and v represent the characteristic mass, separation or size, and velocity within the system under consideration. The smaller this factor, the better is Newtonian gravity as an approximation.

2.1 The Equations of Newtonian Gravity

Most undergraduate textbooks begin their treatment of Newtonian gravity with Newton's second law and the inverse-square law of gravitation:

$$\begin{aligned} m_I \mathbf{a} &= \mathbf{F}\,, \\ \mathbf{F} &= -\frac{G m_G M}{r^2}\,\mathbf{n}\,. \end{aligned} \tag{2}$$

In the first equation, $\mathbf{F}$ is the force acting on a body of inertial mass m_I situated at position $\mathbf{r}$, and $\mathbf{a} = d^2\mathbf{r}/dt^2$ is its acceleration. In the second equation, the force is assumed to be gravitational in nature, and to originate from a gravitating mass M situated at the origin of the coordinate system. The force law involves m_G, the passive gravitational mass of the first body at $\mathbf{r}$, while M is the active gravitational mass of the second body. The force is attractive, it varies inversely with the square of the distance $r := |\mathbf{r}| = (x^2 + y^2 + z^2)^{1/2}$, and it points in the direction opposite to the unit vector $\mathbf{n} := \mathbf{r}/r$.

An alternative form of the force law is obtained by writing it as the gradient of a potential $U = GM/r$, so that

$$\mathbf{F} = m_G \nabla U\,. \tag{3}$$

If the inertial and passive gravitational masses of the body are equal to each other, $m_I = m_G$, then the acceleration of the body is given by $\mathbf{a} = \nabla U$, and its magnitude is $a = GM/r^2$. Under this circumstance, the acceleration is independent of the mass of the body. This statement is known as the *weak equivalence principle* (WEP), and it was a central element in Einstein's thinking en route to the concepts of curved space-time and general relativity. Newton himself deemed this principle so important that he opened his 1687 magnum opus *The Principia* with a discussion of it and the experiments he had performed to test it. Twentieth-century experiments have shown that the two types of mass are equal (or more precisely that the ratio m_G/m_I is independent of the composition of the material) to a few parts in 10^{13} for a wide variety of materials. The chapter, "The Newtonian Gravity and Some of Its Classical Tests (Equivalence Principle, Measure of G, $1/r^2$" by V. Iafolla, discusses these experiments in more detail).

In order to consider the motion of extended bodies made up of continuous matter (solid, fluid, or gas), allowing the bodies to be of arbitrary size, shape, and constitution, and possibly to evolve in time according to their own internal dynamics, it is necessary to generalize the primitive Eq. (2) to a form that applies to a continuous distribution of matter.

To do this, we characterize the matter distribution by a mass-density field $\rho(t, \mathbf{x})$, a pressure field $p(t, \mathbf{x})$, and a velocity field $\mathbf{v}(t, \mathbf{x})$; these quantities depend on time t and position $\mathbf{x}$ within the fluid. The equations that govern the behavior of the matter are the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0\,, \tag{4}$$

which expresses conservation of mass, and *Euler's equation*,

$$\rho \frac{d\mathbf{v}}{dt} = \rho \nabla U - \nabla p\,, \tag{5}$$

which is the generalization of Eq. (2) to continuous matter; here

$$\frac{d}{dt} := \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla, \tag{6}$$

is the convective time derivative associated with the motion of fluid elements. The equation that governs the behavior of the gravitational field is *Poisson's equation*

$$\nabla^2 U = -4\pi G \rho\,, \tag{7}$$

where

$$\nabla^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \tag{8}$$

is the familiar Laplacian operator. Using the Green function for the Laplacian, we can convert this to an integral

$$U(t, \mathbf{x}) = G \int \frac{\rho(t, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}\, d^3x'\,. \tag{9}$$

The integral can be evaluated as soon as the density field $\rho(t, \mathbf{x}')$ is specified, regardless of whether ρ is a proper solution to the remaining fluid equations. As such, Eq. (9) gives U as a *functional* of any arbitrary function ρ. The potential, however, will be physically meaningful only when ρ itself is physically meaningful, which means that it must be a proper solution to the continuity and Euler equations.

To complete the formulation of the theory we must impose a relationship between the pressure and the density of the fluid. This relationship, known as the *equation of state*, takes the general form of

$$p = p(\rho, T, \cdots), \tag{10}$$

in which the pressure is expressed as a function of the density, temperature, and possibly other relevant variables such as chemical composition. The equation of state encodes information about the microphysics that characterizes the fluid, and this information must be provided as an input in most applications of the theory.

A complete description of a physical situation involving gravity and a distribution of matter will be obtained by integrating Eqs. (4), (5), and (7) simultaneously and self-consistently. The solutions must be subjected to suitable boundary conditions, which will be part of the specification of the problem. All of Newtonian gravity is contained in these equations, and all associated phenomena follow as consequences of these equations.

The equations of hydrodynamics give rise to a number of important *global conservation laws*. These refer to global quantities, defined by integrals over the entire fluid system, that are constant in time whenever the system is *isolated*, that is, whenever the system is not affected by forces external to it or by flows of matter out of it. The equation of continuity (4) leads to some useful properties of such integrals. For an arbitrary function $f(t, \mathbf{x})$, use of the equation of continuity and an integration by parts leads to

$$\frac{d}{dt}\int \rho(t, \mathbf{x}) f(t, \mathbf{x})\, d^3x = \int \rho \frac{df}{dt}\, d^3x\,, \tag{11}$$

where, as before, d/dt is the convective (or Lagrangian) time derivative and where the volume of integration is fixed in space and contains all the matter.

For the integral $F(t, \mathbf{x}) := \int \rho(t, \mathbf{x}') f(t, \mathbf{x}, \mathbf{x}')\, d^3x'$,

$$\frac{\partial F}{\partial t} = \int \rho' \left(\frac{\partial f}{\partial t} + \mathbf{v}' \cdot \nabla' f \right) d^3x'\,. \tag{12}$$

The *Lagrangian time derivative* acting on F is $dF/dt = \partial F/\partial t + \mathbf{v} \cdot \nabla F$, and from Eq. (12) and the definition of $F(t, \mathbf{x})$ we find that this can be expressed as

$$\frac{dF}{dt} = \int \rho' \frac{df}{dt}\, d^3x'\,, \tag{13}$$

with

$$\frac{df}{dt} := \frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f + \mathbf{v}' \cdot \nabla' f, \tag{14}$$

denoting a generalized Lagrangian derivative.

With these results it is straightforward, using the equation of continuity and the Euler equation, to show that the following are constant in time:

$$\text{Mass}: \quad M := \int \rho(t, \mathbf{x})\, d^3x, \tag{15a}$$

$$\text{Momentum}: \quad \mathbf{P} := \int \rho(t, \mathbf{x})\mathbf{v}(t, \mathbf{x})\, d^3x, \tag{15b}$$

$$\text{Angular Momentum}: \quad \mathbf{J} := \int \rho\, \mathbf{x} \times \mathbf{v}\, d^3x, \tag{15c}$$

$$\text{Energy}: \quad E := \mathcal{T}(t) + \Omega(t) + E_{\rm int}(t). \tag{15d}$$

In addition, the center of mass, defined by

$$\mathbf{R} := \frac{1}{M} \int \rho(t, \mathbf{x})\mathbf{x}\, d^3x, \tag{16}$$

is a linear function of time. Finally, one can prove the virial theorems

$$\frac{1}{2}\frac{d^2 I^{jk}}{dt^2} = 2\mathcal{T}^{jk} + \Omega^{jk} + P\delta^{jk}, \tag{17a}$$

$$\frac{1}{2}\frac{d^2 I}{dt^2} = 2\mathcal{T} + \Omega + 3P, \tag{17b}$$

where

$$\mathcal{T}(t) := \frac{1}{2} \int \rho v^2\, d^3x, \tag{18a}$$

$$\mathcal{T}^{jk}(t) := \frac{1}{2} \int \rho v^j v^k\, d^3x, \tag{18b}$$

$$\Omega(t) := -\frac{1}{2} \int \rho U\, d^3x = -\frac{1}{2} G \int \frac{\rho\rho'}{|\mathbf{x} - \mathbf{x}'|}\, d^3x' d^3x, \tag{18c}$$

$$\Omega^{jk}(t) := -\frac{1}{2} G \int \rho\rho' \frac{(x - x')^j (x - x')^k}{|\mathbf{x} - \mathbf{x}'|^3}\, d^3x' d^3x, \tag{18d}$$

$$E_{\rm int}(t) := \int \varepsilon\, d^3x, \tag{18e}$$

$$P(t) := \int p\, d^3x, \tag{18f}$$

$$I^{jk}(t) := \int \rho x^j x^k d^3x, \tag{18g}$$

$$I(t) := \int \rho r^2 d^3x, \tag{18h}$$

where ε is the energy density, subject to the first law of thermodynamics for perfect or isentropic fluids, $d(\varepsilon\mathcal{V}) + p d\mathcal{V} = 0$, where the element of volume evolves with the fluid flow according to $\mathcal{V}^{-1} d\mathcal{V}/dt = \nabla \cdot \mathbf{v}$.

2.2 Spherical and Nonspherical Bodies

In spherical symmetry, Poisson's equation (7) takes the form

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) = -4\pi G\rho(t,r)\,. \tag{19}$$

Defining the mass contained inside a sphere of radius r as

$$m(t,r) := \int_0^r 4\pi\rho(t,r')r'^2\,dr'\,, \tag{20}$$

we rewrite Poisson's equation in the form

$$\frac{\partial U}{\partial r} = -\frac{Gm(t,r)}{r^2}\,. \tag{21}$$

If R is the radius at which $\rho = 0$, then, defining $M = m(R)$ (a constant by virtue of the continuity equation), we can write the solution of Eq. (21) that is continuous at $r = R$ and vanishes at infinity in the form

$$U(t,r) = \frac{Gm(t,r)}{r} + 4\pi G\int_r^R \rho(t,r')r'\,dr'\,. \tag{22}$$

Outside the body, $U = GM/r$, a constant in time.

To describe the potential of nonspherical bodies, we return to Eq. (9), and, for a field point outside the body ($|\mathbf{x}'| < |\mathbf{x}|$), we carry out a Taylor expansion of $|\mathbf{x} - \mathbf{x}'|^{-1}$:

$$\begin{aligned}\frac{1}{|\mathbf{x}-\mathbf{x}'|} &= \frac{1}{r} - x'^j\partial_j\left(\frac{1}{r}\right) + \frac{1}{2}x'^{jk}\partial_{jk}\left(\frac{1}{r}\right) - \cdots \\ &= \sum_{\ell=0}^{\infty}\frac{(-1)^\ell}{\ell!}x'^L\partial_L\left(\frac{1}{r}\right)\,,\end{aligned} \tag{23}$$

where we introduce a condensed notation in which an expression like x^{jkn} stands for the product $x^jx^kx^n$, and ∂_{jkn} stands for $\partial_j\partial_k\partial_n$. We adopt the Einstein summation convention and sum over repeated indices in an expression like $x'^{jk}\partial_{jk}r^{-1}$. In the second line, we introduce a *multi-index* notation, in which an uppercase index such as L represents a collection of ℓ individual indices. Thus, x^L stands for $x^{j_1j_2\cdots j_\ell}$, ∂_L stands for $\partial_{j_1j_2\cdots j_\ell}$, and $x^L\partial_L$ involves a summation over all ℓ pairs of repeated indices. Substituting Eq. (23) into Eq. (9) gives

$$U_{\rm ext}(t,\mathbf{x}) = G\sum_{\ell=0}^{\infty}\frac{(-1)^\ell}{\ell!}I^{\langle L\rangle}\partial_{\langle L\rangle}\left(\frac{1}{r}\right)\,, \tag{24}$$

with

$$I^{\langle L\rangle}(t) := \int\rho(t,\mathbf{x}')x'^{\langle L\rangle}\,d^3x'\,, \tag{25}$$

defining a set of *STF multipole moments* for the mass distribution. The STF label denotes *symmetric, trace-free*: an object $A^{\langle ijkl\ldots\rangle}$ is symmetric on all its indices and, for any pair of indices, $\delta_{ij}A^{\langle ijkl\ldots\rangle} = 0$. The gradients of r^{-1} generate STF tensors. Using the fact that

$$\partial_j r = n_j\,, \tag{26a}$$

$$\partial_j n_k = \partial_k n_j = \frac{1}{r}\big(\delta_{jk} - n_j n_k\big)\,, \tag{26b}$$

it is straightforward to show that

$$\partial_j r^{-1} = -n_j r^{-2}\,, \tag{27a}$$

$$\partial_{jk} r^{-1} = \big(3n_j n_k - \delta_{jk}\big) r^{-3}\,, \tag{27b}$$

$$\partial_{jkn} r^{-1} = -\Big[15 n_j n_k n_n - 3\big(n_j\delta_{kn} + n_k\delta_{jn} + n_n\delta_{jk}\big)\Big] r^{-4}\,. \tag{27c}$$

We define the STF tensors

$$n^{\langle jk\rangle} := n^j n^k - \frac{1}{3}\delta^{jk}\,, \tag{28a}$$

$$n^{\langle jkn\rangle} := n^j n^k n^n - \frac{1}{5}\big(\delta^{jk}n^n + \delta^{jn}n^k + \delta^{kn}n^j\big)\,, \tag{28b}$$

and more generally,

$$n^{\langle L\rangle} := \sum_{p=0}^{[\ell/2]} (-1)^p \frac{(2\ell-2p-1)!!}{(2\ell-1)!!}\Big[\delta^{2P} n^{L-2P} + \mathrm{sym}(q)\Big]\,, \tag{29}$$

where δ^{2P} stands for a product of p Kronecker deltas (with indices running from j_1 to j_{2p}), n^{L-2P} stands for a product of $\ell-2p$ unit vectors (with indices running from j_{2p+1} to j_ℓ), and "sym(q)" denotes all distinct terms arising from permuting indices; the total number of terms within the square brackets is equal to $q := \ell!/[(\ell - 2p)!(2p)!!]$. Then, we have

$$\partial_L r^{-1} = \partial_{\langle L\rangle} r^{-1} = (-1)^\ell (2\ell-1)!!\,\frac{n_{\langle L\rangle}}{r^{\ell+1}}\,. \tag{30}$$

The fact that $\partial_L r^{-1}$ is trace-free follows trivially from the fact that $\nabla^2 r^{-1} = 0$.

There is a close connection between $n^{\langle L\rangle}$ and the spherical harmonics. For a chosen unit vector $\mathbf{e}$,

$$e_{\langle L\rangle} n^{\langle L\rangle} = \frac{\ell!}{(2\ell-1)!!} P_\ell(\cos\theta)\,, \tag{31}$$

where $\cos\theta = \mathbf{e}\cdot\mathbf{n}$. In addition,

$$n^{\langle L\rangle} := N_\ell \sum_{m=-\ell}^{\ell} \mathcal{Y}^{\langle L\rangle}_{\ell m} Y_{\ell m}(\theta,\phi)\,, \qquad N_\ell = \frac{4\pi\,\ell!}{(2\ell+1)!!}\,, \tag{32}$$

where $\mathcal{Y}^{\langle L\rangle}_{\ell m}$ is a set of constant, complex STF tensors that satisfy $\mathcal{Y}^{\langle L\rangle}_{\ell,-m} = (-1)^m \mathcal{Y}^{*\langle L\rangle}_{\ell m}$. For $\ell = 1$, for example, Eq. (32) embodies the well-known relation between $(x \pm iy)/r$ and z/r and the ℓ=1 spherical harmonics.

There will be occasions when we need to calculate the average of a quantity $\psi(\theta, \phi)$ over the surface of a sphere:

$$\langle\langle \psi \rangle\rangle := \frac{1}{4\pi} \int \psi(\theta, \phi)\, d\Omega\,. \tag{33}$$

Of particular interest are the spherical average of products $n^j n^k n^n \cdots$ of radial vectors. These are easily computed using the fact that the average of the STF tensor $n^{\langle jkn\cdots\rangle}$ must be zero; this property follows directly from Eq. (32) and the identity $\int Y_{\ell m}(\theta, \phi)\, d\Omega = 0$ (unless $\ell = 0$). In this way we obtain, for example,

$$\langle\langle n^j n^k \rangle\rangle = \frac{1}{3}\delta^{jk}\,, \tag{34a}$$

$$\langle\langle n^j n^k n^n n^p \rangle\rangle = \frac{1}{15}\left(\delta^{jk}\delta^{np} + \delta^{jn}\delta^{kp} + \delta^{jp}\delta^{kn}\right), \tag{34b}$$

and so on. The general expression for such angular averages can be shown to be given by

$$\langle\langle n^L \rangle\rangle = \frac{1}{(\ell+1)!!}\Big[\delta^L + \mathrm{sym}(q)\Big], \tag{35}$$

when ℓ is an even number, and $\langle\langle n^L \rangle\rangle = 0$ when ℓ is odd; we use the same notation as in Eq. (29), in which δ^L stands for a product of $\ell/2$ Kronecker deltas, and sym(q) denotes all distinct terms obtained by permuting indices; the total number of terms within the square brackets is equal to $q = (\ell - 1)!!$.

2.3 *Motion of Extended Fluid Bodies*

We consider a fluid system that is broken up into a number N of separated and isolated bodies, with the size of each body R assumed to be small compared to the interbody distance r. Each body is assigned a label $A = 1, 2, \cdots, N$, and each body occupies a volume V_A bounded by a closed surface S_A. The mass density ρ is assumed to be equal to ρ_A inside V_A, and zero in the vacuum region between bodies. The fluid dynamics inside each body is governed by the continuity and Euler equations (4) and (5) and the gravitational potential U is given everywhere by Eq. (9).

We define the total mass of body A along with its center-of-mass position, velocity, and acceleration by

$$m_A := \int_A \rho(t, \mathbf{x})\, d^3x\,, \tag{36a}$$

$$\mathbf{r}_A(t) := \frac{1}{m_A}\int_A \rho(t,\mathbf{x})\,\mathbf{x}\,d^3x\,, \tag{36b}$$

$$\mathbf{v}_A(t) := \frac{1}{m_A}\int_A \rho(t,\mathbf{x})\,\mathbf{v}\,d^3x\,, \tag{36c}$$

$$\mathbf{a}_A(t) := \frac{1}{m_A}\int_A \rho(t,\mathbf{x})\,\frac{d\mathbf{v}}{dt}\,d^3x\,, \tag{36d}$$

where the domain of integration is a region of space that extends slightly beyond the volume V_A; it is sufficiently small that it contains no other body, but sufficiently large that it continues to contain body A as it moves about in a small interval of time dt. It is easy to show that m_A is time-independent: $dm_A/dt = 0$ and that

$$\mathbf{v}_A = \frac{d\mathbf{r}_A}{dt}\,, \qquad \mathbf{a}_A = \frac{d\mathbf{v}_A}{dt}\,. \tag{37}$$

In addition to these variables, we introduce

$$I_A^{\langle L\rangle}(t) := \int_A \rho(t,\mathbf{x})\,(x - r_A)^{\langle L\rangle}\,d^3x\,, \tag{38}$$

the STF multipole moments of body A, which refer to its center-of-mass position $\mathbf{r}_A(t)$; notice that the dipole moment $I_A^j = \int_A \rho(x-r_A)^j\,d^3x$ vanishes automatically.

Integrating the Euler equation (5) over body A, and noting that $\int_A \nabla p d^3x = 0$ and $\int_A \rho\nabla U_A d^3x = 0$, we obtain

$$m_A\mathbf{a}_A = \sum_{B\neq A}\int_A \rho(t,\mathbf{x})\int_B \frac{\rho(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d^3x\,d^3x'\,. \tag{39}$$

Defining

$$\begin{aligned}\mathbf{x} &= \mathbf{r}_A + \bar{\mathbf{x}}\,,\\ \mathbf{x}' &= \mathbf{r}_B + \bar{\mathbf{x}}'\,,\\ \mathbf{r}_{AB} &= \mathbf{r}_A - \mathbf{r}_B\,,\end{aligned} \tag{40}$$

and carrying out a Taylor expansion in both $\bar{x}$ and $\bar{x}'$, assuming both are small compared to r_{AB}, we obtain, after some rearrangement of terms,

$$\begin{aligned}a_A^j = G\sum_{B\neq A}\Bigg\{&-\frac{m_B n_{AB}^j}{r_{AB}^2}\\ &+\sum_{\ell=2}^{\infty}\frac{1}{\ell!}\Big[(-1)^\ell I_B^{\langle L\rangle} + \frac{m_B}{m_A}I_A^{\langle L\rangle}\Big]\partial_{jL}^A\left(\frac{1}{r_{AB}}\right)\\ &+\frac{1}{m_A}\sum_{\ell=2}^{\infty}\sum_{\ell'=2}^{\infty}\frac{(-1)^{\ell'}}{\ell!\ell'!}I_A^{\langle L\rangle}I_B^{\langle L'\rangle}\partial_{jLL'}^A\left(\frac{1}{r_{AB}}\right)\Bigg\}\,,\end{aligned} \tag{41}$$

where $\mathbf{n}_{AB} := \mathbf{r}_{AB}/r_{AB}$ is a unit vector that points from body B to body A. This expression implies that $\sum_A m_A \mathbf{a}_A = \mathbf{0}$, a statement that reflects Newton's third law or the conservation of total momentum.

The multipole moments of a perfectly spherical body vanish, $I_A^{\langle L\rangle} = 0$ for $\ell \neq 0$, and when all the bodies are spherical, Eq. (41) reduces to the familiar set of "point-mass" equations of motion,

$$a_A^j = -\sum_{B\neq A} \frac{G m_B n_{AB}^j}{r_{AB}^2}\,. \tag{42}$$

When the bodies are not spherical, the motion of body A is affected by the distortion of the gravitational potential caused by the deformation of the other bodies; this influence is described by the terms in Eq. (41) that involve $I_B^{\langle L\rangle}$. It is affected also by the nonspherical coupling of its own mass distribution to the monopole field of each external body; this influence is described by the terms in Eq. (41) that are linear in $I_A^{\langle L\rangle}$. And finally, it is affected by the coupling between its own multipole moments and those of the remaining bodies, as described by the last line in Eq. (41). This last effect is analogous to the dipole–dipole coupling in electrodynamics, except for the fact that there is no dipole moment in gravitation; the leading effect comes from a quadrupole–quadrupole interaction. The presence of terms involving $I_A^{\langle L\rangle}$ in the equations of motion imply that the motion of a body can depend on its internal structure, by virtue of its finite size and the nonspherical coupling of its mass distribution to the external gravitational field. This observation does not constitute a violation of the WEP; such a violation would imply a dependence on internal structure that would be present even when the bodies have a negligible size.

Because each multipole moment $I_A^{\langle L\rangle}$ scales as $m_A R_A^\ell$, each term in the sum scales as

$$\frac{G m_A m_B}{r_{AB}^2}\left(\frac{R_A}{r_{AB}}\right)^{\ell}\left(\frac{R_B}{r_{AB}}\right)^{\ell'}\,,$$

and the assumption that $R_A \ll r_{AB}$ implies that each term gets progressively smaller; the equation is exact, but it is most useful as a starting point for an approximation scheme. For many applications involving a small ratio R_A/r_{AB}, the sums can be safely truncated after just a few terms. For other applications, however, a large number of terms may be required. An example is the motion of a satellite in a low Earth orbit, which is sensitive to many of Earth's multipole moments; in geodesy projects such as the Gravity Recovery and Climate Experiment (GRACE) and the CHAllenging Mini-satellite Payload (CHAMP), Earth's multipole moments up to $\ell \sim 360$ have been measured.

We now specialize to a system of two bodies. The total mass of the binary system is $m := m_1 + m_2$. We define the *relative separation* $\mathbf{r} := \mathbf{r}_1 - \mathbf{r}_2$, and work in *barycentric* coordinates where

$$\begin{aligned}\mathbf{R} &:= (m_1/m)\,\mathbf{r}_1 + (m_2/m)\mathbf{r}_2 = 0\,,\\ \mathbf{r}_1 &= (m_2/m)\mathbf{r}\,,\end{aligned}$$

$$\mathbf{r}_2 = -(m_1/m)\mathbf{r}\,. \tag{43}$$

Defining $r := |\mathbf{r}|$, $\mathbf{n} := \mathbf{r}/r$, $\mathbf{v} := d\mathbf{r}/dt = \mathbf{v}_1 - \mathbf{v}_2$, and $\mathbf{a} := d\mathbf{v}/dt = \mathbf{a}_1 - \mathbf{a}_2$, we obtain the effective one-body equation of motion

$$\begin{aligned} a^j = &-Gm\frac{n^j}{r^2} + Gm\sum_{\ell=2}^{\infty}\frac{1}{\ell!}\left[\frac{I_1^{\langle L\rangle}}{m_1} + (-1)^\ell\frac{I_2^{\langle L\rangle}}{m_2}\right]\partial_{jL}\left(\frac{1}{r}\right) \\ &+ Gm\sum_{\ell=2}^{\infty}\sum_{\ell'=2}^{\infty}\frac{(-1)^{\ell'}}{\ell!\ell'!}\frac{I_1^{\langle L\rangle}}{m_1}\frac{I_2^{\langle L'\rangle}}{m_2}\partial_{jLL'}\left(\frac{1}{r}\right). \end{aligned} \tag{44}$$

Specializing further, we assume that only body 2 has significant multipole moments, and that it is symmetric about an axis aligned with the unit vector $\mathbf{e}$. It is straightforward to show that

$$I_2^{\langle L\rangle} = -m_2 R_2^\ell\,(J_\ell)_2\,e_2^{\langle L\rangle}\,, \tag{45}$$

where J_ℓ is a dimensionless multipole moment, defined by

$$J_\ell := -\sqrt{\frac{4\pi}{2\ell+1}}\frac{1}{MR^\ell}\mathcal{Y}_{\ell 0}^{\langle L\rangle}I_{\langle L\rangle}\,. \tag{46}$$

We inserted the label "2" on all quantities (such as mass, radius, symmetry axis, and multipole moments) to indicate that they belong to body 2. Equation (44) can then be written as

$$\mathbf{a} = \frac{Gm}{r^2}\left[-\mathbf{n} + \sum_{\ell=2}^{\infty}(J_\ell)_2\left(\frac{R_2}{r}\right)^\ell\left(\mathbf{n}\frac{dP_{\ell+1}}{d\mu} - \mathbf{e}_2\frac{dP_\ell}{d\mu}\right)\right], \tag{47}$$

where $P_\ell(\mu)$ is the Legendre polynomial of $\mu = \mathbf{n}\cdot\mathbf{e}_2$. Note that the acceleration has a radial piece $\propto \mathbf{n}$ and a piece parallel to the companion's symmetry axis $\mathbf{e}_2$.

2.4 Newtonian Orbital Dynamics

Kepler's problem is the determination of the motion of two bodies subjected to their mutual gravitational attraction, under the assumption that each body can be taken to be spherically symmetric. This is the simplest problem of celestial mechanics, but also one of the most relevant, because the motion of any planet around the Sun can, to a good first approximation, be calculated while ignoring the effects of the other planets. It is also a problem that can be solved exactly and completely, in terms of simple equations. In this case, dropping all terms involving multipole moments in Eq. (44), we obtain

$$\mathbf{a} = -\frac{Gm}{r^2}\mathbf{n}\,. \tag{48}$$

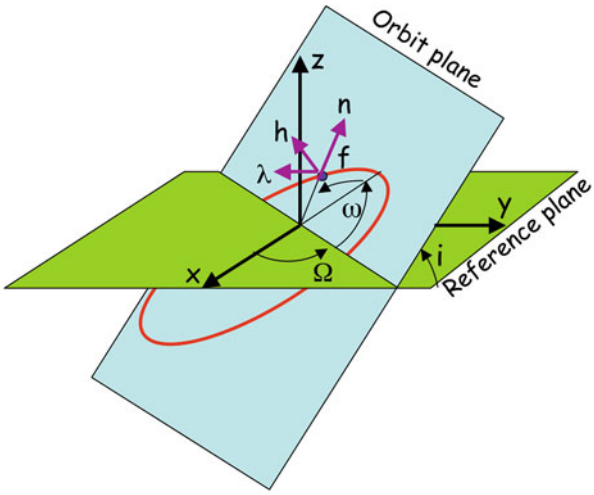


Fig. 1 Keplerian orbit in space

Because the potential is static and spherically symmetric, the equation of motion admits a conserved energy and angular momentum, given by

$$E = \frac{1}{2}\mu v^2 - \frac{G\mu m}{r}\,,$$
$$\mathbf{L} = \mu \mathbf{r} \times \mathbf{v}\,, \tag{49}$$

where $\mu := m_1 m_2/(m_1 + m_2)$ is the *reduced mass*. The constancy of $\mathbf{L}$ implies that the motion lies in a plane perpendicular to $\mathbf{L}$.

In order to write down the general solution of the Kepler problem, we introduce coordinates (X, Y, Z), with the $X - Y$ plane called the reference plane. This is arbitrary, and is often dictated by convention or convenience. For example, in the description of planetary motion in the solar system, the reference plane is chosen to coincide with Earth's own orbital plane (called the *ecliptic*); in the description of satellites orbiting the Earth, the reference plane is Earth's equatorial plane; for binary star systems, the reference plane is the plane of the sky. In each case, the direction of the X-axis is selected by convention.

The situation is represented in Fig. 1, which shows the orbit plane crossing the reference plane at an angle ι called the *inclination*; this is the angle between the z-direction of the orbit frame and the Z-direction of the reference frame. The line of intersection between the two planes is known as the *line of nodes*, and the point at which the orbit cuts the reference plane from below is the *ascending node*. The angle Ω between the X-direction and the line of nodes is the *longitude of the ascending node*. The diagram also shows ω, the *longitude of pericenter*, defined as the angle between the line of nodes and the direction to the pericenter, the point of minimum r, as measured in the orbit plane.

Elementary methods show that the solution to the Keplerian two-body problem is given by

$$\mathbf{r} = r\mathbf{n}\,, \tag{50}$$

where

$$r = \frac{p}{1 + e\cos f}\,, \tag{51}$$

where e is the eccentricity, and p is the semi-latus rectum, related to the angular momentum per unit reduced mass h by $h := |\mathbf{r} \times \mathbf{v}| = \sqrt{Gmp}$. The radial unit vector $\mathbf{n}$, along with the unit vector $\hat{\mathbf{h}} := \mathbf{h}/h$ normal to the orbit plane, and the third vector λ perpendicular to them both, are given by

$$\begin{aligned}
\mathbf{n} = &\left[\cos\Omega\cos(\omega + f) - \cos\iota\sin\Omega\sin(\omega + f)\right]\mathbf{e}_X \\
&+ \left[\sin\Omega\cos(\omega + f) + \cos\iota\cos\Omega\sin(\omega + f)\right]\mathbf{e}_Y \\
&+ \sin\iota\sin(\omega + f)\,\mathbf{e}_Z\,, \qquad (52a)\\
\lambda = &\left[-\cos\Omega\sin(\omega + f) - \cos\iota\sin\Omega\cos(\omega + f)\right]\mathbf{e}_X \\
&+ \left[-\sin\Omega\sin(\omega + f) + \cos\iota\cos\Omega\cos(\omega + f)\right]\mathbf{e}_Y \\
&+ \sin\iota\cos(\omega + f)\,\mathbf{e}_Z\,, \qquad (52b)\\
\mathbf{h} = &\sin\iota\sin\Omega\,\mathbf{e}_X - \sin\iota\cos\Omega\,\mathbf{e}_Y + \cos\iota\,\mathbf{e}_Z\,. \qquad (52c)
\end{aligned}$$

The relation between the angle f, known as the true anomaly, and time is given by integrating the equation

$$\frac{df}{dt} = \frac{h}{r^2}\,. \tag{53}$$

The total listing of orbital elements therefore consists of the size and shape elements p and e, the orientational elements ι, Ω, and ω, and the time element T arising from the constant of integration from Eq. (53); the total number of elements is six, corresponding precisely to the number of initial conditions ($\mathbf{x}(0)$, $\mathbf{v}(0)$) required to select a unique solution to Kepler's problem.

The semimajor axis a is defined by

$$a := \frac{1}{2}(r_{\min} + r_{\max}) = \frac{p}{1 - e^2}\,, \tag{54}$$

and the orbital period P is given by

$$P = 2\pi a^{3/2}(Gm)^{-1/2}\,. \tag{55}$$

Other useful orbital quantities expressed in this language include

$$\begin{aligned}
\mathbf{v} &= \dot{r}\mathbf{n} + (h/r)\lambda\,,\\
\dot{r} &= (he/p)\sin f\,,\\
E &= -G\mu m/2a\,. \qquad (56)
\end{aligned}$$

The constancy of p is directly tied to conservation of angular momentum h, and the constancy of $a := p/(1 - e^2)$ is tied to conservation of energy (so that constancy of e is also assured). In addition, the appearance of T as an integration constant was expected from the fact that the gravitational potential Gm/r does not depend

explicitly on time. The constancy of ι and Ω is related to the spherical symmetry of the problem, which fixed the direction of $\mathbf{L}$ and the orbital plane.

The constancy of ω is a very important property of a Keplerian orbit: it ensures that the orientation of the orbit within its plane stays fixed, that the position of the pericenter does not move, and when the orbit is bound, that the orbit retraces itself after each orbital cycle. The constancy of ω is the result of a hidden, dynamical symmetry of Kepler's problem, associated with the specific $1/r$ nature of the gravitational potential; the symmetry does not exist for other potentials. The symmetry gives rise to a conservation statement for the *Runge–Lenz vector*, defined by

$$\mathbf{A} := \frac{\mathbf{v} \times \mathbf{h}}{Gm} - \mathbf{n}\,. \tag{57}$$

A short computation using $\mathbf{a} = -Gm\mathbf{n}/r^2$ shows that

$$\frac{d\mathbf{A}}{dt} = \mathbf{0}, \tag{58}$$

and the manipulations do indeed reveal that constancy of $\mathbf{A}$ relies on the specific form of the gravitational acceleration. The Runge–Lenz vector can be evaluated explicitly by making use of the Keplerian results obtained previously. The result is

$$\mathbf{A} = e\big(\cos\omega\,\mathbf{e}_x + \sin\omega\,\mathbf{e}_y\big) := e\mathbf{e}_P\,, \tag{59}$$

where $\mathbf{e}_x$ points along the line of nodes, and $\mathbf{e}_y$ and $\hat{\mathbf{h}}$ complete the basis vectors for the orbital plane; it reveals that the vector $\mathbf{A}$ points in the fixed direction of the pericenter $\mathbf{e}_P$. The vector has a length e, and constancy of $\mathbf{A}$ implies that ω is a constant of the motion. This same dynamical symmetry is also responsible for the high degree of degeneracy in the quantum-mechanical energy levels of the nonrelativistic hydrogen atom. However, any deviation from a pure $1/r$ potential will generically cause a variation in the angle ω and a splitting of energy levels in hydrogen. There is also a similar dynamical symmetry for the spherical harmonic oscillator potential $\propto r^2$, both classical and quantum.

Five of the orbital elements that we have described can be expressed directly in terms of the variables $\mathbf{x}$ and $\mathbf{v}$ of the orbit. A simple calculation shows

$$p = \frac{h^2}{Gm}\,, \tag{60a}$$

$$e = |\mathbf{A}|\,, \tag{60b}$$

$$\cos\iota = = \frac{\mathbf{h}\cdot\mathbf{e}_Z}{h}\,, \tag{60c}$$

$$\sin\iota\sin\Omega = \frac{\mathbf{h}\cdot\mathbf{e}_X}{h}\,, \tag{60d}$$

$$\sin\iota\sin\omega = \frac{\mathbf{A}\cdot\mathbf{e}_Z}{e}\,. \tag{60e}$$

2.5 Osculating Orbit Elements and the Perturbed Kepler Problem

In the real world, the Kepler two-body problem does not apply because one or more of the bodies may not be spherically symmetric, because of the presence of other bodies or mass in the system, or because of relativistic effects. Let us assume that the relative acceleration $\mathbf{a} := \mathbf{a}_1 - \mathbf{a}_2$ between two bodies is given by

$$\mathbf{a} = -\frac{Gm}{r^2}\mathbf{n} + \mathbf{f}\,, \tag{61}$$

where $\mathbf{f}$ is a perturbing force per unit mass, which may depend on $\mathbf{r}$, $\mathbf{v}$, and time. The solution of this equation is no longer a conic section of the Kepler problem. However, whatever the solution is, at any given time t_0, for $\mathbf{r}(t_0)$, $\mathbf{v}(t_0)$, there exists a Keplerian orbit with orbit elements e_0, a_0, ω_0, Ω_0, ι_0, and T_0 that corresponds to those values, as we constructed in the previous section. In other words, there is a Keplerian orbit that is *tangent* to the orbit in question at the time t_0, commonly called the *osculating orbit*.

However, because of the perturbing acceleration, at a later time, the orbit will not be the same Keplerian orbit, but will be tangent to a new osculating orbit, with new elements e', a', and so on. The idea then is to study a general orbit with the perturbing acceleration $\mathbf{f}$ by finding the sequence of osculating orbits parametrized by $e(t)$, $a(t)$, and so on. If the perturbing acceleration is small in a suitable sense, then since the orbit elements of the original Kepler motion are constants, we might hope that the osculating orbit elements will vary slowly with time.

Mathematically, this approach is identical to the method of "variation of parameters" used to solve differential equations, such as the harmonic oscillator with a slowly varying frequency.

In this case, we replace our Keplerian solution for the motion with the following *definitions*:

$$\mathbf{r} := r\mathbf{n}\,, \tag{62a}$$

$$r := \frac{p}{1 + e\cos f}\,, \tag{62b}$$

$$\mathbf{v} := \frac{he\sin f}{p}\mathbf{n} + \frac{h}{r}\lambda\,, \tag{62c}$$

$$p := a(1 - e^2)\,, \tag{62d}$$

$$h^2 := Gmp\,, \tag{62e}$$

where the unit vectors $\mathbf{n}$ and λ are given by Eq. (52).

We first decompose $\mathbf{f}$ as

$$\mathbf{f} = \mathcal{R}\,\mathbf{n} + \mathcal{S}\,\lambda + \mathcal{W}\,\hat{\mathbf{h}}\,, \tag{63}$$

in terms of components $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{W}$.

The effect of the perturbing acceleration $\mathbf{f}$ on the vectors $\mathbf{h}$ and $\mathbf{A}$ can be calculated by appealing directly to their definitions. We find that

$$\frac{d\mathbf{h}}{dt} = \mathbf{r} \times \mathbf{f} = -r\mathcal{W}\,\boldsymbol{\lambda} + r\mathcal{S}\,\hat{\mathbf{h}}\,, \tag{64}$$

and

$$Gm\frac{d\mathbf{A}}{dt} = \mathbf{f} \times \mathbf{h} + \mathbf{v} \times (\mathbf{r} \times \mathbf{f}) = 2h\mathcal{S}\,\mathbf{n} - \left(h\mathcal{R} + r\dot{r}\mathcal{S}\right)\boldsymbol{\lambda} - r\dot{r}\mathcal{W}\,\hat{\mathbf{h}}\,. \tag{65}$$

These equations imply that

$$\frac{dh}{dt} = r\mathcal{S}\,, \tag{66}$$

and therefore, that $\mathcal{S}$ produces a change in the magnitude of the angular-momentum vector, while $\mathcal{W}$ produces a change in its direction. Similarly, both $\mathcal{R}$ and $\mathcal{S}$ produce a change in $\mathbf{A}$'s magnitude, as well as a change of direction orthogonal to $\hat{\mathbf{h}}$.

We can now systematically develop equations for the variations with time of the osculating orbit elements. For example, since $\mathbf{h} \cdot \mathbf{e}_Z = h\cos\iota$, then $\dot{\mathbf{h}} \cdot \mathbf{e}_Z = \dot{h}\cos\iota - \sin\iota(d\iota/dt) = r\mathcal{S}\cos\iota - r\mathcal{W}\cos(\omega + f)\sin\iota$, with the result that $d\iota/dt = (r\mathcal{W}/h)\cos(\omega + f)$. Similarly, since $\mathbf{h} \cdot \mathbf{e}_Y = -h\sin\iota\cos\Omega$, then taking the derivative of both sides and substituting our previous results for $\dot{\mathbf{h}}$, $\dot{h}$ and $d\iota/dt$, we obtain $\sin\iota\dot{\Omega} = (r\mathcal{W}/h)\sin(\omega + f)$. To obtain $\dot{e}$, we note that $e\dot{e} = \mathbf{A} \cdot \dot{\mathbf{A}}$, and use the fact that $\mathbf{A} = \mathbf{n}\cos f - \boldsymbol{\lambda}\sin f$. For $\dot{a}$, we use the definition $h^2 = Gma(1-e^2)$, from which $\dot{a}/a = 2\dot{h}/h + 2e\dot{e}/(1-e^2)$. For $\dot{\omega}$, we use the fact that $\mathbf{A} \cdot \mathbf{e}_Z = e\sin\iota\sin\omega$, combined with previous results for $\dot{e}$ and $d\iota/dt$. The final equations for the osculating orbit elements are

$$\frac{da}{dt} = \frac{2a^2}{h}\left(\mathcal{S}\frac{p}{r} + \mathcal{R}e\sin f\right)\,, \tag{67a}$$

$$\frac{de}{dt} = \frac{1-e^2}{h}\left(\mathcal{R}a\sin f + \frac{\mathcal{S}}{er}(ap - r^2)\right)\,, \tag{67b}$$

$$\frac{d\omega}{dt} = -\mathcal{R}\frac{p}{eh}\cos f + \mathcal{S}\frac{p+r}{eh}\sin f - \mathcal{W}\frac{r}{h}\cot\iota\sin(\omega + f)\,, \tag{67c}$$

$$\sin\iota\frac{d\Omega}{dt} = \mathcal{W}\frac{r}{h}\sin(\omega + f)\,, \tag{67d}$$

$$\frac{d\iota}{dt} = \mathcal{W}\frac{r}{h}\cos(\omega + f)\,. \tag{67e}$$

Notice that the orbit elements a and e are affected only by components of $\mathbf{f}$ in the plane of the orbit, while the elements Ω and ι are affected only by the component out of the plane. The pericenter change has both, but this is because of the combination of intrinsic, in-plane perturbations (the first two terms) with the perturbation of the line of nodes from which ω is measured (the third term).

We are missing an equation for the variation of T, the time of pericenter passage, which determines the true anomaly f via Eq. (53). If one is engaged in real-life orbit

determinations, such as navigating a spacecraft, timing a binary pulsar's orbit, or developing ephemerides, then the variation of T is an important quantity. However, if one is interested in elucidating physical effects on an orbit, it is more practical to close the system of equations by providing an expression for df/dt, from which the true anomaly can be obtained directly. Because f is the angle between the (varying) pericenter and the position vector $\mathbf{n}$, we have that $\cos f = \mathbf{n} \cdot \mathbf{e}_P$, and this can immediately be differentiated with respect to time. After some simplification, we find

$$\frac{df}{dt} = \frac{h}{r^2} - \left(\frac{d\omega}{dt} + \cos\iota \frac{d\Omega}{dt} \right), \tag{68}$$

which shows that df/dt differs from the usual Keplerian expression by a term $d\omega/dt + \cos\iota\, d\Omega/dt$ which possesses a direct geometrical meaning. We recall that ω is the angle from the (varying) pericenter to the (varying) line of nodes, as measured in the orbital plane, while Ω is the angle from the line of nodes to the (fixed) X-direction, as measured in the reference X-Y plane. The combination $d\omega + \cos\iota\, d\Omega$ can then be seen to describe the change in the direction to the pericenter relative to the X-direction, as measured entirely in the orbital plane. The non-Keplerian terms in Eq. (68) therefore appear because the true anomaly f is now measured relative to a varying set of directions. In fact, it is common to discuss the variations in the combined orbit "element"

$$d\varpi := d\omega + \cos\iota\, d\Omega\,. \tag{69}$$

The formalism of osculating orbital elements, in the formulation of Eqs. (67) and (68), is *exactly equivalent* to the original formulation of the equations of motion in Eq. (61); no approximations have been introduced in the transcription. The usefulness of the formalism, however, is most immediate when the perturbing force is small, so that the changes in the orbital elements are small. In such a context, one can achieve a very good approximation of the orbital dynamics by inserting the constant, zeroth-order values on the right-hand side of the equations, and integrating with respect to t to get the first-order changes. In such applications, it is usually convenient to use f as an independent variable instead of t, and in this approximate context, one can neglect the non-Keplerian terms on the right-hand side of Eq. (68). The system of osculating equations become

$$\frac{dp}{df} \simeq 2\frac{p^3}{Gm}\frac{1}{(1+e\cos f)^3}\,\mathcal{S}, \tag{70a}$$

$$\frac{de}{df} \simeq \frac{p^2}{Gm}\left[\frac{\sin f}{(1+e\cos f)^2}\,\mathcal{R} + \frac{2\cos f + e(1+\cos^2 f)}{(1+e\cos f)^3}\,\mathcal{S}\right], \tag{70b}$$

$$\frac{d\iota}{df} \simeq \frac{p^2}{Gm}\frac{\cos(\omega+f)}{(1+e\cos f)^3}\,\mathcal{W}, \tag{70c}$$

$$\sin\iota\frac{d\Omega}{df} \simeq \frac{p^2}{Gm}\frac{\sin(\omega+f)}{(1+e\cos f)^3}\,\mathcal{W}, \tag{70d}$$

$$\frac{d\omega}{df} \simeq \frac{1}{e}\frac{p^2}{Gm}\left[-\frac{\cos f}{(1+e\cos f)^2}\,\mathcal{R} + \frac{2+e\cos f}{(1+e\cos f)^3}\sin f\,\mathcal{S} - e\cot\iota\,\frac{\sin(\omega+f)}{(1+e\cos f)^3}\,\mathcal{W}\right], \tag{70e}$$

with

$$\frac{dt}{df} \simeq \sqrt{\frac{p^3}{Gm}}\,\frac{1}{(1+e\cos f)^2} \tag{71}$$

providing the temporal information.

2.6 Non-Keplerian Behavior: Worked Examples

In this section, we work out some real-world examples to illustrate the use of osculating orbit elements and other non-Keplerian effects.

Perturbations by a Third Body

Consider a two-body system in the presence of a distant third body. We assume that the effect of the third body can be treated as a perturbation of the Keplerian two-body orbit of the primary bodies. From the N-body equation of motion (42), we write down the equation of motion of the two primary bodies,

$$\begin{aligned}\mathbf{a}_1 &= -G\frac{m_2\mathbf{r}_{12}}{r_{12}^3} - G\frac{m_3\mathbf{r}_{13}}{r_{13}^3}\,,\\ \mathbf{a}_2 &= G\frac{m_1\mathbf{r}_{12}}{r_{12}^3} - G\frac{m_3\mathbf{r}_{23}}{r_{23}^3}\,.\end{aligned} \tag{72}$$

We now assume that $r_{12} \ll r_{23}$, so that we can expand

$$\frac{r_{13}^i}{r_{13}^3} \approx \frac{r_{23}^i}{r_{23}^3} - \sum_{\ell=1}^{\infty}\frac{1}{\ell!}r_{12}^L\partial^{\langle iL\rangle}\left(\frac{1}{r_{23}}\right). \tag{73}$$

The relative acceleration $\mathbf{a} := \mathbf{a}_1 - \mathbf{a}_2$ can then be written

$$\mathbf{a} = -G\frac{m\mathbf{r}}{r^3} - G\frac{m_3}{R^3}\left[\mathbf{r} - 3\mathbf{N}(\mathbf{N}\cdot\mathbf{r})\right] + O(m_3 r^2/R^4)\,, \tag{74}$$

where we define $R := |\mathbf{r}_{23}|$, and $\mathbf{N} := \mathbf{r}_{23}/|\mathbf{r}_{23}|$, and where we have kept only the $\ell = 1$ term in the sequence. The first term on the right-hand side of Eq. (74) is the normal Newtonian acceleration for the effective one-body problem. As long as $(m_3/m)(r/R)^3 \ll 1$, we can treat the second term in Eq. (74) as a disturbing

acceleration of a Keplerian orbit. From Eq. (74), we obtain the components of the disturbing acceleration

$$\begin{aligned}\mathcal{R} &= -A[1-3(\mathbf{N}\cdot\mathbf{n})^2]\,,\\ \mathcal{S} &= 3A(\mathbf{N}\cdot\mathbf{n})(\mathbf{N}\cdot\lambda)\,,\\ \mathcal{W} &= 3A(\mathbf{N}\cdot\mathbf{n})(\mathbf{N}\cdot\hat{\mathbf{h}})\,,\end{aligned} \tag{75}$$

where $A = Gm_3 r/R^3$.

We now focus on a specific example: the orbit of a planet like Mercury, perturbed by an outer planet, such as Jupiter. Let Mercury be in an orbit with osculating orbit elements a, e, and ω; we will assume that the orbit lies on the reference plane (the ecliptic in this case), so that $i = 0$ and Ω is not defined and can be put equal to zero. The orbit is then described by Eqs. (62a–62e). For simplicity, assume that the perturbing planet is in a circular orbit in the same reference plane, with $R =$ constant, and

$$\mathbf{N} = \mathbf{e}_x \cos F + \mathbf{e}_y \sin F\,, \tag{76}$$

where F is its eccentric anomaly, with constant $dF/dt = (Gm/R^3)^{1/2}$. Using Eqs. (52a) and (52b), we show that $\mathbf{N}\cdot\mathbf{n} = \cos(f - F + \omega)$, $\mathbf{N}\cdot\lambda = -\sin(f - F + \omega)$, and $\mathbf{N}\cdot\hat{\mathbf{h}} = 0$. We are interested in the perihelion advance of our planet, so working to first order in small perturbations, we combine Eqs. (70d), (70e), and (69) to obtain

$$\begin{aligned}\frac{d\varpi}{df} &= -\frac{r^2}{h^2 e}\left[p\mathcal{R}\cos f - (p+r)\mathcal{S}\sin f\right]\\ &= \frac{Gm_3 p^4}{h^2 R^3 e}\left\{\frac{\cos f}{(1+e\cos f)^3}[1-3\cos^2(f-F+\omega)]\right.\\ &\quad\left. -3\frac{(2+e\cos f)\sin f}{(1+e\cos f)^4}\cos(f-F+\omega)\sin(f-F+\omega)\right\}.\end{aligned} \tag{77}$$

In first-order perturbation theory, we can set the orbit elements equal to constants on the right-hand side and then integrate with respect to f to obtain $\varpi(f)$. But to keep things simple, we will further assume that the perturbing planet is far enough away that it barely moves during one orbit of our Mercury, that is, that $dF/dt \ll df/dt$, which follows from $(m_3/m)(r/R)^3 \ll 1$. Hence we set $F =$ constant in (77). We will also expand in powers of the eccentricity e. The solution for $\varpi(f)$ will therefore be proportional to sines and cosines of multiples of f, indicating that the osculating perihelion angle moves back and forth during the orbit as a result of the external perturbation.

Notice that, in the expansion of Eq. (77) in eccentricity, the leading term will be proportional to $1/e$, which might be worrisome in the limit $e \to 0$. However, it can be shown that the actual orbital displacements resulting from a change in ϖ are proportional to $e\delta\varpi$, so those observable displacements are well-behaved in the small-eccentricity limit.

Table 1 Planetary contributions to Mercury's perihelion advance (in arcseconds per century)

Planet	Advance
Venus	277.8
Earth	90.0
Mars	2.5
Jupiter	153.6
Saturn	7.3
Total	531.2
Discrepancy	42.9
Modern measured value	42.98 ± 0.04
General relativity prediction	42.98

But an important question is, after one orbit of Mercury, does the osculating perihelion angle return to its original value, that is, is it purely periodic, or does it suffer a *secular* change over each orbit? To explore this, we integrate Eq. (77) over f from zero to 2π, with the result that the net change $\Delta\varpi$ over one orbit is given by

$$\Delta\varpi = 3\pi \frac{m_3}{m} \left(\frac{p}{R}\right)^3 \frac{1}{(1-e^2)^{5/2}} [5\cos^2(F-\omega) - 2]\,. \tag{78}$$

The secular perihelion advance itself varies as the distant planet orbits the Sun, but if we average over all positions of the distant planet using $\langle \cos^2(F-\omega)\rangle = 1/2$, we obtain

$$\Delta\varpi = \frac{3\pi}{2} \frac{m_3}{m} \left(\frac{a}{R}\right)^3 (1-e^2)^{1/2}\,. \tag{79}$$

Inserting standard orbit elements for Mercury and Jupiter, we see that $\Delta\varpi = 1.79 \times 10^{-6}$. The smallness of this effect is what justifies working only at first order in perturbation theory. It is customary to express the perihelion advance as a rate in units of arcseconds (as) per century, so using the fact that 1 radian $= 2.063 \times 10^5$ as, and that Mercury's orbital period is 0.24 years, we obtain $d\varpi/dt = 154$ as per century. This is very close to the accurately computed value shown in Table 1. However, for perturbations of Mercury's orbit by the Earth, we get 69.2 as per century, quite a bit off from the accurate value of 90. But in this case $r/R \approx 0.38$, so that it is necessary to include the higher-order terms from the expansion in Eq. (74). Also, whereas holding Jupiter's position fixed while integrating over Mercury's orbit and then averaging over Jupiter's position might have been a good approximation since Jupiter's period is 40 times longer than Mercury's, it is not a good approximation in the case of Earth, whose period is only four times longer. Nevertheless, the method of osculating orbit elements allows us to incorporate such details systematically in a straightforward manner, leading to the accurately calculated planetary perturbations listed in Table 1. Still, the total effect of the planetary perturbations on Mercury's orbit falls well short of the observed perihelion advance of 574 as per century. In a later section, we will return to this question once we have obtained the equations of motion including relativistic terms.

Orbits Around a Nonspherical Body

We now consider a two-body problem where one of the bodies has nonzero multipole moments, typically caused by rotation. This could mean, for example, Mercury orbiting the oblate, rotating Sun, a satellite orbiting the Earth with all its multipole moments, or a neutron star orbiting a massive, rotating normal star. In this case, we can use Eq. (47), here restricted to $\ell = 2$:

$$\mathbf{a} = -\frac{Gm}{r^2}\left[\mathbf{n} - \frac{3}{2}(J_2)_2\frac{R_2^2}{r^2}\left(5\mathbf{n}(\mathbf{e}\cdot\mathbf{n})^2 - 2\mathbf{e}(\mathbf{e}\cdot\mathbf{n}) - \mathbf{n}\right)\right], \tag{80}$$

where R_2, $(J_2)_2$, and $\mathbf{e}$ are the radius, dimensionless quadrupole moment and symmetry axis direction of the central body. Hence

$$\begin{aligned}\mathcal{R} &= \frac{1}{2}A[3(\mathbf{e}\cdot\mathbf{n})^2 - 1]\,,\\ \mathcal{S} &= -A(\mathbf{e}\cdot\lambda)(\mathbf{e}\cdot\mathbf{n})\,,\\ \mathcal{W} &= -A(\mathbf{e}\cdot\hat{\mathbf{h}})(\mathbf{e}\cdot\mathbf{n})\,,\end{aligned} \tag{81}$$

where $A = 3(Gm/r^2)(J_2)_2(R_2/r)^2$. We choose $\mathbf{e}$ to be along the z-axis, and consider an orbit with osculating elements a, e, ω, ι, and Ω. Then from Eq. (52), we have $\mathbf{e}\cdot\mathbf{n} = \sin\iota\,\sin(f+\omega)$, $\mathbf{e}\cdot\lambda = \sin\iota\,\cos(f+\omega)$, and $\mathbf{e}\cdot\hat{\mathbf{h}} = \cos\iota$. Working in first-order perturbation theory, we integrate Eqs. (70a–70e), over 2π and obtain the secular changes in the orbit elements, $\Delta a = \Delta e = \Delta\iota = 0$, while

$$\Delta\varpi = 3\pi(J_2)_2\left(\frac{R_2}{p}\right)^2\left(1 - \frac{3}{2}\sin^2\iota\right), \tag{82a}$$

$$\Delta\Omega = -3\pi(J_2)_2\left(\frac{R_2}{p}\right)^2\cos\iota\,. \tag{82b}$$

Notice that, even in the case of zero inclination, where $\mathbf{e}\cdot\mathbf{n} = 0$, and thus $\mathcal{S} = \mathcal{W} = 0$, the quadrupole moment still induces a perihelion advance because it modifies the radial dependence of the acceleration from a pure $1/r^2$ behavior. For inclined orbits, the deviation from spherical symmetry induces advances in both the perihelion and the node. In the latter case, the effect is merely a precession of the orbital angular momentum $\mathbf{h} = \mathbf{r}\times\mathbf{v}$ around the symmetry axis $\mathbf{e}$, as can be obtained directly from Eq. (80),

$$\frac{d\mathbf{h}}{dt} = -3\frac{Gm}{r^2}(J_2)_2\frac{R_2^2}{r^2}(\mathbf{r}\times\mathbf{e})(\mathbf{e}\cdot\mathbf{n})\,. \tag{83}$$

From this, we see that $\mathbf{e}\cdot d\mathbf{h}/dt = 0$, so that $\mathbf{h}$ precesses around the $\mathbf{e}$ axis, inducing a rotation of the node.

For Mercury, the perihelion advance induced by the Sun's J_2 is negligible. Inserting the relevant orbit elements and the value $(J_2)_\odot = 2.2\times10^{-7}$, we obtain $d\varpi/dt = 0.03$ as/century, below the experimental uncertainties (see Table 1). It

was not always so. In fact, it is extremely difficult to measure J_2 for the Sun because at planetary distances, its effects are too small to be measured. The best way is to send a spacecraft relatively close to the Sun, well inside the orbit of Mercury and to determine precisely how J_2 affects its orbit; despite a number of proposals for such a mission, none has come to pass to date.

During the 1960s, Robert Dicke and Mark Goldenberg, attempted to infer a value of J_2 by measuring the visual shape of the Sun. They claimed a J_2 of order of 2.5×10^{-5}, over 100 times larger than the number quoted above. Later observations of the visible shape of the Sun by Henry Hill and others, along with analyses to try to understand how temperature differences across the Sun might contaminate the observations of apparent shape, did not fully resolve this controversy. The resolution came with the advance of "helioseismology." This was the discovery that the Sun actually vibrates in a superposition of thousands of normal modes with an array of frequencies, as reflected in the behavior of Doppler-shifted solar spectral lines. The specific pattern of frequencies proves to depend on how much differential rotation is present in the Sun. Through a systematic program of ground-based and space-based observations of the Sun, it was possible to determine the angular velocity profile over much of the solar interior. The conclusion was that the interior does not rotate much faster than the surface, and the solar models built from that information resulted in the value $J_2 = 2.2 \times 10^{-7}$, which is more or less what one would infer from a Sun that rotates uniformly at its observed surface rate. The bottom line is that, as far as Mercury and general relativity are concerned, the solar quadrupole moment does not play a role.

However, J_2 for the Earth does play an important role in the motion of satellites. Using $(J_2)_\oplus = 1.08 \times 10^{-3}$, and assuming a satellite in a circular orbit with semi-major axis a, with orbital period $P = 2\pi(a^3/Gm)^{1/2} = 83.91(a/R)^{3/2}$ minutes, we can show that the secular rate of advance of the node is given by

$$\frac{d\Omega}{dt} = 3639\left(\frac{R}{a}\right)^{7/2}\cos\iota \quad {}^{\mathrm{o}}\mathrm{yr}^{-1}\,, \tag{84}$$

where R is the Earth's radius. Thus, for example, the orbit of the laser geodynamics satellite LAGEOS I, with $a = 1.93R_\oplus$ and $\iota = 109^{\mathrm{o}}.8$ will precess at a rate of -123° per year. Suppression of this enormous Newtonian effect in order to detect the possible effect of general relativistic frame-dragging (~ 30 milliarcseconds per year) is a major challenge. The chapter "Fundamental Physics with the LAGEOS Satellites" by R. Peron, addresses how this suppression has been carried out in analysing data from the LAGEOS I and II satellites.

3 General Relativity

3.1 Mathematics of Curved Space-Time

The foundation of general relativity is the space-time metric $g_{\alpha\beta}$, which connects the arbitrary coordinates $x^\alpha = (ct, x^j)$ used to label events in space-time to physical

measurements of time and length made by clocks and rods, via the invariant interval

$$ds^2 = g_{\alpha\beta}\, dx^\alpha dx^\beta . \tag{85}$$

Under a change of coordinates described by $x^\alpha = f^\alpha(x'^\mu)$, where f^α are functions of the new coordinates x'^μ, the coordinate displacements change according to

$$dx^\alpha = \frac{\partial f^\alpha}{\partial x'^\mu}\, dx'^\mu , \tag{86}$$

and the space-time interval becomes

$$ds^2 = g_{\alpha\beta} \frac{\partial f^\alpha}{\partial x'^\mu} \frac{\partial f^\beta}{\partial x'^\nu}\, dx'^\mu dx'^\nu . \tag{87}$$

This expression shows that the metric is replaced by

$$g'_{\mu\nu} = g_{\alpha\beta} \frac{\partial f^\alpha}{\partial x'^\mu} \frac{\partial f^\beta}{\partial x'^\nu} \tag{88}$$

in the new coordinate system, so that $ds^2 = g'_{\mu\nu}\, dx'^\mu dx'^\nu$. The coordinate displacements change, and the metric also changes, but ds^2 remains the same during a coordinate transformation. In this way, for example, the proper time $d\tau = \sqrt{-ds^2}/c$ between two events as measured by an atomic clock is the same, regardless of the coordinates used to label the space-time events.

It can be shown that, in the neighborhood of any event $\mathcal{P}$ in space-time (and more generally along any world line), one can always find a coordinate system in which

$$g'_{\mu\nu}|_{\mathcal{P}} = \eta_{\mu\nu} , \qquad (\partial'_\gamma g'_{\mu\nu})|_{\mathcal{P}} = 0 , \tag{89}$$

where $\eta_{\mu\nu} = \mathrm{diag}(-1,\ 1,\ 1,\ 1)$ is the Minkowski metric of special relativity. This corresponds to the local freely falling frame, in which freely moving particles move on straight lines and the laws of nongravitational physics take on their standard special-relativistic, Lorentz invariant forms. This embodies the Einstein equivalence principle, the foundation of the geometrical formulation of gravity.

A coordinate system x^α can be used to define a set of basis vectors $\mathbf{e}_\alpha$, such that a vector describing the displacement from one event in space-time to a neighboring event is given by

$$d\mathbf{x} = \mathbf{e}_\alpha\, dx^\alpha . \tag{90}$$

An arbitrary vector $\mathbf{A}$ can then be expressed as

$$\mathbf{A} = A^\alpha \mathbf{e}_\alpha \tag{91}$$

in terms of the basis vectors and its components A^α.

The inner product between $d\mathbf{x}$ and itself is the space-time invariant ds^2. We have $ds^2 = d\mathbf{x} \cdot d\mathbf{x} = (\mathbf{e}_\alpha \cdot \mathbf{e}_\beta)\, dx^\alpha dx^\beta$, and comparing this with Eq. (85) reveals that

$$g_{\alpha\beta} = \mathbf{e}_\alpha \cdot \mathbf{e}_\beta . \tag{92}$$

If $\mathbf{A}$ is a vector field in space-time, its derivative with respect to coordinate x^β is

$$\partial_\beta \mathbf{A} = (\partial_\beta A^\alpha)\mathbf{e}_\alpha + A^\alpha(\partial_\beta \mathbf{e}_\alpha), \tag{93}$$

in which the first term accounts for the variation of the components, while the second accounts for the variation of the basis vectors. We define

$$\partial_\beta \mathbf{e}_\alpha := \Gamma^\mu_{\alpha\beta}\, \mathbf{e}_\mu, \tag{94}$$

which states the obvious fact that a change in basis vector induced by a coordinate displacement is itself a vector that can be decomposed in terms of basis vectors. Equation (94) provides a definition for the quantities $\Gamma^\mu_{\alpha\beta}$, which are known as *Christoffel symbols*. Because $\partial_\beta \mathbf{e}_\alpha = \partial^2\mathbf{x}/\partial x^\alpha \partial x^\beta = \partial_\alpha \mathbf{e}_\beta$, we have that

$$\Gamma^\mu_{\beta\alpha} = \Gamma^\mu_{\alpha\beta}; \tag{95}$$

the Christoffel symbols are symmetric in their lower indices.

With Eq. (94), our previous expression for $\partial_\beta \mathbf{A}$ becomes $\partial_\beta \mathbf{A} = (\nabla_\beta A^\mu)\mathbf{e}_\mu$, where

$$\nabla_\beta A^\mu := \partial_\beta A^\mu + \Gamma^\mu_{\alpha\beta} A^\alpha \tag{96}$$

is known as the *covariant derivative* of the vector components A^μ.

Differentiating the relation $g_{\alpha\beta} = \mathbf{e}_\alpha \cdot \mathbf{e}_\beta$ with respect to x^γ, substituting Eq. (94) and doing some further algebra, we can show that

$$\Gamma^\mu_{\alpha\beta} = \frac{1}{2} g^{\mu\nu}\left(\partial_\alpha g_{\nu\beta} + \partial_\beta g_{\nu\alpha} - \partial_\nu g_{\alpha\beta}\right). \tag{97}$$

The generalization of the covariant derivative to a tensor is:

$$\nabla_\beta A^{\mu\nu} = \partial_\beta A^{\mu\nu} + \Gamma^\mu_{\alpha\beta} A^{\alpha\nu} + \Gamma^\nu_{\alpha\beta} A^{\mu\alpha}\,. \tag{98}$$

This rule can easily be extended to tensors with an arbitrary number of indices; there is one Christoffel symbol per tensorial index.

Defining $A_\mu := g_{\mu\nu} A^\nu$, we can "lower" indices on vectors and tensors; using the inverse of the metric, identified as $g^{\mu\nu}$, we can raise indices. Then we can show that

$$\nabla_\beta A_\mu = \partial_\beta A_\mu - \Gamma^\alpha_{\mu\beta} A_\alpha\,, \tag{99}$$

and furthermore that

$$\nabla_\gamma g_{\alpha\beta} = 0\,. \tag{100}$$

We now examine a timelike world line $x^\alpha = r^\alpha(\tau)$ in a curved space-time, parameterized by proper time τ, the time measured by a clock moving on this world line. The world line's tangent vector is $u^\alpha = dr^\alpha/d\tau$, and this satisfies the normalization condition

$$g_{\alpha\beta} u^\alpha u^\beta = -c^2\,. \tag{101}$$

On this world line, let there exist a vector field $\mathbf{A}(\tau)$; we wish to evaluate the derivative of this vector with respect to τ:

$$\frac{d\mathbf{A}}{d\tau} = \left(\frac{dA^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} A^\alpha u^\beta\right)\mathbf{e}_\mu \,. \tag{102}$$

The quantity within brackets is the covariant derivative of the component A^μ along the world line. We denote this

$$\frac{DA^\mu}{d\tau} := \frac{dA^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} A^\alpha u^\beta \,, \tag{103}$$

so that $d\mathbf{A}/d\tau = (DA^\mu/d\tau)\mathbf{e}_\mu$. When the vector field $\mathbf{A}$ is defined also in a neighborhood of the world line (and not just directly on the world line), it becomes a function of all the space-time coordinates x^α (instead of just proper time τ); then $dA^\mu/d\tau = u^\beta \partial_\beta A^\mu$ and the covariant derivative can be expressed as $DA^\mu/d\tau = u^\beta \nabla_\beta A^\mu$.

The vector $\mathbf{A}$ is *parallel-transported* along the world line when it stays constant in both direction and magnitude. The mathematical statement of this is $d\mathbf{A}/d\tau = 0$, or

$$\frac{DA^\mu}{d\tau} = 0 \qquad \text{(parallel transport)}. \tag{104}$$

A timelike world line $r^\alpha(\tau)$ is a *geodesic* of the curved space-time when its own tangent vector $\mathbf{u}$ is parallel-transported along the world line. A geodesic, defined in this way, is everywhere *locally straight*. The mathematical statement of the geodesic equation is

$$\frac{Du^\mu}{d\tau} = 0 \,, \tag{105}$$

or

$$\frac{du^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} u^\alpha u^\beta = 0 \,, \tag{106}$$

or

$$\frac{d^2 r^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta} \frac{dr^\alpha}{d\tau}\frac{dr^\beta}{d\tau} = 0 \,. \tag{107}$$

This last form is a system of second-order differential equations for the functions $r^\mu(\tau)$. Given initial conditions $r^\mu(0)$ and $u^\mu(0)$ at some initial time $\tau = 0$, these equations admit a unique solution.

It is useful to note that the geodesic equation (107) can also be obtained on the basis of a variational principle, in which the action functional is the elapsed proper time $\int_1^2 d\tau$ along a parameterized curve $r^\alpha(\tau)$ linking the fixed events 1 and 2. If we write the particle action as

$$S = -mc^2 \int_1^2 d\tau = -mc \int_1^2 \sqrt{-g_{\alpha\beta}\frac{dr^\alpha}{dt}\frac{dr^\beta}{dt}}\, dt \,, \tag{108}$$

then its extremization with respect to world line variations returns the geodesic equation.

The symmetry of the Christoffel symbols in the lower indices implies that the action of two covariant derivatives on a scalar field f is independent of their order:

$$\nabla_\alpha \nabla_\beta f - \nabla_\beta \nabla_\alpha f = 0\,. \tag{109}$$

The same is not true, however, when the covariant derivatives act on a vector field A^μ; in this case

$$\nabla_\alpha \nabla_\beta A^\mu - \nabla_\beta \nabla_\alpha A^\mu = R^\mu_{\nu\alpha\beta} A^\nu\,, \tag{110}$$

and the operations do not commute. This equation defines the *Riemann curvature tensor* $R^\mu_{\nu\alpha\beta}$. A lengthy evaluation of the left-hand side of Eq. (110) shows that the Riemann tensor is given explicitly by

$$R^\alpha_{\beta\gamma\delta} = \partial_\gamma \Gamma^\alpha_{\beta\delta} - \partial_\delta \Gamma^\alpha_{\beta\gamma} + \Gamma^\alpha_{\mu\gamma}\Gamma^\mu_{\beta\delta} - \Gamma^\alpha_{\mu\delta}\Gamma^\mu_{\beta\gamma}\,. \tag{111}$$

The Riemann tensor is evidently antisymmetric in its last two indices. It also possesses additional symmetries that are not immediately revealed by Eqs. (110) and (111). The symmetries are

$$\begin{aligned} R_{\alpha\beta\delta\gamma} &= -R_{\alpha\beta\gamma\delta}\,, \\ R_{\beta\alpha\gamma\delta} &= -R_{\alpha\beta\gamma\delta}\,, \\ R_{\mu\alpha\beta\gamma} + R_{\mu\gamma\alpha\beta} + R_{\mu\beta\gamma\alpha} &= 0\,. \end{aligned} \tag{112}$$

An additional symmetry,

$$R_{\gamma\delta\alpha\beta} = +R_{\alpha\beta\gamma\delta}\,, \tag{113}$$

is an immediate consequence of the preceding ones. By virtue of these symmetries, the Riemann tensor possesses 20 independent components in a four-dimensional space-time.

Another important set of identities satisfied by the Riemann tensor is

$$\nabla_\alpha R_{\mu\nu\beta\gamma} + \nabla_\gamma R_{\mu\nu\alpha\beta} + \nabla_\beta R_{\mu\nu\gamma\alpha} = 0\,. \tag{114}$$

These are known as the *Bianchi identities*, and play a fundamental role in Einstein's general relativity.

The Riemann tensor is important as a measure of *geodesic deviation*. Given a family of nearby geodesics parametrized by τ and a deviation vector $\boldsymbol{\xi}$ that joins neighboring geodesics, one can show that the relative acceleration between geodesics is given by

$$\frac{D^2\xi^\alpha}{d\tau^2} = -R^\alpha_{\beta\gamma\delta} u^\beta \xi^\gamma u^\delta\,. \tag{115}$$

Other curvature tensors can be defined from the Riemann tensor. By contracting the first and third indices of the Riemann tensor, we obtain the *Ricci tensor*

$$R_{\alpha\beta} := R^\mu_{\alpha\mu\beta}\,. \tag{116}$$

The symmetries of the Riemann tensor imply that $R_{\beta\alpha} = R_{\alpha\beta}$, and the Ricci tensor possesses ten independent components. By contracting its indices, we obtain the *Ricci scalar*

$$R := R^{\mu}_{\ \mu} = g^{\alpha\beta} R_{\alpha\beta} \,. \tag{117}$$

Closely related to the Ricci tensor is the *Einstein tensor*

$$G_{\alpha\beta} := R_{\alpha\beta} - \frac{1}{2} g_{\alpha\beta} R \,, \tag{118}$$

which is also symmetric in its indices: $G_{\beta\alpha} = G_{\alpha\beta}$. The Einstein tensor possesses ten independent components, and its trace is given by $G := g^{\alpha\beta} G_{\alpha\beta} = -R$, because $g^{\alpha\beta} g_{\alpha\beta} = \delta^{\alpha}_{\ \alpha} = 4$.

The Bianchi identities of Eq. (114) give rise to

$$\nabla_\beta G^{\alpha\beta} = 0 \tag{119}$$

after two contractions of their indices. Equation (119) is known as the *contracted Bianchi identities*.

3.2 Physics in Curved Spacetime

The physical laws of the standard model of particles and fields are normally formulated in the Lorentz invariant language of special relativity. For example, Maxwell's equations have the form $\partial_\beta F^{\alpha\beta} = \mu_0 j_e^\alpha$, and $\partial_\alpha F_{\beta\gamma} + \partial_\gamma F_{\alpha\beta} + \partial_\beta F_{\gamma\alpha} = 0$, where $F^{\alpha\beta}$ is the antisymmetric field tensor, j_e^α is the electric current vector, and μ_0 is the permeability of vacuum. A freely moving neutral particle moves with unchanging four-velocity, in other words $du^\alpha/d\tau = u^\beta \partial_\beta u^\alpha = 0$.

Since all the fundamental laws of physics are derivable from an action, then, by virtue of the Lorentz invariance of the matter action, the laws imply a conservation statement for the energy-momentum tensor $T^{\alpha\beta} := -2(-\eta)^{-1/2} \delta \mathcal{L}_M / \delta \eta_{\alpha\beta}$ given by $\partial_\beta T^{\alpha\beta} = 0$.

The Einstein equivalence principle tells us that these laws are now to be regarded as being valid in a local freely falling frame, where the Minkowski metric $\eta_{\alpha\beta} = g_{\alpha\beta}|_{\mathcal{P}}$, is the transformed version of the space-time metric. To find the laws in a form valid in any frame or coordinate system, one only has to transform back. As a result of this transformation, $\eta_{\alpha\beta} \to g_{\alpha\beta}$, and $\partial_\alpha \to \nabla_\alpha$. Thus, in curved space-time, the laws mentioned above take the form

$$\begin{aligned}
\text{Maxwell's equations}: \quad & \nabla_\beta F^{\alpha\beta} = \mu_0 j_e^\alpha \,, \\
& \nabla_\alpha F_{\beta\gamma} + \nabla_\gamma F_{\alpha\beta} + \nabla_\beta F_{\gamma\alpha} = 0 \,, && \text{(120a)} \\
\text{Geodesic equation}: \quad & Du^\alpha/d\tau = u^\beta \nabla_\beta u^\alpha = 0 \,, && \text{(120b)} \\
\text{Energy} - \text{momentum conservation}: \quad & \nabla_\beta T^{\alpha\beta} = 0 \,, && \text{(120c)}
\end{aligned}$$

with $T^{\alpha\beta} := -2(-g)^{-1/2}\delta\mathcal{L}_M/\delta g_{\alpha\beta}$.

For most astrophysical applications, we will be more interested in a phenomenological description of hydrodynamics than in the underlying fundamental laws. In a manner similar to Newtonian theory, we define a perfect fluid with the proper mass density ρ, proper density of internal (thermodynamic) energy ε, proper density of total energy $\mu = \rho c^2 + \varepsilon$, and pressure p. The densities are all measured in a freely falling frame that is momentarily comoving with a selected fluid element; the four-velocity of this frame is denoted u^α, and the mass current is $j^\alpha = \rho u^\alpha$. Transforming from the local inertial frame, we can show that the fluid's energy-momentum tensor $T^{\alpha\beta}$, is given by

$$T^{\alpha\beta} = (\mu + p)u^\alpha u^\beta/c^2 + pg^{\alpha\beta}\,. \tag{121}$$

The statement of mass conservation now takes the form of

$$\nabla_\alpha j^\alpha = \frac{1}{\sqrt{-g}}\partial_\alpha\left(\sqrt{-g}\,j^\alpha\right) = 0\,. \tag{122}$$

Substitution of $j^\alpha = \rho u^\alpha$ into Eq. (122) yields

$$\frac{d\rho}{d\tau} + \rho\nabla_\alpha u^\alpha = 0\,. \tag{123}$$

Since $\nabla_\alpha u^\alpha = \mathcal{V}^{-1}d\mathcal{V}/d\tau$, this equation implies $d(\rho\mathcal{V})/d\tau = 0$ as expected. Alternatively, Eq. (122) implies that

$$\frac{\partial\rho^*}{\partial t} + \nabla\cdot(\rho^*\mathbf{v}) = 0\,, \tag{124}$$

where

$$\rho^* := \rho\sqrt{-g}u^0 \tag{125}$$

is sometimes called the *conserved density*. Because it satisfies a continuity equation exactly parallel to Eq. (4), the results gathered in Eqs. (11)–(14) now apply to the analogous integrals using ρ^*.

Substitution of Eq. (121) into Eq. (120c) produces

$$\frac{d\mu}{d\tau} + (\mu + p)\nabla_\beta u^\beta = 0 \tag{126}$$

and

$$(\mu + p)\frac{Du^\alpha}{d\tau} + c^2\left(g^{\alpha\beta} + u^\alpha u^\beta/c^2\right)\nabla_\beta p = 0\,. \tag{127}$$

The first equation gives rise to

$$\rho\frac{d\varepsilon}{d\tau} - (\varepsilon + p)\frac{d\rho}{d\tau} = 0\,, \tag{128}$$

which can be easily seen to be equivalent to the local first law of thermodynamics $d(\varepsilon\mathcal{V}) + pd\mathcal{V} = 0$. The second equation is the curved-space-time version of Euler's

equation. In the first term, we recognize $Du^\alpha/d\tau = u^\beta \nabla_\beta u^\alpha$ as the covariant acceleration of a selected fluid element, which would be zero if the fluid element were moving on a geodesic of the curved space-time. This, however, is prevented by the pressure forces acting within the fluid; the fluid element is not moving freely in the gravitational field.

3.3 Einstein Field Equations

The Einstein field equations relate the curvature of space-time to the distribution of matter within space-time. They read

$$G^{\alpha\beta} = \frac{8\pi G}{c^4} T^{\alpha\beta}, \tag{129}$$

with the Einstein curvature tensor of Eq. (118) on the left-hand side, and the total energy-momentum tensor of all forms of matter and non-gravitational fields on the right-hand side. Taking into account the symmetries of the Einstein and energy-momentum tensors, the Einstein field equations are a set of ten second-order, partial differential equations for the metric tensor $g_{\alpha\beta}$. The equations are all coupled, and they are highly nonlinear in the metric and its first derivatives; they are, however, linear in the second derivatives of the metric tensor.

A naive counting of the number of equations might suggest that given suitable initial and boundary conditions for the metric, the solution to the Einstein field equations should be unique. This suggestion, however, is false, as the freedom to perform coordinate transformations must be retained; two metrics $g_{\alpha\beta}$ and $g'_{\mu\nu}$ related to each other by a coordinate transformation should both be valid solutions to the field equations. This freedom is guaranteed by the contracted Bianchi identities,

$$\nabla_\beta G^{\alpha\beta} = 0\,, \tag{130}$$

which reveal that of the ten field equations, only six are truly independent from each other. The Bianchi identities, together with the field equations, are compatible with the local statement of energy-momentum conservation

$$\nabla_\beta T^{\alpha\beta} = 0\,. \tag{131}$$

The equations of general relativity can be derived from an action principle. The total action S involves a gravitational piece given by the Hilbert action

$$S_{\rm grav} = \frac{c^3}{16\pi G} \int R\, dV, \tag{132}$$

where R is the Ricci scalar and $dV = \sqrt{-g}\, d^4x$ is the invariant volume element, as well as a matter piece given by

$$S_M = \int \mathcal{L}_M\, dV, \tag{133}$$

where $\mathcal{L}_{\mathcal{M}}$ is the Lagrangian density for all the matter (and field) variables. The Einstein tensor results from the functional variation of the gravitational action:

$$\frac{\delta S_{\text{grav}}}{\delta g_{\alpha\beta}} = -\frac{c^3}{16\pi G} G^{\alpha\beta}. \tag{134}$$

The energy-momentum tensor, on the other hand, is defined by

$$\frac{\delta S_M}{\delta g_{\alpha\beta}} = \frac{1}{2c} T^{\alpha\beta}. \tag{135}$$

The Einstein field equations then follow from the requirement that $\delta(S_{\text{grav}} + S_M) = 0$ under an arbitrary variation of the metric tensor.

4 Post-Minkowskian and Post-Newtonian Theory

In this section, we specialize the formalism of general relativity to a description of weak gravitational fields, such as those inside and near the Sun, those inside and near white dwarfs, and those near (but not too near) neutron stars or black holes. This approximation should, of course, reproduce the predictions of Newtonian theory, but we wish to go beyond this and formulate a method of approximation that can be pushed systematically to higher and higher order and generate increasingly accurate descriptions of a weak gravitational field. The foundation for this approach is "post-Minkowskian theory."

4.1 Landau–Lifshitz Formulation of the Field Equations

The post-Minkowskian approach is based on the Landau and Lifshitz formulation of the Einstein equations. In this framework, the main variables are not the components of the metric tensor $g_{\alpha\beta}$, but those of the "gothic inverse metric density"

$$\mathfrak{g}^{\alpha\beta} := \sqrt{-g}\, g^{\alpha\beta}, \tag{136}$$

where $g^{\alpha\beta}$ is the inverse metric and g the metric determinant. Knowledge of the gothic metric is sufficient to determine the metric itself: note first that $\det[\mathfrak{g}^{\alpha\beta}] = g$, so that g can be directly obtained from the gothic metric; then Eq. (136) gives $g^{\alpha\beta}$, which can be inverted to give $g_{\alpha\beta}$.

In the Landau–Lifshitz formulation, the left-hand side of the field equations is built from

$$H^{\alpha\mu\beta\nu} := \mathfrak{g}^{\alpha\beta}\mathfrak{g}^{\mu\nu} - \mathfrak{g}^{\alpha\nu}\mathfrak{g}^{\beta\mu}. \tag{137}$$

This tensor density is readily seen to possess the same symmetries as the Riemann tensor,

$$H^{\mu\alpha\beta\nu} = -H^{\alpha\mu\beta\nu}, \qquad H^{\alpha\mu\nu\beta} = -H^{\alpha\mu\beta\nu}, \qquad H^{\beta\nu\alpha\mu} = H^{\alpha\mu\beta\nu}. \tag{138}$$

It turns out that $H^{\mu\alpha\beta\nu}$ satisfies the remarkable identity

$$\partial_{\mu\nu} H^{\alpha\mu\beta\nu} = (-g)\left(2G^{\alpha\beta} + \frac{16\pi G}{c^4} t_{\rm LL}^{\alpha\beta}\right), \tag{139}$$

where $G^{\alpha\beta}$ is the Einstein tensor, and

$$\begin{aligned}(-g)t_{\rm LL}^{\alpha\beta} := \frac{c^4}{16\pi G}\Big\{ &\partial_\lambda \mathfrak{g}^{\alpha\beta}\partial_\mu \mathfrak{g}^{\lambda\mu} - \partial_\lambda \mathfrak{g}^{\alpha\lambda}\partial_\mu \mathfrak{g}^{\beta\mu} + \frac{1}{2} g^{\alpha\beta} g_{\lambda\mu}\partial_\rho \mathfrak{g}^{\lambda\nu}\partial_\nu \mathfrak{g}^{\mu\rho} \\ &- g^{\alpha\lambda} g_{\mu\nu}\partial_\rho \mathfrak{g}^{\beta\nu}\partial_\lambda \mathfrak{g}^{\mu\rho} - g^{\beta\lambda} g_{\mu\nu}\partial_\rho \mathfrak{g}^{\alpha\nu}\partial_\lambda \mathfrak{g}^{\mu\rho} + g_{\lambda\mu} g^{\nu\rho}\partial_\nu \mathfrak{g}^{\alpha\lambda}\partial_\rho \mathfrak{g}^{\beta\mu} \\ &+ \frac{1}{8}\big(2g^{\alpha\lambda} g^{\beta\mu} - g^{\alpha\beta} g^{\lambda\mu}\big)\big(2g_{\nu\rho} g_{\sigma\tau} - g_{\rho\sigma} g_{\nu\tau}\big)\partial_\lambda \mathfrak{g}^{\nu\tau}\partial_\mu \mathfrak{g}^{\rho\sigma}\Big\}\end{aligned} \tag{140}$$

is the *Landau–Lifshitz pseudotensor*, which (very loosely speaking) represents the distribution of gravitational-field energy. It is called a pseudotensor because it contains ordinary derivatives of the metric, and does not by itself transform as a tensor under general coordinate transformations. The full Eq. (139), of course, is generally covariant, although it is not written in a manifestly covariant form. Equation (139) is valid for any space-time.

We now impose Einstein's equations, $G^{\alpha\beta} = 8\pi G T^{\alpha\beta}/c^4$, to obtain

$$\partial_{\mu\nu} H^{\alpha\mu\beta\nu} = \frac{16\pi G}{c^4}(-g)\big(T^{\alpha\beta} + t_{\rm LL}^{\alpha\beta}\big), \tag{141}$$

which is therefore equivalent to Einstein's equations, combining as it does Einstein's equations and an identity.

By virtue of the antisymmetry of $H^{\alpha\mu\beta\nu}$ in the last pair of indices, we have that the equation

$$\partial_{\beta\mu\nu} H^{\alpha\mu\beta\nu} = 0 \tag{142}$$

holds as a trivial identity. This, together with Eq. (141), implies that

$$\partial_\beta\Big[(-g)\big(T^{\alpha\beta} + t_{\rm LL}^{\alpha\beta}\big)\Big] = 0\,. \tag{143}$$

These are conservation equations for the total energy-momentum pseudotensor (which includes a contribution from the matter and another contribution from the gravitational field), expressed in terms of a partial-derivative operator. These equations are strictly equivalent to the usual expression of energy-momentum conservation, $\nabla_\beta T^{\alpha\beta} = 0$, which involves only the matter's energy-momentum tensor and a covariant-derivative operator.

This is an *exact* reformulation of the standard form of the theory. No approximations are involved, and no restrictions are placed on the choice of coordinates. It has

to be acknowledged, however, that the usefulness of the formalism is largely limited to situations in which (i) the coordinates $x^\alpha = (ct, x^j)$ are modest deformations of the Lorentzian coordinates of flat space-time, and (ii) $\mathfrak{g}^{\alpha\beta}$ deviates only moderately from the Minkowski metric $\eta^{\alpha\beta}$.

Because they involve a partial-derivative operator, the differential identities of Eq. (143) can immediately be turned into integral identities. We consider a three-dimensional region V, a fixed (time-independent) domain of the spatial coordinates x^j, bounded by a two-dimensional surface S. We assume that V contains at least some of the matter (so that $T^{\alpha\beta}$ is nonzero somewhere within V), but that S does not intersect any of the matter (so that $T^{\alpha\beta} = 0$ everywhere on S).

We formally define a total momentum four-vector $P^\alpha[V]$ associated with the region V by the three-dimensional integral

$$P^\alpha[V] := \frac{1}{c}\int_V (-g)\big(T^{\alpha 0} + t_{\mathrm{LL}}^{\alpha 0}\big)\, d^3x\,. \tag{144}$$

This total momentum includes a contribution from the matter's momentum density $c^{-1}T^{\alpha 0}$, and a contribution from the gravitational field represented by $c^{-1}t_{\mathrm{LL}}^{\alpha 0}$.

The momentum four-vector can be broken down into a time component $P^0[V]$ and a spatial three-vector $P^j[V]$. The time component can be used to define an energy $E[V] := cP^0[V]$ associated with the region V,

$$E[V] := \int_V (-g)\big(T^{00} + t_{\mathrm{LL}}^{00}\big)\, d^3x\,, \tag{145}$$

and a corresponding mass $M[V] := E[V]/c^2$. The three-momentum is given by

$$P^j[V] := \frac{1}{c}\int_V (-g)\big(T^{j0} + t_{\mathrm{LL}}^{j0}\big)\, d^3x\,. \tag{146}$$

The total angular-momentum four-tensor $J^{\alpha\beta}[V]$ is defined by

$$J^{\alpha\beta}[V] := \frac{1}{c}\int_V \Big[x^\alpha\,(-g)\big(T^{\beta 0} + t_{\mathrm{LL}}^{\beta 0}\big) - x^\beta\,(-g)\big(T^{\alpha 0} + t_{\mathrm{LL}}^{\alpha 0}\big)\Big]\, d^3x, \tag{147}$$

$$:= \frac{2}{c}\int_V x^{[\alpha}\,(-g)\big(T^{\beta]0} + t_{\mathrm{LL}}^{\beta]0}\big)\, d^3x\,, \tag{148}$$

and we note that this tensor is antisymmetric in its indices. The interpretation of $J^{\alpha\beta}[V]$ is easier to extract once it is decomposed into time and spatial components. The antisymmetry of the tensor implies that $J^{00}[V] = 0$. The time-space components can be expressed in the form

$$c^{-1}J^{0j}[V] = P^j[V]t - M[V]R^j[V], \tag{149}$$

where

$$R^j[V] := \frac{1}{M[V]c^2}\int_V (-g)\big(T^{00} + t_{\mathrm{LL}}^{00}\big)x^j\, d^3x \tag{150}$$

represents the position of the center-of-mass of the region V. Equation (149) reveals that, if $c^{-1}J^{0j}[V]$ is constant, it fixes the location of the center-of-mass at $t = 0$. The spatial components of the angular-momentum tensor are

$$J^{jk}[V] = \frac{1}{c}\int_V \Big[x^j\,(-g)\big(T^{k0} + t_{\rm LL}^{k0}\big) - x^k\,(-g)\big(T^{j0} + t_{\rm LL}^{j0}\big)\Big]\,d^3x, \tag{151}$$

and this is best recognized in its equivalent angular-momentum vectorial form as

$$J_j[V] := \frac{1}{2}\varepsilon_{jpq}J^{pq}[V] = \frac{1}{c}\int_V \varepsilon_{jpq}x^p\,(-g)\big(T^{qo} + t_{\rm LL}^{q0}\big)\,d^3x\,, \tag{152}$$

where ε_{jpq} is the completely antisymmetric permutation symbol.

To obtain the conservation statements satisfied by $P^\alpha[V]$, we differentiate its defining expression with respect to x^0 and use the local conservation identity of Eq. (143). Starting with Eq. (144), we get

$$\frac{d}{dx^0}P^\alpha[V] = \frac{1}{c}\int_V \partial_0\Big[(-g)\big(T^{\alpha 0} + t_{\rm LL}^{\alpha 0}\big)\Big]\,d^3x = -\frac{1}{c}\int_V \partial_k\Big[(-g)\big(T^{\alpha k} + t_{\rm LL}^{\alpha k}\big)\Big]\,d^3x\,.$$

Converting this to a surface integral, and recalling our previous assumption that S does not intersect the matter distribution, so that $T^{\alpha\beta} = 0$ on S, we arrive at

$$\dot{P}^\alpha[V] = -\oint_S (-g)t_{\rm LL}^{\alpha k}\,dS_k\,, \tag{153}$$

in which an overdot indicates differentiation with respect to $t := x^0/c$. The rate of change of $P^\alpha[V]$ is therefore expressed as a flux integral over S, and the flux is measured by the Landau–Lifshitz pseudotensor.

Equation (153) gives rise to the individual statement about the total energy,

$$\dot{E}[V] = -c\oint_S (-g)t_{\rm LL}^{0k}\,dS_k\,. \tag{154}$$

Similar conservation statements can be derived for linear momentum, angular momentum, and the center-of-mass.

In the limit in which V is taken to include all of three-dimensional space, $P^\alpha[V]$ is known to coincide with the Arnowitt–Deser–Misner four-momentum of an asymptotically flat space-time, and its physical interpretation as a measure of total momentum is robust. This statement is true whenever the coordinates x^α coincide with a Lorentzian system at infinity; the coordinates do not have to be Lorentzian (and indeed, they could not be) at finite spatial distances.

4.2 The Relaxed Einstein Equations

It is advantageous at this stage to impose the four conditions

$$\partial_\beta \mathfrak{g}^{\alpha\beta} = 0 \tag{155}$$

on the gothic inverse metric. These are known as the *harmonic coordinate conditions*. It is also useful to introduce the potentials

$$h^{\alpha\beta} := \eta^{\alpha\beta} - \mathfrak{g}^{\alpha\beta}\,, \tag{156}$$

where $\eta^{\alpha\beta} := \mathrm{diag}(-1, 1, 1, 1)$ is the Minkowski metric expressed in Lorentzian coordinates. In terms of these potentials, the harmonic coordinate conditions read

$$\partial_\beta h^{\alpha\beta} = 0\,, \tag{157}$$

and in this context they are usually referred to as the *harmonic gauge conditions*.

The introduction of the potentials $h^{\alpha\beta}$ and the imposition of the harmonic gauge conditions simplify the appearance of the Einstein field equations. It is easy to verify that the left-hand side becomes

$$\partial_{\mu\nu} H^{\alpha\mu\beta\nu} = -\Box h^{\alpha\beta} + h^{\mu\nu}\partial_{\mu\nu} h^{\alpha\beta} - \partial_\mu h^{\alpha\nu}\partial_\nu h^{\beta\mu}\,, \tag{158}$$

where $\Box := \eta^{\mu\nu}\partial_{\mu\nu}$ is the flat-space-time wave operator. Isolating the wave operator on the left-hand side, and putting everything else on the right-hand side gives us the formal wave equation

$$\Box h^{\alpha\beta} = -\frac{16\pi G}{c^4}\tau^{\alpha\beta} \tag{159}$$

for the potentials $h^{\alpha\beta}$, where

$$\tau^{\alpha\beta} := (-g)\big(T^{\alpha\beta}[\mathsf{m}, g] + t^{\alpha\beta}_{\mathrm{LL}}[h] + t^{\alpha\beta}_{\mathrm{H}}[h]\big) \tag{160}$$

is the *effective energy-momentum pseudotensor* for the wave equation. We have introduced

$$(-g)t^{\alpha\beta}_{\mathrm{H}} := \frac{c^4}{16\pi G}\Big(\partial_\mu h^{\alpha\nu}\partial_\nu h^{\beta\mu} - h^{\mu\nu}\partial_{\mu\nu} h^{\alpha\beta}\Big) \tag{161}$$

as an additional (harmonic-gauge) contribution to the effective energy-momentum pseudotensor. The harmonic conditions slightly simplify the form of the Landau–Lifshitz pseudotensor, as can be seen from Eq. (140).

In our expression for $\tau^{\alpha\beta}$, we have indicated that the matter's energy-momentum tensor $T^{\alpha\beta}$ is a functional of matter variables m, in addition to being a functional of the metric tensor $g_{\alpha\beta}$ (which is obtained from the gravitational potentials). As an example, when the matter consists of a number of isolated fluid bodies, m collectively denotes fluid variables such as the mass density ρ, pressure p, and velocity field u^α. We have also indicated that the Landau–Lifshitz and harmonic pseudotensors are functionals of $h^{\alpha\beta}$.

The imposition of the gauge conditions (157) is equivalent to imposing the conservation equations

$$\partial_\beta \tau^{\alpha\beta} = 0\,. \tag{162}$$

It is easy to verify that $(-g)t^{\alpha\beta}_{\mathrm{H}}$ is separately conserved, in that it satisfies the equation $\partial_\beta[(-g)t^{\alpha\beta}_{\mathrm{H}}] = 0$ automatically.

The wave equation of Eq. (159) is the main starting point of post-Minkowskian theory. It is worth emphasizing the fact that this equation, together with Eq. (157) or (162) are an *exact formulation* of the Einstein field equations; no approximations have been introduced at this stage.

For a metric $g_{\alpha\beta}$ to satisfy the complete set of Einstein field equations, it is necessary for the potentials $h^{\alpha\beta}$ to satisfy *both* the wave equation *and* the gauge condition/conservation statement; it is the *union* of Eqs. (159) and (157) or (162) that is equivalent to the original form of the Einstein field equations, $G^{\alpha\beta} = (8\pi G/c^4)T^{\alpha\beta}$. The two sets of equations play different roles, however. The wave equation (159) determines the gravitational potentials $h^{\alpha\beta}[\mathsf{m}]$ (and therefore the metric) as functions of the harmonic coordinates x^α, in terms of the matter variables m; these, however, remain undetermined until we also invoke the conservation Eq. (162). It is this equation that determines the behavior of the matter variables in a curved space-time whose metric is extracted from $h^{\alpha\beta}[\mathsf{m}]$. Solving both sets of equations amounts to integrating the full set of Einstein field equations, and this procedure simultaneously determines the metric *and* the matter variables.

One is entirely free to solve the wave equation (159) without also enforcing the gauge condition of Eq. (157) or the conservation statement of Eq. (162). Solving the wave equation independently of the gauge condition/conservation statement amounts to integrating only a subset of the Einstein field equations, and the outcome of this procedure is ten gravitational potentials $h^{\alpha\beta}[\mathsf{m}]$ that are expressed as functionals of undetermined matter variables m. The metric obtained from these potentials is also a functional of m, and it is not yet a solution to the Einstein field equations; it becomes a solution only when the gauge condition/conservation statement is imposed as an additional condition on the matter variables. The wave equation (Eq. (159)), taken by itself independently of Eq. (157), is known as the *relaxed Einstein field equation*.

4.3 Solution of the Wave Equation

The wave equation (159) admits the immediate formal solution

$$h^{\alpha\beta}(t,\mathbf{x}) = \frac{4G}{c^4}\int \frac{\tau^{\alpha\beta}(t-|\mathbf{x}-\mathbf{x}'|/c,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x'. \tag{163}$$

This is the *retarded solution* to the wave equation, and the domain of integration extends over $\mathcal{C}(x)$, the *past light cone* of the field point $x = (ct,\mathbf{x})$ (see Fig. 2).

To evaluate the integral, we partition the integration domain into two pieces, the *near-zone domain* $\mathcal{N}(x)$ and the *wave-zone domain* $\mathcal{W}(x)$. We place the boundary of the near/wave zones at an arbitrarily selected radius $\mathcal{R}$, with $\mathcal{R}$ imagined to be of the same order of magnitude as $\lambda_c \sim c\,t_c$, the characteristic wavelength of the radiation emitted by the source, where t_c is the characteristic time scale for noticeable changes to occur within the source. For slowly moving sources, $\lambda_c \sim cr_c/v_c \gg r_c$, where r_c and v_c are the characteristic size and internal velocity of the material source.

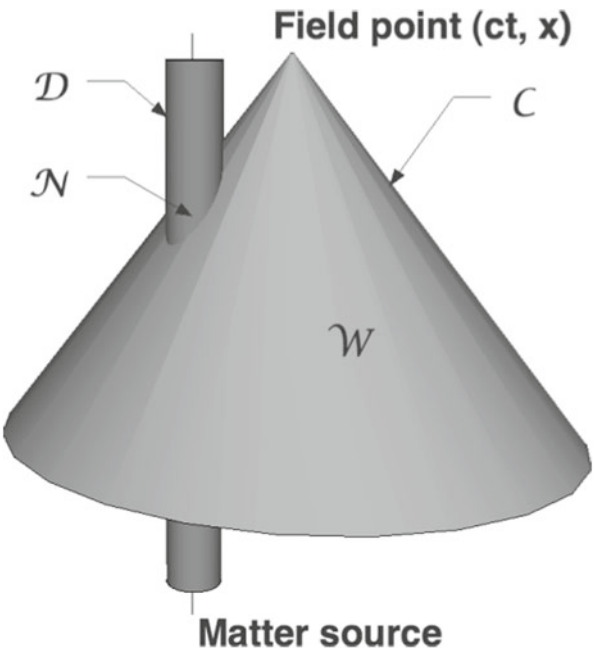


Fig. 2 Integration domains for the retarded solution of the wave equation

Thus, the near zone is the region of three-dimensional space in which $r := |\mathbf{x}|$ is small compared with a characteristic wavelength λ_c, while the wave zone is the region in which r is large compared with this length scale. The potential behaves very differently in the two zones: in the near zone, the difference between $t - r/c$ and t is small (the field retardation is unimportant), and derivatives with respect to $x^0 = ct$ are small compared with spatial derivatives; in the wave zone, the difference between $t - r/c$ and t is large, and x^0-derivatives are comparable to spatial derivatives. These properties are shared by all generic solutions to the wave equation.

In space-time, this sphere of radius $\mathcal{R}$ traces a near-zone *world tube*, $\mathcal{D}$. We let $\mathcal{N}(x)$ be the intersection between $\mathcal{C}(x)$ and the near zone $\mathcal{D}$, formally defined as the spatial region such that $r' := |\mathbf{x}'| < \mathcal{R}$. Similarly, we let $\mathcal{W}(x)$ be the intersection between $\mathcal{C}(x)$ and the wave zone, formally defined as the spatial region such that $r' > \mathcal{R}$, or $\mathcal{C}(x) - \mathcal{N}(x)$. We write Eq. (163) as

$$h^{\alpha\beta}(x) = h^{\alpha\beta}_{\mathcal{N}}(x) + h^{\alpha\beta}_{\mathcal{W}}(x)\,. \tag{164}$$

In electromagnetism, there is generally no contribution from the wave zone integration, because the source—the charges and currents—are confined to the near zone. But because of the nonlinearity of gravity, gravity itself acts as a source, so that, even though $T^{\alpha\beta}$ may be confined to the near zone, $t^{\alpha\beta}_{\rm LL}$ and $t^{\alpha\beta}_{\rm H}$ are not. Special integration techniques must be used to carry out the integrals over $\mathcal{W}$, but for most physical systems of interest, the wave-zone integrals generate higher-order corrections to the

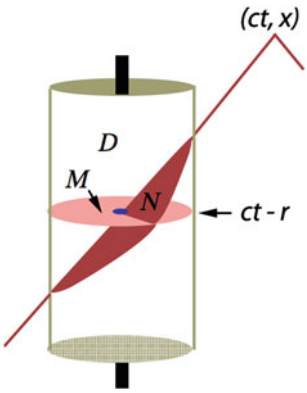


Fig. 3 Near-zone integration, far-zone field point

dominant terms, which come from the near-zone integrals. In this chapter, we will henceforth concern ourselves only with calculating $h^{\alpha\beta}_{\mathcal{N}}(x)$.

We first evaluate $h^{\alpha\beta}_{\mathcal{N}}(x)$ when the field point x is situated in the wave zone, that is, when $r \gg \mathcal{R}$. Knowing that $\mathbf{x}'$ lies within the near zone, we treat it as a small vector, and we Taylor-expand the variable $|\mathbf{x} - \mathbf{x}'|$ in $h^{\alpha\beta}_{\mathcal{N}}(x)$ about $\mathbf{x}' = \mathbf{0}$, with the result

$$h^{\alpha\beta}_{\mathcal{N}}(t, \mathbf{x}) = \frac{4G}{c^4} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} \partial_L \left[\frac{1}{r} \int_{\mathcal{M}} \tau^{\alpha\beta}(\tau, \mathbf{x}') x'^L \, d^3x' \right], \tag{165}$$

where

$$\tau := t - r/c \tag{166}$$

is a retarded-time variable. Notice that the temporal dependence of the source function no longer involves $\mathbf{x}'$, the variable of integration. The integration domain has therefore become a surface of constant time (the constant being equal to $\tau = t - r/c$) bounded externally by the sphere $r' = \mathcal{R}$. This domain is denoted by $\mathcal{M}$ in Eq. (165), and is illustrated in Fig. 3. The partial derivatives ∂_L operate both on $1/r$ and on the r embedded in τ.

Equation (165) is valid everywhere within the wave zone. It simplifies when $r \to \infty$, that is, when $h^{\alpha\beta}_{\mathcal{N}}$ is evaluated in the *far-away wave zone*. In this limit, we retain only the dominant, r^{-1} term in $h^{\alpha\beta}_{\mathcal{N}}$, and we approximate Eq. (165) by

$$h^{\alpha\beta}_{\mathcal{N}} = \frac{4G}{rc^4} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} \int_{\mathcal{M}} \partial_L \tau^{\alpha\beta}(\tau, \mathbf{x}') x'^L \, d^3x' + O(r^{-2}) . \tag{167}$$

The dependence of $\tau^{\alpha\beta}$ on x^j is contained in $\tau = t - r/c$, so, using the fact for any function $f(\tau)$ that $\partial_j f = -c^{-1} \partial_j r (df/d\tau) = -c^{-1} n_j (df/d\tau)$, we find that Eq. (167) becomes

$$h^{\alpha\beta}_{\mathcal{N}}(t, \mathbf{x}) = \frac{4G}{rc^4} \sum_{\ell=0}^{\infty} \frac{1}{\ell! c^\ell} n_L \left(\frac{d}{d\tau} \right)^\ell \int_{\mathcal{M}} \tau^{\alpha\beta}(\tau, \mathbf{x}') x'^L \, d^3x' + O(r^{-2}) . \tag{168}$$

This is a *multipole expansion* for the potential $h^{\alpha\beta}_{\mathcal{N}}$, in which each ℓ-pole moment $\int_{\mathcal{M}} \tau^{\alpha\beta} x^L \, d^3x$ is differentiated ℓ-times with respect to τ. Notice that $n_L x'^L = n_{j_1} n_{j_2} \cdots n_{j_\ell} x'^{j_1} x'^{j_2} \cdots x'^{j_\ell} = (\mathbf{n} \cdot \mathbf{x}')^\ell$.

We next evaluate $h^{\alpha\beta}_{\mathcal{N}}(x)$ when x is situated in the near zone, that is, when $r = |\mathbf{x}| < \mathcal{R}$. In this situation, both $\mathbf{x}$ and $\mathbf{x}'$ lie within the near zone, and $|\mathbf{x} - \mathbf{x}'|$ can be treated as a small quantity. To evaluate the integral, we simply Taylor-expand the time-dependence of the source with the result

$$h^{\alpha\beta}_{\mathcal{N}}(t, \mathbf{x}) = \frac{4G}{c^4} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell! c^\ell} \left(\frac{\partial}{\partial t} \right)^\ell \int_{\mathcal{M}} \tau^{\alpha\beta}(t, \mathbf{x}') |\mathbf{x} - \mathbf{x}'|^{\ell - 1} \, d^3x', \tag{169}$$

which is valid everywhere within the near zone. Notice that once more, the domain of integration is $\mathcal{M}$, a surface of constant time bounded externally by the sphere $r' = \mathcal{R}$, but here evaluated at the unretarded time t.

4.4 Iteration of the Relaxed Field Equations

We have cast the solution of the wave equation(Eq. (159)) as a retarded integral, but we really cannot call it a "solution", because $h^{\alpha\beta}$ appears inside the integral. In order to construct solutions for a particular choice of matter variables, we proceed by successive approximations, or iterations.

In the *zeroth iteration* of the relaxed field equations, we set $h^{\alpha\beta}_0 = 0$ and immediately get $g^0_{\alpha\beta} = \eta_{\alpha\beta}$, the metric of Minkowski space-time. From this, we construct $T^{\alpha\beta}[\mathsf{m}, g] = T^{\alpha\beta}[\mathsf{m}, \eta]$, $t^{\alpha\beta}_{\mathrm{LL}}[h] = t^{\alpha\beta}_{\mathrm{LL}}[h_0] = 0$, and $t^{\alpha\beta}_{\mathrm{H}}[h] = t^{\alpha\beta}_{\mathrm{H}}[h_0] = 0$. From all this, we obtain $\tau^{\alpha\beta}_0 = T^{\alpha\beta}[\mathsf{m}, \eta]$; this is the energy-momentum tensor of the matter variables m, and in the zeroth iteration these live in Minkowski space-time.

In the *first iteration* of the relaxed field equations, we solve the wave equation $\Box h^{\alpha\beta} = -(16\pi G/c^4)\tau^{\alpha\beta}_0$ for $h^{\alpha\beta}_1$. Because the source $\tau^{\alpha\beta}_0$ is known from the zeroth iteration, the wave equation can be integrated without difficulty (at least in principle), and this returns the potentials $h^{\alpha\beta}_1$ as functionals of the matter variables m, which are yet to be determined. From the potentials, we form the metric $g^1_{\alpha\beta}$ and construct $\tau^{\alpha\beta}_1$, an improved version of the effective energy-momentum pseudotensor. This involves the material contribution $T^{\alpha\beta}[\mathsf{m}, g_1]$, as well as the field contributions $t^{\alpha\beta}_{\mathrm{LL}}[h_1]$ and $t^{\alpha\beta}_{\mathrm{H}}[h_1]$.

In the *second iteration* of the relaxed field equations, we solve the wave equation $\Box h^{\alpha\beta} = -(16\pi G/c^4)\tau^{\alpha\beta}_1$ for $h^{\alpha\beta}_2$, an improved version of the gravitational potentials. Because the source $\tau^{\alpha\beta}_1$ is known from the first iteration, the wave equation can once more be integrated, and $h^{\alpha\beta}_2$ are again functionals of the undetermined matter variables m. From the new potentials, we form the metric $g^2_{\alpha\beta}$ and construct $\tau^{\alpha\beta}_2$, the latest version of the effective energy-momentum pseudotensor. The stage is ready

for the next iteration. One continues iterating until the desired precision is reached for determining the equations of motion for the matter and the resulting space-time metric.

5 Post-Newtonian Theory

Post-Newtonian theory can be considered as a special case of post-Minkowski theory in which one (a) imposes a slow-motion condition on top of the weak-field condition and (b) focusses on the near zone. Because the main physical applications of interest will be gravitationally self-interacting systems, such as binary systems, we know from Newton's equations that $U \sim v^2$, and from the equations of structure that $p/\rho \sim U$. In units of c^2, these are all small quantities for weak fields, and thus we assign a dimensionless parameter ε to characterize this smallness, so that

$$\varepsilon \sim \frac{U}{c^2} \sim \frac{v^2}{c^2} \sim \frac{p}{\rho c^2}\,. \tag{170}$$

In addition, since, in the near zone, quantities vary by virtue of their motion, one can also argue that a partial time derivative $\partial/\partial t$ is of order v compared to a spatial gradient, and thus that $\partial/\partial x^0 \sim (v/c)\nabla$, or

$$\frac{\partial/\partial x^0}{\partial/\partial x^j} \sim \varepsilon^{1/2}\,. \tag{171}$$

This approximation modifies to a certain extent how one treats the different iterations of the relaxed Einstein equations. For example, in the first iteration, with $\tau^{00} = T^{00}[\mathsf{m}, \eta]$, we find

$$h_1^{00} = \frac{4G}{c^4}\int \frac{T^{00}[\mathsf{m}, \eta]}{|\mathbf{x} - \mathbf{x}'|} d^3x'\,. \tag{172}$$

But the special-relativistic $T^{00} \sim \rho c^2[1 + O(v/c)^2]$, and hence h_1^{00} has the schematic form

$$h_1^{00} \sim 4U/c^2 + O(v^2 U/c^4) \sim \varepsilon + \varepsilon^2\,. \tag{173}$$

At the second iteration, contributions of $h_1^{\alpha\beta}$ to τ^{00} will induce contributions of order

$$h_2^{00} \sim \frac{4G}{c^4}\int \frac{T^{00}[\mathsf{m}, \eta] h_1^{00}}{|\mathbf{x} - \mathbf{x}'|} d^3x' \sim \frac{G^2}{c^4}\left(\frac{m}{r}\right)^2 \sim \frac{U^2}{c^4} \sim \varepsilon^2\,, \tag{174}$$

which are of the same order as the $O(v/c)^2$ corrections to the first iteration. Thus, to obtain the metric to the first *post-Newtonian* order, one must iterate the relaxed equations twice, as well as retain suitable $(v/c)^2$ correction terms. We will make the requirements more explicit shortly.

5.1 General Structure of the Fields

It is instructive to examine the general structure of the fields $h^{\alpha\beta}_{\mathcal{N}}$ in the near zone. The first few terms in the expansion look like

$$h^{\alpha\beta}_{\mathcal{N}}(t,\mathbf{x}) = \frac{4G}{c^4}\left[\int_{\mathcal{M}} \frac{\tau^{\alpha\beta}(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,d^3x' - \frac{1}{c}\frac{\partial}{\partial t}\int_{\mathcal{M}} \tau^{\alpha\beta}(t,\mathbf{x}')\,d^3x' + \frac{1}{2c^2}\frac{\partial^2}{\partial t^2}\int_{\mathcal{M}} \tau^{\alpha\beta}(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|\,d^3x' - \ldots\right]. \tag{175}$$

Taking into account the conservation equation $\partial_\beta \tau^{\alpha\beta} = 0$, we can convert certain terms into surface integrals that either vanish in the absence of gravitational radiation, or that can be shown to contribute very small, higher-order corrections. To simplify the notation somewhat, it is useful to define $N := h^{00}$, $K^i := h^{0i}$, $B^{ij} := h^{ij}$, and $B := \delta_{ij}h^{ij}$. Then

$$N_{\mathcal{N}} = \frac{4G}{c^2}\left[\int_{\mathcal{M}} \frac{c^{-2}\tau^{00}(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' + \frac{1}{2c^2}\frac{\partial^2}{\partial t^2}\int_{\mathcal{M}} c^{-2}\tau^{00}(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|d^3x' - \frac{1}{6c^3}\overset{(3)}{\mathcal{I}}{}^{kk}(t) + \frac{1}{24c^4}\frac{\partial^4}{\partial t^4}\int_{\mathcal{M}} c^{-2}\tau^{00}(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|^3 d^3x' - \frac{1}{120c^5}\left\{(4x^{kl}+2r^2\delta^{kl})\overset{(5)}{\mathcal{I}}{}^{kl}(t) - 4x^k\overset{(5)}{\mathcal{I}}{}^{kll}(t) + \overset{(5)}{\mathcal{I}}{}^{kkll}(t)\right\} + h^{00}_{\partial\mathcal{M}} + O(c^{-6})\right], \tag{176a}$$

$$K^i_{\mathcal{N}} = \frac{4G}{c^3}\left[\int_{\mathcal{M}} \frac{c^{-1}\tau^{0i}(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' + \frac{1}{2c^2}\frac{\partial^2}{\partial t^2}\int_{\mathcal{M}} c^{-1}\tau^{0i}(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|d^3x' + \frac{1}{18c^3}\left\{3x^k\overset{(4)}{\mathcal{I}}{}^{ik}(t) - \overset{(4)}{\mathcal{I}}{}^{ikk}(t) + 2\varepsilon^{mik}\overset{(3)}{\mathcal{J}}{}^{mk}(t)\right\} + h^{0i}_{\partial\mathcal{M}} + O(c^{-4})\right], \tag{176b}$$

$$B^{ij}_{\mathcal{N}} = \frac{4G}{c^4}\left[\int_{\mathcal{M}} \frac{\tau^{ij}(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' - \frac{1}{2c}\overset{(3)}{\mathcal{I}}{}^{ij}(t) + \frac{1}{2c^2}\frac{\partial^2}{\partial t^2}\int_{\mathcal{M}} \tau^{ij}(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|d^3x' - \frac{1}{36c^3}\left\{3r^2\overset{(5)}{\mathcal{I}}{}^{ij}(t) - 2x^k\overset{(5)}{\mathcal{I}}{}^{ijk}(t) - 8x^k\varepsilon^{mk(i}\overset{(4)}{\mathcal{J}}{}^{m|j)}(t) + 6\overset{(3)}{M}{}^{ijkk}(t)\right\} + h^{ij}_{\partial\mathcal{M}} + O(c^{-4})\right], \tag{176c}$$

where

$$\mathcal{I}^{Q} \equiv \int_{\mathcal{M}} c^{-2}\tau^{00}x^{Q}d^{3}x\,, \tag{177a}$$

$$\mathcal{J}^{iQ} \equiv \varepsilon^{iab}\int_{\mathcal{M}} c^{-1}\tau^{0b}x^{aQ}d^{3}x\,, \tag{177b}$$

$$M^{ijQ} \equiv \int_{\mathcal{M}} \tau^{ij}x^{Q}d^{3}x\,, \tag{177c}$$

and where $h^{\alpha\beta}_{\partial\mathcal{M}}$ denote the surface terms, which will not be displayed here. The notation (n) over quantities such as $\mathcal{I}^{jk}$ denotes the number of time derivatives. In obtaining these forms, we have made good use of a number of identities that follow from the conservation statement $\partial_\beta\tau^{\alpha\beta}=0$, namely,

$$\tau^{0j} = \partial_0\big(\tau^{00}x^{j}\big) + \partial_k\big(\tau^{0k}x^{j}\big)\,, \tag{178a}$$

$$\tau^{jk} = \frac{1}{2}\partial_{00}\big(\tau^{00}x^{j}x^{k}\big) + \frac{1}{2}\partial_p\big(\tau^{jp}x^{k}+\tau^{kp}x^{j}-\partial_q\tau^{pq}x^{j}x^{k}\big)\,, \tag{178b}$$

$$\begin{aligned}\tau^{jk}x^{n} = {}& \frac{1}{2}\partial_0\big(\tau^{0j}x^{k}x^{n}+\tau^{0k}x^{j}x^{n}-\tau^{0n}x^{j}x^{k}\big)\\ &+\frac{1}{2}\partial_p\big(\tau^{jp}x^{k}x^{n}+\tau^{kp}x^{j}x^{n}-\tau^{np}x^{j}x^{k}\big)\,.\end{aligned} \tag{178c}$$

The first term in N clearly leads off with the Newtonian potential$/c^2$, of order ε. Embedded within that term are corrections of order ε and higher, both from $(v/c)^2$ corrections and from contributions from higher iterations of the relaxed equations. The second term in N is known as a superpotential; because of the two time derivatives, it clearly leads off at order $(v/c)^2$ compared to the Newtonian term, and is thus already a post-Newtonian correction (and higher). (The superpotential itself is of order mc^2r, but since $\partial/\partial t \sim v/r$, the result is of order $(v/c)^2$ compared to the first term.) Notice that there was no term involving a single time derivative. By virtue of the fact that $\partial_0\tau^{00} = -\partial_k\tau^{0k}$, that term in fact is converted purely to a surface term within $h^{00}_{\partial\mathcal{M}}$, and contributes only at a very high post-Newtonian order via a flux of gravitational radiation.

The third term in N contains three time derivatives, and one might be tempted to give it the name of a (post)$^{3/2}$-Newtonian correction. However, it can be shown that, because this term depends only on time, it can always be absorbed to lowest order into a redefinition of time, and hence removed by a coordinate transformation. Alternatively, since the leading contribution of N to the equations of motion arises from a gradient, this term will make no contribution to leading order.

The fourth term is called a superduperpotential term, and because of the four time derivatives, contributes at second post-Newtonian (2PN) order and higher.

The fifth term looks schematically like $mc^2(\partial/\partial ct)^5r^4 \sim c^2(v/c)^5(m/r) \sim \varepsilon^5\times$ Term 1. This a 2.5PN correction, and represents the leading effects of gravitational

radiation reaction. Because it involves an odd number of time derivatives, it is asymmetric under the transformation $t \to -t$; as such, it represents the effects of an irreversible, dissipative process, the loss of energy into gravitational waves.

The other field components have a similar structure: K^i leads at order $\varepsilon^{3/2}$, while B^{ij} and B lead at order ε^2.

We now need to establish the order in ε to which each of these components must be calculated to reach a desired post-Newtonian order. To do this, we return to the expression for the action for a test particle moving in the space-time of a metric $g_{\alpha\beta}$, Eq. (108), and express it in terms of separated time and space components, namely

$$S = -mc^2 \int_1^2 \sqrt{-g_{00} - 2g_{0j}\frac{v^j}{c} - g_{ij}\frac{v^j v^k}{c^2}}\, dt\,. \tag{179}$$

Newtonian gravity is clearly given by the approximations $g_{00} \approx -1 + 2U/c^2$, $g_{0i} \approx 0$, and $g_{ij} \approx \delta_{ij}$, in other words, by evaluating the action to order ε. Post-Newtonian gravity is then given by evaluating the action to order ε^2, which requires evaluating the metric coefficients to the following orders of approximation:

$$\begin{aligned} g_{00} &\text{ to } O(\varepsilon^2)\,,\\ g_{0i} &\text{ to } O(\varepsilon^{3/2})\,,\\ g_{ij} &\text{ to } O(\varepsilon)\,. \end{aligned} \tag{180}$$

The orders needed to descend because of the additional factors of v/c in the action. From this, we can generalize to the requirement that to work to nPN order, we must determine the metric components to the orders

$$\begin{aligned} g_{00} &\text{ to } O(\varepsilon^{n+1})\,,\\ g_{0i} &\text{ to } O(\varepsilon^{n+\frac{1}{2}})\,,\\ g_{ij} &\text{ to } O(\varepsilon^{n})\,. \end{aligned} \tag{181}$$

To determine the order to which we must then determine the components of the field, we must expand in powers of ε the relation between the physical metric $g^{\alpha\beta}$ and the fields using Eq. (136), The result is

$$g_{00} = -(1 - \frac{1}{2}N + \frac{3}{8}N^2 - \frac{5}{16}N^3) + \frac{1}{2}B(1 - \frac{1}{2}N) + \frac{1}{2}K^j K^j + O(\varepsilon^4)\,, \tag{182a}$$

$$g_{0i} = -K^i(1 - \frac{1}{2}N) + O(\varepsilon^{7/2})\,, \tag{182b}$$

$$g_{ij} = \delta^{ij}(1 + \frac{1}{2}N - \frac{1}{8}N^2) + B^{ij} - \frac{1}{2}B\delta^{ij} + O(\varepsilon^3)\,, \tag{182c}$$

$$(-g) = 1 + N - B + O(\varepsilon^3)\,. \tag{182d}$$

This expansion is sufficient to determine the metric through 2.5PN order.

5.2 The Post-Newtonian Limit of General Relativity

To illustrate the use of the post-Minkowskian framework, we will derive the post-Newtonian limit of general relativity. In going to post-Newtonian order, we wish to express the density in terms of ρ^* instead of ρ, because the former density satisfies a continuity equation to all orders. This makes a number of calculations simpler.

The Newtonian limit is trivial: inserting the zeroth iteration of the relaxed equations, $h_0^{\alpha\beta} = 0$ in the source terms, it is clear that $\tau_0^{00} = \rho^* c^2$, and thus that the first iteration of the equations yields $N_1 = 4U/c^2$, and $g_{00} = -1 + 2U/c^2$, $g_{0i} = 0$, and $g_{ij} = \delta_{ij}$, where U is the "Newtonian" potential derived from $\nabla^2 U = -4\pi G\rho^*$. From the first iteration, we also obtain expressions for K^i and B^{ij}, but these are not needed for the Newtonian limit of $g_{\alpha\beta}$; we will return to these shortly. Notice that in determining N, we also get the bonus that $g_{ij} = \delta_{ij}(1 + 2U/c^2)$; this also is not needed for the Newtonian limit, but will be needed in order to obtain $(-g)$ for the next iteration.

Using the definition (125), together with the approximations $u^0/c \approx 1 + v^2/2c^2 + U/c^2$, and $-g \approx 1 + 4U/c^2$ we find to the order needed for the next iteration that

$$\rho^* = \rho\left(1 + \frac{v^2}{2c^2} + \frac{3U}{c^2}\right). \tag{183}$$

Substituting these results into Eq. (121), and carefully evaluating the contributions of $t_{\rm LL}^{\alpha\beta}$ and $t_{\rm H}^{\alpha\beta}$ needed to this order, we obtain the first iterated version of $\tau^{\alpha\beta}$, given to the required post-Newtonian order by

$$\tau_1^{00} = \rho^* c^2\left(1 + \frac{v^2}{2c^2} + \frac{3U}{c^2} + \frac{\Pi}{c^2}\right) - \frac{7}{8\pi G}\nabla U \cdot \nabla U\,, \tag{184a}$$

$$\tau_1^{0i} = \rho^* c v^i\,, \tag{184b}$$

$$\tau_1^{ii} = \rho^* v^2 + 3p - \frac{1}{8\pi G}\nabla U \cdot \nabla U\,, \tag{184c}$$

where $\Pi := \varepsilon/\rho^*$ is the internal energy per unit mass within the fluid.

Recalling the identity

$$\nabla U \cdot \nabla U = \frac{1}{2}\nabla^2 U^2 + 4\pi G\rho^* U\,, \tag{185}$$

we substitute Eqs. (184) into (176) and keep all contributions needed for the metric through 1.5PN order. The result is

$$N_2 = \frac{4U}{c^2} + \frac{1}{c^4}\left(7U^2 + 2\Phi_1 - 2\Phi_2 + 4\Phi_3 + 2\ddot{X}\right) - \frac{2G}{3c^5}\overset{(3)}{\mathcal{I}}{}^{kk}(t) + O(\varepsilon^3)\,, \tag{186a}$$

$$K_2^i = \frac{4U^i}{c^3} + O(\varepsilon^{5/2})\,, \tag{186b}$$

$$B_2 = \frac{1}{c^4}\left(U^2 + 4\Phi_1 - 2\Phi_2 + 12\Phi_4\right) - \frac{2G}{c^5}\overset{(3)}{\mathcal{I}}{}^{kk}(t) + O(\varepsilon^3)\,, \tag{186c}$$

where we define the auxiliary gravitational potentials

$$\begin{aligned}
U^i &:= G\int \frac{\rho^*(t,\mathbf{x}')v^i(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x'\,,\\
\Phi_1 &:= G\int \frac{\rho^*(t,\mathbf{x}')v'^2}{|\mathbf{x}-\mathbf{x}'|}d^3x'\,,\\
\Phi_2 &:= G\int \frac{\rho^*(t,\mathbf{x}')U(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x'\,,\\
\Phi_3 &:= G\int \frac{\rho^*(t,\mathbf{x}')\Pi(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x'\,,\\
\Phi_4 &:= G\int \frac{p(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x'\,,\\
X &:= G\int \rho^*(t,\mathbf{x}')|\mathbf{x}-\mathbf{x}'|d^3x'\,.
\end{aligned} \tag{187}$$

Substituting these results into Eq. (182), we obtain finally the metric for general relativity through 1.5PN order,

$$g_{00} = -1 + \frac{2}{c^2}U + \frac{2}{c^4}\left(\Psi - U^2\right) - \frac{4G}{3c^5}\overset{(3)}{\mathcal{I}}{}^{kk}(t) + O(\varepsilon^3)\,, \tag{188a}$$

$$g_{0i} = -\frac{4}{c^3}U^i + O(\varepsilon^{5/2})\,, \tag{188b}$$

$$g_{ij} = \delta_{ij}\left(1 + \frac{2}{c^2}U\right) + O(\varepsilon^2)\,, \tag{188c}$$

where

$$\Psi = \frac{3}{2}\Phi_1 - \Phi_2 + \Phi_3 + 3\Phi_4 + \frac{1}{2}\ddot{X}\,. \tag{189}$$

By making a coordinate transformation

$$t = t' - \frac{2G}{3c^5}\ddot{\mathcal{I}}^{kk}(t)\,, \tag{190}$$

we can eliminate the $O(G/c^5)$ term from g_{00} to lowest order, indicating that the term is purely a gauge or coordinate artifact. The leading nontrivial odd-order terms in g_{00} will occur at $O(\varepsilon^{7/2})$, corresponding to 2.5PN order.

Calculating the Christoffel symbols from the metric (188), inserting them into the energy-momentum conservation equation $\nabla_\beta T^{\alpha\beta} = 0$, one obtains the post-Newtonian Euler equation of hydrodynamics,

$$\rho^*\frac{dv^j}{dt} = \rho^*\partial_j U - \partial_j p$$

$$+\frac{1}{c^2}\left[\left(\frac{1}{2}v^2+U+\Pi+\frac{p}{\rho^*}\right)\partial_j p - v^j\partial_t p\right]$$
$$+\frac{1}{c^2}\rho^*\Big[(v^2-4U)\partial_j U - v^j\left(3\partial_t U+4v^k\partial_k U\right)$$
$$+4\partial_t U_j+4v^k\left(\partial_k U_j-\partial_j U_k\right)+\partial_j\Psi\Big]$$
$$+O(c^{-4}). \tag{191}$$

The post-Newtonian Euler equation, together with an equation of state $p(\rho)$ can be solved consistently, leading to a solution for the space-time metric, valid to post-Newtonian order.

5.3 Non-Newtonian Behavior: Worked Examples

Using the post-Newtonian metric derived above, we can work out some important examples of the effects of relativistic gravity. For simplicity, we will specialize to a single, spherically symmetric body of mass M at rest in our chosen coordinate system. This is a good approximation for studying the motion of a planet or a ray of light in the solar system. In this case, the metric of Eq. (188) reduces to

$$g_{00} = -1+2\frac{GM}{c^2 r}-2\left(\frac{GM}{c^2 r}\right)^2, \tag{192a}$$
$$g_{0i} = 0, \tag{192b}$$
$$g_{ij} = \delta_{ij}\left(1+2\frac{GM}{c^2 r}\right), \tag{192c}$$

where

$$M := \int \rho^*\left[1+\frac{1}{c^2}\left(\frac{3}{2}v^2-U+\Pi+3\frac{p}{\rho^*}\right)\right]d^3x$$
$$= m+\frac{1}{c^2}\left[3\mathcal{T}+2\Omega+E_{\rm int}+3P\right]$$
$$= m+\frac{1}{c^2}\left[\mathcal{T}+\Omega+E_{\rm int}\right], \tag{193}$$

where $m = \int \rho^* d^3x$ is the total conserved or baryonic mass of the body, and $\mathcal{T}$, Ω and $E_{\rm int}$ are the internal kinetic, gravitational, and thermal energy of the body, defined as in Eq. (18) (but using ρ^* instead of the Newtonian ρ). We have used the Newtonian virial theorem (17b) in the static limit to simplify the post-Newtonian corrections to the mass. Note that the final answer, not surprisingly, is the total of rest mass and the mass equivalent of the internal energies.

The geodesic equation (107) can be rewritten using coordinate time as the parameter; this is a useful step because the motion of planets or light is normally referred

to an external "time" coordinate, such as atomic time measured on Earth, not proper time along the particle's world line. The result is

$$\frac{d^2r^i}{dt^2} + \left(\Gamma^i_{\alpha\beta} - \Gamma^0_{\alpha\beta}\frac{v^i}{c}\right)v^\alpha v^\beta = 0\,, \tag{194}$$

where $v^\alpha := (c, v^i)$. Evaluating the Christoffel symbols to post-Newtonian order using Eq. (192), we obtain the equation of motion for a test body

$$\frac{d^2\mathbf{r}}{dt^2} = -\frac{GM}{r^2}\mathbf{n} + \frac{GM}{c^2r^2}\left\{\mathbf{n}\left[4\frac{GM}{r} - v^2\right] + 4\dot{r}\mathbf{v}\right\}\,, \tag{195}$$

where $\dot{r} = \mathbf{n}\cdot\mathbf{v}$.

This equation of motion is ready-to-use for analysis of perturbed orbit elements, since it is in the form of the Newtonian two-body acceleration plus a disturbing acceleration. The relevant components of the disturbing acceleration (Eq. (63)) are given by

$$\begin{aligned}\mathcal{R} &= \frac{GM}{c^2r^2}\left[4\frac{GM}{r} - v^2 + 4\dot{r}^2\right],\\ \mathcal{S} &= 4\frac{GM}{c^2r^2}\dot{r}\mathbf{v}\cdot\lambda\,,\\ \mathcal{W} &= 0\,.\end{aligned} \tag{196}$$

From the orbit element perturbation equations (67) we immediately conclude that the inclination ι and angle of nodes Ω are constants, consistent with the motion staying in a fixed plane. We can choose this plane to be the $X-Y$ plane to simplify matters. We now substitute the osculating orbit formulae for the variables that appear in $\mathcal{R}$ and $\mathcal{S}$, such as $r = p/(1+e\cos f)$, obtained from Eq. (62). In the perturbation equations (67), we convert from d/dt to d/df using the relation $df/dt = (h/r^2)$, valid to first order in perturbation theory (see Eq. (68)) . Integrating over one orbit, we find the net change in the orbit elements,

$$\Delta a = 0\,, \tag{197a}$$

$$\Delta e = 0\,, \tag{197b}$$

$$\Delta\varpi = \frac{6\pi GM}{c^2p}\,. \tag{197c}$$

The first two results express the fact that energy and angular momentum are conserved to post-Newtonian order. The second is the famous advance of the pericenter.

Einstein derived this for a test particle in a geodesic around a point object of mass M, using an approximate solution of his equations equivalent to Eq. (192). The same result, to 1PN order is obtained from the Schwarzschild geometry. It turns out that it also applies to an arbitrary binary system of masses m_1 and m_2; in this case, $M = m_1 + m_2$. Since the mass of Mercury is only one 6 millionth of that of

the Sun, its contribution is negligible, and it can be treated as a massless test particle moving on a geodesic of a fixed space-time. This cannot be done for a binary star system because both masses move about the common center of mass; nevertheless, the answer for $\Delta\varpi$ is the same, and depends only on the total mass.

The advance per orbit can be converted to a rate of advance by dividing by the orbital period. We can also eliminate the semimajor axis appearing in p by using Kepler's third law, $a = (GM)^{1/3}(P/2\pi)^{2/3}$, where P is the orbital period, valid here because the effect is already of 1PN order. The result is

$$\frac{d\varpi}{dt} = \frac{3}{1-e^2}\frac{2\pi}{P}\left(\frac{GM/c^3}{P/2\pi}\right)^{2/3}$$
$$= 716.25\frac{1}{1-e^2}\left(\frac{M}{M_\odot}\right)^{2/3}\left(\frac{P}{1\,\mathrm{day}}\right)^{-5/3}\ \mathrm{as\,yr}^{-1}\,. \qquad (198)$$

Note that GM/c^3 is the light-crossing time of half the Schwarzschild radius corresponding to mass M, which is generally much smaller than the orbital period. Substituting the values for Mercury, $e = 0.2056$, $P = 87.97$ days, we obtain 42.98 as per century; as we saw in Table 1, the modern difference between the measured advance and that predicted by Newtonian N-body perturbations is 42.98 ± 0.04 as per century.

The discovery of binary pulsar systems with total masses of order 2–3 solar masses, and with orbital periods as small as fractions of a day resulted in the observation of huge periastron advances of several degrees per year. But in these systems, the relativistic periastron advance played an interesting role. In the solar system, GM for the Sun is known to high precision from the measured orbital period and orbital radius of the Earth, combined with Kepler's third law. By contrast, the masses of the neutron stars in these binary systems are not known, apart from the general expectation that they should be around the Chandrasekhar limit, $1.4\,M_\odot$, based on models of how such systems might have formed. In the famous Hulse–Taylor binary pulsar 1913+16, the first such system to be discovered, it was possible to measure the pericenter advance, the eccentricity and the orbital period very accurately; the current values are $\dot\omega = 4.226595(5)\,\mathrm{deg\,yr}^{-1}$, $P = 0.322997448930(4)\,\mathrm{day}$, $e = 0.6171338(4)$, where the number in parentheses denotes the error in the final digit. Assuming that there is no other source of periastron advance, we can turn Eq. (198) around and use it to measure the total mass of the system. The result is $M = 2.828296(5)\,M_\odot$. This is a remarkably accurate measurement of an astrophysical quantity, and general relativity played a central role in the analysis. The recently discovered "double pulsar" J0737-3039A/B, in which both stars were observed as pulsars, has $e = 0.0877775(9)$, $P = 0.10225156248(5)\,\mathrm{day}$, and $\dot\omega = 16.8995(7)\,\mathrm{deg\,yr}^{-1}$, so the total mass is $2.5871(2)\,M_\odot$. The relativistic advance is so large in these systems that we could easily see the evolving orientation of the orbit month by month with the naked eye, that is if we could see pulsars with the naked eye! For further details on binary pulsars, see the chapter, "The Role of Binary Pulsars in Testing Gravity Theories" by A. Possenti and M. Burgay.

The motion of light near the sun can also be analyzed, either by considering the equation for a null geodesic, or equivalently by taking the $v^i \to cN^i$ limit of Eq. (195). The result is

$$\frac{d^2\mathbf{r}}{dt^2} = -\frac{GM}{r^2}\left[\mathbf{n} - 2\mathbf{N}(\mathbf{N}\cdot\mathbf{n})\right] . \tag{199}$$

This equation can be integrated by setting $\mathbf{N}$ equal to the unperturbed direction of the light ray to obtain the deflection of a light ray that passes close to the sun, given by

$$\delta\theta = \frac{4GM_\odot}{c^2 d} = 1.7504\frac{R_\odot}{d}\,\text{as}\,, \tag{200}$$

where d is the distance of closest approach of the ray.

The first successful measurement of the bending of light by the Sun was carried out by British astronomer Arthur Stanley Eddington and colleagues during the total solar eclipse of May 29, 1919. It was a differential measurement: photographs of the stars near the Sun taken during the eclipse were compared with photographs of the same stars taken at night from the same locations later in the year, and the *changes* in angles between pairs of stars were carefully measured. Eddington's announcement in November 1919 of the measurement of the bending helped make Einstein an international celebrity. However, the experiments of Eddington and his coworkers had only 30 % accuracy, and succeeding eclipse measurements were not much better. The results were scattered between one half and twice the Einstein prediction, and the accuracies were low.

However, the advent of radio astronomy, and in particular the development of radio interferometry, and later of very-long-baseline radio interferometry (VLBI), produced greatly improved determinations of the deflection. These techniques now have the capability of measuring angular separations and changes in angles to accuracies of 10–100 microarcseconds. Early measurements took advantage of the fact that a few strong quasistellar radio sources (quasars) pass very close to the Sun (as seen from the Earth), including the group 3C273, 3C279, and 3C48, and the group 0111+02, 0119+11, and 0116+08. As the Earth moves in its orbit, changing the lines of sight of the quasars relative to the Sun, the angular separation $\delta\theta$ between pairs of quasars varies. A number of measurements of this kind over the period 1969–1975 yielded results in agreement with general relativity at a few parts in 10^3. In recent years, transcontinental and intercontinental VLBI observations of quasars and radio galaxies have been made primarily to monitor the Earth's rotation and to establish a highly accurate "reference" frame for astronomy and navigation. These measurements are sensitive to the deflection of light over almost the entire celestial sphere. A 2004 analysis of almost 2 million VLBI observations of 541 radio sources, made by 87 VLBI sites over a 20-year period verified Einstein's prediction to a few parts in 10^4. Analysis of observations made by the Hipparcos optical astrometry satellite yielded a test at the level of 0.3 %, and a future orbiting observatory called GAIA has the capability of testing the deflection to parts per million.

The deflection of light has become a cornerstone of the empirical edifice that supports general relativity. But in 1979, the phenomenon became much more than

that. That year, astronomers Dennis Walsh, Robert Carswell, and Ray Weymann discovered the "double quasar" Q0957+561, which consisted of two quasar images about 6 as apart, with almost the same redshift ($z = 1.41$) and very similar spectra. Given that quasars are thought to be among the most distant objects in the universe, the probability of finding two so close together was low. It was soon realized that there was just one quasar but that intervening matter in the form of a galaxy or cluster of galaxies was bending the light from the quasar and producing two separate images.

Since then, over 60 lensed quasars have been discovered. But more importantly, gravitational lensing has become a major tool in efforts to map the distribution of mass around galaxies and clusters, in the search for dark matter, dark energy, and compact objects, and in the search for extrasolar planets. Major subtopics in lensing have been developed to cover different astronomical realms: microlensing for the search for dim compact objects and extrasolar planets, the use of luminous arcs in the effort to map the distribution of mass and dark matter, and weak/statistical lensing in the effort to measure the properties of dark energy. Lensing has to be taken into account in interpreting certain aspects of the cosmic microwave background radiation, and in extracting information from the gravitational-wave signal from sources at large redshift.

Equation (199) can also be used to derive the Shapiro time delay, a retardation of the propagation of light past a massive object. For light emitted at time t_e at $\mathbf{x}_e$ and received by an observer at t_O at $\mathbf{x}_O$, the propagation time is given by

$$c(t_O - t_e) = |\mathbf{x}_O - \mathbf{x}_e| + \frac{2GM}{c^2} \ln\left[\frac{(r_O + \mathbf{x}_O \cdot \mathbf{n})(r_e - \mathbf{x}_e \cdot \mathbf{n})}{d^2}\right]. \qquad (201)$$

For a ray that propagates in a round-trip from the Earth to a planet or spacecraft on the far side of the Sun, the total delay can be as large as 200 μs.

Radio astronomer Irwin I. Shapiro discovered that this effect was a prediction of general relativity in 1964, and called it the "fourth" test of relativity. Shapiro and colleagues then carried out the first measurement of the time delay in 1967, by bouncing radar signals off the surface of Mercury. Later experiments involved radar echos from Venus and tracking of the Mars exploration spacecraft, Mariners 6, 7, and 9, and the Viking landers and orbiters.

The most recent measurement of the Shapiro delay involved tracking the Cassini spacecraft while it was on its way to Saturn. Several circumstances made this mission particularly favorable. One was the ability to do tracking measurements using both X-band (7175 MHz) and Ka-band (34316 MHz) radar, thereby significantly reducing the dispersive effects of the solar corona. In addition, during the 2002 superior conjunction of Cassini, the spacecraft was at 8.43 astronomical units from the Sun, and the distance of closest approach of the radar signals to the Sun was only 1.6 $R_\odot$. The result was a test in agreement with general relativity to two parts in 10^5. This and other solar system tests are discussed in the chapter, "Space-based Tests of Relativistic Gravitation" by V. Turyshev.

The Shapiro delay now figures in range of astrophysical phenomena. It has been measured in a number of binary pulsar systems, most notably in the double pulsar

J0737-3039A/B, where the orbit is seen almost edge-on, and the pulsed radio signals from each pulsar pass close to the other once per orbit. Analyses of the spectra and time variations of high-energy emissions (X-rays) from accretion disks around black holes must take into account both the strong bending of light and the Shapiro delay, but now using full general relativity, not just the post-Newtonian limit.

5.4 The Parametrized Post-Newtonian Formalism

An important application of the post-Newtonian limit of general relativity has been its extension to encompass a range of alternative metric theories of gravity. In many cases, such as the scalar-tensor theory of Jordan, Brans, and Dicke and generalizations thereof, the post-Newtonian limit has a form quite similar to that of general relativity, except that the coefficients in front of some terms are different, and a few additional post-Newtonian potentials may be present.

This is a very fortunate circumstance, because the post-Newtonian limit is sufficient to describe the solar system and the experimental tests one can perform there, and to a limited degree can also describe binary pulsar systems. Therefore, if we simply replace the numerical coefficients in the post-Newtonian limit of general relativity with arbitrary parameters, and add a few new potentials with their own parameters, we will have a framework that encompasses a wide range of alternative theories, and that can be used to calculate a wide range of testable phenomena.

This framework is called the parametrized post-Newtonian (PPN) framework. Eddington was the first to consider such a framework. In his classic 1922 textbook, *The Mathematical Theory of Relativity*, he parametrized the post-Newtonian limit of the Schwarzschild metric (in isotropic coordinates) in the form

$$ds^2 = -\left[1 - 2\alpha\frac{GM}{c^2 r} + 2\beta\left(\frac{GM}{c^2 r}\right)^2\right]c^2 dt^2 + \left[1 + 2\gamma\frac{GM}{c^2 r}\right](dx^2 + dy^2 + dz^2)\,, \tag{202}$$

where $\alpha = \beta = \gamma = 1$ in general relativity. Actually, the α parameter is redundant, since it can always be absorbed into GM, which is by definition the Kepler-measured mass, or into Newton's constant G as measured in a Cavendish-type experiment. The modern version of the PPN framework was developed by Kenneth Nordtvedt, Jr. in 1968, using a collection of N point masses as the physical system rather than a single mass; later, Will generalized the framework to post-Newtonian hydrodynamics, and in 1972 Nordtvedt and Will unified the two approaches into the PPN framework used today.

In a gauge that parallels the harmonic gauge of general relativity, the PPN metric has the form

$$g_{00} = -1 + \frac{2}{c^2}U + \frac{2}{c^4}\left(\Psi - \beta U^2\right) + O(\varepsilon^3)\,, \tag{203a}$$

$$g_{0i} = -\frac{1}{c^3}\left[2(1+\gamma) + \frac{1}{2}\alpha_1\right]U^i + O(\varepsilon^{5/2})\,, \tag{203b}$$

$$g_{ij} = \delta_{ij}\left(1 + \frac{2}{c^2}\gamma U\right) + O(\varepsilon^2)\,, \tag{203c}$$

where

$$\Psi := \frac{1}{2}(2\gamma + 1 + \alpha_3)\Phi_1 - (2\beta - 1 - \zeta_2 - \xi)\Phi_2 + (1 + \zeta_3)\Phi_3 + \left[3(\gamma + \zeta_4) - \zeta_1\right]\Phi_4 + \frac{1}{2}(\zeta_1 - 2\xi)\Phi_5 - \xi\Phi_W + \frac{1}{2}(1 + \alpha_2)\ddot{X}\,. \tag{204}$$

The potentials

$$\Phi_5 := G\int \rho^{*\prime}\nabla' U(t, \mathbf{x}')\cdot\frac{\mathbf{x} - \mathbf{x}'}{|\mathbf{x} - \mathbf{x}'|}d^3x'\,,$$
$$\Phi_W := G^2\int\int \rho^{*\prime}\rho^{*\prime\prime}\frac{(x - x')_j}{|\mathbf{x} - \mathbf{x}'|^3}\left(\frac{(x' - x'')^j}{|\mathbf{x} - \mathbf{x}''|} - \frac{(x - x'')^j}{|\mathbf{x}' - \mathbf{x}''|}\right)d^3x'd^3x''\,, \tag{205}$$

do not appear in the post-Newtonian limit of general relativity in this gauge.

The choice of parameters is made in part so that the values in general relativity are particularly simple, but more importantly so that specific meanings or implications can be attached to them. These heuristic meanings are summarized in Table 2. The parameters γ and β are directly related to Eddington's original parameters. Roughly speaking, γ measures the amount of curvature of space generated by a body, compared to what general relativity would predict; specifically, a calculation of the Riemann curvature tensor of the three-dimensional subspace defined by $dt = 0$ gives a result proportional to $\gamma GM/c^2r^3$. Roughly speaking, β measures how nonlinear gravity is, in that it multiplies the U^2 term in g_{00}. This is *very* rough, because it is a coordinate-dependent statement. In general relativity for example, the Schwarzschild metric has no $(GM/c^2r)^2$ term in Schwarzschild coordinates; the post-Newtonian limit of the Schwarzschild geometry takes the Eddington form of Eq. (202) with $\gamma = \beta = 1$ *only* in isotropic coordinates. The lesson is that these interpretations of the PPN parameters must be considered heuristic at best.

The parameters α_i are linked to violations of local Lorentz invariance in gravitational physics, which some theories predict. Suppose that all three parameters vanish in a given theory. Then the PPN metric in the above form is valid in any chosen coordinate system, with all velocities defined relative to that system. However, if any of them is nonzero, then the theory in question singles out a preferred universal rest frame (in most cases, it is the frame in which the cosmic background radiation is isotropic), and the PPN metric above is valid *only* when written in that rest coordinate system. To obtain the metric in a different frame, such as the barycentric frame of the solar system, one must perform what is called a post-Galilean transformation of the metric: this is a low-velocity Lorentz transformation evaluated to the appropriate post-Newtonian order, together with a post-Newtonian gauge transformation to put the final metric in a simple form. Such a transformation generates additional terms in the metric that depend *explicitly* on the velocity of the coordinate system relative to the preferred frame, and on the PPN parameters α_i. Theories of

Table 2 The PPN parameters and their significance (note that α_3 has been shown twice to indicate that it is a measure of two effects)

Parameter	What it measures relative to GR	GR value	Semiconservative theories	Fully conservative theories
γ	How much space-curvature produced by unit rest mass?	1	γ	γ
β	How much "nonlinearity" in the superposition of gravity?	1	β	β
ξ	Preferred-location effects?	0	ξ	ξ
α_1	Preferred-frame	0	α_1	0
α_2	effects?	0	α_2	0
α_3		0	0	0
α_3	Is total	0	0	0
ζ_1	momentum	0	0	0
ζ_2	conserved?	0	0	0
ζ_3		0	0	0
ζ_4		0	0	0

PPN parametrized post-Newtonian, *GR* general relativity

this type predict *preferred-frame* effects, many of which have been tested, leading to strong upper bounds on these parameters. The ζ_i parameters indicate whether the theory admits conservation laws for energy and momentum of the kind defined by Eq. (144); any theory that is based on an action predicts that these parameters vanish.

A number of well-known relativistic effects can now be expressed in terms of these PPN parameters (compare with Eqs. (198), (200), and (201)):
Perihelion advance:

$$\begin{aligned}\frac{d\varpi}{dt} &= \left(\frac{2+2\gamma-\beta}{3}\right)\frac{3}{1-e^2}\frac{2\pi}{P}\left(\frac{GM/c^3}{P/2\pi}\right)^{2/3}\\ &= \left(\frac{2+2\gamma-\beta}{3}\right)\times 42.98\,\text{arcsec}/100\,\text{yr}\,,\end{aligned} \tag{206}$$

where P and e are the period and eccentricity of the orbit; the second line is the value for Mercury.
Deflection of light:

$$\begin{aligned}\Delta\theta &= \left(\frac{1+\gamma}{2}\right)\frac{4GM}{c^2 d}\\ &= \left(\frac{1+\gamma}{2}\right)\times 1.7504\frac{R_\odot}{d}\,\text{as}\,.\end{aligned} \tag{207}$$

Shapiro time delay:

$$c(t_O - t_e) = |\mathbf{x}_O - \mathbf{x}_e| + \left(\frac{1+\gamma}{2}\right)\frac{2GM}{c^2}\ln\left[\frac{(r_O + \mathbf{x}_O\cdot\mathbf{n})(r_e - \mathbf{x}_e\cdot\mathbf{n})}{d^2}\right]. \quad (208)$$

Nordtvedt effect:

$$\frac{m_G - m_I}{m_I} = \left(4\beta - \gamma - 3 - \frac{10}{3}\xi - \alpha_1 - \frac{2}{3}\alpha_2 - \frac{2}{3}\zeta_1 - \frac{1}{3}\zeta_2\right)\frac{|E_g|}{m_I c^2}, \quad (209)$$

where m_G and m_I are the gravitational and inertial masses of a body such as the Earth or Moon, and E_g is its gravitational binding energy. A nonzero Nordtvedt effect would cause the Earth and Moon to fall with a different acceleration toward the Sun. This effect does not occur in general relativity, a result of the *Strong Equivalence Principle*, satisfied by that theory. The chapter "Probing Gravity with Next Generation Lunar Laser Ranging" by S. Dell'Agnello et al., discusses how lunar laser ranging has set a strong bound on this effect.

Precession of a gyroscope:

$$\begin{aligned}\Omega_{FD} &= -\frac{1}{2}\left(1+\gamma+\frac{\alpha_1}{4}\right)\frac{G}{r^3c^2}\left[\mathbf{J} - 3\mathbf{n}(\mathbf{n}\cdot\mathbf{J})\right]\\ &= \frac{1}{2}\left(1+\gamma+\frac{\alpha_1}{4}\right)\times 0.041\ \text{arcsec yr}^{-1},\\ \Omega_{Geo} &= -\frac{1}{2}(1+2\gamma)\mathbf{v}\times\frac{Gm\mathbf{n}}{r^2c^2}\\ &= \frac{1}{3}(1+2\gamma)\times 6.6\ \text{arcsec yr}^{-1},\end{aligned} \quad (210)$$

where Ω_{FD} and Ω_{Geo} are the precession angular velocities caused by the dragging of inertial frames (Lense–Thirring effect), and by the geodetic effect, a combination of Thomas precession and precession induced by spatial curvature; J is the angular momentum of the Earth, and $\mathbf{v}$, $\mathbf{n}$, and r are the velocity, direction, and distance of the gyroscope. The second line in each case is the corresponding value for a gyroscope in polar Earth orbit at about 650 km altitude.

A wide range of experiments have led to tight bounds on the PPN parameters, all consistent with general relativity. The current values are listed in Table 3.

Gravity Probe B

On May 4, 2011, NASA announced the long-awaited results of Gravity Probe B (GP-B). Over 47 years and 750 million $ in the making, GP-B was an orbiting physics experiment, designed to measure the frame-dragging and geodetic effects of Eq. (210). Josef Lense and Hans Thirring first pointed out the existence of the frame-dragging phenomenon in 1918, but it was not until the 1960s that George Pugh in the Defense Department and Leonard Schiff at Stanford independently pursued the idea of measuring it with gyroscopes.

Table 3 Current limits on the PPN parameters

Parameter	Effect	Limit	Remarks
$\gamma - 1$	(i) Time delay	2.3×10^{-5}	Cassini tracking
	(ii) Light deflection	4×10^{-4}	VLBI
$\beta - 1$	(i) Perihelion shift	3×10^{-3}	Assumes $J_2 = 10^{-7}$ from helioseismology
	(ii) Nordtvedt effect	2.3×10^{-4}	$\eta = 4\beta - \gamma - 3$ assumed
ξ	Earth tides	10^{-3}	Gravimeter data
α_1	Orbital polarization	10^{-4}	Lunar laser ranging
		4×10^{-5}	PSR J1738+0333
α_2	Solar spin precession	4×10^{-7}	Alignment of Sun and ecliptic
α_3	Pulsar acceleration	2×10^{-20}	Pulsar $\dot{P}$ statistics
η^{a}	Nordtvedt effect	9×10^{-4}	Lunar laser ranging
ζ_1	–	2×10^{-2}	Combined PPN bounds
ζ_2	Binary motion	4×10^{-5}	$\ddot{P}_p$ for PSR 1913+16
ζ_3	Newton's 3rd law	10^{-8}	Lunar acceleration
ζ_4	–	–	Not independent

[a]Here $\eta = 4\beta - \gamma - 3 - 10\xi/3 - \alpha_1 - 2\alpha_2/3 - 2\zeta_1/3 - \zeta_2/3$

PPN parametrized post-Newtonian, *VLBI* very-long-baseline radio interferometry

GP-B started officially in late 1963 when NASA funded the initial research and development (R&D) work that identified the new technologies needed to make such a difficult measurement possible. Francis Everitt became Principal Investigator of GP-B in 1981, and the project moved to the mission design phase in 1984. Following a major review of the program by a National Academy of Sciences committee in 1994, GP-B was approved for flight development, and began to collaborate with the Lockheed-Martin and Marshall Space Flight Center. The satellite launched on April 20, 2004 for a planned 16-month mission, but another five years of data analysis were needed to tease out the effects of relativity from a background of other disturbances of the gyros[1].

There were four gyroscopes aboard the GP-B satellite, launched into a polar orbit with an altitude of 640 km above the Earth's surface. Each gyroscope was a fused silica rotor, about the size of a ping-pong ball, machined to be spherical and homogeneous to tolerances better than a part per million, and coated with a thin film of niobium. The gyroscope assembly, which sat in a dewar of 2440 l of superfluid helium, was held at 1.8 K. At this temperature, niobium is a superconductor, and the supercurrents in the niobium of each spinning rotor produce a "London" magnetic moment parallel to its spin axis. Extremely sensitive magnetometers (superconducting quantum interference devices, or SQUIDs) attached to the gyroscope housing

[1] From 1998 to 2011, CMW chaired the NASA Science Advisory Committee for GP-B.

were capable of detecting minute changes in the orientation of the gyros' magnetic moments and hence the precessions predicted by general relativity.

At the start of the mission, the four gyros were aligned to spin along the symmetry axis of the spacecraft. This was also the optical axis of a telescope directly mounted on the end of the structure housing the rotors. Spacecraft thrusters oriented the telescope to point precisely toward the star IM Pegasi (HR 8703) in our galaxy (except when the Earth intervened, once per orbit). In order to average out numerous unwanted torques on the gyros, the spacecraft rotated about its axis once every 78 s.

Almost every aspect of the spacecraft, its subsystems, and the science instrumentation performed extremely well, some far better than expected. Still, success relied on figuring out the sources of error. In particular, having an accurate calibration of the electronic readout from the SQUID magnetometers with respect to the tilt of the gyros was essential. The plan for calibrating the SQUIDs was to exploit the aberration of starlight, which causes a precisely calculable misalignment between the rotors and the telescope as the latter shifts its pointing toward the guide star by up to 20 as to compensate for the orbital motion of the spacecraft and the Earth. However, three important, but unexpected, phenomena were discovered during the experiment that affected the accuracy of the results.

First, because each rotor is not exactly spherical, its principal axis rotates around its spin axis with a period of several hours, with a fixed angle between the two axes. This is the familiar "polhode" motion of a spinning top. In fact, this polhoding was essential in the calibration process because it led to modulations of the SQUID output via the residual trapped magnetic flux on each rotor (about 1 % of the London moment). But the polhode period and angle of each rotor were observed to decrease monotonically with time, implying the presence of some damping mechanism, and this significantly complicated the calibration analysis. In addition, each rotor was found to make occasional, seemingly random "jumps" in its orientation—some as large as 100 milliarcseconds. Some rotors displayed more frequent jumps than others. Without being able to continuously monitor the rotors' orientation, the GP-B team could not fully exploit the calibrating effect of the stellar aberration in their analysis. Finally, during a planned 40-day, end-of-mission calibration phase, the team discovered that when the spacecraft was deliberately pointed away from the guide star by a large angle, the misalignment induced much larger torques on the rotors than expected. From this, they inferred that even the very small misalignments that occurred during the science phase of the mission induced torques that were probably several hundred times larger than the designers had estimated.

What ensued during the data analysis phase following the mission was worthy of a detective novel. The critical clue came from the calibration tests. Here, they took advantage of the residual trapped magnetic flux on the gyroscope. (The designers used superconducting lead shielding to suppress stray fields before they cooled the niobium coated gyroscopes, but no shielding is ever perfect.) This flux adds a periodic modulation to the SQUID output, which the team used to figure out the phase and polhode angle of each rotor throughout the mission. This helped them to conclude that interactions between random patches of electrostatic potential fixed to the

Table 4 Final results of Gravity Probe B

Effect (mas/yr)	Measured	Predicted
Geodetic precession	6602 ± 18	6606
Frame-dragging	37.2 ± 7.2	39.2

surface of each rotor, and similar patches on the inner surface of its spherical housing, were causing the extraneous torques. In principle, the rolling spacecraft should have suppressed these effects, but they were larger than expected.

Fortunately, the patches are fixed on the various surfaces, and so it was possible to build a parametrized model of the patches on both surfaces using multipole expansions, and to calculate the torques induced by those interactions when the spin and spacecraft axes are misaligned, as a function of the parameters. One prediction of the model was that the induced torque should be perpendicular to the plane formed by the two axes, and this was clearly seen in the data. Another prediction of the model was that, when the slowing decreasing polhode period crosses an integer multiple of the spacecraft roll period, the torques fail to average over the roll period, whereupon the spin axis precesses about its initial direction in an opening Cornu spiral, then migrates to a new direction along a closing Cornu spiral. This is known as a loxodromic path, familiar to navigators as a path of fixed bearing on the Earth's surface. Detailed observation of the orientation of the rotors during such "resonant jumps" showed just such loxodromic behavior. In the end, every jump of every rotor could be identified by its "mode number," the integer relating its polhode period to the spacecraft roll period.

The original goal of GP-B was to measure the frame-dragging precession with an accuracy of 1 %, but the problems discovered over the course of the mission dashed the initial optimism that this was possible. Although the GP-B team was able to model the effects of the patches, they had to pay the price of the increase in error that comes from using a model with so many parameters. The experiment uncertainty quoted in the final result—roughly 20 % for frame dragging—is almost totally dominated by those errors (see Table 4). Nevertheless, after the model was applied to each rotor, all four gyros showed consistent relativistic precessions. Gyro 2 was particularly "unlucky"—it had the largest uncertainties because it suffered the most resonant jumps. Numerous cross-checks were carried out, including estimating the relativity effect during different segments of the 12-month science phase (various events, including computer reboots and a massive solar storm in January 2005, caused brief interruptions in data-taking), increasing and decreasing the number of parameters in the torque model, and so on.

The most serious competition for the results from GP-B comes from the LAGEOS experiment, in which laser ranging accurately tracked the paths of two laser geodynamics satellites orbiting the Earth. Relativistic frame dragging induces a small precession (around 30 milliarcseconds per year) of the orbital plane of each satellite in the direction of the Earth's rotation. However, the competing Newtonian effect of the Earth's nonspherical shape (Eq. (84)) had to be subtracted to very high precision using a model of the Earth's gravity field. The first published result

from LAGEOS in 1996 quoted an error for the frame-dragging measurement of 20–30 %, though this result was likely too optimistic given the quality of the gravity models available at the time. Later, the GRACE geodesy mission offered dramatically improved Earth gravity models, and the analysis of the LAGEOS satellites has finally yielded tests at a quoted level of approximately 10 %. In the chapter "Fundamental Physics with the LAGEOS Satellites" by R. Peron, the LAGEOS measurements are described in more detail.

5.5 Gravitational Radiation Reaction

We return to general relativity, and focus our attention on the odd-post-Newtonian-order terms in $N_{\mathcal{N}}$, $K^i_{\mathcal{N}}$, and $B^{ij}_{\mathcal{N}}$. We have already argued that the 1.5PN term proportional to $d^3\mathcal{I}^{kk}(t)/dt^3$ in $N_{\mathcal{N}}$ is a pure coordinate artifact to lowest order. However, the next odd-order terms in these potentials are of 2.5PN order and are nontrivial. They are responsible for the loss of energy, momentum, and angular momentum in a dynamical system as the result of the radiation of gravitational waves. After some work, it can be shown that these terms generate an additional radiation-reaction force to the post-Newtonian Euler equation of hydrodynamics (191), given by

$$\rho^* \frac{dv^j}{dt} = \rho^* \partial_j U - \partial_j p + O(c^{-2}) + O(c^{-4}) + c^{-5} f^j_{\rm RR} + O(c^{-6}), \tag{211}$$

where

$$f^j_{\rm RR} = \frac{G}{c^5}\Big[-\rho^* \overset{(3)}{\mathcal{I}}{}^{pq}\, \partial_{jpq} X + 2\left(\rho^* \partial_k U - \partial_k p\right) \overset{(3)}{\mathcal{I}}{}^{jk} + \frac{2}{3}\left(2\rho^* \partial_j U - \partial_j p\right) \overset{(3)}{\mathcal{I}}{}^{pp} + 2\rho^* \overset{(4)}{\mathcal{I}}{}^{jk}\, v_k + \frac{3}{5}\rho^* \overset{(5)}{\mathcal{I}}{}^{\langle jk\rangle}\, x^k - \frac{2}{15}\rho^* \overset{(5)}{\mathcal{I}}{}^{jpp} + \frac{2}{3}\rho^* \varepsilon^{pjq} \overset{(4)}{\mathcal{J}}{}^{pq} \Big]. \tag{212}$$

The $O(c^{-2})$ symbol represents the post-Newtonian terms shown in Eq. (191), and the $O(c^{-4})$ symbol represents 2PN terms not shown here. The radiation reaction force is itself coordinate dependent; one can change the coordinates by 2.5PN-order corrections so that the 1PN and 2PN terms are unaffected, but the radiation-reaction force is transformed into the simple form

$$f^j_{\rm RR} = -\frac{2G}{5c^5} \rho^* \overset{(5)}{\mathcal{I}}{}^{\langle jk\rangle}\, x^k . \tag{213}$$

This is known as the Burke–Thorne gauge.

The Newtonian, 1PN, and 2PN contributions to the equations of motion are *conservative*, in that they enforce the conservation of total energy, momentum, and angular momentum for an isolated system. The energy is given by

$$E = \mathcal{T} + \Omega + E_{\rm int} + O(c^{-2}) + O(c^{-4}), \tag{214}$$

where $\mathcal{T}$, Ω, and $E_{\rm int}$ are defined just as in Eq. (18), but using ρ^*. Now, however, including the radiation reaction force, one finds that

$$\frac{dE}{dt} = \int \rho^* v^j f^j_{\rm RR} d^3x = -\frac{2G}{5c^5} \overset{(5)}{\mathcal{I}}{}^{\langle jk\rangle} \int \rho^* x^k v^j d^3x = -\frac{G}{5c^5} \overset{(5)}{\mathcal{I}}{}^{\langle jk\rangle} \dot{\mathcal{I}}^{jk} . \quad (215)$$

However this last expression can be written as

$$\frac{dE}{dt} = -\frac{G}{5c^5} \overset{(3)}{\mathcal{I}}{}^{\langle jk\rangle} \overset{(3)}{\mathcal{I}}{}^{\langle jk\rangle} + \frac{d}{dt}\frac{G}{c^5}(\dots) . \quad (216)$$

But because E is only well-defined to even post-Newtonian orders, the time-derivative term on the right-hand side of Eq. (216) can be moved to the left-hand side and absorbed into a modified energy that has been corrected by meaningless 2.5PN order terms. The final result is

$$\frac{dE}{dt} = -\frac{G}{5c^5} \overset{(3)}{\mathcal{I}}{}^{\langle jk\rangle} \overset{(3)}{\mathcal{I}}{}^{\langle jk\rangle} . \quad (217)$$

We will see this exactly equals the energy loss as measured in the far zone. Similar arguments can be made for the effects of radiation reaction on linear momentum and angular momentum.

6 Far-zone Fields and Gravitational Radiation

6.1 The Quadrupole Formula

We now turn to the faraway wave-zone regime of our formal solution $h^{\alpha\beta}_{\mathcal{N}}(t, \mathbf{x})$ of the relaxed Einstein's equation, given by Eq. (168). We first notice that this has the general form of $h \sim r^{-1} f(\tau)$, where $\tau = t - r/c$. As a consequence, it is simple to deduce that $\partial_j h = -(n_j/c)\dot{h} + O(r^{-2})$. Combining this fact with the harmonic gauge condition $\partial_\beta h^{\alpha\beta} = 0$, we can show that

$$\begin{aligned} \dot{h}^{0j} &= n_k \dot{h}^{jk} + O(r^{-2}) , \\ \dot{h}^{00} &= n_j n_k \dot{h}^{jk} + O(r^{-2}) , \end{aligned} \quad (218)$$

and consequently, in the faraway wave zone, a knowledge of h^{jk} is enough to determine the other components of the field, modulo terms that are constant in time. For gravitational waves, of course, this is all one needs.

The leading $\ell = 0$ term in the expression (168) for h^{jk} is given by

$$h^{jk}_{\mathcal{N}} = \frac{4G}{c^4 r} \int_{\mathcal{M}} \tau^{jk}(\tau, \mathbf{x}') d^3x' + O(r^{-2}) . \quad (219)$$

Making use of the identity (178b), and discarding the surface term, we can rewrite this in the form

$$h^{jk}_{\mathcal{N}} = \frac{2G}{c^4 r}\ddot{\mathcal{I}}^{jk}(\tau) + O(r^{-2})\,, \tag{220}$$

where $\mathcal{I}^{jk}$ is given by Eq. (177a), evaluated at the retarded time $\tau = t - r/c$.

This is the famous "quadrupole formula" of general relativity, that gives the gravitational waveform in terms of two time derivatives of the source quadrupole moment. During the middle 1970s, there was considerable debate over the validity of the derivation of this formula, largely because of the fact that the effective source $\tau^{\alpha\beta}$ extends over all space-time. The questions raised included the possibility that higher moments of the effective source $\tau^{\alpha\beta}$ could diverge if the integrals were taken to infinity, or that various surface integrals, such as the one involved in going from Eqs. (219) to (220), could actually fail to vanish. Ultimately, these concerns were laid to rest. One approach involved recalling that the formal solution is actually given by Eq. (164). Since $h^{\alpha\beta}_{\mathcal{N}}$ is an integral over a finite spatial volume—the near zone—it is by definition finite. Furthermore, it could be shown that the integration over the far zone $h^{\alpha\beta}_{\mathcal{W}}$ was likewise finite. In addition, any dependence of each integral on the arbitrarily chosen radius $\mathcal{R}$ of the boundary between the two zones was shown to cancel term-by-term, order-by-order, in the sum $h^{\alpha\beta}_{\mathcal{N}} + h^{\alpha\beta}_{\mathcal{W}}$. Similar arguments were applied to the near-zone expressions for the fields, used in the equations of motion.

The field h^{jk} in the faraway wave zone is a central ingredient in determining the response of a gravitational wave detector. From the equation of geodesic deviation (115), one can show that the relative acceleration of a pair of freely moving masses that are nearly at rest in a local freely falling frame is given by

$$\frac{d^2\xi^j}{dt^2} = \frac{1}{2}\ddot{h}^{jk}_{TT}\xi_k\,, \tag{221}$$

where ξ^j is the physically measured displacement between the two bodies, and t is proper time as measured at the origin of the local frame. The subscript "TT" denotes the *transverse-traceless* part, given by

$$h^{jk}_{TT} := h^{lm}\left(P^j_l P^k_m - \frac{1}{2}P^{jk}P_{lm}\right)\,, \tag{222}$$

where P^i_j is a projection tensor, given by

$$P^i_j := \delta^i_j - n^i n_j\,. \tag{223}$$

Gravitational-wave detection is covered in the chapter "The Detection of Gravitational Waves" by S. Braccini and F. Fidecaro.

The field h^{jk} also determines the energy flux radiated to infinity, as expressed by Eq. (154). Examining the expression for $\tau^{\alpha\beta}$ (Eq. (160)) in the faraway zone, and making use of Eq. (218), it is straightforward to show that

$$\tau^{0k}_{\rm LL} = n^k \tau^{00}_{\rm LL} + O(r^{-3})\,. \tag{224}$$

This is not surprising, since for a field propagating in a null direction, the energy flux should be the energy density times a unit vector. Substituting this into Eq. (154), with $dS_k = r^2 n_k d\Omega$ and taking the surface to infinity, we obtain

$$\dot{E} = -c \oint_{\infty} (-g)\tau_{\rm LL}^{00} r^2 d\Omega \,. \tag{225}$$

After some algebra using Eq. (218), it is also possible to show that

$$(-g)\tau_{\rm LL}^{00} = \frac{c^2}{32\pi G} \dot{h}_{TT}^{jk} \dot{h}_{TT}^{jk} + O(r^{-3}) \,. \tag{226}$$

Using the fact that $P_k^i P_j^k = P_j^i$, and $P_i^i = 2$, together with Eq. (220), we obtain

$$\dot{E} = -\frac{G}{2c^5} \overset{(3)}{\mathcal{I}}{}^{lm} \overset{(3)}{\mathcal{I}}{}^{pq} \langle\langle P_{lp} P_{mq} - \frac{1}{2} P_{lm} P_{pq} \rangle\rangle \,, \tag{227}$$

where $\langle\langle \ldots \rangle\rangle$ denote the angular average, defined in Eq. (33). Making use of Eq. (34), we arrive finally at

$$\frac{dE}{dt} = -\frac{G}{5c^5} \overset{(3)}{\mathcal{I}}{}^{\langle jk \rangle} \overset{(3)}{\mathcal{I}}{}^{\langle jk \rangle} \,, \tag{228}$$

which is precisely the same as Eq. (217) for the energy lost as a consequence of radiation reaction forces.

6.2 Energy Flux and Inspiraling Compact Binaries

For a binary star system, $\mathcal{I}^{jk}$ is given to lowest order by

$$\mathcal{I}^{jk} = \mu r^j r^k \,, \tag{229}$$

where $\mu = m_1 m_2/(m_1 + m_2)$ is the reduced mass, and $\mathbf{r}(t) = \mathbf{x}_1 - \mathbf{x}_2$. Calculating the three time derivatives, and using the Newtonian equation of motion in place of $d\mathbf{v}/dt$, we obtain

$$\frac{dE}{dt} = -\frac{8G^3}{15c^5} \eta^2 \left(\frac{m}{r}\right)^4 \left(12v^2 - 11\dot{r}^2\right) \,, \tag{230}$$

where $\eta := \mu/m$. Notice that the fractional change in energy over one orbital period, $\Delta E/E \sim (dE/dt)(P/E)$, can be written, using the fact that $P \sim r^{3/2}/(Gm)^{1/2}$, and that $E \sim \mu v^2 \sim G\mu m/r$, in the schematic form

$$\frac{\Delta E}{E} \sim \frac{r^{3/2}}{(Gm)^{1/2}} \left(\frac{r}{G\mu m}\right) \frac{8G^3}{15c^5} \eta^2 \left(\frac{m}{r}\right)^4 \frac{Gm}{r} \sim \eta \left(\frac{Gm}{c^2 r}\right)^{5/2} \,, \tag{231}$$

indicating clearly that radiation damping is a 2.5PN-order effect. Using our Keplerian orbit formulae from Sect. 2, and averaging the energy loss over one orbit of the binary system, we obtain the average rate of energy loss,

$$\frac{dE}{dt} = -\frac{32G^4}{5c^5}\eta^2\left(\frac{m}{a}\right)^5 F(e)\,, \tag{232}$$

where a is the semimajor axis, e is the eccentricity, and $F(e)$ is given by

$$F(e) := \frac{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4}{(1-e^2)^{7/2}}\,. \tag{233}$$

It is useful to express these results in terms of the dimensionless variable x, given by

$$x := \left(\frac{2\pi}{P}\frac{Gm}{c^3}\right)^{2/3}\,. \tag{234}$$

The quantity in parentheses is the ratio of the light travel time across a distance given by Gm/c^2 to the orbit period; it is easy to see that $x = Gm/ac^2 \sim (v/c)^2$. In terms of x, we obtain

$$\frac{dE}{dt} = -\frac{32c^5}{5G}\eta^2 x^5 F(e)\,. \tag{235}$$

Since $E \propto a^{-1}$ and $P \propto a^{3/2}$, we have that $2P^{-1}dP/dt = -3E^{-1}dE/dt$. Thus we can obtain an expression for the rate of decrease of the orbital period,

$$\frac{dP}{dt} = -\frac{96}{5}\eta x^{5/2} F(e)\,. \tag{236}$$

This prediction of the damping of an orbit caused by gravitational radiation reaction has been confirmed spectacularly, first using the Hulse–Taylor binary pulsar PSR 1913+16, discovered in 1974, and later in a number of similar binary pulsar systems, notably the "double pulsar," J0737-3039, discovered in 2003. For more on binary pulsar observations, see the chapter "The Role of Binary Pulsars in Testing Gravity Theories" by A. Possenti and M. Burgay.

The Hulse–Taylor binary and the double-pulsar binary are destined to have their orbits shrink to such a state that the two stars will merge to form a black hole, on a timescale that is short compared to the age of the galaxy. If systems such as this exist in the final stage of inspiral and merger today within several hundred megaparsecs, the waves they emit should be detectable by the advanced ground-based laser interferometers in the LIGO-Virgo-Geo network. In a similar vein, the waves emitted during the inspiral and merger of a pair of supermassive black holes should be detectable by a space-based interferometer such as LISA, out to very large redshifts.

However, in order to enhance the detection confidence, and to extract useful astrophysical information from the detected wave trains, it has been shown that predictions for the motion and gravitational wave field will be needed that go far

beyond the lowest, Newtonian, and post-Newtonian approximations that we have discussed in this chapter. With this motivation, many groups have calculated the higher-order corrections, mainly using techniques based on the post-Minkowskian formalism presented here. The equations of motion for spinless binaries are now known through 3.5PN order, and the effects of spin have also been calculated to rather high post-Newtonian order. Similarly, the gravitational waveform and energy flux have been calculated through 3.5PN order *beyond* the quadrupole formula shown here; again many effects of spin have also been calculated. Work has also begun on the extraordinarily difficult 4PN contributions to both the motion and the waveform. To illustrate what these results look like, we quote here the energy flux, expressed as an expansion in powers of the variable x, for spinless binaries in circular orbits, through 3.5PN order:

$$\begin{aligned}\frac{dE}{dt} = -\frac{32c^5}{5G}\eta^2 x^5 \Bigg\{ & 1 - \left(\frac{1247}{336} + \frac{35}{12}\eta\right) x + 4\pi x^{3/2} \\ & - \left(\frac{44711}{9072} - \frac{9271}{504}\eta - \frac{65}{18}\eta^2\right) x^2 - \left(\frac{8191}{672} + \frac{583}{24}\eta\right)\pi x^{5/2} \\ & + \left[\frac{6643739519}{69854400} + \frac{16}{3}\pi^2 - \frac{1712}{105}C - \frac{846}{105}\ln(16x)\right. \\ & \left. - \left(\frac{134543}{7776} - \frac{41}{48}\pi^2\right)\eta - \frac{94403}{3024}\eta^2 - \frac{775}{324}\eta^3\right] x^3 \\ & - \left(\frac{16285}{504} - \frac{214745}{1728}\eta - \frac{193385}{3024}\eta^2\right)\pi x^{7/2} + O(c^{-8})\Bigg\}, \end{aligned} \tag{237}$$

where $C = 0.577\ldots$ is the Euler constant. The first term in this sequence is the leading quadrupole term of Eq. (235). Results like these will play a key role in the development of gravitational wave astronomy.

Acknowledgements This work has been supported by the National Science Foundation under grant PHY 09-65133. Much of this work was completed while the author was at Washington University in St. Louis. I am also grateful to the Institut d'Astrophysique de Paris, where part of this work was completed. This chapter is based in part on the forthcoming book *Gravitation: Newtonian, post-Newtonian, Relativistic*, by Eric Poisson and Clifford Will. I am grateful to my coauthor and our publisher Cambridge University Press for permission to use parts of that text in the present chapter.

References

1. Blanchet L. Gravitational radiation from post-Newtonian sources and inspiralling compact binaries. Living Rev Relativ. 2014;17:2 (cited on 1 July 2014): http://www.livingreviews.org/lrr-2014-2.
2. Merkowitz SM. Tests of gravity using lunar laser ranging. Living Rev Relativ. 2010;13:7. (cited on 1 July 2012): http://www.livingreviews.org/lrr-2010-7.

3. Poisson E, Will CM. Gravity: Newtonian, post-Newtonian, relativistic. Cambridge: Cambridge University Press; in 2014.
4. Sathyaprakash BS, Schutz BF. Physics, astrophysics and cosmology with gravitational waves. Living Rev Relativ. 2009;12:2. (cited on 1 July 2012): http://www.livingreviews.org/lrr-2009-2.
5. Soffel MH. Relativity in astrometry, celestial mechanics and geodesy. Berlin: Springer-Verlag; 1989.
6. Stairs IH. Testing general relativity with pulsar timing. Living Rev Relativ. 2003;6:5. (cited on 1 July 2012): http://www.livingreviews.org/lrr-2003-5.
7. Turyshev SG. Experimental tests of general relativity. Ann Rev Nucl Particle Sci. 2008;58:207. (arXiv:0806.1731).
8. Will CM. Theory and experiment in gravitational physics. Cambridge: Cambridge University Press; 1993.
9. Will CM. The confrontation between general relativity and experiment. Living Rev Relativ. 2014;17:4 (cited on 1 July 2014): http://www.livingreviews.org/lrr-2014-4
10. Will CM. Resource letter PTG-1: precision tests of gravity. Am J Phys. 2010;78:1240. (arXiv:1008.0296).
11. Will CM. On the unreasonable effectiveness of the post-Newtonian approximation in gravitational physics. Proc Natl Acad Sci U S A. 2011;108:5938. (arXiv:1102.5192).

The Newtonian Gravity and Some of Its Classical Tests

Valerio A. Iafolla

Dedicated to Franco Fuligni, who has been a dear friend and colleague, in memory of time spent together working at the main part of the activities described in this chapter.

Abstract In this chapter, the historical evolution of the concept of gravitation, with its connections to the observations of natural phenomena and to the related experiments, is presented. It starts with a brief description of the first stage of the experiments and continues giving details on the work related to it. The smallness of the gravitational effects to be detected, along with the spurious effects (presence of Earth's gravity and environmental noise) that make the measurement difficult and the techniques most suitable for their detection are shown. In the last part of this chapter, some gravitational experiments performed in space and on ground are described.

1 Introduction

The studies related to gravitation, and the sequence of theories that have been developed to explain the physical phenomena observed in nature or produced in the laboratory are a good example of how the scientific process enables us to elaborate the concepts and the general relationships that allow us to understand an enormous variety of experiences and to make them accessible to mathematical methods.

In this chapter, we present a review of the techniques concerning classical experiments related to gravitation that, together with the observation of events in nature and from a priori hypotheses, lead to the evolution of the concept of gravitation toward more and more complex theories that are able to explain all these physical phenomena and to foresee new ones.

V. A. Iafolla (✉)
Istituto di Astrofisica e Planetologia Spaziali (IAPS), Istituto Nazionale di Astrofisica (INAF), Via Fosso del Cavaliere 100, Roma, Italy
e-mail: valerio.iafolla@iaps.inaf.it.

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_3

In the first section, we give a brief presentation of the historical evolution of the concept of gravitation, with its connections to the observations of natural phenomena and the earlier experiments to test the assumptions made; also presented are the experiments conducted by Galileo Galilei (GG) to support his a priori assumptions concerning the universality of free fall, in connection to his idea of the scientific method.

In the second section, we go into the details of the work related to experimental gravitation. We discuss the small size of the effects to be detected, along with the spurious effects (presence of Earth's gravity and environmental noise) that make their measurement difficult. The techniques most suitable for the detection of the gravitational signals are described, along with their maximization, in order to avoid the noise that would prevent the measurement itself. In this section, we also present a survey of possible sites on which to run experiments of gravitation in the most profitable mode (underground laboratories, free-fall systems, on board satellites).

In the third section, the techniques adopted to calibrate apparatus for the gravitational experiments are described, taking as reference the calibration of an accelerometer for space use.

In the following final three sections are described three main experiments to test gravitational theories.

In the fourth section, we describe an experimental test of the weak equivalence principle (WEP). This experiment, named General Relativity Accuracy Test (GReAT), aims to obtain an accuracy of 5 x 10^{-15} in free-fall conditions from stratospheric altitude.

In the fifth section, we describe the measurement of G—the universal constant of gravitation. G is the oldest known constant of nature; it was introduced at the end of the 1600s by Newton in his law of gravitation, but it was measured for the first time only in 1798 by Henry Cavendish, and it remains still now the least well-known constant of nature, due to the very high difficulty of this measurement.

In the sixth section, the radio science experiments (RSE) are described. These are intended for the European Space Agency (ESA) cornerstone mission denominated "BepiColombo" for the exploration of Mercury. The RSE will use tracking techniques to measure the position and velocity of the satellite orbiting around Mercury; these measurements, together with the measure of the perturbing acceleration noise performed with a high precise accelerometer, are used in a high accurate orbit determination of the satellite with a simultaneous evaluation of gravitational field parameters, useful to determine the internal structure of the planet, its rotational state, and to test the theories of gravitation.

2 Historical Part

In the past, the concept of gravitation was assimilated to the idea of Aristotle (384−322 BC) in which each of the four elements (Earth, Water, Air, and Fire), already introduced by Empedocles, tends to return to its natural place when it is moved far away from it. In particular, "the land" with its qualities of cold and dry,

representative of the modern concept of solid, tends toward the center of the universe, which is also the point where the Earth seen as a planet is placed. There is no doubt that many of the Aristotelian concepts such as the concept of cosmos, of the movement, fate, time, space, etc... have a close connection with the modern concepts, in particular, bringing back to cosmology and to gravitation the power to regulate everything.

In particular, in Aristotle's discussion the movement of things is well differentiated from the movement of planets and stars that move in an unchanging aether that pervades the cosmos.

The Aristotelian concept of the cosmos and the elements that compose it, together with their interactions persisted for several millennia, in particular the Ptolemaic structure (Ptolemy (100–175 AD)) of the geocentric solar system, as he expressed it in the Almagest, stood out until the work of Copernicus (1473–1543) did not bring away the Earth out of its position at the center of the universe. To the dissemination of this new revolutionary idea of the cosmos contributed Galileo (1564–1642) with his astronomical observations, and through his writings (Dialogue Concerning the Two Chief World Systems, and Discourses and Mathematical Demonstrations Relating to Two New Sciences [21]), giving a substantial input to the subsequent work of Isaac Newton (1642–1727) who, with his laws of motion and that of universal gravitation, determined one of the greatest works in the unification of concepts that physicists already knew but had not related to each other. A non-exhaustive list of physical aspects explained as solutions of his famous equations could be as follows:

- Copernican ideas
- Kepler's Laws
- Universality of free fall
- Quadratic proportionality between space and time for the free-fall bodies
- Ballistics problem
- Tides

In particular, he translated the observation about the universality of free fall of bodies made by Galilei in the equivalence between inertial and gravitational mass, leaving to Albert Einstein (1879–1955) the task to formulate his principle of equivalence between a gravitational field and an accelerated reference frame, which has at its basis the principle of equivalence of Newton.

In this simple summary of the way toward an understanding of the ideas of the world around us, we should not miss the salient aspects of the processes that led to the formulation of theories of gravitation, which, starting from the classical Newtonian ideas have been evolving toward Einstein's General Relativity (currently the most accepted theory) and the alternatives to it, as well as attempts to formulate quantum theories of gravitation and theories of unification of the gravitational force to the other three known forces of nature. The process has been a trend at jumps caused by so many insights that can come out only by having a base consisting of information from previous work done by a large group of scientists. Galileo's insight expressed in his phrase "I fell in opinion that if the resistance of medium will be removed totally all the bodies fall down with equal speed," is a process analogous to that made by Newton in its observation of the fall of his famous apple and

not dissimilar from what must have happened in the mind of Einstein in transforming the gravitational force in a deformation of space-time, determined by the masses themselves.

In the following section, we include a brief description of the research on universality of free fall of bodies made by Galilei, starting from his a priori intuition and then passing to the description of his forgotten experiment aimed at experimental verification of his intuition, performed with the use of pendulums (and not throwing two bodies from the tower of Pisa).

2.1 Galilei and the Universality of the Free Fall

This section shows the original work done by Galilei in relation to his intuition of the universality of free fall of bodies and the subsequent experimental work, aimed at verifying this intuition (see also Refs. [1, 2]). In particular, it describes the experiment conducted using two pendulums of different materials and observing the free fall of these masses along the trajectory of oscillation, in what could be the first experiment of "zero," or comparison of two physical systems (two pendulums) in the history of science. This experiment was subsequently repeated by Newton and Bessel. The historical value of the pages herein reported is to give importance to Galilei and his scientific method, by removing a historical error, inherent in the scientific community that attributes to him, as evidence of his insight, the experiment of falling bodies from the tower of Pisa, but he clearly indicates that it is not feasible. What has been said is completely clarified by the following three excerpts taken from its original work "Discourses and Mathematical Demonstrations Concerning Two New Sciences," in particular, they are taken from the English edition, translated by H. Crew and A. De Salvio [21].

In the first phrase, there is a clear definition of the hypotheses with respect to the universality of free fall and the Galilean idea about the scientific method to be used for testing the hypotheses (Figs. 1, 2, and 3).

Simple estimates allow us to affirm that the verification of the universality of free fall of bodies, as performed by Galilei in his experiment with pendulums, made in terms of equivalence of inertial and gravitational mass, led to an accuracy of one part in 10^3.

SALV. The facts set forth by me up to this point and, in particular, the one which shows that difference of weight, even when very great, is without effect in changing the speed of falling bodies, so that as far as weight is concerned they all fall with equal speed: this idea is, I say, so new, and at first glance so remote from fact, that if we do not have the means of making it just as clear as sunlight, it had better not be mentioned; but having once allowed it to pass my lips I must neglect no experiment or argument to establish it.

Fig. 1 Statement of the universality of free fall of all bodies

"The experiment made to ascertain whether two bodies, differing in weight will fall from a given height with the same speed offers some difficulty; because, if the height is considerable, the retarding effect of the medium,...., will be greater in the case of the small momentum of the very light body than in the case of the great force of the heavy body; so that, in a long distance, the light body will be left behind;..."

Fig. 2 Exclusion of the possibility of execution of the verification of the universality of the bodies by means of free fall, if in the presence of air

Accordingly I took two balls, one of lead and one of cork, the former more than a hundred times heavier than the latter, and suspended them by means of two equal fine threads, each four or five cubits long. Pulling each ball aside from the perpendicular, I let them go at the same instant, and they, falling along the circumferences of circles having these equal strings for semi-diameters, passed beyond the perpendicular and returned along the same path. This free vibration [*per lor medesime le andate e le tornate*] repeated a hundred times showed clearly that the heavy body maintains so

[129]

nearly the period of the light body that neither in a hundred swings nor even in a thousand will the former anticipate the latter by as much as a single moment [*minimo momento*], so perfectly do they keep step. We can also observe the effect of the medium which, by the resistance which it offers to motion, diminishes the vibration of the cork more than that of the lead, but without altering the frequency of either; even when the arc traversed by the cork did not exceed five or six degrees while that of the lead was fifty or sixty, the swings were performed in equal times.

Fig. 3 Description of the experiment performed by means of pendulums

3 Experimental Gravitation

In this section are reported considerations related to the difficulties encountered in carrying out tests on gravitation; emphasized the smallness of the signals to be detected and the problems introduced by the presence of the Earth's gravity and to the vibrational and thermal noise, which act on the measurement equipment. In particular, the problem of vibrational noise is dramatic because it acts on the test mass in a manner indistinguishable from gravitational effects, just because of the equivalence of inertial and gravitational mass. From the above it is obvious that an experiment to measure the gravitational effects is equally suitable for the measurement of accelerations acting on the reference in which the experiment is installed. For further information on this topic see [3–9].

The basic concept of an experiment on gravitation to be carried out in the laboratory is the measurement of the effect produced by a source mass on a test mass, determining the force which it undergoes in a reference rigidly connected with the source mass. In the case of a source mass M_T equal to 1 kg, placed at a distance of 1 m with respect to the test mass of the same weight, they attract each other with a force equal to 6.67×10^{-11} N, corresponding at an acceleration equal to 6.67×10^{-11} m/s^2, in the case the test mass is a free mass; in these conditions, the relative displacement of the test mass would be proportional to the acceleration which acts on it and to the time squared. In a time of the order of 1 s, this displacement would be equal to 3.335×10^{-11} m.

In general, the specific measure of the gravity effect is made through the measurement of the displacement of the test mass, and this would suggest that, while waiting for a sufficiently long time, the mass shift becomes large enough to be easily measurable. The difficulty of this type of experiment lies in the fact that it is not easy to put a test mass in the condition of a free mass, since on any ground laboratory there is the presence of Earth's gravity, notably by equal to about 9.8 m/s^2, therefore far superior to the acceleration to be measured. The second problem is related to the need to carry out the measures with respect to a fixed reference and, as well known, is not easy to refer a free mass with respect to a fixed reference. As we shall see, these problems are solved by connecting the mass to the reference with springs extremely soft, so as to constitute a harmonic oscillator of very low frequency, such as to allow large displacements of the test mass and the simultaneous maintenance of the reference. The mass can be suspended by means of a spring so to have the main degree of freedom with a low frequency, and used for the measurement of the gravity, while another degree of freedom, with a higher frequency, that acts against the gravity, so as to cancel its effects. These systems give an idea of how it can be solved the problem of the presence of gravity, but remains open the one tied to the smallness of the signals to be detected. In the following paragraph, the methods and techniques used for the detection of signals are shown; in particular, the procedures to maximize the signal to be detected with respect to acceleration or gravitational force acting on the experimental apparatus are described. Then it follows the description of techniques to reduce and/or circumvent the noise which disturbs the measurement itself and a review of the characteristics of laboratories and systems that allow these measurements in conditions of micro- and pico-gravity. Much of the general descriptions of the detection techniques presented here are closely related to the development activities of the accelerometer called Italian Spring Accelerometer (ISA), performed by the "Experimental Gravitation" team of the IAPS/INAF. The name of the indicated accelerometer is suggested by the fact that its proof mass is connected to its reference frame by means of a soft mechanical spring so to constitute a mechanical oscillator with a single degree of freedom, and therefore able to detect a single component of an acceleration or of a gravitational effect.

3.1 Techniques of Signals Detection

The main constituent parts of each single-axis accelerometer that will be more deeply described in the following section (and in Refs. [10, 11]) are as follows:

- The mechanical oscillator
- The actuation and control
- The signal detection

3.1.1 Mechanical Oscillator

A common problem to different fields of physics, disregarding whether the experiments are carried out in ground-based laboratories or in space-borne enclosures, consists in the detection of small forces or accelerations that act on the proof mass of a harmonic oscillator, producing exceedingly small displacements. When it is possible, harmonic oscillators having a very low resonance frequency are used to magnify this effect (a free mass is the ideal tool); this displacement magnification is in spite of the frequency range of the instrument. The oscillator can be regarded as a test mass connected to the reference frame through a spring with low elastic constant; the accelerations acting on the system can be regarded as inertial accelerations acting on the test mass in the reference frame of the experimental apparatus. To detect these accelerations, it is necessary to measure the consequent displacement of the proof mass with respect to the reference frame (Fig. 4).

In principle, a single test mass can be used to detect the three components of a gravitational force acting directly on the proof mass or a linear acceleration acting on its reference frame, but in the ISA concept there are three different test masses, one for each axis.

Figure 4 shows the mechanical part of the accelerometer. It consists of a central plate, in which the proof mass is connected to an external rigid frame by means of

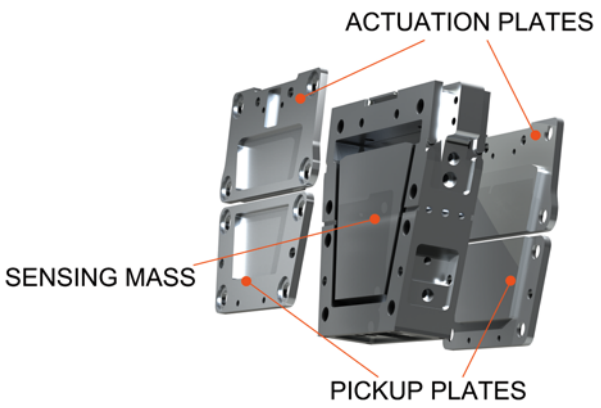


Fig. 4 Accelerometer mechanical structure

an elastic element (flexural or torsional) constituting a harmonic oscillator. Usually the sensitivity axis is perpendicular to the face of the proof mass.

Four additional plates are connected on the opposite sides of the central one and electrically separated by insulating washers, forming plain capacitors. A couple of these capacitors provide the reading of the signal. The other couple has more than one function. It could be used to (i) lower the electromechanical frequency of the oscillator; (ii) bring the capacitive bridge to its equilibrium position, by the application of a constant voltage, as an actuator; and (iii) excite the mechanical oscillator with known electrical signals.

The harmonic mechanical oscillator is described by the equation:

$$I\ddot{\vartheta} + \eta_t \dot{\vartheta} + \kappa_t \vartheta = M, \tag{1}$$

where k_t represents the elastic constant, I is the proof mass moment of inertia, η_t is its dissipation coefficient, and M is the magnitude of the external torque. The formula for the evaluation of the mechanical resonance frequency is as follows:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{\kappa_t}{I}}. \tag{2}$$

As we will see, one important characteristic of the mechanical oscillator, besides its resonance frequency, is represented by its mechanical quality factor, connected to the oscillator Brownian noise.

3.1.2 Pickup and Actuation Systems

Figure 5 shows the electric scheme of the accelerometer. It can be divided into two sections: The first one for the control, where the capacitors are with direct current bias and the second one for the detection of the signal.

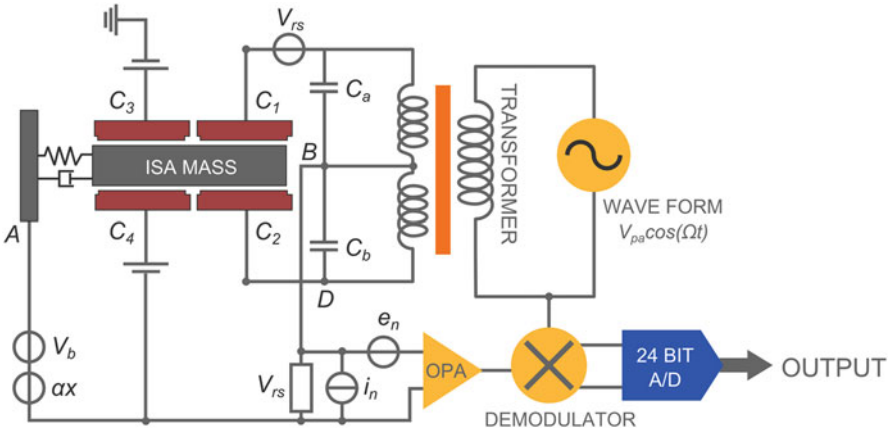


Fig. 5 Schematic drawing of the accelerometer electric scheme

The bridge is biased at a frequency $f_p = 10$ kHz so that accelerations at frequency f_s, acting on the proof mass, cause an unbalance of the capacitive bridge and a modulation of the bias voltage: At the output of the capacitive bridge the signal is seen in frequency at the two side bands, $f_\pm = f_p \pm f_s$. The transfer of the signal to a high frequency allows the amplifier to work at a frequency of 10 kHz where its temperature noise is lower, avoiding its $1/f$ noise. The transducer must have a high electromechanical transducer factor β (ratio between the mechanical energy of the oscillator and the electric power to be measured) and the amplifier must have a temperature noise lower than the energy in the transducer, due to the signal to be measured.

3.1.3 Electrical Lowering of the Frequency of the Mechanical Oscillator

By applying a constant voltage V across one of the control capacitors, a torque is produced that, for small ϑ, is given by

$$M = -\frac{1}{2}V^2\left(-\frac{C_0}{d_0}\frac{b}{2} + \frac{C_0}{{d^2}_0}\left(\frac{b}{2}\right)^2\vartheta\right), \tag{3}$$

where b is the distance between the center of mass (COM) of the central plate and the rotation axis, ϑ is the rotation angle, C_0 is the capacity of the capacitor, when the plate is in its equilibrium position, V is the voltage on it, and d_0 is the distance between the two faces of the planar capacitors. The term depending on ϑ has the effect of a negative elastic constant that reduces the torsion elastic constant of the oscillator, causing a lowering of the system resonance frequency. The new frequency is given by

$$\omega^2 = \frac{k - \frac{V^2 C_0}{{d_0}^2}\left(\frac{b}{2}\right)^2}{I} = {\omega_0}^2\left(1 - \frac{4\beta}{5}\right), \tag{4}$$

where we have introduced the electromechanical transducer factor β:

$$\beta = \frac{C_0 V^2}{\frac{5}{4}m{\omega_0}^2{d_0}^2}. \tag{5}$$

The simultaneous application of the constant voltage on two opposite capacitors gives a factor of 2 in the previous formula and allows maintaining the proof mass in its equilibrium position. We must say that this procedure, essential to increase the accelerometer sensitivity, reduces its frequency band.

3.1.4 Electrical Variation of the Equilibrium Conditions of the Capacitive Bridge

The term in Eq. (3), that does not depend on ϑ, represents a constant torque that changes the position of the proof mass and therefore the four capacitors capacitance

values. Simple evaluations indicate that the value of the capacitor on which the constant voltage V is applied and that of the adjacent capacitor, increase according to the equation:

$$C(V) = C_0\left(1 + \frac{1}{2}\frac{C_0 V^2 b^2}{4k_t {d_0}^2 - V^2 C_0 b^2}\right) = C_0\left(1 + \frac{1}{2}\frac{\frac{4\beta}{5}}{1 - \frac{4\beta}{5}}\right). \tag{6}$$

On the other hand, the capacitors capacity decreases. It is clear that if we want to have a low frequency and get large variations of capacity, we need to have high values of β. We must notice that if $\beta = 1$, the frequency becomes zero, two capacities go to infinity and two to zero (this in principle). This means that the system becomes unstable; the attraction force, determined by the electric field between the two faces of the capacitor, is equal to or larger than the elastic restoring force.

3.1.5 Electromechanical Actuator

To characterize the system, it is useful to have the possibility to excite it at known frequencies. This possibility is obtained using the control capacitors as an actuator. Superimposing an alternate voltage $v = v_0 \cos(2\pi f_e t)$ to the constant voltage V, from Eq. (3) we can see that, besides the terms seen before, a torque M_1 results at frequency f_e given by:

$$M_1 = \frac{C_0}{d_0}\frac{b}{2} v V. \tag{7}$$

At 2 f_e, the torque is equal to:

$$M_2 = \frac{b}{2}\frac{C}{d_0} v^2. \tag{8}$$

Terms causing small variations of the elastic constant are not considered. Such an actuator could be used to calibrate using an electrical signal.

3.1.6 Electromechanical Detection System

The two opposite sensing capacitors, C_1 and C_2, with the fixed external capacitors, C_a and C_b, inserted in a bridge configuration, provide the extraction of the signal. The bridge is driven by the transformer with an alternate voltage:

$$V_p = V_{po} \cos(2\pi f_p t). \tag{9}$$

The mechanical signal causes the variation of the capacity of the two detector capacitors and the subsequent modulation of the residual output voltage at the signal frequency. This signal is seen as an unbalance of the bridge. The output of the capacitive bridge is sent to a low-noise amplifier, characterized by input impedance

Table 1 List of elements appearing in the electric scheme

$\alpha \cdot x$	Voltage generator associated with signal
v_b	Generator associated with Brownian noise
e_n	Generator associated with amplifier noise
i_n	Current generator associated with amplifier noise
Z_i	Amplifier input impedance
V_{rs}	Voltage generator associated with loss capacitors

Z_i with very high value, an equivalent generator of voltage noise e_n, and an equivalent generator of current noise i_n. A list of the elements that appear in the scheme of Fig. 5, with their meaning, are shown in Table 1.

Now, we analyze the different contributions to the noise of the system, starting with an indication on the signal level, represented in the scheme with the $\alpha \cdot x$ generator.

3.1.7 Signal Level

The residual voltage at the output of the capacitance bridge, in the case $Z_i >> 1/(2\pi f_p C)$, C being the capacity between the points A and B of Fig. 5, is:

$$V_{AB} = V_{CD}\left(\frac{C_1(\vartheta)}{C_1(\vartheta)+C_2(\vartheta)} - \frac{C_a}{C_a+C_b}\right). \tag{10}$$

Considering that

$$C_1(\vartheta) \approx C_o + \frac{C_o}{d_1} b\vartheta \quad C_2(\vartheta) \approx C_o - \frac{C_o}{d_2} b\vartheta, \tag{11}$$

if the distance between the two faces of the capacitors is $d_1 \approx d_2 = d$, we have:

$$v_s = \alpha b \theta_s. \tag{12}$$

This formula establishes the relationship between the amplitude value of the mechanical signal at frequency f_s and the value of the signal at the bridge output, at frequency $f_p + f_s$ or $f_p - f_s$.

The transducer factor between output voltage and the central mass angle of deflection is:

$$\alpha = \frac{V_{CD}}{2d}. \tag{13}$$

The transducer factor represents the maximum value of the electrical field inside each single detection capacitor.

For the following evaluations, we will refer to the middle shifts of the harmonic oscillator COM and not to its rotations ($\theta_s b = x_s$). In our scheme, Eq. (1) can be written in the following way:

$$\frac{5}{4} m\ddot{x} + \frac{\eta_t}{b^2}\dot{x} + \frac{k_t}{b^2} x = F. \tag{14}$$

If we set:

$$m_r = \frac{5}{4}m, \eta = \frac{\eta_t}{b^2}, k = \frac{k_t}{b^2}, \tag{15}$$

we have:

$$m_r\ddot{x} + \eta\dot{x} + kx = F. \tag{16}$$

With this assumption, the mechanical quality factor becomes:

$${Q_m}^{-1} = \frac{\eta}{m_r\omega_o}. \tag{17}$$

The accelerometer works at frequencies lower than the resonance frequency of the mechanical oscillator and at these frequencies, the transfer function between the acceleration of the sensitive mass and its displacements is:

$$x(\omega) \approx \theta_s(\omega)b \approx a(\omega)/{\omega_o}^2. \tag{18}$$

3.1.8 Brownian Noise Associated with the Harmonic Oscillator

The dissipation of the mechanical oscillator is associated with a generator of Brownian noise that causes acceleration in the unit band, equal to:

$${a_b}^2(\omega) = \frac{F^2(\omega)}{{m_r}^2} = \frac{4k_bT\eta}{{m_r}^2} = \frac{4k_bT\omega_o}{m_rQ_m}. \tag{19}$$

In this equation, k_b is the Boltzmann constant, T is the thermodynamic temperature of the oscillator, ω_o is its resonance frequency, Q_m is the mechanical quality factor, m_r is its mass, and η is the dissipation factor (given in Eq. 15). It is clear that in order to have a low level of Brownian noise, we need to have a mechanical system with a high mechanical quality factor.

3.1.9 Preamplifier Noise

As we underlined before, an equivalent voltage noise generator and an equivalent current noise generator can model the amplifier noise contribution (see Ref. [12]):

$$\begin{cases} {e_n}^2 = 4k_bT_nZ_n \\ {i_n}^2 = 4k_bT_n/Z_n, \end{cases} \tag{20}$$

where Z_n is the noise impedance, and T_n is the noise temperature of the amplifier.

$$Z_n = e_n/i_n. \tag{21}$$

From the previous assumptions we have:

$$T_n = \frac{e_n i_n}{4k_b}. \tag{22}$$

This noise generators act in two different ways: (a) They directly produce a voltage at the input of the ideal amplifier; (b) They produce a fluctuation of charge across the detection capacitors and a consequent acceleration in the harmonic oscillator.

3.1.10 Total System Noise

If Q is the total quality factor of the system, including the dissipations in the mechanical oscillator and the thermal noise associated with the transducer loss, given by the equation:

$$\frac{1}{Q} = \frac{1}{Q_m} + \frac{1}{Q_{de}}, \tag{23}$$

then the total noise in the acceleration, from the point of view of the mechanical oscillator, can be expressed by the formula:

$${a_t}^2(\omega) \approx \frac{4k_b\omega_o}{m_r}\left[\frac{T}{Q} + T_n\frac{4Z_nC\omega_o}{\beta}\right]\Delta f, \tag{24}$$

where $\Delta f = 1/\Delta t$ is the inverse of the acquisition time for each single measurement.

The factor β, with the new positions, is given by:

$$\beta = \frac{\alpha^2 C}{m_r {\omega_o}^2}. \tag{25}$$

The electrical dissipation is represented by:

$$\frac{1}{Q_{de}} = 4\frac{\omega_o}{\Omega_p}\frac{tg\delta}{\beta}, \tag{26}$$

where $tg\delta$ is the angle of loss of the electrical part of the transducer.

3.2 Techniques to Circumvent the Noise

In this section, the information on the methodologies commonly used in experimental gravitation in order to circumvent the various types of noise otherwise limiting the accuracy of measurements are given (see also Ref. [13]). We begin with a description of the noise present in the various laboratories or sites, mostly located in underground locations or at the bottom of a borehole or in a very deep seabed,

and away in general from population centers, so as to avoid anthropogenic noise. It follows the description of the technique according to which low-frequency signals that have to be detected are transferred to a higher frequency, making less troublesome the effects due to thermal variations and to "aging." It presented a brief review of methods to implement the damping of mechanical vibrations, where this is possible (i.e., for signals with a frequency higher than a few Hz). We continue with the description of the cryogenic techniques useful for the reduction of the Brownian noise, to improve some particular physical parameters of the experimental systems (mechanical quality factor, coefficient of thermal stability, and temperature stabilization), and to implement preamplifiers for the pickup signals with a very low noise. Systems for the rejection of the common mode noise are also described. These systems are useful in the case in which the gravitational signal (or the inertial acceleration) can be seen as a differential mode. Finally, the section concludes by giving information about the techniques of the so-called quantum nondemolition (QND) or back-action evading (BAE), pushing the measurements of gravitation "beyond" the quantum limit.

3.2.1 Choice of the Laboratory

The execution of an experiment in the field of gravitation undoubtedly requires performing measurements in a laboratory chosen so as to ensure an environment with minimal noise ([14, 15]). The types of environmental noise are as follows:

- Vibrational noise induced by the local seismicity
- Thermal noise determined by the temperature variations which induce changes in buoyancy of the experiment and a consequent variation of its response to gravitational and inertial forces (mainly determined by the modulation of the local gravity acceleration)
- Thermal noise due to temperature variations that cause a spurious signal in the response of the experimental apparatus
- Noise due to variations in the environmental pressure which determine effects in the experimental apparatus positioning, with consequent modulation of the signals (also in this case the main effect is in the modulation of local gravity acceleration), and directly on the test mass, by means of buoyancy force.

The plot of the spectral density of the acceleration noise that the geophysicists consider as the minimum (possible) noise that can be measured in the quietest laboratory on ground (particular sites in which the anthropic and thermal noise are at their minimum) is shown in Fig. 6. This means that any experiment on gravitation will be subject to a microseismic noise whose effects must be overcome.

From this figure, it results that the minimum obtainable noise in the quietest on ground laboratories is in a frequency band in the range $10^{-4} - 3 \times 10^{-2}$ Hz, where the environmental *rms* noise due to the Earth seismicity, is less than $10^{-10} \frac{\mathrm{m}}{\mathrm{s}^2}$ per octave $\equiv \frac{1}{\sqrt{2}} \times 10^{-10} \frac{\mathrm{m}}{\mathrm{s}^2} / \sqrt{\mathrm{Hz}}$. This value is very low and would

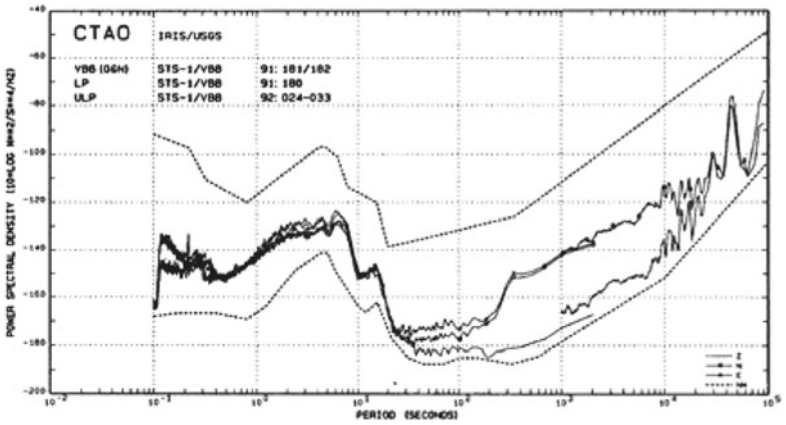


Fig. 6 Plot of the spectral density of the acceleration noise components *x, y*, and *z* that the geophysicists take as the minimum noise possible to measure in the quietest laboratory (particular sites in which the anthropic and thermal noise are at their minimum)

allow the execution of a good test in gravitation; in particular, relying on the possibility of integrating the measurements for a long time in the case of experiments in which we should reveal sinusoidal signals with a frequency within the indicated band and in stable conditions. Obviously, it is clear that a laboratory which is useful to perform tests of gravitation is also a suitable place for geophysical measurements intended to determine the specific frequencies of the earthquakes, especially those of large intensity that excite the free oscillations of the Earth and whose determinations provide information on its internal structure. Of course, in the laboratories where such low levels of seismic noise are measured, also the thermal variations as well as those of the atmospheric pressure need to be highly contained. The experimental data collected over the years by our group in term of acceleration noise, temperature, and pressure recorded at several sites in Italy are shown below. These values give a good indication about the real possibility of being able to operate in a laboratory where the noise level is really low and to run good measurements of gravitational signals.

Seismic noise measurement at the underground LNGS/INFN laboratory (Gran Sasso, L'Aquila)

In the following are shown some geophysical measurements performed by our team with one of our accelerometer prototypes installed in the Gran Sasso underground laboratory of the Istituto Nazionale di Fisica Nucleare (INFN). The spectral density of the horizontal component of the seismic noise measured is shown in Fig. 7. The plot has been produced using two files containing data collected with different sampling times; the values are in complete agreement with data recorded by classical geophysical instruments (Fig. 8).

Figure 8 shows the horizontal component of the solid tide of the Earth, recorded by the accelerometer used as an inclinometer, in the Gran Sasso laboratory. The

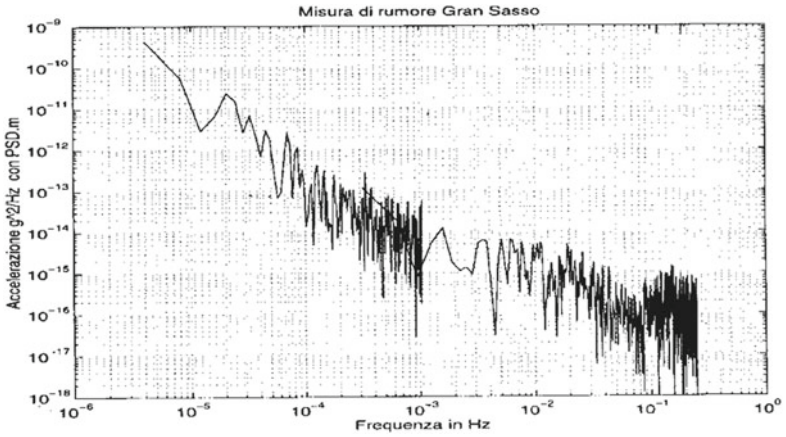


Fig. 7 Spectral density of the seismic noise (horizontal component)

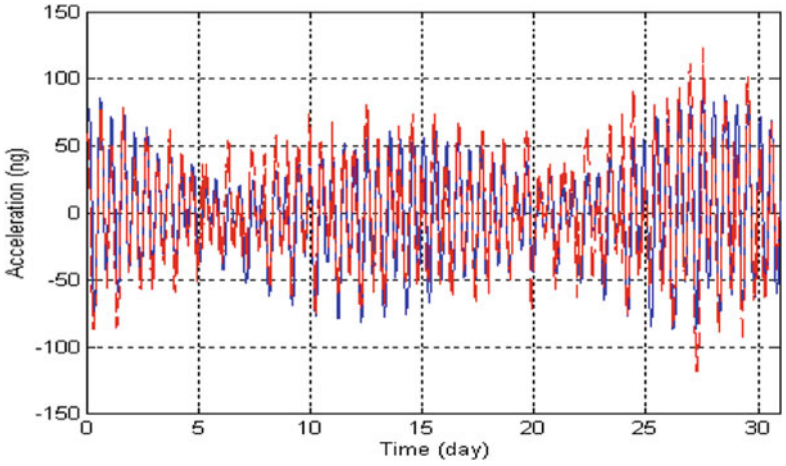


Fig. 8 Experimental Earth tides, compared with the theoretical ones, measured at the Gran Sasso laboratory, L'Aquila

data have been collected during the month of August, a period when people are on holidays and the induced anthropic seismic noise reaches its minimum.

Figure 9 shows an earthquake event located close to Japan and recorded at the Gran Sasso laboratory by the ISA accelerometer.

Finally, as shown in Fig. 10, the power spectral density of seismic noise measured at the Gran Sasso laboratory in a very quiet period, when its level is lower than $10^{-9} g_{\oplus}/\sqrt{\text{Hz}}$ at frequency of 4×10^{-2} Hz.

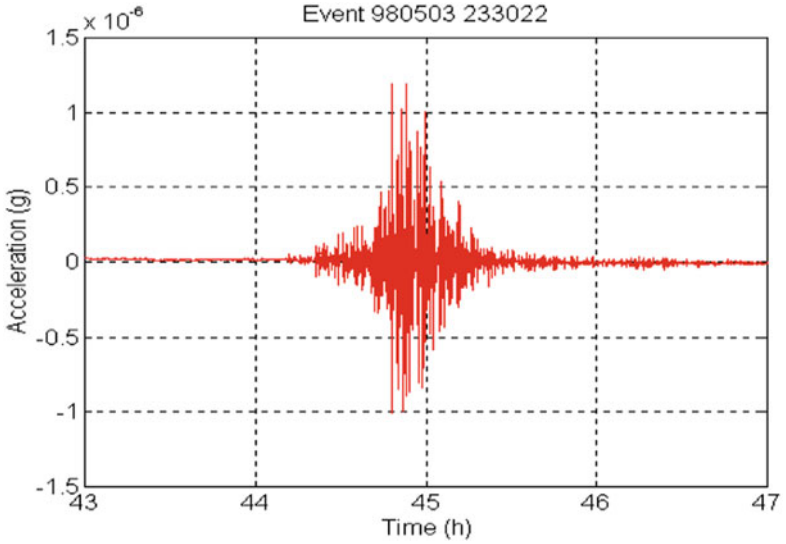


Fig. 9 Earthquake event 980503 recorded at the Gran Sasso laboratory by the ISA accelerometer

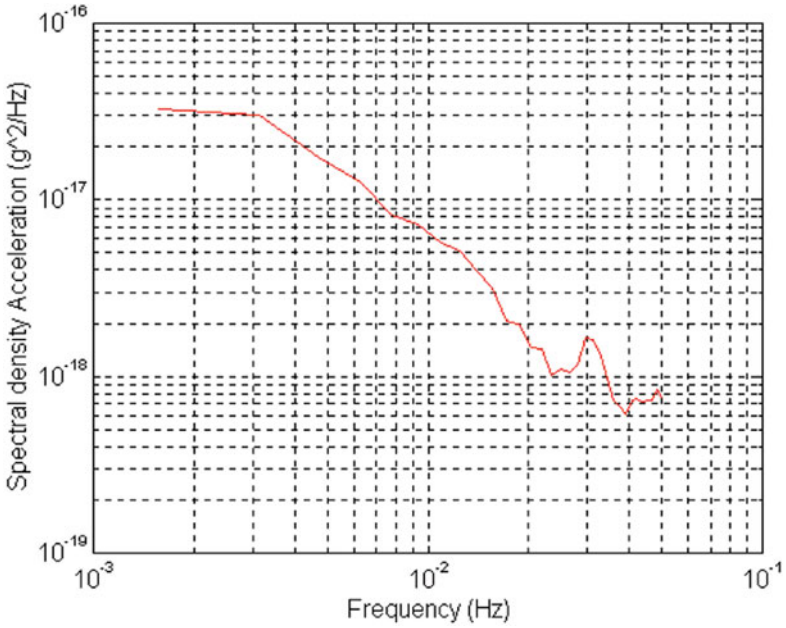


Fig. 10 Power spectral density of the seismic noise measured at the Gran Sasso laboratory, at the level of $10^{-9} g_{\oplus}/\sqrt{\text{Hz}}$

3.2.2 Transfer of Signal in a Frequency Range Where Noise Is Lower

One of the techniques used to circumvent the environmental noise is to transfer the frequency of the signal in areas where the ambient noise is lower; examples of this are shown below:

Direct Transfer of the Signal

- Experiment to test the universality of free fall performed with pendulums: The signal, otherwise, in continuous, is transferred to the oscillation frequency of the pendulum.
- Experiments to test the WEP in free fall in the gravitational field of the Sun: The signal is transferred from the continuous to the period of revolution of the Earth.
- Space experiments to test the WEP in the field of the Earth: The signal is transferred from the continuous to the orbital frequency of the satellite.

Transfer of the Caused Effect

- Polarization of high-frequency transducers in order to transfer the signal around the frequencies of polarization such to avoid the $1/f$ noise of the amplifiers.

Direct Production of the Signal at a Frequency Where the Environmental Noise Is Lower

- Measurement of G where the mass sources can be made to rotate at a suitable frequency in order to produce a signal in the area of minimum environmental noise.

3.2.3 Suspension Systems and Mechanical Insulation

The possibility to isolate an experiment relative to gravitation from the vibrational noise induced by the environment is given by the condition of being able to put the experiment in a suitable environment capable of reducing such noise. The vibrational noise propagates through the medium interposed between the noise source and the experiment. Therefore, the technique for the isolation must necessarily provide the insertion, between the source of noise and the system, of a suitable insulating system. The basic concept of these dampers is in the response of a mechanical oscillator able to provide an attenuation of 40 dB/dec for noise frequencies above the resonance frequency of the system itself. Assuming we can make a resonant frequency of the order of 1 Hz, and in the case in which the signal to be detected is at a frequency of 100 Hz, the attenuation of the signal will be 80 dB, that is, equal to a factor of 100. It is therefore clear that a good attenuator must have a resonance frequency as low as possible, so as to have the highest attenuation. The problem associated with the implementation of these dampers is precisely in the realization

of low-frequency damping. As mentioned above, the basic element of the dampers is the mechanical oscillator which is realized with a spring or a spring system, with low spring constant K, interposed between the "laboratory" and the experiment that determines its mass M. In most cases, the experiment is under the action of gravity and for the equilibrium condition the following equation must be valid:

$$K \cdot \Delta x = M \cdot g. \tag{27}$$

From this equation it follows that, if an attenuator of frequency equal to 1 Hz is implemented, the mechanical spring must be able to compress or stretch, in an elastic way, by an amount equal to about:

$$\Delta x = \frac{g}{\omega^2} = 25 \text{ cm}. \tag{28}$$

This simple example shows that it is not easy to set the mechanical frequency of the attenuator below 1 Hz; the problem can be solved using nonmechanical springs such as pneumatic systems, using multiple attenuation (more oscillators in cascade) whose transfer function provides a much higher attenuation of 40 dB/dec, adding elements that introduce negative elastic constants (e.g., electric fields) which determine the lowering of the frequency of the system. This approach leaves to the original spring the task of supporting the mass in the field of the Earth's gravity. It should be noted that, to avoid that the disturbances can reach the experiment through the gas present in the environment, the experiment must generally be placed under vacuum conditions in a suitable chamber. In a practical case of isolation of a cryogenic gravitational antenna, cooled to liquid helium temperatures, its operation was ensured by isolating the antenna also from the noise introduced by the boiling of liquid helium; this last damping stage ensured an attenuation equal to 135 dB at the resonance frequency of the antenna, equal to 1790 Hz, allowing to appreciate its thermal noise (energy equal to 1 KT), at a temperature of 1.5 K.

3.2.4 Cryogenic Techniques

The temperature can act under several different ways; a first list can be the following:

- Thermal contraction of the sizes of the experimental apparatus, causing a variation of the gravitational signal and/or response to the same solicitation, with noise signals.
- By varying the structure of the laboratory and with it the support plane of the experimental apparatus and also the local component of the gravity acceleration viewed from the apparatus.
- Change of the properties of the experimental system (pickup systems, values of the preamplifier, electrical frequency of voltage bias of the transducers and propriety of the demodulation systems).

- Introducing noise directly through the force fluctuations associated with the dissipation of the mechanical oscillator, whose mass is the experimental test mass.

To avoid the indicated problems, many experiments of gravitation make use of the cryogenic techniques for the reduction of different kind of noise; in particular, the lowering of the thermodynamic temperature means that there is a drastic reduction of the Brownian noise directly linked to it, see the Formula _19, Sect. 2.1. As already said, the cryogenic techniques are useful, together with the reduction of the Brownian noise, also for the improvement of particular physical parameters of the experimental systems:

- Increasing the mechanical quality factor: Most physical systems when cooled show an increase in their quality factor, the lowering of the temperature makes it to "crystallize" the structure of the materials, resulting in a reduction of dissipation in their vibrational modes.
- Stabilization of the temperature: The temperature stabilization has a direct consequence in the stabilization of the structure of the experimental apparatus itself, avoiding that some effects of noise, once modulated by the thermal contractions of the structure, to produce a change of the frequency of the signal to be detected (in the case of use of cryogenic liquids this thermal stability is determined by the temperature of evaporation of cryogenic liquids themselves).
- Increase of coefficients of thermal stability of the systems: Even in this case, the majority of the physical systems once cooled show an increase of their thermal stability, and this effect gives a benefit to the experiment, being added to the previous effect.

In the case of a Weber-type gravitational resonant antenna, the problem is to cool a cylinder several tons heavy at temperatures lower than those of liquid helium, using dilution chiller.

3.2.5 Low-Noise Amplifiers

In modern experiments of gravitation, a fundamental role is played by the preamplifier to which is delegated the task of amplifying the weak signals that are present at the output of the transducer, so as to make possible their further amplification and therefore their possible demodulation, conversion, and acquisition; all this without the introduction of electronic noise in excess with respect to the signal to detect. These so demanding benefits are obtained using the best devices offered by technology (as Field Effect Transistor (FET) and Superconducting Quantum Interference Device (SQUID)) and in optimal condition of impedance matching, and by making them work to the thermodynamic temperature at which their voltage and current noise are to a minimum.

The experimental measurements (carried out at the IAPS laboratories) of the electronic noise, reported in acceleration, for two special electronic devices cooled to a temperature of 120 K are shown in Table 2.

Table 2 Summary of noise values, reported in acceleration, for two active components using FET technology and cooled down to 120 K, obtained in no-adapted and adapted condition and for a superconducting transformer in air

FET	120 K Values estimated from measured data No-adapted $g/\sqrt{Hz}$	120 K Values estimated from measured data Adapted $g/\sqrt{Hz}$
2N4416	1e-13	4.9e-15
BF817	4.5e-14	1e-15

These evaluations, arising from real measurements performed on the active elements, indicate the possibility to reach a quite high sensitivity in acceleration, using the BF817, with appropriate impedance adaptation and increasing the transducer electrical field at 10^5 V/m.

3.2.6 Common Mode Rejection (CMR)

The CMR is indicated as a solution in many experiments of gravitation: The gravitational signal to be detected acts simultaneously on two test masses, and the experiments are carried out looking at the gravity gradient signal; the high level of common mode noise, acting on the two test masses, is rejected making the difference between the two signals. Obtaining high level of rejection is connected to the fact that the two responses are equal for the two detectors, and it is clear that the higher the CMR the more performing is the apparatus. Usually the apparatus can be constituted by two different mechanical and electronic systems, in this case they must be equal; otherwise, there is the possibility to adjust their parameters in order to equalize their response, or evaluating experimentally a coefficient that makes their answer equal. The requirement is that the evaluated parameter must be useful at the appropriate frequencies where there is the signal, and this parameter must remain fixed over the entire duration of the experiment. In some experiments, the apparatus is able to perform a "self-rejection" in the sense that in the same mode as in the previous case the gravitational signal to be detected acts on two different test masses; again, the result of the experiment is connected to the gravity gradient between the two test masses, but the system is sensitive directly to the gravity gradient and makes the self-rejection of the noise common mode (a torsion balance is an example of this kind of device). In this case, there is no possibility to "calibrate the system" for this factor, but at the same time it is possible to adjust the apparatus mechanically. One major advantage is the fact that it makes the self-rejection of noise and the signal to be amplified is not so high as if also noise must be amplified, eliminating the problem of amplifier saturation.

As an example of these techniques and to explain also what we mean with "zero experiment" for the Galileo's pendulums experiments, we can assimilate this at a differential experiment; the simultaneity of the observation for the two pendulums eliminates the majority of the thermal effect variations and environmental conditions in general, and makes the measurement of time unique.

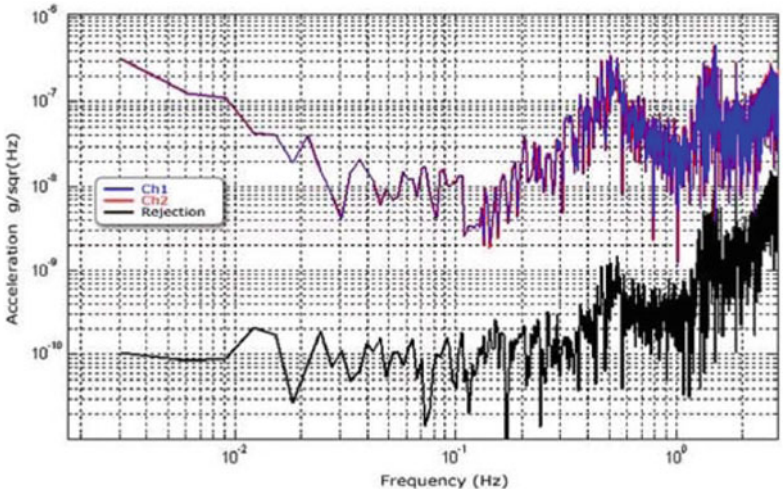


Fig. 11 Difference of seismic signals recorded by two ISA elements arranged with their sensitive axes parallel

As shown in Fig. 11, the results of an experimental activity concerning the CMR, performed at the IAPS/INAF laboratory. It shows the difference of signal recorded by two ISA accelerometer elements arranged with their sensitive axes parallel. It is possible to see that the level of the residual signal is equal to $10^{-10}\mathrm{g}/\sqrt{\mathrm{Hz}}$ in a frequency band in the range 10^{-3} – 1 Hz; this level is obtained adjusting the parameters for the comparison of the two accelerometers answer so to obtain a CMR bigger than 10^4. In the same time, this low level gives us the indication that each accelerometer has a level of intrinsic noise less than $7 \times 10^{-10}\mathrm{g}/\sqrt{\mathrm{Hz}}$, well below the sensitivity required for the accelerometer in the case of the BepiColombo mission to Mercury, in about the entire frequency band.

This characterization has been performed in our laboratory; new calibrations are planned to be performed in another laboratory, where a less level of seismic noise can allow reaching a low level of residual noise in the entire interesting frequency band.

3.2.7 Quantum Nondemolition and Back-Action Evading

The need to push the sensitivity of experiments of gravitation, in particular of experiments for the detection of gravitational waves, beyond the limit imposed by the quantum theory for a linear transducer and by the instrumental performance, has prompted the search for new schemes of measurement on the basis of which it is possible, at least in principle, to measure repeatedly a dynamic variable of the system (harmonic oscillator), with arbitrary precision. These schemes, called QND or

BAE, are based on the possibility of measuring a dynamic variable, without the disturbance exerted by the measuring system on the physical system under observation, disrupting the physical variable under measurement (see Ref. [16]). Sufficient condition in order to say that an observable is a QND observable is that it is a constant of motion in the absence of external forces. In some cases, the schemes allow only the determination of the modulus of the force, while in others they also allow the determination of its dependence over time; the first case is connected to the determination of the quantum number N of the harmonic oscillator, while the second is connected to the measurement of the real part X_1 or the imaginary part X_2, of the complex amplitude of its state of oscillation. For this last case, in the proposed schemes the oscillator under test is coupled to the apparatus of measurement in such a way that their Hamiltonian of interaction contains only the QND variable under observation, $\hat{h}$, and some variable of the measuring system, $\hat{q}$.

In its simplest form it can be written as

$$\hat{H}_I = k\hat{h}\hat{q}, \tag{29}$$

where k represents the intensity of the coupling.

A simple possibility would appear to be the direct measurement of $\hat{X}_1$ (or $\hat{X}_2$) for which the request is a Hamiltonian of the type:

$$\hat{H}_I = k\hat{X}_1\hat{q} = k\left(\hat{x}cos\,(\omega_0 t) - \frac{\hat{p}}{m\,\omega_0}\sin(\omega_0 t)\right)\hat{q}, \tag{30}$$

which presents difficulties in the implementation, either by the modulated form that its coefficient k must have, either for the impossibility to implement a transducer of momentum.

In practice, it is possible to achieve a measure of $\hat{X}_1$, using a position transducer with coupling coefficient modulated at the frequency of the oscillator, together with a mechanical low-pass filter:

$$\hat{H}_I = k\cos(\omega_0 t)\;\hat{x}\;\hat{q} = \frac{k\hat{q}}{2}\left(\hat{X}_1 + \hat{X}_1\cos(2\omega_0 t) + \hat{X}_2\sin(2\omega_0 t)\right). \tag{31}$$

The low-pass filter has the function to filter away the component at $2\omega_0$, leaving only the component at zero frequency, $\hat{X}_1$. To avoid having to detect the signal at zero frequency, it is introduced a new modulation of the interaction at frequency ω_e, precisely to the frequency of an electrical oscillator (LC) at the frequency ω_e, which also acts as a filter. The new Hamiltonian thus becomes:

$$\hat{H}_I = k\cos(\omega_0 t)\cos(\omega_e t)\;\hat{x}\;\hat{q}. \tag{32}$$

The modulation ω_e transfers the signal relative to $\hat{X}_1$ previously in continuous at this frequency, plus contamination of $\hat{X}_2$ at frequency $\omega_e \mp \omega_0$.

It is therefore clear as it works this so-constructed Hamiltonian of interaction, one can sample the signal relative to the $\hat{X}_1$ component, while having only a

residual component of the disturbance in $\hat{X}_2$. Through this nonlinear interaction, introduced by the electric field present in the capacitive transducer, it is possible to observe a physical variable, causing the effect that the perturbation introduced by the interaction disturbs only the component not observed.

It can be shown that, if τ is the electric oscillator relaxation time, the minimum energy detectable in an optimized time of integration it is equal to (Thorne):

$$E_s = \frac{\hbar\omega_0}{\sqrt{2\omega_0\tau}}, \tag{33}$$

where E_s is the energy deposited by a classical force acting on a harmonic oscillator initially at rest. In a scheme used for the experimental activities in IAPS, the coupling was carried out by biasing the pickup capacitive bridge using simultaneously two "pump" voltages at frequency $\omega_e + \omega_0$ and $\omega_e - \omega_0$ which operate simultaneously a parametric down-conversion and an up-conversion of the signal, providing a signal coherent conversion from mechanical to electrical frequency. The biggest problem is the maintenance of the relative stability between the electrical and mechanical frequency of the system so to ensure long-term measurement stability; a beat shuffles the components $\hat{X}_1$ and $\hat{X}_2$ canceling the effect of noise reduction of the system. The problem can be solved by using a different detection scheme, which leads to a quadratic type of detector (Braginsky and Unruh). In essence, in this last scheme, we observe the variation of energy of a harmonic oscillator due to the action of a force of short duration; the change is associated with the initial condition of the oscillator; the bigger it is, the more excited is the oscillator (eigenstate E_0). The possibility of being able to perform measurements with precision in energy (in principle indefinitely big) is due to the fact that the phase of the oscillator remains almost indefinite, thus guaranteeing the uncertainty principle. It can, in principle, measure the variation of a single quantum of energy of the harmonic oscillator, compared to $\mp\sqrt{N}$ as suggested by standard techniques (where N is the number of quanta).

In a possible scheme, used by our group, a capacitive sensor biased in continuous provides a voltage $V_u = kx$ that is used to bias a pickup capacitive bridge for the same mechanical oscillator. The output of this second system will then be (quadratic detector):

$$V = \frac{V_u}{d}x = \frac{k}{d}x^2, \tag{34}$$

where d is the distance between the two faces of the second transducer. For this system, the Hamiltonian of interaction will be:

$$\hat{H}_I = \frac{k}{d}\hat{x}^2 q\left[\hat{X}_1^{\,2} + \hat{X}_2^{\,2} + \left(\hat{X}_1^{\,2} - \hat{X}_2^{\,2}\right)\cos(2\omega_0 t) + \left(\hat{X}_1\hat{X}_2 + \hat{X}_2\hat{X}_1\right)\sin(2\omega_0 t)\right]. \tag{35}$$

In the last equation, $\hat{q}$ is the charge in the capacitor.

As in the previous case, a low-pass filter will enable to detect only the energy $E_0 = \hat{X}_1^2 - \hat{X}_2$. The sensitivity for this scheme can be written in term of energy limit: The amplifier limit value $kT_n = \sqrt{e_n^2 + i_n^2}$ come to $kT_n = \hbar\omega_a$, where ω_a is the operation frequency of the amplifier.

Also in this case it can be shown that the minimum detectable energy will be given by a formula analogous to the previous case:

$$E_s = \frac{\hbar\omega_0}{\sqrt{2\omega_0\tau}}. \tag{36}$$

3.2.8 Facility for Micro- and Pico-gravity Experiments

In this section, it made the description of the main facilities for microgravity experiments, implemented in different parts of the world; the section ends with the description of the facility "GiZero," considered on behalf of ASI, which should allow the execution of experiments in conditions of pico-gravity, including the GReAT experiment, which is described in a later section of this chapter.

Drop Towers A large number of drop towers were implemented in several parts of the world (Spain, USA, etc.); as an example, we illustrate the drop tower at the University of Bremen, Germany, which is by far the most representative in this area.

The Bremen drop tower is 146 m high and the free fall of the capsule occurs inside a tube in which it is made the vacuum, for a height of 110 m. The duration of the free fall is of 4.7 s and the residual acceleration that is obtained is 10^{-6} g, after that the transients introduced at the stage of release are damped. The system of deceleration is implemented with a 10 m high container filled with polystyrene balls. A catapult system can launch the capsule from the bottom of the tower; the capsule in "free fall" reaches the top of the tower and then falls, in a time of the order of 9.5 s. In this case, the recovery system is positioned after the passage of the capsule.

Japan Microgravity Center (JAMIC) This structure is located in Kamisunagawa, north of Japan, and it is implemented in an old mine whose entrance is constituted by a vertical borehole about 710 m deep. Inside this borehole, an evacuated capsule is dropped, simultaneously a platform is released inside of it on which the experiments are arranged. Thrusters, external to the capsule, compensate the resistance of the medium during the 490 m fall. The braking is obtained with a pneumatic system, by dropping the capsule for 200 m inside an appropriate tube. The residual acceleration on the platform is 10 μg and the fall time about 10 s.

Aircraft in Parabolic Flight (Ilyushin 76-MDK) In this case, the experiment is housed in an aircraft capable of taking a parabolic flight during which the gravity is cancelled. The time of free fall is about 20 s, during which residual accelerations of the order of 10^{-2} g are obtained.

Another category of facility to obtain the microgravity conditions is represented by the sounding rockets; among many, here we mention two of them:

Suborbital Rockets (Consort Mission—Maser) In these "facilities," the experiment is allocated aboard a suborbital rocket-type Starfire. In a time of the order of 1.2 min the rocket—that receive the thrust from the propulsor—reaches the maximum altitude of about 300 km, and then free fall for a distance of about 80 km from the starting point. The time in which there are the microgravity conditions is about 8.3 min, with residual accelerations of the order of 10^{-5} g. The space available for the experiments is a cylinder with a diameter of 0.44 m and a length of 3.65 m.

Orbital Rockets (Maxus—Wolna) In this case, the probe reaches 800–1200 km of height, with residual accelerations in its return to the ground of the order of $10^{-2} - 10^{-4}$ g, for a time of the order of 14 min.

Space Shuttle This type of "facility" is very well known and does not need a description. In this case, the shuttle orbits at a height of the order of 300 km and allows conditions of residual acceleration for times of the order of the week. The accelerations to which the experiment is subject during launch and reentry are extremely reduced.

Orbiting Stations In this case, the experiment is brought aboard the station with the use of a rocket and the conditions of microgravity may extend from 1 week to several months. The presence of a crew on board means that the experiment can be followed throughout its execution.

Drag-Free Satellites A satellite in orbit around the Earth is actually a system in free fall, whose free-fall time is enormously long. It is clear the big interest of the scientific community dealing with gravitation, for space activities and in particular for those involving the so-called drag-free satellites. Onboard accelerometers measure the accelerations that disturb their dynamics and give the appropriate signal to a set of thrusters that compensate for the measured effect. For these systems, the residual accelerations are of the order of 10^{-12} g.

The comparison between the different facilities for micro- and pico-gravity must be made taking into account the conditions of microgravity that have been reached, for the time span in which these conditions are kept, for the volume, mass, and power available for the experiments and to the transient acceleration to which the experiments are subject at the times of release and recovery. A particular interest is in the possibility of being able to perform the measurements of the dynamics of the capsule, so as to be able to assess and remove the inertial effects. Below it is described the GiZero facility which should be able to ensure the conditions of microgravity comparable with those of a drag-free satellite, although for shorter times, but with the advantage of being able to perform a large number of tests, so to have the chance to adjust, whenever needed, the parameters of the experiment.

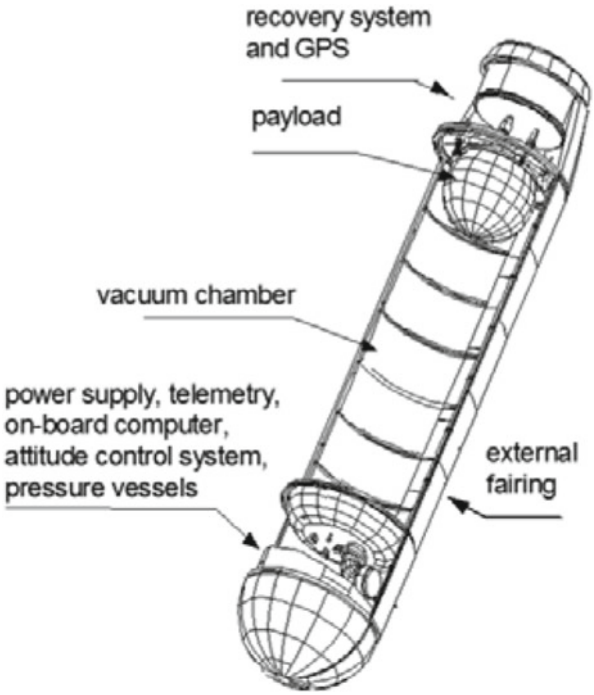


Fig. 12 The GiZero structure

3.2.9 Ballooning Facility for Microgravity (GiZero)

The GiZero facility (Refs. [17–20]) has been implemented to perform experiments in free fall from stratospheric (i.e., at an altitude of about 40 km) balloon; its level of residual acceleration will be less than $10^{-12}\mathrm{g}/\sqrt{\mathrm{Hz}}$, and the conditions of free fall maintained for a time longer than 20 s (Fig. 12).

The facility, shown in Fig. 12, is composed of a vacuum capsule that shields from drag a platform containing the experiment which free falls inside it. The facility provides a new opportunity to perform scientific and technological experiments that need very low residual acceleration conditions. In particular, it seems a good tool to test new apparatus developed for space and to perform precise fundamental physics experiments.

Precisely it is composed by three parts: An external fairing that provides a good aerodynamics, an internal vacuum chamber where the platform containing the experiments falls, and a recovery system. In the structure, the following parts are also allocated: the power supply, the telemetry, an on-board computer, the attitude and control system, the pressure vessel, and a global positioning system (GPS) receiver.

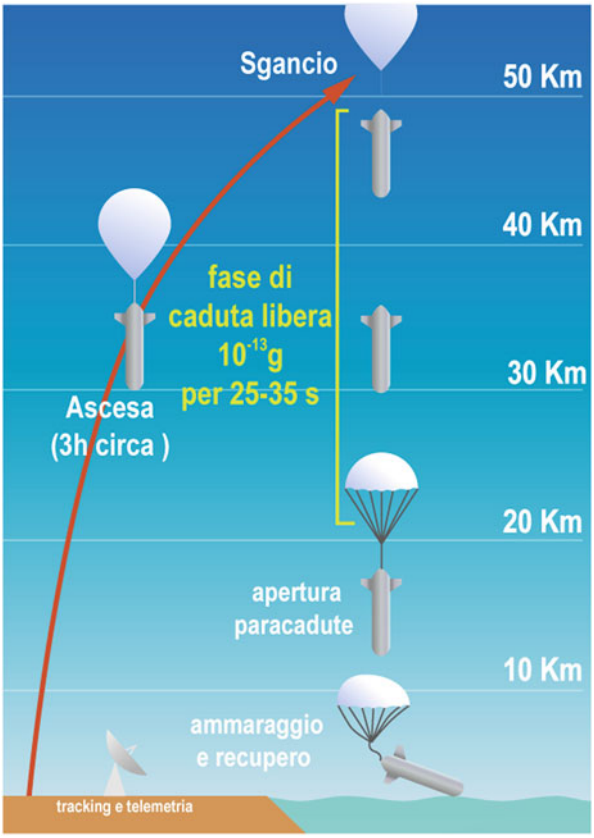


Fig. 13 Mission profile

The free-fall condition is determined by the time during which the internal platform, in free fall inside the capsule, reaches the bottom of it. This time it depends on the aerodynamic line of fairing.

Figure 13 shows the mission profile for GiZero. Table 3 shows the value of the main figures of GiZero and its performance.

In about 3 h the capsule is lifted at 42 km. The oscillations of the gondola are controlled and it is released in an almost steady state. With a delay of a few seconds, the platform is released inside it and free falls under vacuum conditions until it reaches the bottom of the capsule, slowed by the atmospheric drag. In $20 \div 25$ s of free-fall, the capsule covers a distance of about $1960 \div 3060$ m , respectively. At this point, an automatic system catches the platform. The capsule continues its free fall until it reaches an altitude of 20 km where a parachute system will be activated to recover it. After the splash down the capsule is recovered; suitably refurbished, it can be used again for another mission.

Table 3 GiZero characteristics and performance

Capsule	
Total length	6.5 m
External diameter	1.1 m
Total weight	1000 kg
Payload	
External diameter (max)	0.7 m
Weight (max)	60 kg
Performances	
Free-fall time	20–25 s
Ascent time	≈ 3 h
Maximum speed	≈ mach 1
Maximum load at the parachute opening	4 g
Splash down speed	7 m/s

The platform in free fall is not an inertial system. Gravitational and inertial accelerations act all over it; the formula to evaluate these different contributions is the following:

$$g(\vec{r},t) = -\vec{\nabla}V(\vec{r},t) - \vec{\Omega}\wedge(\vec{\Omega}\wedge\vec{r}) - 2\vec{\Omega}\wedge\vec{v} - \dot{\vec{\Omega}}\wedge\vec{r} - \vec{\alpha}(t). \qquad (37)$$

In this formula, $V(\vec{r},t)$ represents the gravitational potential, $\vec{\Omega}$ and $\overrightarrow{\dot{\Omega}}$ are the platform angular velocity and angular acceleration, respectively, $\vec{r}$ is the distance of the considered point on the platform between a reference origin, $\vec{v}$ is the velocity of the considered point, and $\vec{\alpha}(t)$ is the linear acceleration that acts on the platform. The acceleration terms are respectively: gravitational, centrifugal, Coriolis, a term depending on the variation of angular velocity and linear one. During the free-fall condition, the gravitational effects that can be observed on the reference system of the platform are only gravity gradients (Fig. 14).

Figure 14 shows the main contributions at the residual noise on the platform, during the free fall, in terms of gradients of acceleration, evaluated with a numerical simulation.

Curve labeled "1" (in green) shows the spectral density of the gravity gradient of the Earth. The curve labeled "2" (in red) is the contribution of masses placed on the capsule. The curve labeled "3" (yellow) is the inertial gradient, which arises from the vibrations of the capsule wall, transmitted to the platform through the residual gas present in the capsule. The curve labeled "4" (blue) indicates the level of $10^{-2}\mathrm{EU}/\sqrt{\mathrm{Hz}}$, where 1 Eötvös Units (EU) is equal to 10^{-9} s^{-2}. For two points on the platform distant half a meter from each other, the same plot gives the differential acceleration in unit of g, multiplying the scale for the factor 5×10^{-11}. The gravitational effects can be reduced taking out the low-frequency components (if the structure of the capsule is not complicated) and the inertial effects, coming from the vibrations of the capsule walls, can be easily reduced increasing the vacuum level

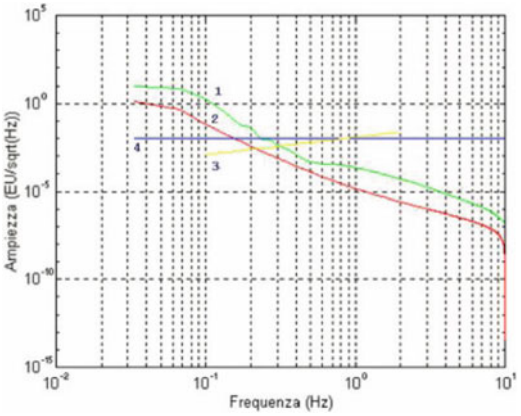


Fig. 14 Power Spectral Density (PSD) of the contributions at the residual gradient of acceleration on the experimental platform

in it. In this numerical evaluation, the CMR has not been considered and the vacuum value is set to $P_i = 10^{-4}$ m bar. To these terms it is necessary to add this noise term a continuous component equal to Ω^2 and a component due to the precession velocity. If $\Delta I/I$ indicates the percentage difference in the principal moments of inertia of the platform, the precession velocity during the free fall is given by: $\Omega_p = \frac{\Delta I}{I}\Omega$. Experimental tests indicate the possibility to obtain values of $\Omega = 10^{-3}$ rad/s. Taking $\Delta I/I$=10^{-2} as a possible value, a continuous component equal to $\Omega^2 = 10^{-6}s^{-2} = 10^3$EU and a component ${\Omega_p}^2 = 10^{-10}\mathrm{s}^{-2} = 10^{-1}$EU at the precession frequency act on the platform.

In the following, we give a list of experiments which can be carried out with GiZero:

Test and Calibration of High Sensitivity Space-Borne Gravity Gradiometers The calibration is one of the main problems in the development of such instruments due to the presence of Earth's gravity and vibrational noise present in laboratory on-ground. The gravity gradients of the capsule could be used for inferring calibration signals and additional gauge masses could be placed along the capsule.

Fundamental Gravity Experiments Adding a cryogenic part, the facility allows performing experiments to test the WEP. The experiment consists on the measurement of differential accelerations acting on two test masses of different materials (e.g., aluminum and gold) free falling in the Earth's gravitational field. The falling test masses are part of a "differential detector" with zero baseline (the two centers of mass are coincident). Sensitivities of about $1.5 \times 10^{-13}\mathrm{g}/\sqrt{\mathrm{Hz}}$ (at the liquid nitrogen temperature of 77 K) and $1.5 \times 10^{-14}\mathrm{g}/\sqrt{\mathrm{Hz}}$ (at the liquid helium temperature of 4 K) can be reached for these experiments. The rejection factor for inertial accelerations must be considered.

The detector is spun about a horizontal axis at a frequency of1 Hz in order to modulate the gravity signals during the free fall. The estimated accuracies, within

95 % confidence level, in testing the WEP in a 25 s integration time are 5×10^{-14}, at the temperature of liquid nitrogen, and 5×10^{-15}, at the temperature of liquid helium. For more details, we refer to Sect. 4

General Microgravity Experiments In a multi-payload configuration the system allows to perform experiments in material science, biology, and fluid mechanics, being able to accommodate up to five payloads with a total weight of 500 kg.

4 Calibration of the Apparatus for Measurements of Gravitation and Acceleration

As already told, the difficulties to calibrate an apparatus for gravitational measurements on-ground are mainly due to the presence of environmental noise and Earth's gravity; in this section, the activities devoted to the on-ground test of a high-sensitivity accelerometer of ISA kind will be described; these operations can be envisaged as examples for more general experiments in gravitation.

We start remembering that for the accelerometer here considered, of the ISA kind, the frequency band is $(3 \times 10^{-5} \div 10^{-1}$ Hz). Any calibration must be performed trying to circumvent the unavoidable seismic noise, or forcing it at a level of excitation well above it. A first way to calibrate one instrument is by means of the measurements of all the fundamental parameters that determine its characteristic in terms of sensitivity and to evaluate it by means of calculation. In Sect. 2.1 of this chapter, we had already introduced a formula that makes a compendium of all the main noise elements of both electrical and mechanical nature for every single axis of the accelerometer; as shown in Tables 4 and 5, the main mechanical and electrical parameters experimentally measured for one of the three ISA accelerometer elements (we remember that ISA is a three-axis accelerometer implemented using three different elements with sensitive axes orthogonal to each other).

Let us recall that the mechanical quality factor Q_m is tightly tied up to the vacuum conditions in which the sensor is located, being strongly dependent on the

Table 4 Mechanical parameters of the oscillators

M	Proof mass	0.22 kg
I	Inertia	4×10^{-4} kg/m^2
f_o	Mechanical resonance frequency	3.5 Hz
Q_m	Mechanical quality factor (environment vacuum)	$0.3 - 10^{-4}$
k_t	Elastic constant of torsion	1.9×10^{-1} N m
G	Module of rigidity	2.6×10^{10}
L	Arms length	45×10^{-3} m
B	Ray of proof mass (for a circular proof mass shape)	43×10^{-3} m

Table 5 Electrical parameters of the accelerometer

$C_1 = C_2 = C_3 = C_4$	Detection capacitor	300 pF
$tg\delta_{C1}$	Angle of electrical loss in C_1	4×10^{-4}
$C_a = C_b$	External fixed capacitor	300 pF
$tg\delta_{Ca}$	Angle of electrical loss in C_a	3×10^{-4}
v_n	Equivalent voltage noise generator	$3 \times 10^{-9}\text{V}/\sqrt{\text{Hz}}$
i_n	Equivalent current noise generator	$7\times10^{-15}\text{A}/\sqrt{\text{Hz}}$
T_n	Amplifier noise temperature	0.38 K
Z_i	Amplifier input impedance	$8 \times 10^5\ \Omega$
A	Amplification	50
α	Transducer factor	10^5 V/m
β	Electromechanical transducer factor	3×10^{-2}

gas trapped in the gap between the faces of the plain capacitors. The two values indicated in Table 4 for Q_m are related to the pressure of 10^3 *and* 10^{-4} mbar.

From the values reported in Tables 4 and 5, it is possible to evaluate the sensitivity of the accelerometer; this gives a value equal to $3.3 \times 10^{-12}\text{g}_\oplus/\sqrt{\text{Hz}}$ in vacuum and $3 \times 10^{-10}\text{g}_\oplus/\sqrt{\text{Hz}}$ at atmospheric pressure (limits due to the Brownian noise). For the BepiColombo ISA accelerometer, we chose a Q_m equal to 100, so that the resulting sensitivity is $10^{-9}\text{m/s}^2/\sqrt{\text{Hz}}$.

4.1 Ground Testing and Calibration

The analysis of the noise sources and the theoretical evaluation of the performance of the accelerometer are not satisfactory and, consequently, direct experimental tests are necessary. A prototype of accelerometer already implemented as shown in Fig. 15. The main parts of this instrument are the three mechanical units (with their own preamplifiers) arranged in order that their sensitive axes are perpendicular to each other. In the same box, the microcontroller and all the electronics for power supply, demodulation of signal, and acquisition of data are also included.

For on-ground calibrations, the base is equipped with three support points; two of them are electronically controlled with micrometer screws. By adjusting these screws, the gravity acceleration can be set parallel to a sensitive axis (e.g., z) and perpendicular to the other two axes (x, y). In these conditions, the unit with sensitive axis along the vertical is subjected to acceleration equal to 1 $g_\oplus$; the other two units are sensitive to a component of $g_\oplus$ depending on the angles ϑ_x, ϑ_y, from the horizontal plane. For small variations of ϑ_x, ϑ_y, their value in radians corresponds to the acceleration in $g_\oplus$ to which the unit is subject (Fig. 16).

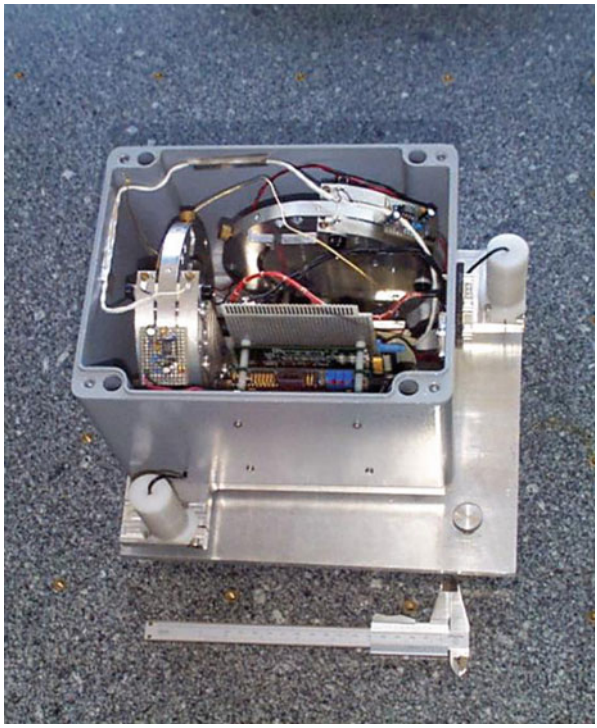


Fig. 15 Prototype of ISA accelerometer

4.2 Measurement of Transducer Factor and Linearity

The calibration of each single unit of the accelerometer is performed applying to it a known component of the gravity acceleration by varying the proper angle $(\vartheta_x, \vartheta_y)$ and by reading the related output. In this case, the connection between the input acceleration and output voltage is direct and enables an evaluation of the transducer factor. To perform an evaluation also of the system linearity, the micrometer screws are replaced with a piezoelectric, able to give a very small change of the angle. Figure 16 shows the output of a single unit of the accelerometer, excited with DC accelerations, step by step. The x-axis represents the acceleration measured by the accelerometer, while the y-axis represents the voltage given to the piezoelectric. It also represents the inclination in radians (multiplied by a conversion factor equal to 2×10^{-8} rad/V). The plot shows the linearity of the accelerometer response, for a wide range of accelerations.

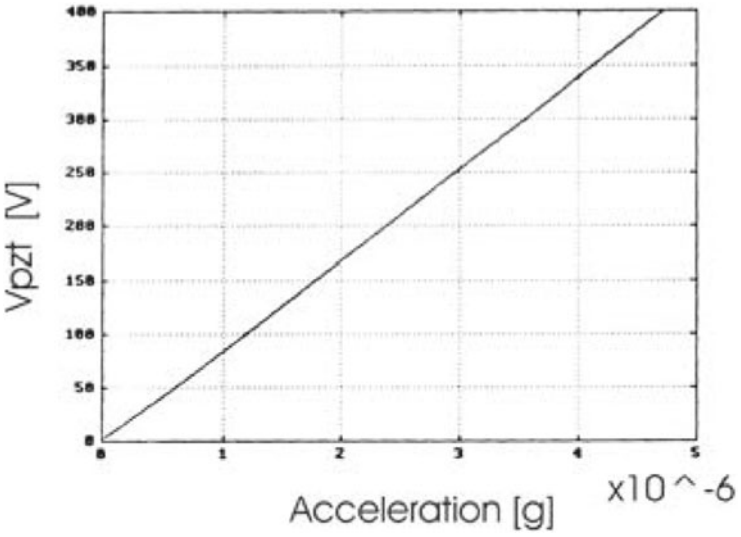


Fig. 16 Output of the accelerometer excited with DC accelerations

4.3 Measurement of the Electromechanical Actuation Factor

In Sect. 2, we introduced the formula concerning the electromechanical actuator. Biasing one of the control capacitors with a constant voltage V, and superimposing an alternate voltage $v = v_0 \cos(\omega_e t)$, a torque equal to $M_1 = \frac{C_0}{d_0}\frac{b}{2} vV = \alpha_{att}.v$ at frequency f_e will arise. From the experimental measurement of the parameters that appear in the formula, it is possible to evaluate the actuation factor α_{att}. Another possibility is to measure it experimentally. The measure is performed applying to the accelerometer elements to be calibrated a well-known torque by changing the component of gravity acting on them, by means of a piezoelectric system and measuring the voltage necessary to produce the same torque. We have to remember that in this case the torque is given by $M_1 = \frac{C_0}{d_0}\frac{b}{2} V^2 = \alpha_{att} V$.

4.4 Measurement of the Accelerometer Transfer Function (Resonance Frequency and Mechanical Quality Factor)

Usually an instrument for gravitational measurements (and especially an accelerometer) is implemented in order to have a flat transfer function in its frequency band; remembering that this transfer function is primarily related to the mechanical harmonic oscillator, while the whole electronics actuation and transducer element give a contribution flat in frequency, this characteristic is obtained using the instrument at frequencies well below the mechanical frequency of the harmonic oscillator, where

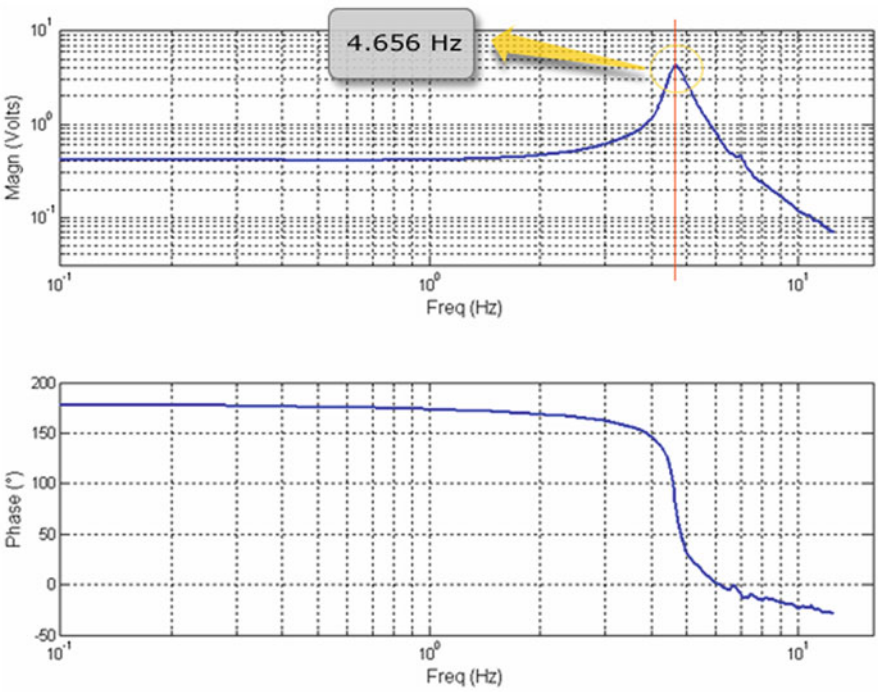


Fig. 17 Transfer function of the accelerometer

the transfer function is flat. The measurement of the transfer function can be performed in two ways: exciting the system using the electromechanical actuator with a constant amplitude voltage swept in the frequency range where the transfer function must be evaluated, or exciting the system with a direct acceleration, using the micrometer screw controlled by the piezoelectric so to change the component of $g_{\oplus}$ acting on the specific element with sensitive axis initially in the horizontal plane.

Figure 17 shows the transfer function of one single axis of the accelerometer. In this case, Q_m is kept low avoiding to make a high vacuum, so to allow the use of a dynamic signal analyzer, which does not work correctly if Q_m is high. As it can be seen, above the fundamental frequency, equal to 3.5 Hz, the system has a second harmonic at 16.6 Hz.

From the previous arguments, it follows that the parameters necessary to characterize the transfer function are the mechanical frequency, the mechanical quality factor, and the response of the system at zero frequency. The correct determination of the transfer function is evaluated with a precise determination of these parameters. The measurement of the mechanical frequency and high mechanical Q (100 in this case) is performed placing the accelerometer in a vacuum chamber, exciting the mechanical oscillator at its resonance at high level and looking at its free exponential decay.

4.5 Measurement of the Intrinsic Noise of the System

The test and calibration previously described, give information on the linearity and level of the output in response to an input excitation, but they did not give information on the sensitivity of the instrument. We stated that, due to the impossibility to filter the seismic noise at low frequencies, the evaluation of the sensitivity of the system and of its intrinsic noise is quite difficult. The only possibility to perform such a calibration is to find a place where the seismic noise is very low in itself and perform measurements with two elements with their sensitive axes aligned so that the seismic noise can be regarded as a true signal and their difference, representing the differential acceleration acting on the system, can be regarded as an upper limit for the sensitivity of each single axis.

Results of such kind of measurements have been already presented in Sect. 2.2. In Fig. 11, the difference of the signals recorded by two ISA's elements, arranged with their sensitive axes parallel is shown.

4.6 Measurement of the ISA Thermal Stability

One of the main characteristics for the gravitational measurement apparatus and accelerometers is their thermal stability. In the ISA accelerometer type developed for the BepiColombo mission, due to the high thermal temperature variations inside the Mercury Planetary Orbiter (MPO) satellite, especially at half of the sidereal period of Mercury around the Sun, equal to 44 days, the immunity of the accelerometer to the thermal variations must be very high; ISA thermal stability is 5×10^{-7} m/s^2/°C. The instability figure means that, a temperature change of one degree, on the electronics or mechanical parts, produces a spurious accelerometer output equivalent to an accuracy of 5×10^{-7} m/s^2. To reduce this effect down to the required accuracy, it is necessary to reduce the thermal variations at the level of 2×10^{-2}°C. The attenuation of the thermal variations can be obtained using an active control loop or a passive one. The advantage of the active control resides in the possibility to attenuate the temperature variations at very low frequency, but this requires additional power dissipation. The passive control method consists in the employment of a system with a very large time constant for its thermal inertial, in order to have a system with a very low cutoff frequency, ensuring the required thermal attenuation inside the measurement frequency band. The calibration of this parameter is performed inserting the device into a thermo-vacuum chamber in order to change its temperature and control the corresponding signal at the output of the accelerometer due to the change in temperature.

5 Test of the WEP—GReAT

The universality of free fall, postulated by G. Galilei [13] and expressed in terms of equality between inertial and gravitational mass by I. Newton [2], is at the basis of Einstein's General Relativity, in terms of the local identity of the gravitational field with a non-inertial reference frame and it is also indicated as WEP, to be distinguished from the strong equivalence principle, which states the covariance of all the laws of nature with respect to all continuous coordinate transformations. The L. Schiff hypothesis implies that every theory that satisfies WEP is a metric theory, and vice versa, that is, every metric theory satisfies WEP [33, 35]. On the basis of such a hypothesis, a WEP experiment allows to test the viability of the metric theories as opposed to the nonmetric ones. No experimental deviation from the *equivalence* has yet been found, even by experiments of very high sensitivity. Every WEP experiment is essentially the test of the universality of the free fall of bodies, as it was described in the famous experiment that G. Galilei performed and already shown in Sect. 1.1.

The accuracy of a WEP experiment can be expressed in terms of the Eötvös parameter η which denotes the ratio between the differential acceleration that acts on two test masses of different material, divided by the mean acceleration acting on them: $\Delta a = \eta \cdot a$.

The difficulty of such an experiment, as already expressed by G. Galilei itself, is due to the drag acceleration acting on the free-fall bodies.

A better accuracy can be achieved by comparing the oscillation of two pendulums and by considering that the two masses are in free fall along the tangent to the trajectory of their respective oscillation. For such kind of experiments (G. Galilei, I. Newton, and F. Bessel) [16], an accuracy $\eta < 10^{-3}$ can be achieved. The use of a torsion balance (R. von Eötvös) allows obtaining $\eta < 10^{-8}$. In such an experiment, the inertia is determined by the centrifugal force originated by the rotation of the Earth, and the gravitational acceleration is the g component necessary for compensating it. A possible WEP violation causes a rotation of the torsion balance. The difficulty of such an experiment relies on the fact that the violation signal occurs at zero frequency. The effect can be detected by exchanging the position of the two masses. By using a torsion balance, and by locking it on the gravitational field of Sun at equilibrium with the inertia of the Earth that rotates around it, the (possible) violation signal would be modulated by the rotation of the Earth around its axis with a 24 h sidereal period. The accuracy of such kind of experiments is (R.H. Dicke and V.B. Braginsky) $\eta = 10^{-11}$. Better results can be achieved by means of space experiments, whenever a satellite is in "free-fall" around the Earth. The conceptual arrangement is always the same: A very sensitive system detects the position of the two test masses, and in order to avoid the effects of the gravitational gradient, the two centers of mass have to be coincident. The sensitivity of the system is increased by holding the two masses by a very soft spring, by which a small acceleration acting on them causes a large displacement that can be easily detected

Table 6 Performance comparison between ongoing space WEP experiments and GReAT

Name	Accuracy	Type	Funding institution(s)
STEP	10^{-18}	Drag-free satellite	NASA
GG	10^{-17}	Drag-free satellite	INFN/ASI
MICROSCOPE	10^{-15}	Drag-free satellite	CNES/ESA
GReAT	5×10^{-15}	Drag-shielded capsule	ASI/NASA

(under ideal conditions they ought to be free). In the last decades, several experiments were set up to test the validity of WEP (for a review see [36]). Apart from Lunar laser ranging and free-fall experiments, the majority of them were performed on-ground, where several sources of noise (among them seismic noise) ultimately limit their accuracy. For this reason, the environment offered by space is suitable for substantial improvements. According to the international proposal of Satellite Test of the Equivalence Principle (STEP) [37], it seems possible to attain an accuracy $\eta \approx 10^{-18}$. In such a case, the system works at liquid He temperature (4.2 K), and uses a drag-free satellite. An Italian non-cryogenic satellite named GG [19] is under study, with a target accuracy $\eta \approx 10^{-17}$ and a French experiment named MICROSCOPE is under development to be flown in the next years [34]. Of course, space experiments are very costly and need a rather long time to be developed. The GReAT experiment that will be described in this section has the advantage of combining the relative low-noise environment given by free fall, with the repeatability, resulting in a competitive performance with respect to other ongoing projects (see Refs. [22–27]). It aims at reaching an accuracy of five parts in 10^{15}. In Fig. 18, the improvement provided by GReAT (two orders of magnitude) is shown in terms of the Eötvös ratio η, while in Table 6 its performance is compared with those of ongoing space experiments.

5.1 Generality of the GReAT Experiment

On-ground and in-space tests had already verified the WEP validity with an accuracy of 5×10^{-13}. In spite of so high accuracy, grand unification and quantum gravity theories impose a violation of this symmetry at higher level, demanding for a more accurate experiment. GReAT aims to obtain an accuracy better than 5 parts in 10^{-15}; this accuracy is intermediate between the accuracy obtained with the past experiments and what is predicted for the future space experiments (STEP: 10^{-18}, GG: 10^{-17}, and MICROSCOPE: 10^{-15}).

The accuracy of ground-based tests of the WEP is limited by the Earth's seismic noise and by the small strength of the gravitational signal source. Increase of signal strength and isolation from external noise are essential ingredients to improve the experimental accuracy. By performing the experiment in free fall under the action of the Earth's gravity acceleration, the signal is three orders of magnitude stronger

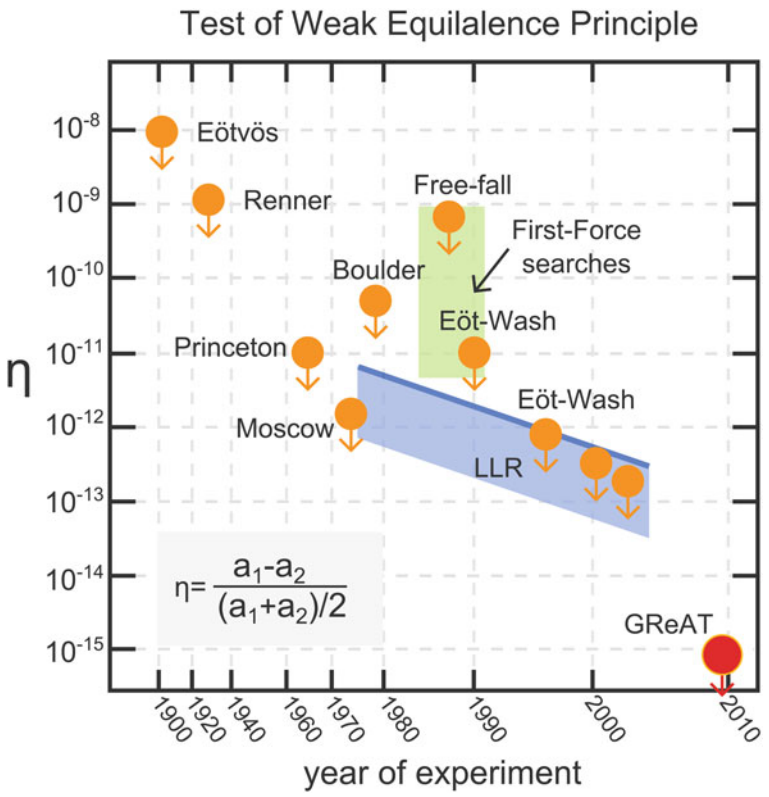


Fig. 18 Tests of WEP in terms of Eötvös ratio η (from [36], modified)

than that in the field of the Sun, used by ground-based experiments (Eötvös type), and by using the drag-shielded free-fall technique of GiZero (this facility has been introduced in the Sect. 2.2), the vibrational acceleration noise can be reduced at a level under 10^{-12} g. The vertical free fall, compared with the orbital free fall ("drag-free" satellite), has the drawback of having a much shorter duration—that means short integration time—but it has the advantages of being much less expensive and offers the possibility to be repeated at intervals of a few months, allowing to adjust the experimental set-up, if necessary. The experiment will be repeated with different material pairs for the test masses of the detector; and the first flight will be carried out with test masses of the same material in order to exclude any possible spurious signal that can mimic a WEP violation. See also Refs. [28–30, 32].

A system of two masses of different material represents the main part of a differential accelerometer which is cooled down to the temperature of liquid helium and spun up to a maximum rate of 0.5 Hz about the horizontal axis during a free fall. The free fall is performed inside the GiZero evacuated capsule, released from an altitude of about 40 Km. We have to remember that the experiment is also cooled down using liquid helium (4.2 K). The possible WEP violation signal will be detected at

Table 7 List of the main parts constituting the facility for the free fall

Helium balloon
Gondola attached to the balloon with the mechanism to release the capsule and other housekeeping equipment
Shielding capsule with a large vacuum chamber
Liquid helium, evacuated cryostat
Instrument package which houses the detector inside the high vacuum chamber
Spin/release mechanism for spinning the instrument package and releasing it into the capsule at the start of the fall
Differential acceleration detector
Two video cameras at the top of the capsule for monitoring the motion of the instrument package falling inside the capsule for post-flight
Telemetry system for the downlink from the capsule to the ground
Transonic parachute for decelerating the capsule at the end of the fall

the spin frequency, while the gravity gradients (due to the capsule mass and to the Earth's mass) produce components at twice the spin frequency. The strength of the gravity gradient components is proportional to the distance between the centers of mass of the two proof masses. The components at the spin frequency can be further canceled out by keeping the spin axis of the detector as close as possible to the horizontal plane.

When compared to future proposed satellite experiments (which could reach higher accuracy) and to classic on-ground experiments, GReAT results as a good compromise, potentially able to improve by two orders of magnitude the accuracy of the test.

In Table 7, it is reported a list of the main parts constituting the GiZero facility, specialized for this particular experiment, while in its general structure and in the profile mission it remains the same as the one already described in Sect. 2.2.

To perform the GReAT experiment, it is necessary that the balloon reaches its floating altitude so to move at the speed of the local wind, that is, the capsule will be at relative speed equal to zero with respect to the local wind; if its vertical profile is constant, the capsule and the instrument package in free fall would move laterally during the fall with the same initial lateral velocity and hence maintain the same lateral distance with respect to each other. We have to note that also the Coriolis acceleration on the capsule and on the falling package are the same and, consequently, they do not alter their relative position. However, if the wind vertical profile changes, the capsule will experience a lateral force that will change its lateral speed while the instrument package will not experience such a force. A diameter of 1.5 m is large enough to guarantee that the instrument package does not touch the walls of the capsule, with a lateral wind gradient up to 50 knots over the 4.3 km of vertical drop. If the balloon is launched during the periodically occurring wind reversal times the vertical wind gradient is much smaller than the value indicated above.

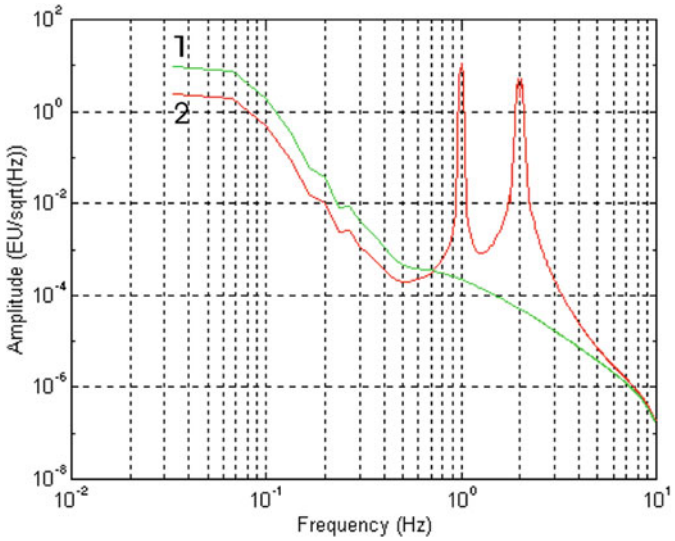


Fig. 19 Numerical evaluation of the gravity gradient due to the mass of the capsule (line-1) and the signal on the rotating accelerometer during the free-fall plus a possible signal of WEP violation with $\eta \approx 10^{-15}$

Due to the low relative velocity of the experimental package and to the high level of vacuum ($< 10^{-6}$ m bar) present inside the capsule, a disturbance from the capsule wall to the free-floating detector produces on it a residual acceleration less than $10^{-12}\text{g}/\sqrt{\text{Hz}}$. The free-falling capsule reduces the acceleration noise to values unmatched by any other Earth-based drop facility and comparable to values achieved on board drag-free satellites [De Bra]. This acceleration noise acts in the same mode on the two test masses (common-mode noise) and it can be rejected by the common-mode rejection factor (CMRF) of the differential accelerometer. For a conservative value of 10^4 for the CMRF, the influence of these accelerations on the differential measurement is made negligible. It is also important to note that the acceleration noise components produced by the residual gas are proportional to the pressure inside the chamber. This means that the pressure can be reduced in successive flights if, for any unanticipated reasons, its influence on the measurement proves to be greater than estimated. It is, in fact, well within the state-of-the-art to obtain pressures at room temperature as low as 10^{-8} m bar in large volumes.

Figure 19 shows a numerical evaluation of the spectral density of the gravity gradient Γ_{zx} (z and x are the vertical and horizontal axis, respectively) from two point 1 μm apart along the central axis of the GiZero capsule and originated by the capsule mass distribution (green line); the red line represents a numerical evaluation of the gravity gradient that acts on the differential accelerometer when it falls along the capsule axis, while rotating at 1 Hz around the y-axis. The peak at 2 Hz is the possible signal due to WEP violation at level of $\eta = 10^{-15}$. It is important to stress that the rotation clearly allows to separate the gravity gradient effects and a possible WEP violation effect.

Several prototypes of accelerometers have been implemented and tested in laboratory conditions; the experimental activities were focused in the development of a mechanical oscillator with high quality factor, low noise electronics, and very high stability versus temperature variations. A prototype of accelerometer with sensitivity equal to $3.3 \times 10^{-12} \mathrm{g}/\sqrt{\mathrm{Hz}}$, evaluated measuring separately its parameters, was implemented and tested in laboratory for a long time period. A value close to $10^{-10} \mathrm{g}/\sqrt{\mathrm{Hz}}$ due to the seismic noise was obtained in a quite environmental place. A sensitivity of $4 \times 10^{-14} \mathrm{g}/\sqrt{\mathrm{Hz}}$ can be obtained for the same prototype, increasing the mechanical quality factor and using a low noise temperature preamplifier matched to the capacitive transducer. A cryogenic version could reach a sensitivity equal to $5.7 \times 10^{-15} \mathrm{g}/\sqrt{\mathrm{Hz}}$. In this section, we want to focus our attention on the description of experimental activities required to obtain the experiment goal.

An outline of the mechanical structure of a prototype of differential accelerometer is shown in Fig. 20. Two sensing masses of different materials are connected to a rigid frame by means of a couple of flexural elements so that their two centers of mass are made coincident. Acceleration acting on them can give a rotation around the common rotation axis passing through the center of the flexural elements. Such a rotation can be detected by means of two capacitive bridges that in principle work as described in Sect. 2. Each bridge is composed of two sensing capacitors (the faces of each sensing mass and two fixed faces) and two external fixed capacitors. The capacitive bridges are biased at 10 kHz in order to transfer the signal from low to high frequency and to allow the preamplifier get rid of its 1/f noise. The differential value is attained by subtracting the two signals. A precise mechanical machining determines the coincidence of the two centers of mass. A possible difference in the position of the two centers of mass, as well as the gravity gradient between two such points, produce a signal at two times the rotation frequency, as previously explained.

The design of the differential acceleration detectors capitalizes on the experience gained in experimental activities and on several numerical simulations. The two sensing masses (1 and 2) consist of solid hollow right cylinders with spherical ellipsoid of inertia so as to cancel the second-order gravity-gradient torque (quadrupole moments). The design of the proposed capacitive detector can accommodate a variety of sensing masses with different dimensions and materials. The differential detector must be designed as much as possible in a way that allows modifications from one flight to the next, based on the experience gained from the previous flight. The centers of mass of the sensing masses have been made as close as possible, as allowed from the technical point of view, in order to minimize the effect of gravity-gradient forces, rotational motion, and linear accelerations upon the differential output signal. The prototype implemented at IAPS/INAF has the two sensing masses of the same material (e.g., aluminum-aluminum). This prototype could be flown in a test balloon flight so to characterize the noise environment during the free fall.

The two sensing masses are constrained by the flexural springs to rotate about a common axis and their resonant frequencies are electrostatically controlled for frequency matching. The lower is the resonant frequency of the two oscillators, the more sensitive is the detector, but with a smaller dynamic range. On the other hand,

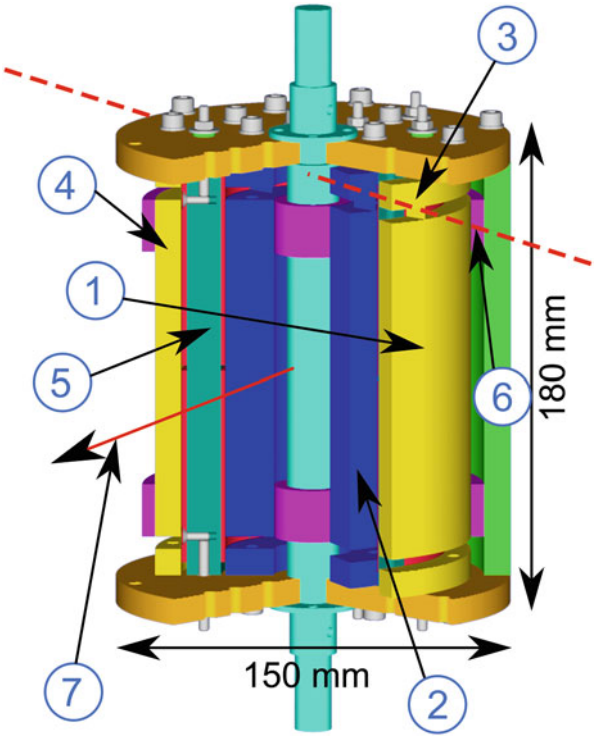


Fig. 20 Outline of a differential accelerometer prototype built at IAPS/INAF

the higher the resonant frequency, the larger the dynamic range and the shorter the time constant of the transient oscillations. The value of the resonant frequency stems from a trade-off between sensitivity on one side and fast transient response and large dynamic range (and also tolerance of centrifugal forces) on the other side. A value of the resonant frequency in the range 2–5 Hz strikes a balance between the above competing requirements. Once the instrument is built with a specific mechanical resonant frequency, this frequency can be lowered by supplying a constant voltage to the feedback capacitor fixed plates. All the other modal frequencies of the instrument are at least two orders of magnitude higher than the controlled flexural frequency. This wide frequency separation allows most of the signal energy to excite the degree of freedom of interest. The sensing masses of this detector are not subject to electrostatic charging because they are grounded to the instrument's case through the flexural springs. A high Q factor is obtained by means of liquid helium refrigeration and by eliminating dissipation sources, that is, the flexural springs, the instrument's case and the capacitor moving plates are machined from the same block of material. We have to remember that this differential accelerometer is using for the pickup system and for the preamplifier the same techniques as described in Sect. 2.1 and Sect. 2.2.

5.2 Intrinsic Noise of the Differential Accelerometer

Taking into account what has been shown in previous section and in the case of the matched preamplifier, the formula to evaluate the acceleration spectral density of acceleration noise for one single accelerometer, in the case where $\omega_s < \omega_0$, can be written as

$$S_a = \left(\frac{\omega_0 k}{m_{eff}} \left(\frac{4T}{Q} + 2T_A \frac{\omega_0}{\omega_s + \omega_P} \right) \right)^{1/2} \text{ms}^{-2}/\sqrt{\text{Hz}}. \tag{38}$$

The two terms in inner parentheses correspond to the Brownian noise and to the preamplifier noise, respectively; ω_0 is the detector resonant frequency, ω_S is the signal frequency, ω_P is the pumping frequency of the bridge, k is the Boltzmann's constant, T is the temperature of the sensing masses, T_A is the preamplifier noise temperature, Q is the quality factor, and m_{eff} is the effective mass of the sensing element. The effective mass is linked to the real mass m, and it simply converts a translational degree of freedom in a rotational one. For this detector prototype, $m_{eff} \approx 1.8\text{m}$.

Clearly, from Eq. 38, the sensitivity of the detector increases by decreasing the resonant frequency and the temperature, and by increasing the effective mass of the sensing mass and the Q factor. Liquid helium refrigeration will be used to provide low Brownian noise, high thermal stability, low thermal gradients, and a high Q factor which are necessary to attain the desired instrument sensitivity.

5.3 Damping of the Transient Induced by the Release of the Experiment

At the time of release of the package, each accelerometer is subject to an acceleration of 1 g that excites the mechanical oscillators. Due to the high mechanical quality factor, which is necessary to keep the Brownian noise at a low level, and to the low resonance frequency of the mechanical oscillators, the two oscillators remain excited for a very long time, which is not tolerable in this kind of experiment. We describe the experimental activity concerning the possibility to damp the transient in a few seconds introducing an extra damping provoked by a resistance R coupled to the mechanical oscillator by means of a control capacitor biased at a DC voltage V_0. It is worth reminding that the decay of the proof mass oscillations is driven by the total quality factor of the electromechanical system which is as follows:

$$\frac{1}{Qt} = \frac{1}{Qm} + \frac{1}{Qe} \quad \text{with} \quad \frac{1}{Qe} = \beta \frac{\omega_0 RC}{1 + (\omega_0 RC)^2}, \tag{39}$$

where β is the electromechanical coupling factor given by

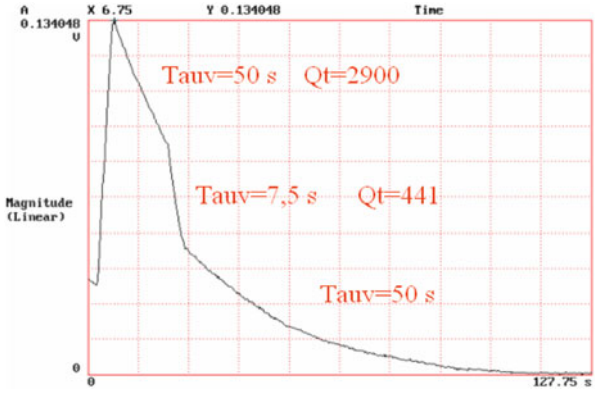


Fig. 21 Decaying amplitude of the oscillator after an excitation of the accelerometer, with a resistance coupled and decoupled to the mechanical part

$$\beta = \frac{CE^2}{m{\omega_0}^2}, \tag{40}$$

that is, the ratio of the electrical energy to the mechanical energy.

Figure 21 shows the decaying amplitude of the oscillator after an excitation of the accelerometer with the resistance coupled and decoupled using the switch. The accelerometer has a resonant frequency of 18.5 Hz and an (undamped) quality factor $Q_m = 2900$. The coupling of a resistance $R = 5010^6\ \Omega$ trough an electromechanical coupling $\beta = 0.01$, obtained with a voltage $V_0 = 1400\ V$, reduces the value of the total quality factor at $Q_t = 441$. The flight instrumentation will have a much lower resonance frequency and higher value of the capacitance C with a consequent higher electromechanical coupling and a total quality factor close to unity, enabling a damping of the initial transient in a few seconds. It is very important to note (see the plot) that the switching on and off of the resistance does not introduce new transients of electrical origin.

5.4 Common-Mode Rejection

During the free fall, the differential accelerometer is under the action of residual accelerations that act as common mode. As an example we indicate the noise induced by the vibrations of the capsule's walls that propagate trough the residual gas present in it, or the sinusoidal noise due to the precession of the experiment platform. The CMRF quantifies the ability of the device to perform such a rejection. A theoretical evaluation has indicated in 10^4, the required value for this parameter. To perform an experimental estimate of this parameter, we built a rotating system whose axis could be tilted away from the vertical, see Fig. 22. The accelerometer is mounted onto the system with its sensitive axis orthogonal to the spin axis. The spin axis is tilted of 10^{-3} rad, so that each accelerometer element is under the common action of a component of g equal to 10^{-3} g; this value is made variable at 0.15 Hz, rotating the device at this frequency. To evaluate the CMRF, the two single

Fig. 22 Experimental arrangement for the measurement of the CMRF

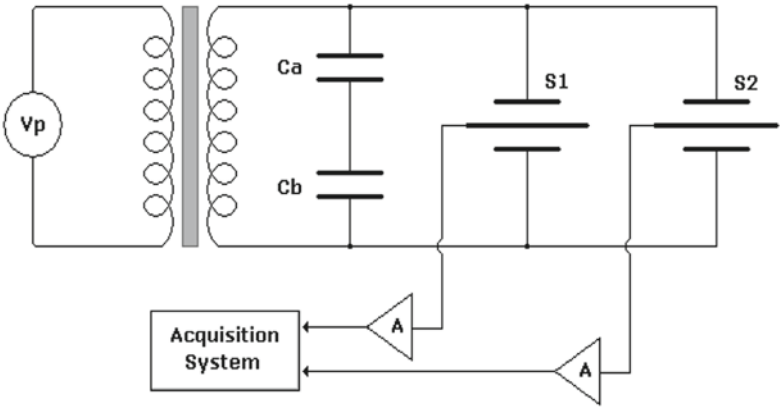


Fig. 23 Detection system for the measurement of the CMRF

accelerometers output are detected separately. In Fig. 23, it is possible to see the detection system. The capacitor detectors of the two single accelerometers, S1 and S2, are connected in a bridge configuration using a single couple of fixed capacitor and biased at 10 kHz trough a transformer. In Fig. 24a, the two separate sinusoidal output of the accelerometer (with amplitude of 10^{-3} g) and their difference are shown, while in Fig. 24b the difference in a more expanded scale is shown.

The spectra of the signal presented in the previous plots are shown in Fig. 25. In this case, it is possible to see that at the rotation frequency the attenuation, or

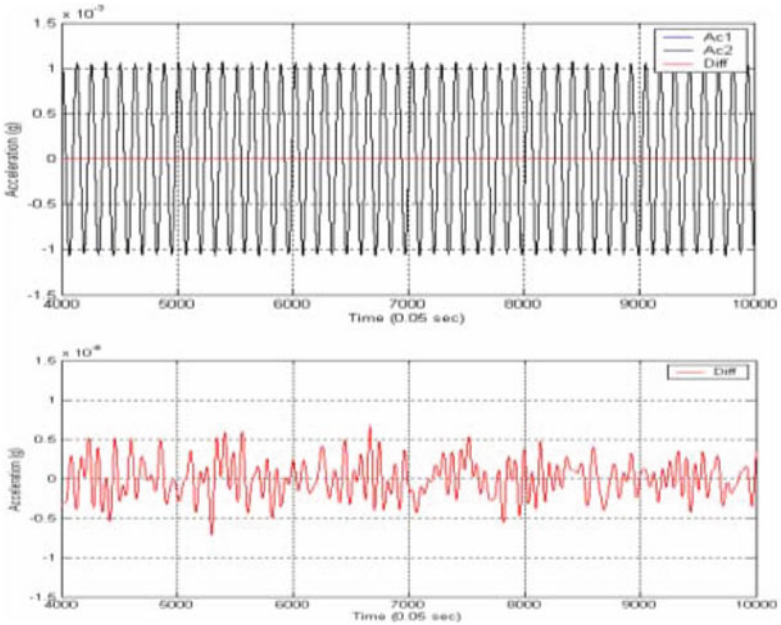


Fig. 24 **a** Sinusoidal output of the two accelerometers in the time domain and their difference. **b** Sinusoidal output of the two accelerometers in the time domain and their difference

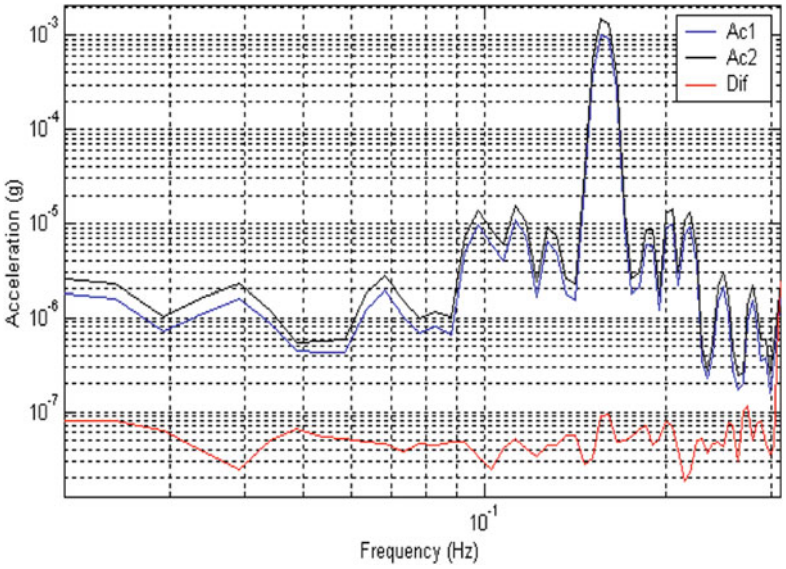


Fig. 25 Spectra of the signal plotted in the previous plot

Table 8 Characteristics of IFSI precursor accelerometers and WEP detector requirements

Item	Characteristics of IFSI prototype accelerometers	Requirements for WEP detector
Temperature	Ambient	Liquid He
Q factor	1000[a]	$> 10^5$
Resonant frequency, ω_0	4 Hz	2–5 Hz
Sensing mass	0.2 kg	> 1 kg
Amplifier noise temperature	10 K	≤ 100 mK
Frequency range	Wide band 10^{-5} – 1 Hz	Monochromatic signal at 0.5 Hz
External acceleration noise	seismic $\left(10^{-8} - 10^{-10}\text{g}/\sqrt{\text{Hz}}\right)$	$\leq 10^{-12}\text{g}/\sqrt{\text{Hz}}$
Common-mode rejection, χ	10^{-4}	$\leq 10^{-4}$
Linearity range	10^6	$\geq 10^6$
Acceleration noise (differential) spectral density	$10^{-10}\text{g}/\sqrt{\text{Hz}}$	$\leq 10^{-14}\text{g}/\sqrt{\text{Hz}}$

[a]The *Q factor* is artificially kept low to prevent saturations by using a low vacuum inside the instrument

CMRF, is equal to 10^4. This value is obtained adjusting the amplitude and phase of one of the two outputs. The adjustment of the phase is at level of 4.4×10^{-3}s.

Table 8 compares the characteristics of the precursor accelerometers to the requirements of the differential accelerometer for the proposed WEP test.

Following the experimental results and aware of the strong improvements that refrigeration brings to temperature stability and Q factor, we are confident that the experience developed with the precursor accelerometer will enable us to develop a differential accelerometer of the characteristics required to carry out the free-fall WEP test.

5.5 Experiment Error Budget

Error sources are both internal and external to the detector. The most important internal sources are: the amplifier noise, the thermal noise (Braginsky 1974; Giffard 1976), and the viscous drag due to residual gas inside the capacitors (Worden et al. 1990). The most important external noise sources are: the force and torque due to Earth's gravity gradient; the Earth's magnetic field interaction with the ferrous impurities in the sensing masses; and similarly paramagnetism of the proof masses coupled to the magnetic moment of the capsule-fixed electrical equipment. The proposed experimental technique is such that the hypothetical WEP-violating signal has either a frequency well separated from narrow-band noise sources or a strength much larger than the broad-band noise sources.

Table 9 Error budget for the WEP test

Noise source	Maximum differential acceleration	Frequency content
Brownian noise	$1 \times 10^{-14}\text{g}/\sqrt{\text{Hz}}$	white
Amplifier noise	$4 \times 10^{-15}\text{g}/\sqrt{\text{Hz}}$	white
Capsule's vibrations	$10^{-17}\text{g}/\sqrt{\text{Hz}}$	white
Drag in capsule	6×10^{-17} g	$1/t_{\text{fall}}$
Proof masses magnetic disturbances	$< 10^{-17}$ g	f_S
Radiometer effect	2×10^{-16} g	f_S
Earth's gravity gradient torques	10^{-16} g 10^{-12} g	f_S $2f_S$
Higher order gravitational coupling to capsule mass	$< 10^{-16}$ g	$f_S, 2f_S, 3f_S, \ldots$
Others	$< 10^{-17}$ g	Various
Error sum (rms), $t_{\text{int}} = 20$ s	2.4×10^{-15} g	f_S

t_{fall} free-fall time, t_{int} integration time, f_S signal frequency

Table 9 provides a summary of the most important noise sources and their frequency content for an instrument with a quality factor $Q > 10^5$. Our error analysis shows that a proposed detector that meets the characteristics specified here will be capable to carry out a test of the WEP with an accuracy of five parts in 10^{15}, with 95 % confidence level, in a 20 s integration time.

5.6 Release System Overview

The release mechanism is designed to reduce the spurious components of the spin velocity about the other axes to less than 0.1 deg/s. A (non-cryogenic) prototype of the spin/release mechanism (see Fig. 26) was built at the Harvard-Smithsonian Center for Astrophysics (CfA) and has already achieved the desired accuracy at release on tests carried out in the laboratory on a mock-up of the instrument package.

5.7 Instrument Package

The instrument package, with a mass of 20 kg, will be virtually an autonomous system during the experiment duration. It will be in free fall within the vacuum chamber, relying only on internal power. Its only connection to the outside world will be an infrared link to the on-board data system. Up to the point of release, the package relies on external power for heating and for operating its electronics. A small vacuum chamber contains the detector, and it is held at a higher vacuum (10^{-11} m Bar) than the cryostat in which it resides. A superconducting magnetic

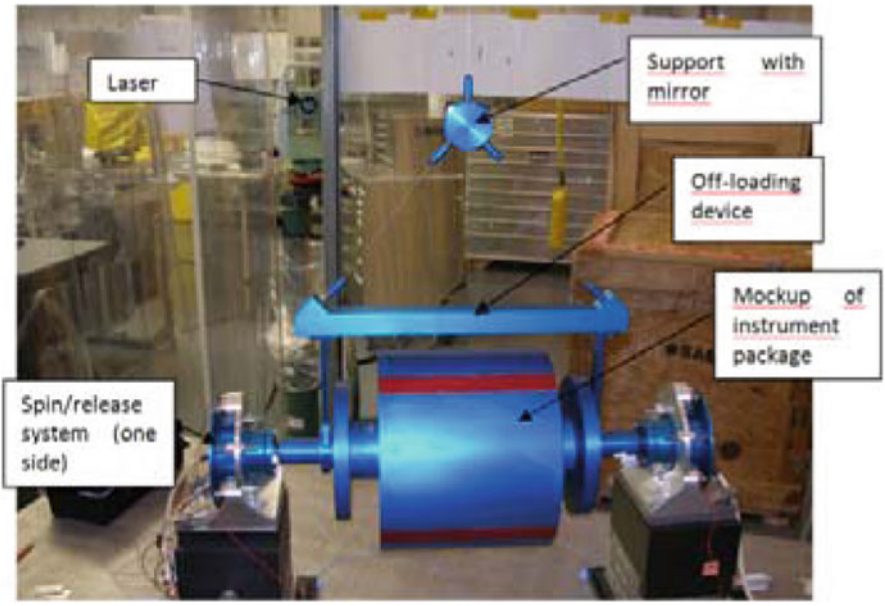


Fig. 26 Picture of spin/release prototype system in the CfA laboratory

shield, within the small vacuum chamber, surrounds the detector in order to attenuate the effects of external magnetic fields on the instrument. A small conical trap, made from a wire net at the bottom of the cryostat captures the instrument package at the end of free fall. The instrument package will wedge itself into the trap and will become more tightly bound along with each bounce (parachute snap, etc.). Since the net is flexible, it further isolates the instrument chamber from impulse loads.

5.8 Balloon Launch Facility and Launch Campaign

The drop system can be operated from any balloon launch facility that enables the operations of a drop system. A possibility is to use a launch base close to the coast so to have chance to recover the capsule in water. Past experiences demonstrated the possibility to lunch and recover a capsule of 3500 kg after its released at stratospheric altitude and using a parachute to decelerate it from Mach = 0.8. Once the experiment is operational, we envision a series of balloon launches separated by time intervals of a few months. One possibility is to follow the wind time reversal which occurs twice per year and schedule one or two launches for each wind reversal period. The times between launches will be devoted to data analysis, refurbishment of the flight hardware, and correction of any problem which may occur.

Each launch (or couple of launches) will capitalize on the experience gained during the previous launch(es) and, if necessary, it will be possible to improve the experiment performance.

Among the ongoing experiments to test the WEP, GReAT offers an improvement in accuracy with respect to current limits on Eötvös ratio η with moderate cost and has the advantage of reusability. Its development is based on a well-established expertise on acceleration sensors and an overall design study conceived for a strong isolation of detector with respect to all the main sources of noise. The experimental activity and the simulations performed so far show that a hypothetical WEP-violating signal greater than the accuracy level could be neatly distinguished.

6 Measurement of the Newton's Gravitational Constant G Using a One-Axis Gravity Gradiometer

Every theory of gravitation can be expressed in terms of a set of equations by which it is possible to evaluate the motion of a test particle in the gravitational field of a system of masses. In spite of the simplicity of this requirement, the large amount of theories alternative to general relativity (GR) indicates the complexity of the problem. One of the basic criteria to test the different theories consists in the verification of the hypothesis made upon the nature of the gravitational constant G. In GR, the components of the Ricci tensor are connected to the energy–momentum tensor by the equation (Einstein equation):

$$R_{ij} - \frac{R}{2} g_{ij} = -kT_{ij}, \tag{41}$$

where k is a coupling constant, to be determined through the experimental verifications. In the approximation of weak field and in the static case, it follows the connection between k and G: $k = 8\pi G$. In this equation, the speed of light is take as unity; in the more general form: $k = 8\pi G/c^4$. From the above indication, it clearly appears the importance of a precise measurement of G in the formulation of an expression for the gravitational field connected to the experimental values. We can assert that in the Einstein equation, G establishes a relation between the space curvature and the energy–momentum density. Recent experiments for the determination of G provided values with an accuracy $\Delta G \cong 10^{-4}$, but they differ from each other more than a factor 40 with respect to their relative errors. These discrepancies are an apparent proof of systematic errors in the different methods and of the difficulty of the measurement. Accredited values are:

CODATA value: $6.6726 \times 10^{-11} m^3 Kg^{-1} s^{-2}$
German PTB results: $6.7154 \times 10^{-11} m^3 Kg^{-1} s^{-2}$

The difficulties of the measurement are connected with the smallness of the gravitational effects, to the presence of spurious effects, such as electromagnetic and thermal, and to the necessity of obtaining the necessary metrological precision.

In the following, a list of different methods and experiments to measure the gravitational constant is provided.

- Torsional balance; Henry Cavendish (1798) performed the first laboratory measurements of this quantity.
- Measurement of the oscillation period of a pendulum; Luther and Towler (1982).
- Compensated torsional balances. An electrostatic force nulls the effects of a source mass; Fitzgerald and Armstrong.
- Michaelis et al. have used a torsion balance whose mass is suspended inside mercury.
- Perturbation of the free fall (acceleration) of a body (test mass), produced by another mass *M*. The motion is monitored by means of a laser. In this case, there is not the elastic force of a spring that usually introduces systematic errors. The *perturbing* mass is arranged in order to decrease or increase the acceleration of the *test mass*; in this way, the *common mode* effects will be eliminated. This experiment is analogous to the experiments involving satellite tracking, but the possibility to measure *M* enables to determine separately the value of *G*.

The method described here consists in the measurement of the gravitational effects by means of a torsional oscillator that makes up a single-axis gravity gradiometer for offline components of the gravity gradient tensor. The gradient at frequency $2f_r$ is produced by a system of two masses rotating at a frequency f_r. The advantage of such a system resides in the possibility to produce the signal at a frequency in which the seismic noise is low, avoiding the thermal effects and effects depending on time, for DC signals.

6.1 Description of the Experimental Apparatus

As we told, the measure of the gravitational constant G presented here has been made using a single-axis gravity gradiometer. The gravitational effect produced by a system of two masses, rotating at a frequency f_r, excites a torsional mechanical oscillator at frequency $2f_r$. Such a torsional oscillator is essentially a single-axis gravity gradiometer able to detect the out-of-diagonal components of the gravity gradient tensor. The low mechanical oscillator frequency, together with the high sensitive capacitive bridge detection system, enables the detection of a gravitational signal with precision better than one part in 10^4. Such a precision, that we experimentally obtained, can be translated into the same precision in the measurement of G, at the condition to determine the physical parameters of the whole system with the same relative error. This experiment has strong analogy with the experiments involving the classic torsion balance, but it differs for some extremely important peculiarities which make it particularly interesting: The oscillator system is in neutral equilibrium under the action of the Earth's gravity field, so that it can be installed in any place, independently from the local vertical; it can be used also in absence of gravity (free fall and space experiment); the capacitive bridge pickup system enables to obtain a high transducer factor to convert the mechanical signal into an electric one, avoiding the preamplifier noise (Figs. 27–29).

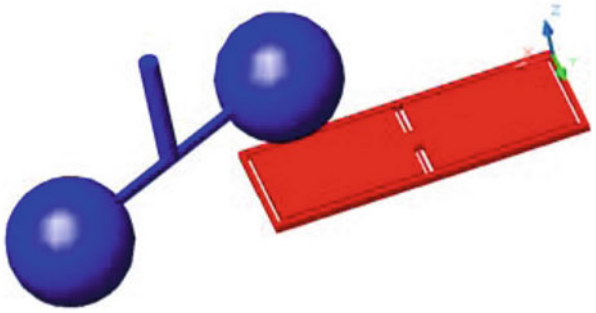


Fig. 27 Scheme of the experimental apparatus

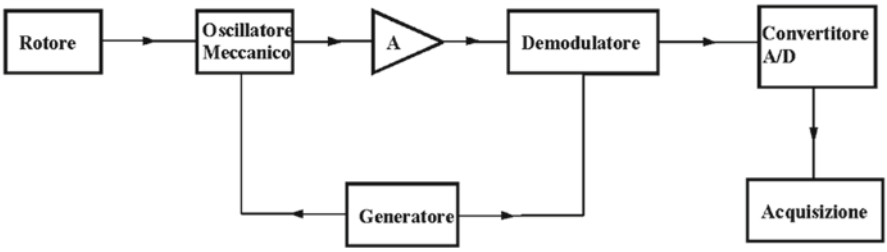


Fig. 28 Block scheme of the whole experimental system

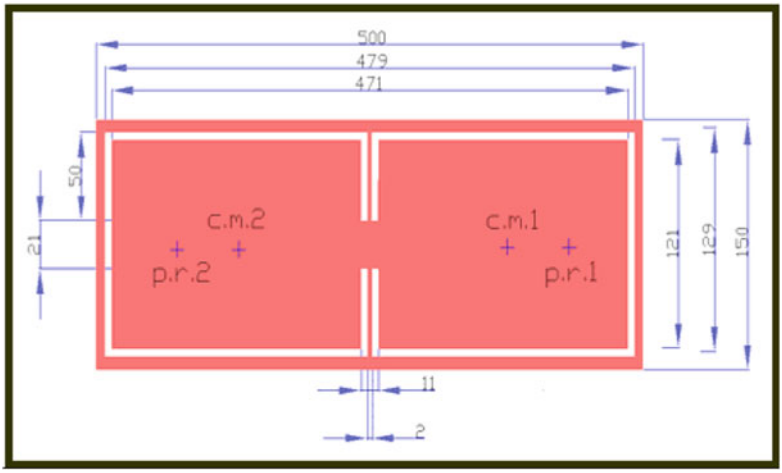


Fig. 29 Mechanical oscillator; the points c.m.1 and c.m.2 identify the COM of the two sides of the proof mass, while the points p.r.1 and p.r.2 indicate the geometrical center position of the two detector capacitors

The scheme of the experimental apparatus used to perform the measurements is shown in Fig. 27. It consists of a rotor which produces the gravitational excitations, and of a torsional oscillator sensitive to the gradient produced by these excitations. More specifically, a rotation of the rotor around the a_r axis, parallel to the z-axis, produces a gravity gradient that acts on the torsional harmonic oscillator, determining its rotation around the a_g axis, lying along the two torsional arms.

In Fig. 28, the block scheme of the whole experimental system is shown. The gravity gradient produced by the rotor forces the torsional oscillator whose displacements are detected by means of a capacitive transducer. The transducer is biased at a frequency $f_p = 10$ KHz by means of an AC generator, so that the signal is seen like a modulation of this bias frequency. After the AC amplifier, the signal is demodulated using a reference frequency coming from the same generator, then converted and read-out.

Figure 29 shows the details of the torsional oscillator. A test mass is connected to an external reference frame by means of two torsional arms. The torsion axis is also the symmetry axis of the test mass and divides it in two sides. The oscillator is sensitive to a gradient of force which produces a consequent momentum. The gravity gradient produces a rotation of the test mass around its arms and is balanced by the torsion momentum of the arms themselves.

As already described in Sect. 2, the harmonic oscillator is described by the following equation:

$$I\ddot{\vartheta} + \eta_t \dot{\vartheta} + \kappa_t \vartheta = \vec{M}, \tag{42}$$

where κ_t represents the elastic constant of torsion, I is the momentum of inertia of the test mass, and η_t is the coefficient of dissipation. The formula connecting these parameters to the mechanical resonance frequency is the following:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{\kappa_t}{I}}. \tag{43}$$

We have to remember that the elastic torsional constant is given by the following formula:

$$\kappa_t = \frac{2GI_p}{l}, \tag{44}$$

where G represents the rigidity modulus, I_p is the inertial polar momentum, and l is the length of the torsional element. G is related to the torsional modulus by the relation $\mu = \pi G/2$. For arms having a rectangular section, indicating with a their width and by h their thickness, one has

$$I_p = \frac{a^* h}{12}\left(a^2 + h^2\right). \tag{45}$$

From the above formulas follows that a low frequency is obtained making thin and long arms and a big momentum of inertia. Manufacturing limitations do not permit to produce oscillators with frequency below some Hz. Another important parameter

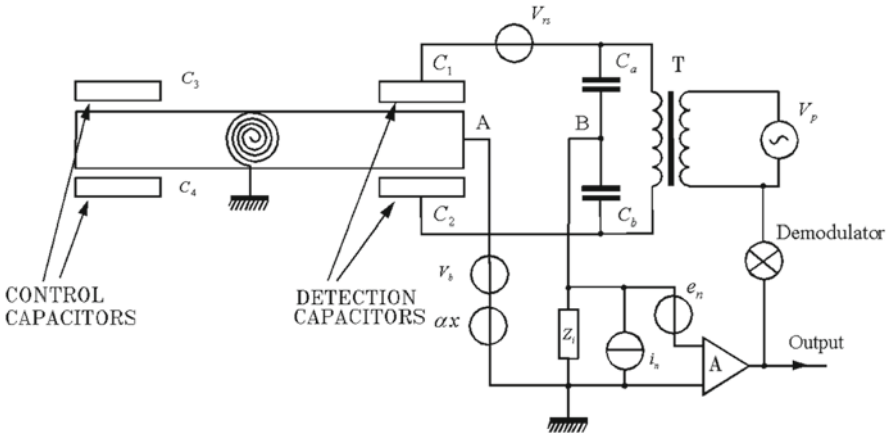


Fig. 30 Scheme for the signal detection

which characterizes the oscillator is its mechanical merit factor, connected to the Brownian noise. This parameter is given by $Q = \omega_0 \cdot I/\eta_t$ and it is connected to the decay constant of the oscillations: $\vartheta(t) = \vartheta_0 e^{-t/\tau_v} \cos(\omega_0 t + \varphi)$, where $\tau_v = \frac{2I}{\eta_t}$ represents the time after which the oscillations are reduced of a factor equal to $1/e$. Clearly, the oscillator transfer function is that of a usual second-order system. In our case, the excitations are at a frequency lower than the resonance frequency, where the transfer function is flat. The relation between the momentum acting on the oscillator and the angular displacements is given by $\vartheta \approx \frac{M}{I\omega_0{}^2}$. The physical characteristics of the oscillator made in aluminum (Al 5056) are the following:

$M = 2.2638[\text{kg}]$; $I = 4.24 \times 10^{-2}[\text{kgm}^2]$; $I_p = 2.6 \times 10^{-12}[\text{m}^4]$; $G = 2.6 \times 10^{10}\,[\text{Nw/m}^2]$; $\omega_0 = 8.2[\text{rad/s}]$; $f_0 = 1.01[\text{Hz}]$; $k_t = \omega_0{}^2 \times I = 1.7\left[\frac{N \times m}{\text{rad}}\right]$.

The signal is detected by means of a capacitive bridge transducer as described in Sect. 2.3; in particular, for this experiment, the pickup scheme has shown in Fig. 30.

The output signal at frequency f_r, due to the bridge unbalance, is seen like a modulation of the bias voltage. A high transducer factor α, between the displacement of the test mass and voltage at the output of bridge is obtained by means of high values of the driving voltage V_p. The bridge's attenuation avoids this high voltage to be present at the input of the amplifier. An alternative configuration to that indicated is obtained using the two capacitors C_1, C_3 to detect the signal. In this case, the system is able to perform a rejection of the flexural oscillation produced by linear accelerations which act perpendicular to the faces of the test mass and do not transfer a momentum on it. The output signal of the capacitive bridge is sent to a low-noise amplifier, working at 10 kHz, where its noise is low. After this, the signal is demodulated and read-out.

The remaining two capacitors, C_3, C_4, are used to control the system: A constant voltage applied across it produces a momentum able to produce a change in the equilibrium position of the sensing mass and a consequent change of the bridge balance conditions. Applying a suitable voltage it is possible to bring the bridge

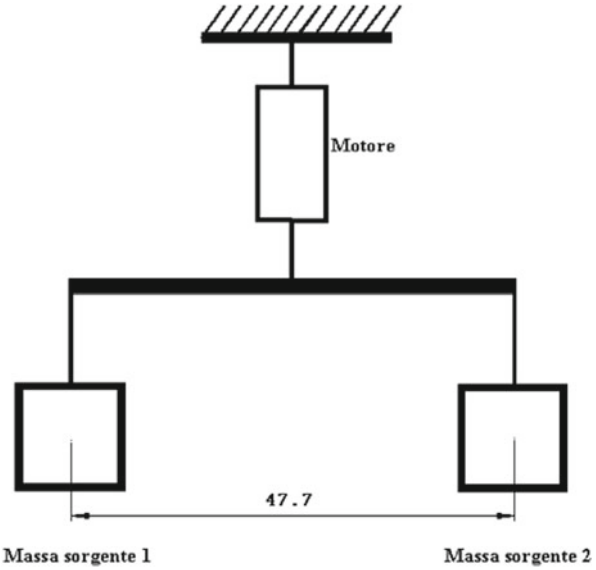


Fig. 31 Scheme of the rotor for the generation of the gravitational signal

to the equilibrium. High bridge attenuation is essential to avoid the amplifier driven voltage and for the reduction of the noise introduced by the driven voltage generator.

In this preliminary measurement, the rotor is made by suspending a system of two masses by means of a wire at a pivot point; an electrical engine produces the desired rotation. The schematic view of the system is shown in Fig. 31.

Two cylindrical masses, whose dimensions are $\Phi = 18$ cm, height $= 11.5$ cm, are placed in a distance of 30 cm apart and connected by means of a rigid structure.

The fundamental frequency of the gravity gradient produced by this system is two times the rotation one, to avoid the possible noise present at this frequency.

6.2 Estimation of the Gravitational Gradient

In this section, the numerical evaluations of the gravitational effects produced by the rotor on the gradiometer are reported. The evaluations are performed in two different ways: concentrated masses and distributed masses.

The first evaluation considers the source masses and the two sides (each one at the opposite side of the torsional arms axis) of the test masses, as concentrated in their COM (see Figs. 32 and 33).

Figure 33 shows the momentum (as a function of time) acting on the torsional oscillator, numerically evaluated. The signal is not perfectly sinusoidal; it shows a pronounced maximum when one of the source masses is over one of the test masses, while the minimum, after a quarter of the period of rotation T_{rot}, is not

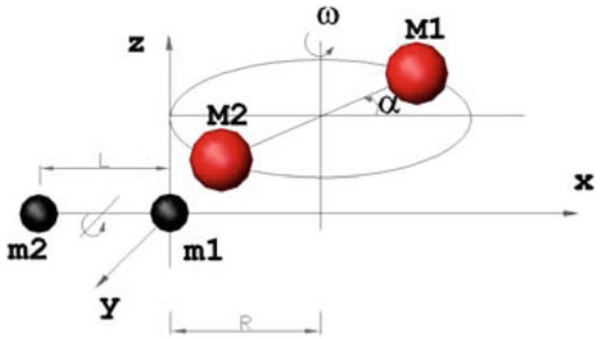


Fig. 32 Scheme for the evaluation of the gravitational gradient in the case of concentrated masses

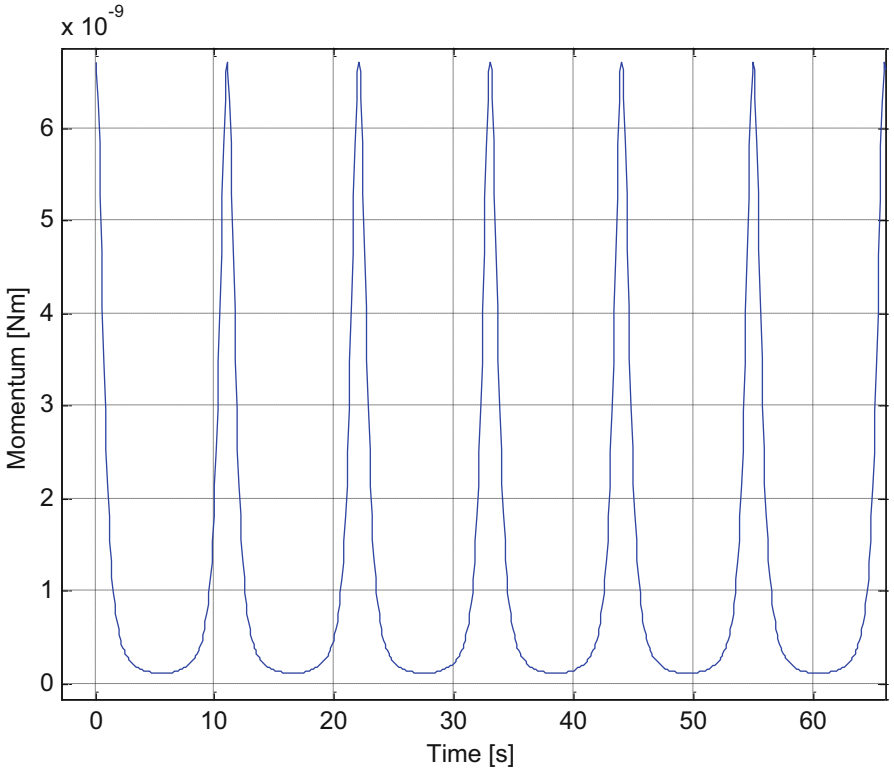


Fig. 33 Signal in the time domain evaluated numerically for concentrated masses

so pronounced. Figure 34 shows the fast Fourier transform (FFT) of the signal, renormalized in order to give the amplitude of the different harmonic components.

The momentum acting on the torsional oscillator, obtained with a numerical evaluation for distributed masses is shown in Fig. 35. The numerical evaluation is

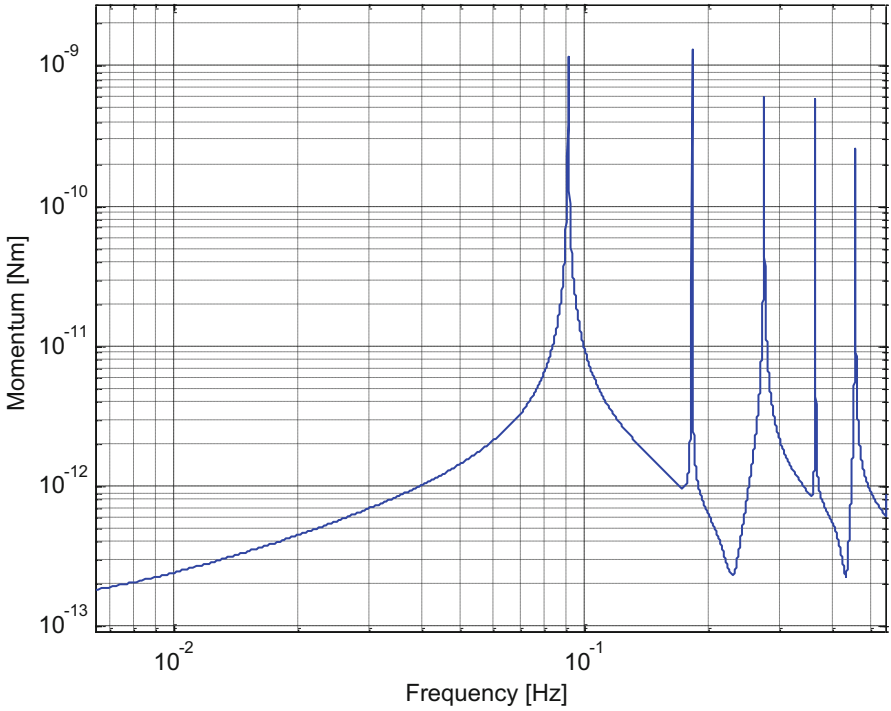


Fig. 34 The FFT of the gravitational momentum (concentrated masses); notice *the different harmonic components*

repeated several times increasing each time the number of elements for the source and test masses. The asymptotic value for the evaluated momentum, as a function of the number of elements considered, gives the sought-for value. The error assigned to this evaluation is taken as the difference between this asymptotic value and the last value obtained. The FFT of the considered signal is shown in Fig. 36.

6.3 Signals and Noise Analysis

The voltage at the output of the capacitive bridge, in the case of $Z_i >> 1/(2\pi f_p C)$ *(*C is the equivalent capacity between the point AB*)*, is given by the formula:

$$V_{AB} = V_{CD}\left(\frac{C_1(\vartheta)}{C_1(\vartheta)+C_2(\vartheta)} - \frac{C_a}{C_a+C_b}\right). \tag{46}$$

Taking into account that $C_1(\vartheta) \approx C_o + \frac{C_o}{d_1} b\vartheta \quad C_2(\vartheta) \approx C_o - \frac{C_o}{d_2} b\vartheta$, we obtain, in the case that the distances between the faces of the capacitors are $d_1 \approx d_2 = d : v_s = \alpha b \theta_s$. In this formula, b represents the distance between the rotation axis

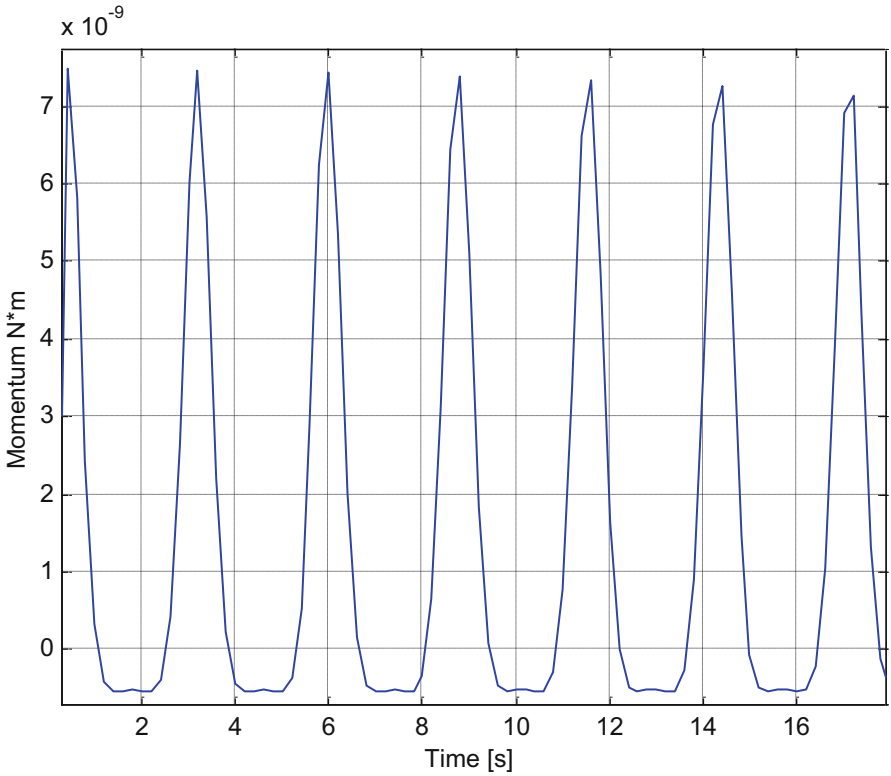


Fig. 35 Numerical evaluation of the gravitational momentum for finite elements evaluation

and the midpoint of the detector capacitors. This expression establishes a relation between the mechanical signal at frequency f_s and the amplitude of the bridge output signal at the frequencies $f_p + f_s$ and $f_p - f_s$, where $\alpha = \frac{nQ_eV_{po}}{2d}$ is the transducer factor between the output voltage and deflection angle of the test mass. It represents also the maximum value of the electrical field inside every detection capacitor.

Concerning the total momentum noise, it can be expressed by the following formula:

$$M_t{}^2(\omega) \approx \frac{4k_b\omega_o}{m_r}\left[\frac{T}{Q_t} + T_n\frac{2Z_nC\omega_o}{\beta}\right]\frac{I^2}{b^2}\Delta f, \tag{47}$$

obtained from some of the considerations described in Sect. 2, where instead of the linear oscillator a flexural one is considered.

6.4 Seismic Noise and Vibrational Noise Rejection

As it will be seen in the results of the performed experimental measurements, the limit in the precision of the experiment is determined by the value of the seismic

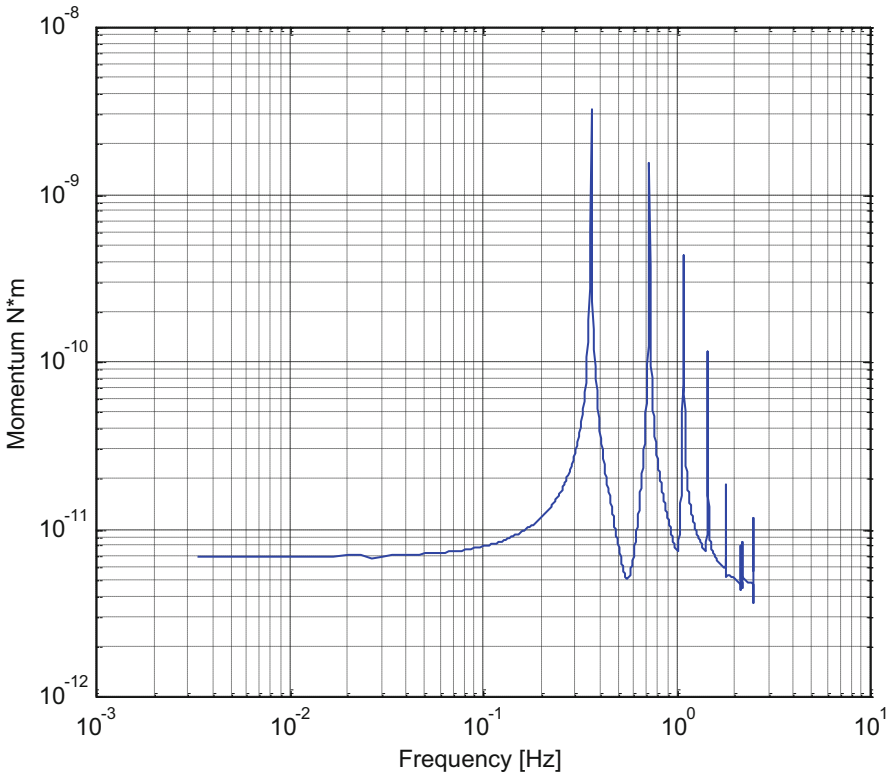


Fig. 36 The FFT of the gravitational momentum for finite elements evaluation

noise that acts on the harmonic oscillator around the frequencies of the selected gravitational signal.

The measurements are usually performed with the harmonic oscillator lying in a horizontal plane with its sensitive axis parallel to the vertical direction. It is clear that the system performs a rejection of the seismic noise part, not changing the momentum of the system.

The execution of these measurements requires, in every case, a place in which seismic noise is low. The preliminary measurements have been performed in the Experimental Gravitation Laboratory of our IAPS-INAF.

6.5 Thermal Noise

The symmetry of the instrument and the choice of the electric and mechanic components with which it has been built, make it stable under temperature variations.

Generally, the measurements are performed in an environment in which the variations of temperature are sufficiently low and a thermal control system is used. In particular, the effects of the temperature variations effect should be avoided near the frequencies of the gravitational signal. Usually such frequencies are of the order of the tenth of Hz. At these frequencies, the environmental thermal variations are reduced and the time thermal constant of the system contributes to a further reduction of their effects.

Variations of temperature over longer periods can give rise to some problems (as big signals at the output of the bridge, with consequent saturation of the amplifier).

6.6 Integration

The particularity of this experiment resides in the fact that the gravitational signal to be detected is periodic with a good stability in frequency; this fact allows performing a long integration, with a consequent reduction of the seismic and thermal environmental noise and of the intrinsic noise of the instrument. The limit in the integration time is determined by the frequency stability of the gravitational signal and by the stability of the read-out system. Besides, as already said, all the effects that produce signal variation with time (also at low frequency) that can determine the saturation of the instrument must be eliminated or reduced. Preliminary estimates point out the necessity to perform measurements for time periods of the order of 10^4 s.

6.7 Calibration and Metrological Errors

The metrological errors are those caused by the not perfect determination of the quantities involved in the theoretical evaluation of the effect, but for which it is possible, at the beginning, to establish the entity of their weight.

The calibration of the system is performed by exciting it with a known acceleration. Such acceleration is produced using small test masses which are set on one side of the harmonic oscillator test mass so that to produce, under the action of the terrestrial gravity, a known momentum. We have to remember that the harmonic oscillator is always in neutral equilibrium in presence of gravity. This method introduces an error in the measurement of the mass, which is done with a balance calibrated in a different position and at different times and therefore, with a different value of g acting on it. There is also an error in the evaluation of the position of the small masses with respect to the torsional arms. The following issues are also to be considered: nonlinearity of the signal and possible presence of high-order terms, stability of the calibration with time. The source and oscillator masses must be known with high precision in their geometric structure, homogeneity and density. Their shape has, moreover, to be simple, possibly spherical, so that to easily be manageable and to allow the estimate of the gravitational acceleration produced by it. The structure of

the source masses has to be such (cylindrical or spherical) not to introduce multipole moments of higher order, so that to facilitate the theoretical calculation of the gravitational effect produced by them. It must be remembered that the procedure adopted by us foresees a numerical calculation with the automatic insertion of the higher order multipoles.

6.8 Systematic Errors

This type of error is connected to effects on the measure which are not easily predictable. Among these errors those due to wrong evaluations of the experimental quantities, values measured under different environment conditions (pressure and temperature) and in different positions (measure of the masses with different local g values) can be considered. Such errors are responsible for the variety of estimates of G done in different experiments, with differences that exceed of a factor 40 the pointed out precision.

6.9 Estimation of the Sensitivity

From the results of the experimental measurements the signal-to-noise ratio is of the order of 10^{-4}; such value also points out the precision that can be obtained from the measure: The seismic noise being predominant in comparison to the intrinsic one due to the experimental apparatus, being possible, in principle, to reduce the metrological errors to a very low level.

6.10 Experimental Measurements

In this section, the experimental measurements of the system parameters are reported.

6.10.1 Measurement of the Mechanical Parameters

The measurement of the resonance frequency of the system is done by setting the torsional oscillator in the same position that will be used during the measurements of the gravitational effects. The local vertical is perpendicular to the face of the oscillator test mass. In this way, the torsional arms are stressed in traction and they suffer a small deformation. This deformation is such that the fundamental frequency is not only given by a torsion of the arms but also from their traction. This condition determines an increase of the system frequency. In the sensitivity calculations, the

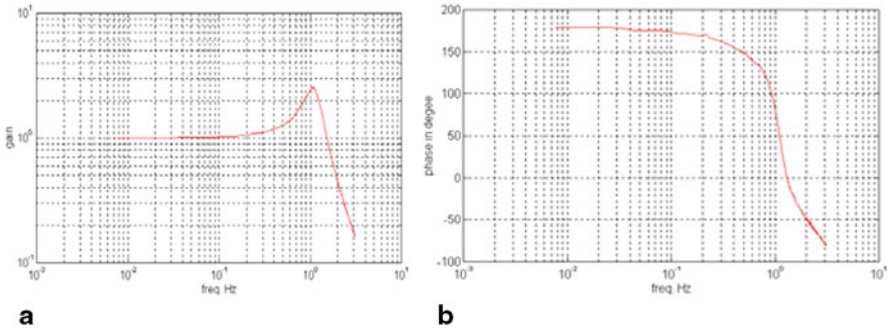


Fig. 37 **a** Transfer function of the system (amplitude). **b** Transfer function of the system (phase)

experimentally measured frequency value is used. The experimental measurement is performed by exciting the system with a mechanical solicitation; the signal is read-out with the capacitive bridge; after the amplifier, the signal is sent to a spectrum analyzer. A more refined measure is performed by setting the oscillator under vacuum so that to increase its mechanical quality factor, but leaving it low enough so as to be able to excite it with an electric signal; the same one, together with the response of the system, is sent to a spectrum analyzer. The system transfer function in amplitude and phase is shown in Fig. 37a and 37b.

The experimental frequency is $f_{0sper} = 1.01\text{Hz}$. The measurement of the mechanical quality factor is performed with the torsional oscillator in the same position as in the case of the measurement of the frequency, in this case the analyzer is used for visualizing the signal in the time domain. A mechanical solicitation produces the oscillation to the resonance of the system, whose free decay provides the value of the decay constant (value connected to the factor Q_m).

6.10.2 Measure of the Transfer Function in Environmental Conditions

Considering the fact that the experimental measurements are performed at environment pressure (condition in which the harmonic oscillator behaves as an overdamped system, because of the presence of gas trapped between the faces of the detection capacitors), some measurements of characterization of the system have been performed under these conditions (Fig. 38a and 38b).

Also for these measurements, the two control capacitors are used, with which the mechanical system is excited with an electric signal whose frequency is made to vary in the range required for the transfer function measurement. The transfer function in amplitude and phase measured experimentally is shown in Fig. 38a and 38b.

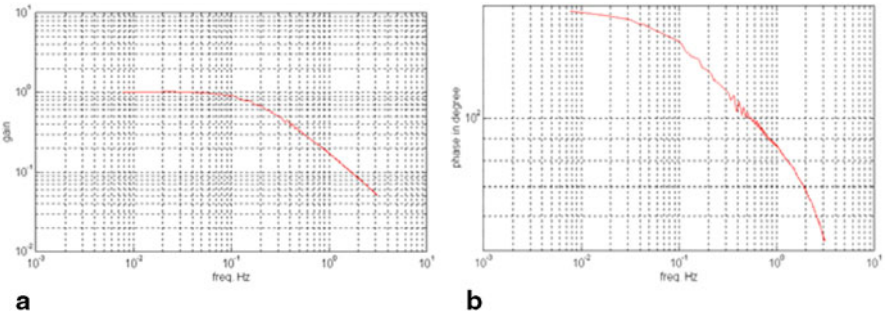


Fig. 38 **a** Transfer function of the system in environmental conditions (amplitude). **b** Transfer function of the system in environmental conditions (phase)

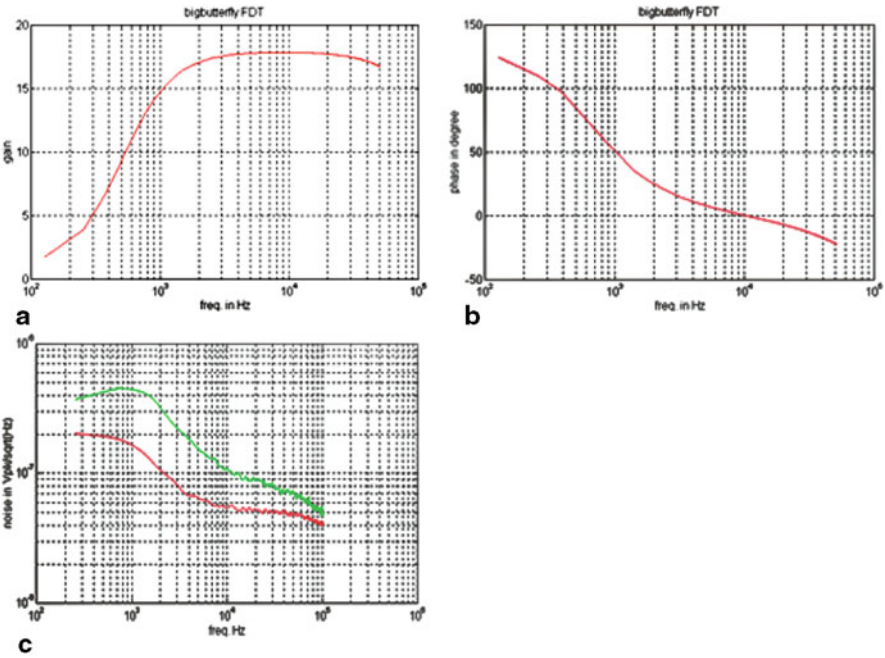


Fig. 39 **a** Characterization of the amplifier: gain versus frequency. **b** Characterization of the amplifier: phase versus frequency. **c** Noise introduced by the preamplifier. *Green: 4pF/44MOhm; Red: Zi = 22pF/44MOhm*

6.10.3 Measurement of the Electrical Parameters

In this section, the characterization of the amplifier is reported. The amplification as a function of frequency is shown in Fig. 39a, while the phase between the input signal and the output of the amplifier is shown in Fig. 39b.

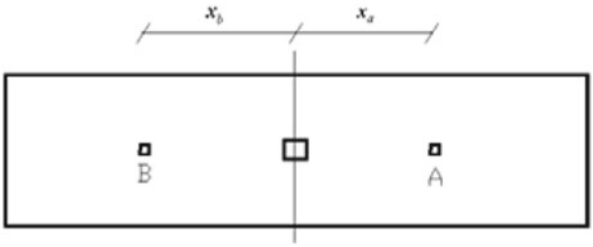


Fig. 40 Configuration for the calibration of the system

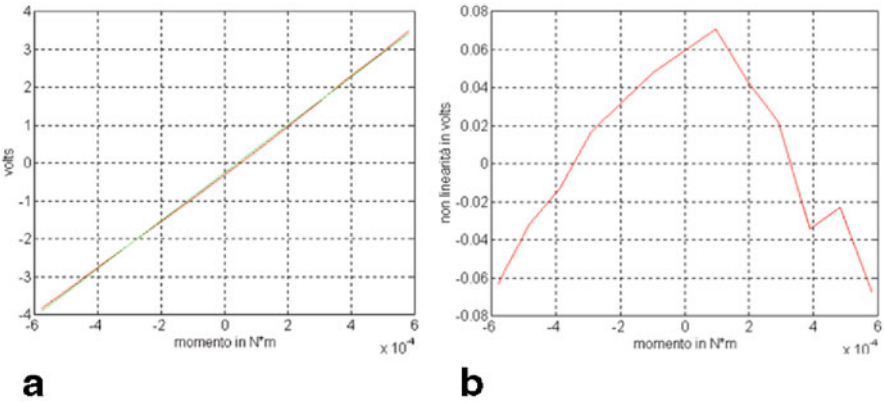


Fig. 41 **a** Output signal versus the momentum. **b** Difference between the slopes shown in the previous figure, this result shows the nonlinearity of the system

In Fig. 39c, the spectral density of the amplifier noise with its input impedance close to two different impedances, whose values are reported in the picture caption, is reported. From the plot it is possible to infer the values of the voltage and current of the amplifier at 10 kHz: $e_n = 50$ nV/sqrt(Hz), $i_n = 18$ fA/sqrt(Hz).

6.10.4 Measurement of the Transducer Factor

The transducer factor of the system represents the relationship among the momentum applied to it and the output voltage of the system. The momentum is applied by placing some small calibration masses (weight of 80 mg) in a suitable position so that to determine a known momenta (Fig. 40).

The calibration configuration is shown in Fig. 40. The calibration masses were placed in the points A and B, where $x_a \approx x_b \approx 121 \pm 2$mm.

The output signal versus the momentum corresponding to the number of calibration masses arranged in A and B is shown in Fig. 41a. In the same figure, it is also shown the linear fit of the data.

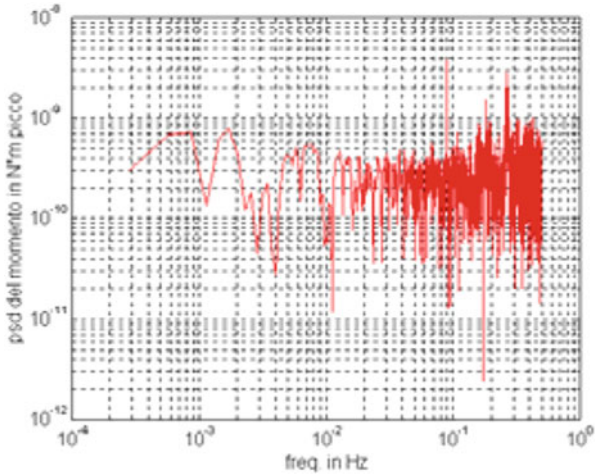


Fig. 42 The FFT of the seismic noise recorded in the laboratory, in absence of the gravitational signal (rotor in steady position)

As shown in Fig. 41b, the difference between the two precedent slopes, indicating the nonlinearity of the system. The obtained transducer factor is: $\alpha = \frac{M}{V} = 1.58 \times 10^{-4} \frac{N \times m}{V}$.

The nonlinearity produces a maximum displacement between the two curves equal to 0.06/7.5, equivalent to 1.6 %. To increase the precision it is necessary to use a transducer factor function of the output values. To avoid the effects depending on the variation of g, the weighting of the calibration masses must be performed in the place where the experiment is executed.

6.10.5 Measurement of the Seismic Noise

The value of the seismic noise, particularly at the frequencies to which the gravitational signal is produced, is one of the main effects that limit the measurement precision.

The FFT of the noise recorded for a period of 1 h in absence of gravitational signal is shown in Fig. 42. In the frequency band of interest, the differential noise seen by the oscillator is under 10^{-9} N m.

6.10.6 Measurement of the Gravitational Effect

The measurements of the gravitational effects are performed by setting the torsional accelerometer inside a metallic box. The centers of mass of the two sides of the proof mass were made close as much as possible to the upper wall of the box, so that to maximize the produced gravitational effect. The metallic box assures the electric

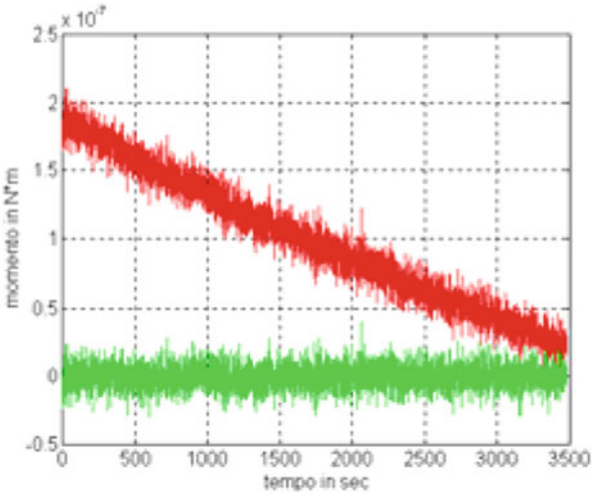


Fig. 43 In *red*, it is shown the gravitational signal produced by the rotation of the masses. In *green*, it is shown the same signal after a detrend

insulation of the system. Outside the box it is positioned the rotor, composed of two masses of the same weight (29 kg each). The frequency of the gravitational signal produced in this way is two times the rotation frequency, to avoid noise effects induced by the rotation of the system.

For a further check some tests are performed using a simulacrum of the masses, of the same volume but of 1 kg in mass. The low-noise amplifier is set inside the box, while the rest of the electronics is set in a separate box. An important issue is the equilibrium of the capacitive bridge that handles the reduction of the output bridge for the driven voltage. The equilibrium is obtained with two variable ceramic capacitors. The experimental results are shown in the following plots that respectively represent: Figure 43 shows the time slope of the signal; Fig. 44 is a detail of the previous plot in which the gravitational signal is visible, superimposed to the same signal filtered at high frequency, it shows the sinusoidal form of the fundamental frequency of the gravitational signal; Fig. 45 is the fast Fourier transform (FFT) of the signal shown in Fig. 44 that emphasizes the different harmonics; Fig. 46 is analogous of Fig. 45 but in a log-log scale; a comparison of the FFT of the signals produced by the rotor with masses of 29 kg and using a simulacrum with masses of 1 kg is shown in Fig. 47.

6.10.7 Evaluation of G

Taking into account the peak value of the first harmonic from the simulation and the value obtained with the experimental measurements, we deduced a value for the gravitational constant equal to $G = 8.2 \times 10^{-11} \frac{\mathrm{Nm}^2}{\mathrm{kg}^2}$. This value differs by

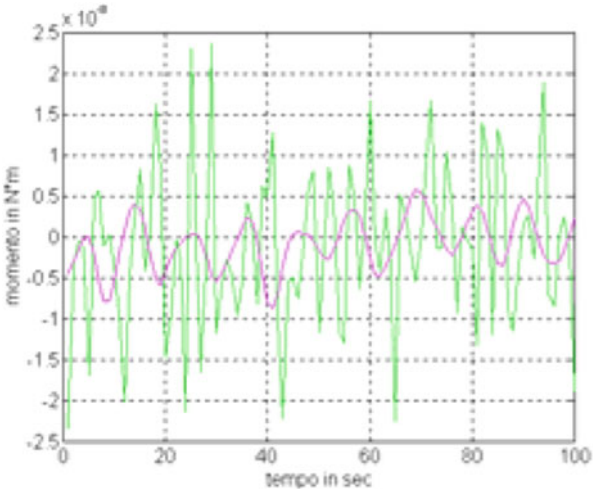


Fig. 44 Detail of the previous plot where it is possible to see the gravitational signal (*green*), superimposed to the same signal filtered at high frequency (*cyan*)

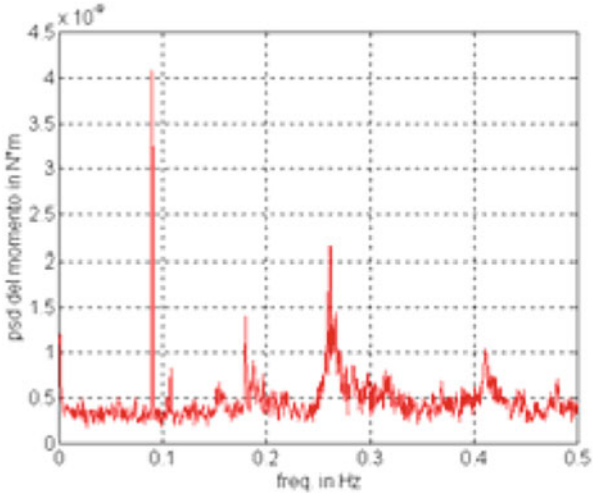


Fig. 45 FFT of the gravitational signal showing the harmonics of the gravitational signal

about 160 % with respect to the CODATA value. The difference among the two values is substantial, but it is compatible with the error that can be attributed to our measurement; this error, keeping track of the various relative errors, is in fact of the order of $\frac{\Delta G}{G} = 20/100$. The experimental error has a value equal to $\frac{\Delta G}{G} = 0.195 = 19.5/100$.

Such a deviation is substantial besides with the error relative associate to the same measurement. Of course, with these first evaluations we do not certainly claim

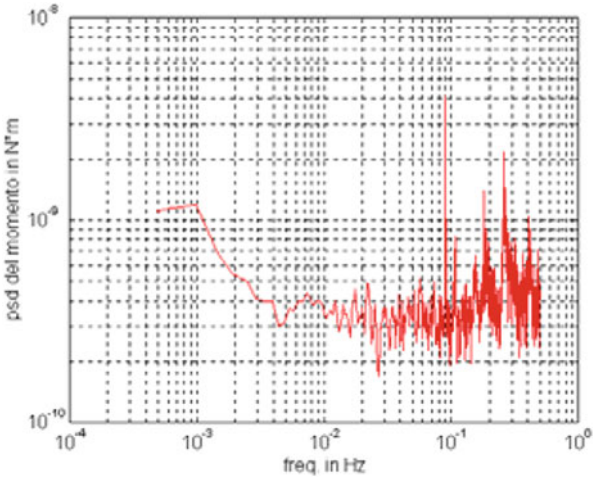


Fig. 46 FFT of the signal in a log-log scale

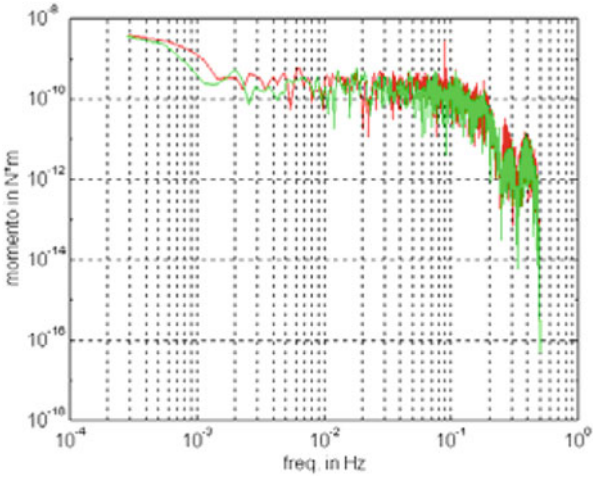


Fig. 47 Comparison between the FFT of the signal produced by the action of the rotor with the source masses (*red*) and the signal obtained by using the simulacrum with mass equal to 1 kg (*green*)

to point out a value of G, but rather to describe a method and a new tool for the measurement of G. The tool and the described procedure seem to be very promising for future refined measures of G. Compared with the classical torsion balances the proposed instrument seems unfavorable, since its frequency is higher than with respect to the first one; but it must be remembered that the proposed instrument is excited in a frequency zone where the rotational seismic noise takes on its lower level. A last consideration for this apparatus is in its possibility to be used in free

fall and therefore on-board a satellite, unlike the torsion balances in which gravity is essential to determine the tension of the suspension thread of the test mass. The proposed instrument, by its nature, performs the rejection of the common mode noise; in the laboratory, it is subject to a very low differential seismic noise and it is, therefore, possible to use it for the characterization of the electronic noise of the signal of pickup systems at the levels of maximum sensitivity. In this respect, it is necessary to consider that the system has a very low frequency and therefore its transducer factor is higher compared to the accelerometer that uses a harmonic oscillator with resonance frequency of the order of 3 Hz. From what results from a first investigation, the instrument can be considered as a new tool for geophysics; in fact with it, it is possible to perform measurements of rotation of the plan on which it is set. Measurements executed in this way seem to be complementary to the gravimetric and clinometric measurements. The activity seems very promising, but it emerges clearly that subsequent refined measurements require big efforts, either financial and of time.

7 Space Experiments Using Satellite Tracking: ISA for BepiColombo

In this section, we describe a different class of experiments consisting in the use of tracking techniques to measure the position and velocity of a satellite; these measurements are used in a high precise orbit determination of the satellite with a simultaneous evaluation of gravitational field parameters, useful to determine the structure of the planets, their rotational state and to test the theories of gravitation. In particular, we focus on the Radio Science Experiment (RSE) conceived for the ESA Cornerstone mission to Mercury named *BepiColombo*, giving major attention to the use of an accelerometer that enables to measure the acceleration acting in the specific position where it is installed inside the satellite. This allows us to refer the precise orbit determination (POD) with respect to this point, that can be regarded as a free-fall point, if the linear acceleration acting on it and measured simultaneously by the accelerometer and the tracking system are removed. See also Ref. [24, 26].

BepiColombo is the first ESA interplanetary mission direct to a planet (Mercury) close to the Sun; the mission is especially challenging because Mercury's orbit is so close to our star, and difficult to reach due to the enormous gravity of the Sun. Only the NASA missions Mariner 10 and MESSENGER have visited Mercury so far.

The BepiColombo mission, consisting of two orbiters, will be launched in July 2016 with an Ariane 5 rocket from ESA Spaceport in Kourou, French Guiana, and it will arrive at Mercury in 2024 (Fig. 48). In its journey, the spacecraft will use several gravity assists around Earth, Venus, and Mercury in combination with solar electric and traditional chemical rocket propulsion. During the voyage to Mercury, the two orbiters and a transfer module, consisting of electric propulsion and traditional chemical rocket units, will form one single composite spacecraft.

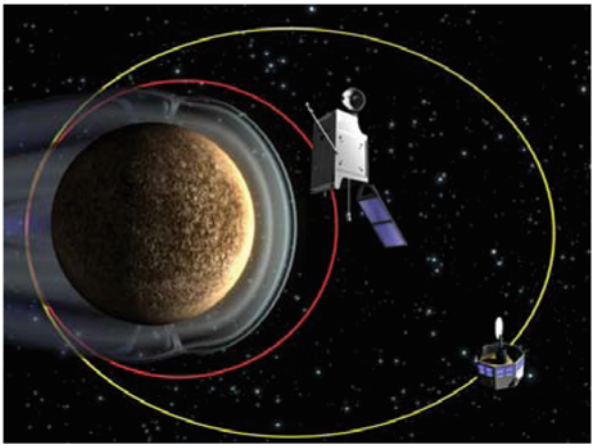


Fig. 48 Artist impression of the Bepi Colombo spacecraft

Once in orbit, the two spacecrafts will start their specific mission with observations and measurements for at least one Earth year, with a possible extension of another year.

Mercury Planetary Orbiter The MPO is a three-axis stabilized spacecraft with a 400×1500 km polar orbit for coverage of the whole planet. The on-board payload is devoted to experiments in remote sensing and radio science, among other things. The main payloads of the MPO are:

- *Wide-angle and narrow-angle camera* used to map the planet surface and for the Libration experiment. The cameras require a relative pointing error ≤ 10 μrad/s and an attitude determination better than 3 μrad (post-flight).
- IR spectrometer, UV spectrometer, X-ray spectrometer, γ-ray spectrometer, neutron spectrometer, to study the rock composition of the planet, to detect the presence of ice under the surface of the planet, for the atmosphere composition, and the Sun activity.
- *Accelerometer* to detect the inertial accelerations acting on the MPO.
- *Ka-band transponder* to detect the MPO and Mercury positions referred to the Earth.

The last two instruments are used for global mapping of Mercury gravity and for tests of fundamental aspects of the GR theory.

Mercury Magnetospheric Orbiter (MMO) MMO will be released from the MPO in a $400 \times 12{,}000$ km polar orbit; the spin-stabilized spacecraft will rotate at 15 rpm. The magnetospheric payloads are *magnetometer, ion spectrometer, ion/electron analyzer, wave analyzer, cold plasma detector, energetic particle detector.*

The Mercury magnetic field and its interactions with the wind of particles coming from the Sun will be observed. The vicinity of the Sun gives rise to a magnetosphere with important differences with respect to the Earth one.

7.1 Radio Science Experiments

The RSE are aiming at four scientific goals, to be achieved by means of four different measurements.

The scientific goals are as follows:

1. Measure the rotation state of Mercury, so allowing to accurately constrain the size and physical state of the core of the planet.
2. Measure the global structure of the gravity field of Mercury, so allowing to accurately constraining its internal structure.
3. Measure the local gravitational anomalies of Mercury, so allowing to accurately constrain the structure of the mantle and the crust/mantle interface.
4. Measure the orbit of Mercury around the Sun and the propagation of electromagnetic waves between the Earth and Mercury, in order to improve the determination of the parameterized post-Newtonian (PPN) parameters β and γ, and test GR to an unprecedented level of accuracy.

The measurements are as follows:

1. The range and range rate between the ground stations and the spacecraft, having removed the effects of the plasma along the path by means of a multi-frequency link in X and Ka band.
2. The non-gravitational perturbations acting on the spacecraft, by means of an on-board accelerometer.
3. The absolute attitude of the spacecraft, by means of a star tracker.
4. The angular displacement, with respect to previous passages, of portions of the Mercury surface, by means of pattern matching among camera images.

The scientific results can be obtained only by combining data from four different pieces of equipment. Of the latter, only one (the accelerometer) is dedicated to these experiments (although it could also be used to monitor the performance of the ion propulsion).

The scientific goals of the RSE are an integral part of the BepiColombo science value, making it a cornerstone mission, and requires to study and simulate these experiments in a global way, with special attention to the exceptionally complex spacecraft/instrument interfaces.

Mercury Centric Orbit Determination To obtain the goals 2 and 3, it is necessary to determine the position of the spacecraft in a Mercury centric frame, with cm-level accuracy, for the time span of the effective high-accuracy (multi-frequency) tracking. The solution has to be obtained in the context of the multi-arc method: The orbit is split in several arcs, and for each arc the initial conditions are determined. This allows us to absorb the systematic long-term effects of the unmodeled perturbations, while the use of the accelerometer ensures medium-term (10^4 s) stability of the dynamic model at the cm level. At the same time, global parameters such as the Mercury gravitational potential coefficients are determined. This is conceptually not too different from satellite geodesy around the Earth, but in practice a software

ready to perform this class of experiments for gravimetry around Mercury has to be used.

Mercury Orbit Anomalies For goal 4, the deviations of the orbit of Mercury from existing models (JPL ephemerides) have to be determined, with cm accuracy. This can be obtained by a linear propagation of small deviations, provided the time span is not too long. The additional unknown parameters would then appear in the equation for observation (range and range rate). The solution for β and γ would then be obtained by a fit of a number of these orbital anomalies.

Orbit Propagation Goal 1 requires also the knowledge of the Mercury centric spacecraft position, but the accuracy required is at the level of a few meters. This implies that the orbit can be propagated for a time span covering the gap between two tracking sessions.

Integrated Radio Frequency System Range and range rate are the essential observables to be used for orbit reconstruction and therefore it is essential to devise a radio system (both on-board and at the ground stations) capable to ensure the measurement accuracy. An advanced Doppler system using multi-frequency links at X and Ka band has been developed for the Cassini mission. The goal to which DSN engineers are committed is for a frequency stability (Allan deviation) of 3×10^{-15} for an integration time between 1000 and 10,000 s. Such high frequency stability is probably not necessary for achieving the scientific objectives of the RSE, whose Allan deviation target is 10^{-14} for the same integration times. It is, however, important to remark that a Cassini-like configuration allows us to achieve this goal in the Mercury environment for a large range of operational conditions.

7.2 The ISA Accelerometer

The accelerometer selected for the BepiColombo mission—the ISA—has been developed at IAPS/INAF and is now under development for space use by Thales Alenia Space–Italia. A prototype with a sensitivity of $10^{-9}\mathrm{ms}^{-2}/\sqrt{\mathrm{Hz}}$, one order of magnitude better of what required for the RSE, in the required frequency band ($3 \times 10^{-5} \div 10^{-1}$ Hz), has already been implemented and tested for a long time.

The implementation work of the ISA accelerometer has to take into account several requirements and conditions, among them we can recall the following ones:

- Right sensitivity and frequency band of the accelerometer, taking into account the satellite environment (temperature and vibrational noise).
- Thermal and mechanical interfaces with the MPO, including constraints on vibrations, accommodation with respect to the MPO center of mass (COM) (position, position accuracy, and stability), interface temperature excursion and short-term stability, and the thermal design of the accelerometer, with the goal of minimizing its mass.

- Minimizing of the effects of any disturbances unlikely to be eliminated by spacecraft design (small-amplitude, low-frequency vibrations).
- Identification of suitable calibration methods and strategies.

7.2.1 ISA and the Scientific Performance Requirements

The ISA accelerometer will not directly measure none of the physical quantities connected with the previously cited goals. Hence, we are not able to link with a 1:1 correspondence the instrument performances with a given physical characteristic of Mercury or to a parameter of GR. Anyway the benefits of an on-board accelerometer have been well evidenced by the RSE simulations performed by Milani et al. 2003. The accelerometer role is just to measure the non-gravitational accelerations produced on the MPO structure by the incoming solar visible radiation and by the planet albedo (Mercury Environment Specification (Part I), SCI–PF/BC/TN01, 4 July 2002). These perturbing accelerations are then removed from the list of unknowns (practically, some of the unknown parameters are replaced by the accelerometer calibration coefficients), in such a way that the MPO spacecraft (i.e., the vertex of ISA at which all the measurements are referred) should be considered, *a posteriori*, as a drag-free satellite falling—along a geodesic—in the pure gravitational fields of Mercury and the Sun (and, to a lesser extent, of the other planets, in particular of Jupiter).

This target is essential for an accurate orbit determination of the MPO around Mercury and, as a consequence, of Mercury's COM around the Sun. Consequently, in the case of the gravimetry experiment, the Mercury gravitational anomalies are well isolated with respect to any other perturbation which can act with similar effects. The harmonic coefficients of Mercury gravity field are solved through a global least squares fit, using all the data, together with the corrections to the orbit of the planet, the initial conditions (state vector) for each orbital arc and, of course, the accelerometer calibration constants. In particular, the accelerometer thermal stability with respect to the long-period thermal effects represents a strong requirement for a full success of the RSE. From the simulation results, it is clear that the thermal effect with a 44 days periodicity (at half of the sidereal period of Mercury around the Sun) must be reduced, using an active control, at a level below 10^{-8} m/s^2, considering a time variation of more than 9 h (connected to the minimum of the ISA frequency band). It will be therefore possible to determine the coefficients of the spherical harmonics of the planet gravity field, up to a degree $l = 30$, with a signal-to-noise ratio still greater than 1 (and larger than 10 to the degree $l = 20$). Concerning the GR experiment (Milani et al. 2003), the same thermal condition enables the determination of the PPN parameter γ, at the level of 2×10^{-6}.

As shown in Fig. 49, the envelope of the total noise considered for the RSE, coming from the accelerometer and the tracking, in terms of equivalent accelerations. The red line represents the ISA noise, considered as white and at a level of $10^{-9} m/s^2/\sqrt{Hz}$. The green line represents the thermal noise due to a possible white noise at a level of $4°C/\sqrt{Hz}$, present at the mechanical interface between the MPO

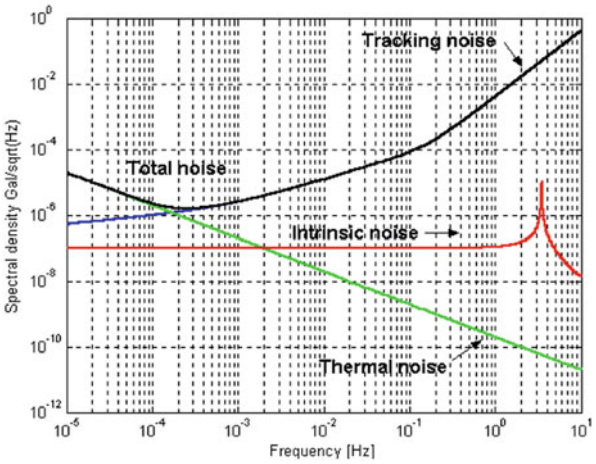


Fig. 49 Total noise considered for the RSE, coming from the accelerometer and the tracking, in terms of equivalent accelerations

and ISA, attenuated by the passive thermal filter of ISA. The blue line represents the total noise for the tracking expressed in acceleration. The black line represents the total noise (quadratic sum of the previous ones). This plot is a good guideline to set the limits of the vibrational noise allowed on the MPO. It is necessary to remark that this limit has been obtained in case of a passive thermal control. In the same hypothesis of thermal variations and using an active thermal control (now the baseline), the attenuation to consider is 700. In this case, the green line is flat at level of $2.8 \times 10^{-9} \mathrm{m/s^2}/\sqrt{\mathrm{Hz}}$ at frequency below approximately 10^{-3} Hz; at higher frequencies, the attenuation increase of 20 dB per decade.

7.2.2 Instrument System Description

The ISA instrument configuration is based on two units: The "ISA Detector Assembly" (IDA) and the "ISA Control Electronics" (ICE) which interfaces electrically the MPO as shoved in Figs. 50 and 51. The measurement concept is the same as described in Sect. 2. As shown in Fig. 52, the general architecture of the instrument, where it is possible to see the three sensors with the relative sections for biasing the transducers and to pickup the signal, amplify, and read-out it after the demodulation. Also indicated are the modules for accelerometer calibration, thermal control, reference, processing and control.

Mechanical Arrangement

The implementation of the three-axis accelerometer is made by three elements arranged with their centers of mass along the same axis (which, if possible, will

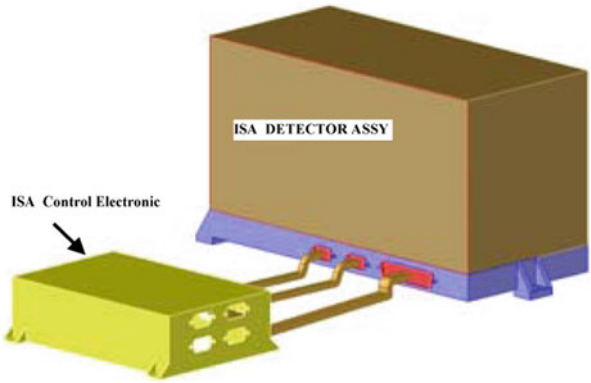


Fig. 50 ISA experiment concept sketch

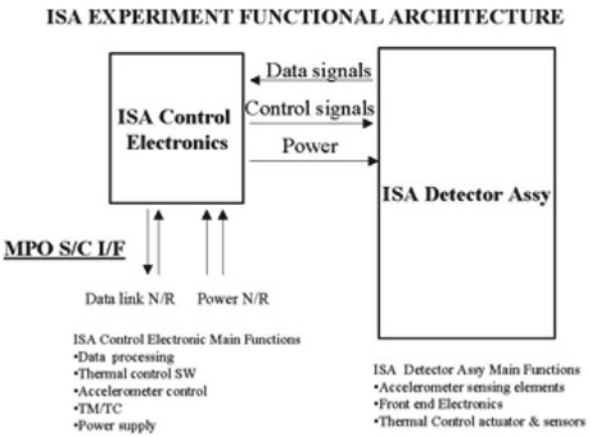


Fig. 51 ISA experiment high level block diagram

be made coincident with the MPO rotation axis). This choice allows avoiding the inertial effects due to the nominal angular velocity and angular acceleration, equal to zero for points lying along the rotation axis; the gravity gradient effects will be minimized too, especially if the sensitive element with axis parallel to the rotation one is positioned with its COM coincident with the spacecraft COM. The three accelerometer elements with the sensitive axis of the central element aligned along the satellite rotation axis (Z-axis) are shown in Fig. 53; the origin of the reference frame is the spacecraft nominal COM.

In Sect. 2.2 of this chapter, the indications concerning the noise analysis for each element of the accelerometer are reported. In Fig. 54, the drawing of the mechanical oscillator with its dimensions is reported. In Fig. 55, the three mechanical elements are shown in their new arrangement, installed inside a box necessary to guarantee

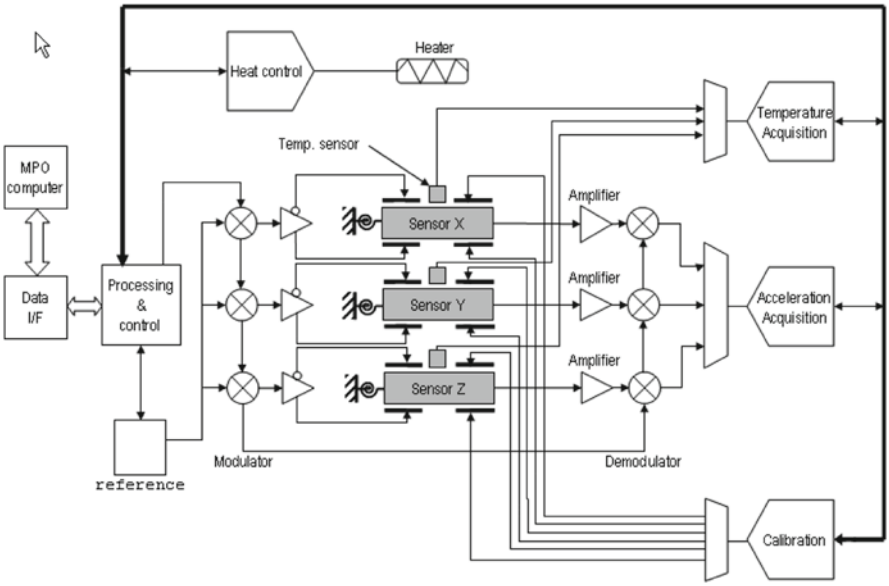


Fig. 52 General functional instrument architecture

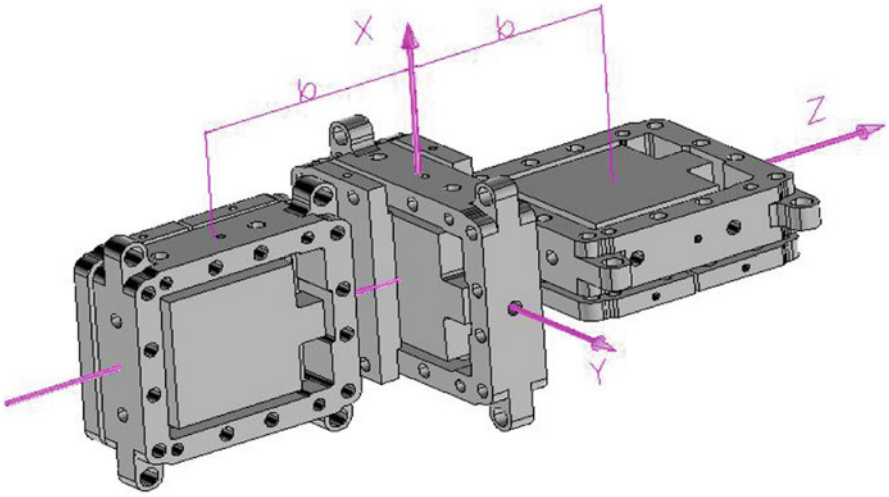


Fig. 53 Arrangement of the three sensitive axes

the alignment of the sensitive axes and the conditions for a proper thermal insulation (Fig. 56).

In Fig. 56, the external aspect of the box, with the indication of the attach points to the other external boxes is shown.

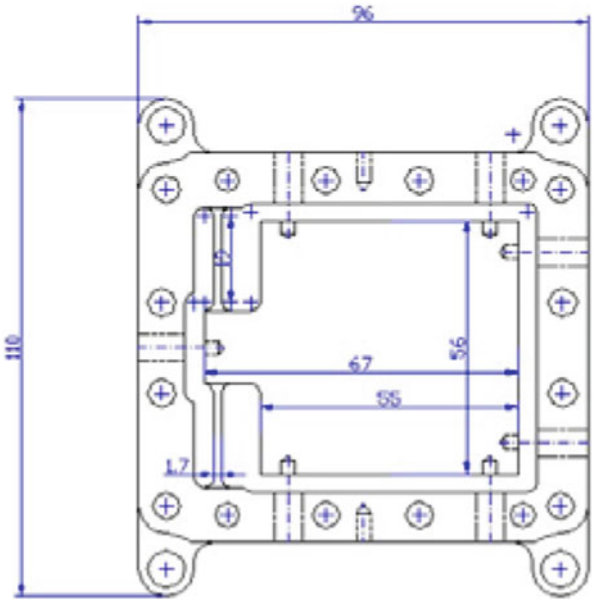


Fig. 54 Dimensions of the new ISA mechanical oscillator prototype

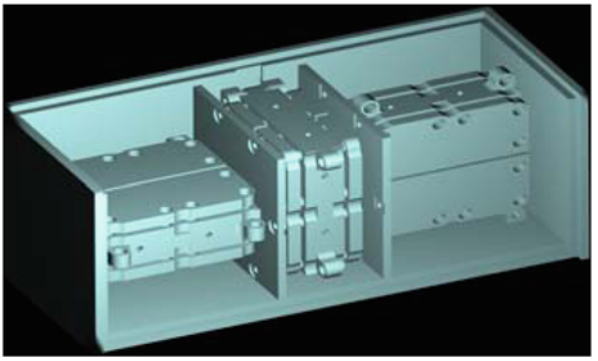


Fig. 55 New ISA mechanical elements installed inside the box

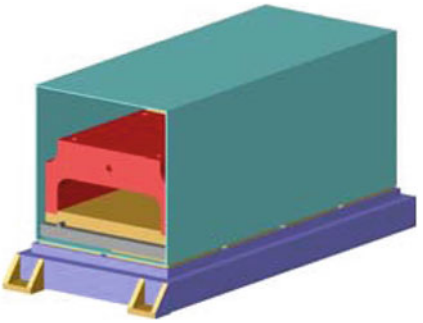


Fig. 56 Box containing the three mechanical elements

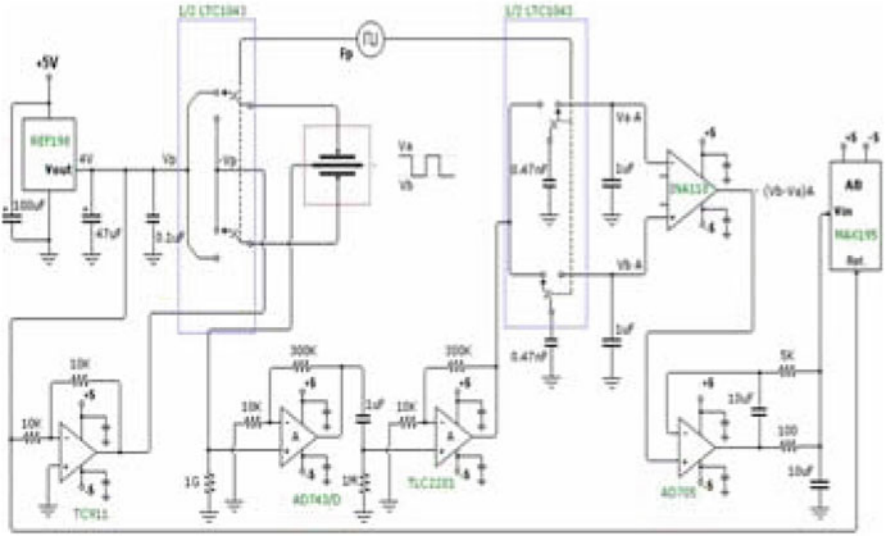


Fig. 57 Analog section of electronics prototype

Low Dissipation Electronics

The criteria used to implement the electronics for the accelerometer are the following ones:

- Obtain the right sensitivity of the apparatus
- Be very compact
- Have the lowest level of power dissipation
- Be very stable with respect to temperature changes

A very stable reference is used by a modulator to generate an alternate voltage that biases the capacitive bridge. The modulated signal is sent to a low-noise amplifier followed by a demodulator; after the demodulation the signal is sent to an analog-to-digital converter (A/D) that uses the same reference signal. A microcontroller commands the demodulation, the averaging, the filtering and the other functionality of the accelerometer.

The analog section of the electronics for one element of the accelerometer is shown in Fig. 57. It has reference, modulator, low-noise amplifier and demodulator (Fig. 58).

The digital section of the new ISA electronics is shown in Fig. 58. It consists of a microcontroller that takes the three signals and makes some operation on them and controls also the output system.

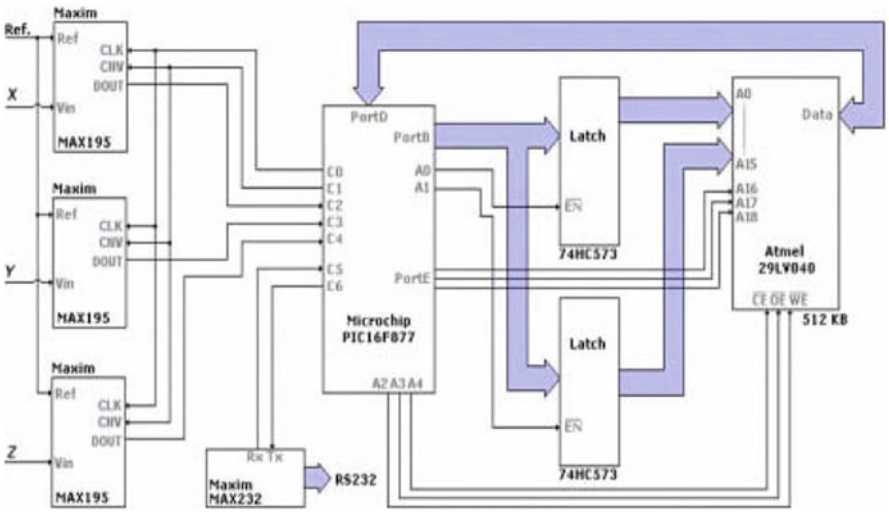


Fig. 58 Digital section of the new ISA electronics of prototype

7.2.3 Instrument Operations Concept and Its Error Budget

The accelerometer continuously measures the three components of the acceleration acting on the MPO as well as the inertial and gravitational gradient components acting on its proof masses. The concept of the instrument operation is to detect all the accelerations that perturb its pure gravitational orbit. In fact, the MPO in orbit around Mercury is a not an inertial system; in every points on it (in particular, in the COM of each proof mass) acts an acceleration given by the known formula_37, that we report here for commodity:

$$g\left(\vec{r},t\right)=-\vec{\nabla}V(\vec{r},t)-\vec{\Omega}\wedge\left(\vec{\Omega}\wedge\vec{r}\right)-2\vec{\Omega}\wedge\vec{v}-\vec{\Omega}\wedge\vec{r}-\vec{\alpha}(t), \qquad (48)$$

where $V\left(\vec{r},t\right)$ is the gravitational potential, $\vec{\Omega}$ is the angular velocity, $\vec{r}$ is the position vector on the satellite from the reference origin to a point on the MPO (COM of each proof mass), $\vec{v}$ is its velocity vector, and $\vec{\alpha}\left(t\right)$ is any linear acceleration acting on the MPO. The first three terms on the right-hand side can be easily recognized as gravitational, centrifugal and Coriolis accelerations respectively. The last term depends on the variation of the angular velocity. On the reference system of the "free falling" MPO it is possible to measure also the acceleration gradients.

Inertial Effects These effects are related with the last four terms of the previous formula_49. We can divide them in linear and angular accelerations.

Linear Accelerations Every point in the satellite is under the same action of linear accelerations; these can be external to the satellite, as the solar radiation pressure, and internal to the satellite as in the case of the motion of the mechanical parts

(reaction wheels among them), fuel sloshing, and mechanical vibrations with high mechanical merit factor. These accelerations are the main components that need to be measured in order to reduce the effect of perturbations on the gravitational orbit of the MPO. It is necessary to remember that ISA is not an absolute accelerometer and so, it is not able to measure constant accelerations, and the reliability of its output is only for measurements in its own dynamic range.

Angular Accelerations For the angular accelerations the effect on the satellite is different on points at different distances from the rotation axis; to avoid these effects the accelerometer must be installed with its COM very close to the MPO COM (possibly coincident). Also in this case it is necessary to underline that ISA cannot detect constant angular accelerations, but only terms varying with time and, clearly, if the MPO COM is away from the COM of the ISA proof masses.

Coriolis Acceleration This effect arises only if there is a motion of the proof masses. Recalling that the accelerometer is hard mounted on the MPO and each proof mass con move only with one degree of freedom, the Coriolis acceleration, acting perpendicularly to this direction has no effect on it.

Gravitational Effects The gravitational effects on the ISA proof masses can be induced by a mass in the satellite and by a mass external to it. The gravitational effects induced by the satellite mass are, in general, constant and so, they are not detected by ISA. Gravitational effects induced by masses, which are in motion on the satellite, produce effects that are variable with time and likely to be detected. Such effects can be due to reaction wheels, fuel sloshing, level change of the fuel, and so on. Gravitational effects induced by masses external to the satellite are mainly due to Mercury and to the Sun.

Thermal Effects Thermal variations inside the MPO have very strong implications on the performance of the accelerometer, acting directly on it and inducing perturbations through the thermoelastic deformation of MPO structure. Recalling that the accelerometer stability is $5 \times 10^{-7} \mathrm{m/s^2/\,^\circ C}$, this means that to avoid the impact of the thermal effects on it, the thermal variations must be less than $2 \times 10^{-2} ^\circ \mathrm{C}$ in the frequency range of the accelerometer.

Inertial Effects Induced by the Thermal Variations ISA is not an absolute instrument and cannot detect constant signals (gravitational and inertial), but depending on the temperature change, inertial and gravitational constant effects can be "modulated" by the change of the COM position of the proof mass and by the change of the sensitive axis direction.

Intrinsic Accelerometer Noise The last noise term that we like to mention here is the intrinsic accelerometer noise. It is necessary to recall that this noise is flat in the ISA frequency range, and at the level of $10^{-9} \mathrm{m/s^2/\sqrt{Hz}}$. In Table 10, we report a list

Table 10 Signals and noise acting on the accelerometer

Effect	Type	Characteristic
Solar radiation pressure	Signal	Modulated at orbital frequency and spin
Albedo and infrared emission from Mercury	Signal	Modulated at spin frequency
Attitude maneuvers and spacecraft buffeting	Signal	Random
Electromagnetic effects	Signal	Random + modulated at orbital frequency
Orbiter spin rotation	False signal	Modulated at spin frequency
Changing of orbiter mass	False signal	Colored noise during maneuvers
Mercury gravity gradient on the accelerometer	False signal	Modulated at orbital frequency
Thermal variations in the instrument mounting plate	Noise	Colored noise, 1/f
Sinusoidal thermal variations	Noise	Modulated at orbital frequency
Differential heating of the MPO	Signal	Modulated at orbital frequency
Instrumental noise	Noise	White noise

of these effects, indicating their type and characteristic. The effects are divided into three categories: Signal: all the inertial effects that produce a real displacement of the MPO COM, measured by ISA. These measurements give useful information to recover the MPO gravitational trajectory by tracking the MPO from the Earth. False signal: gravitational effects inducing signals on the accelerometer, but not useful to recover the gravitational trajectory of MPO, because the tracking takes them into account. Noise: effects on the accelerometer not related to displacements of the MPO COM.

7.3 ISA_BC On-Ground Calibration

In Sect. 2, we have already introduced the accelerometer on-ground calibration here we only remember that for the ISA accelerometer proposed for BepiColombo the sensitivity is $10^{-9}\mathrm{m/s^2}/\sqrt{\mathrm{Hz}}$.

7.4 ISA_BC In-Cruise and In-Orbit Calibrations

The ISA in-cruise and in-orbit calibrations will be performed forcing each of its three elements with a known acceleration signal using the respective actuators. The actuators will be previously calibrated on ground. The electrical force corresponding to a given voltage applied to the actuators is not connected with the mechanical frequency but only to the distances between the two plates of the actuators capacitors.

Table 11 Error budget for the pseudosinusoidal noise. Here, n represents the MPO mean motion ($\cong 7.52 \times 10^{-4}$ rad/s), ω_0 ($\omega_0 \propto n$) and $\dot{\omega}_0$ $\left(\dot{\omega}_0 \propto n^2\right)$ are, respectively, the MPO angular rate and angular acceleration, $\Delta\vec{R}_0$ and $\Delta\vec{R}_t(t)$ are the displacement errors (see the text), 2 °C represents the amplitude of the thermal effect (that will be attenuated by a factor 700 by the ISA active thermal control), finally $\Delta\alpha$ represents the error (orbital period component) of the time-independent misalignment angle of the accelerometer. The MPO orbital period P is about 8355 s. The last column gives the maximum percentage of error over one orbital period of the MPO with respect to the requested accuracy A_0 (with $A_0 = 10^{-9} g_{\oplus}$)

Type	Due to	Spectral content	requirement on	Error % A_0
Gravity-gradients	n	Orbital period P and ½ P	$\Delta\vec{R}_0; \Delta\vec{R}_t\,(t)$	85
Apparent forces	$\omega_0; \dot{\omega}_0$	Orbital period P and ½ P		
Thermal effects	2 °C	Orbital period P	ΔT	*15*
Components coupling	Misalignment angle	Orbital period P	$\Delta\alpha$	*Negligible*
Total				*100*

An appropriate electrical force applied on them enables to determine the transducer factor and the ISA transfer function. A particular calibration can be performed by using dedicated maneuvers of the MPO in order to determine the alignment of the ISA sensitive axes with respect to the MPO body reference system and so with respect to an inertial reference system.

7.5 ISA_BC Error Budget

In this section, we report schematically the error budget of ISA, with the purpose to introduce the main parameters of interest and to deduce the requirements imposed by the on board accelerometer to the MPO spacecraft. Such analysis has given indications to the Prime in order to identify the best solution for the MPO attitude control system. The ISA error budget can be estimated taking into account separately the two kinds of noise:

I. Sinusoidal or pseudosinusoidal contributions
II. Random contributions

In Tables 11 and 12, we specify the contribution of each type of error with respect to the accelerometer accuracy $A_0 = 10^{-9}$ g. The noise of the first kind is mainly due to Mercury's gravity gradients (i.e., its tidal forces) and to the apparent forces on the rotating spacecraft, while the noise of the second kind is mainly due to the angular-rate and angular-acceleration noise due to the MPO attitude control, necessary to guarantee the spacecraft nadir pointing. The displacement $\vec{R} = (\Delta X, \Delta Y, \Delta X)$ of each proof mass with respect to the spacecraft COM can be seen as the sum of four different terms:

$$\vec{R} = \vec{R}_0 + \Delta\vec{R}_0 + \vec{R}_t(t) + \Delta\vec{R}_t(t), \tag{49}$$

Table 12 Error budget for the random noise. The vectors $\vec{R}_0$ and $\vec{R}_t(t)$ define, respectively, the ISA position matrix and the MPO COM movements due to the HGA and fuel consumption. The quantities $\delta\omega$ and $\delta\dot{\omega}$ are due to the attitude control necessary to keep the MPO at its nominal angular rate and nominal angular acceleration, $4°C/\sqrt{Hz}$ represents the amplitude of the thermal effect (that will be attenuated by a factor of 700 by the ISA active thermal control), $\Delta\ddot{R}_t$ is the random acceleration noise arising from the error part of displacement of the COM as function of time, finally $\delta\alpha$ represents the error (random component) of the misalignment angle of the accelerometer. The last column gives the maximum percentage of error with respect to the requested level of noise A (with $A = 10^{-9}\mathrm{g}_{\oplus}/\sqrt{\mathrm{Hz}}$), here assumed equal to the accelerometer sensitivity. The errors are assumed not correlated

Type	Due to	Spectral content	Requirement on	Error %A
Apparent forces	$\vec{R}_0;\vec{R}_t\ (t)$	Random	$\delta\omega;\ \delta\dot{\omega}$	*60*
Thermal effects	$4°\mathrm{C}/\sqrt{\mathrm{Hz}}$	Random	ΔT	*30*
Noise on the MPO	Movements due to the HGA, fuel sloshing, ...	Random		*10*
MPO COM displacement	Movements due to the HGA and fuel consumption, ...	Random	$\Delta\vec{R}_t\ (t)$	*70*
Components coupling	Misalignment angle	Random	$\delta\alpha$	*Negligible*
ISA intrinsic noise		Random		*Negligible*
Total (not correlated noise)				*< 100*

where $\vec{R}_0$ represents the proof mass position in the MPO frame with respect to the nominal position of the spacecraft COM, $\Delta\vec{R}_0$ represents the errors in knowledge of such position. The spacecraft COM is not fixed in position in the MPO frame because of the fuel consumption and sloshing and the high gain antenna (HGA) movements. The $\vec{R}_t(t)$ term accounts for such displacements that therefore impact on the accelerometer measurements. The last term $\Delta\vec{R}_t(t)$ accounts for the error in the knowledge of the MPO COM movements: the errors in the determination of $\vec{R}_t(t)$. In other words, $\vec{R}_0$ and $\Delta\vec{R}_0$ define the time-independent part of the proof masses positions with respect to the spacecraft COM, while, $\vec{R}_t(t)$ and $\Delta\vec{R}_t(t)$ are the time-dependent part of the proof masses positions with respect to the COM of the spacecraft. The objective of this analysis is to determine the value of the main parameters involved, so to keep the error budget for the pseudosinusoidal terms below 10^{-8} m/s^2 and that of the random contributions below $10^{-8}\mathrm{m/s^2}/\sqrt{\mathrm{Hz}}$.

As we can see, the first kind of error is directly connected to the determination of the ISA positions $\left(\Delta\vec{R}_0 \text{ and } \Delta\vec{R}_t\right)$, imposing requirements on them, while the second kind of error is connected directly to the ISA positions $\left(\Delta\vec{R}_0 \text{ and } \Delta\vec{R}_t\right)$, imposing requirements on these positions as well as in the attitude control errors ($\delta\omega$, $\delta\omega$).

We define nominal effects or nominal noise the acceleration errors coming from the sum of the gravity-gradients effects with those produced by the apparent forces. The maximum disturbing accelerations produced by these nominal effects are function of the accelerometer proof masses displacements ΔX, ΔY, and ΔZ with respect

to the MPO COM ($\Delta\vec{R}_0$ and $\Delta\vec{R}_t(t)$ in Table 16). The 85 % impact for the maximum error from this nominal noise arises from the 15 % constraint for the thermal effect. This is dictated by the thermal stability of the accelerometer to the temperature variations (about $5 \times 10^{-8} g_{\oplus}/°C$) and to the thermal attenuation that we are able to reach inside the accelerometer band, in particular in the low-frequency part of the accelerometer bandwidth, that is, in the $10^{-4} \div 10^{-3}$ Hz band. The ISA active thermal control guarantees attenuation by a factor of 700. As we can see from Table 12, the thermal effect at the MPO orbital period has amplitude of about 2 °C; we, therefore, obtain for the thermal disturbing acceleration:

$$A_{therm.}|_{orb.} \cong 2°C \times 5 \times 10^{-8} \frac{g}{°C} \times \frac{1}{700} = 1.4 \times 10^{-10} g, \tag{50}$$

which is about 15 % of the requested accuracy. With regard to the misalignment angle, estimated to be 1/3000 rad, its contribution to the error budget is negligible. Indeed, the maximum non-gravitational acceleration at the MPO orbital period is due to the direct solar radiation pressure, and it has been estimated to be $a_{sun} \cong 10^{-6} m/s^2$; we, therefore, obtain:

$$A_{mis.ang.}|_{orb.} \cong 1 \times 10^{-6} m/s^2 \times \frac{1}{3000} = 3.3 \times 10^{-10} m/s^2 \tag{51}$$

that is about 3 % of the requested accuracy, and therefore negligible with respect to the other contributions.

With regard to the results shown in Table 12, we assumed the random noise sources not correlated and we add their contributions in a root-sum-square fashion. The acceleration noise due to the thermal effects has been estimated as in the previous case, see Eq. (50). The random component of the misalignment angle is much smaller than that estimated in Table 12, therefore its contribution to the error budget is negligible.

7.5.1 Instrument Requirements to Spacecraft

In this section, the requirements imposed to the spacecraft and to other instruments, connected to the use of ISA on board the MPO and to its level of precision inside the frequency band are identified. In the following, it is reported a list of these requirements on the basis of the accelerometer error budget previously reported.

ISA Noise Requirement Inside the Frequency Band

The total noise in terms of acceleration undergone by the ISA accelerometer must be below the ISA required noise in the measurement frequency band ($3 \times 10^{-5} \div 10^{-1}$ Hz). In Fig. 59, the required noise as function of frequency is plotted.

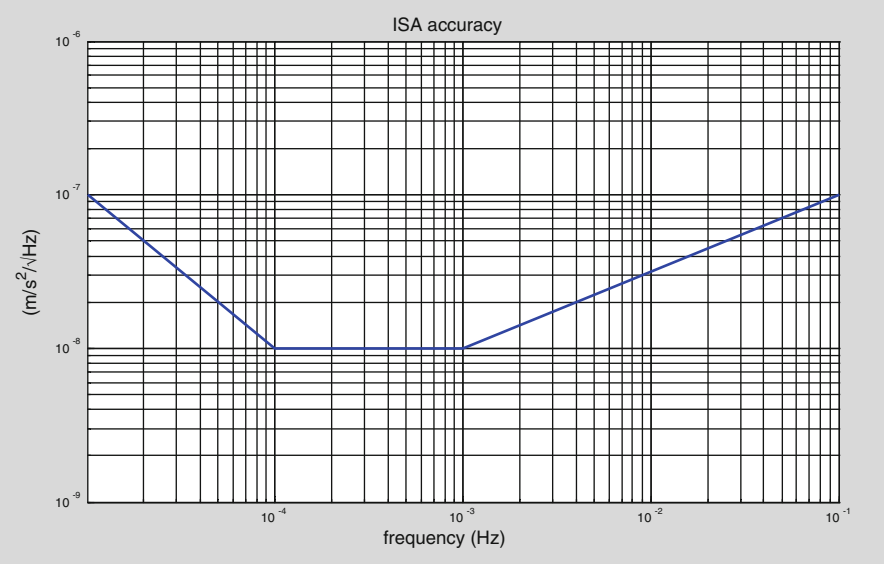


Fig. 59 ISA noise requirement in the measurement frequency range

Table 13 New values for ISA required noise

Frequency (Hz)	3×10^{-5}	$10^{-4} - 10^{-3}$	10^{-1}
Acceleration values ($m/s^2/\sqrt{Hz}$)	3×10^{-8}	10^{-8}	10^{-7}

Table 14 Microvibration requirements inside the measurement frequency band: $3 \times 10^{-5} - 10^{-1}$ Hz

Frequency (Hz)	3×10^{-5}	$10^{-4} - 10^{-3}$	10^{-1}
Acceleration values ($m/s^2/\sqrt{Hz}$)	3×10^{-9}	10^{-9}	10^{-8}

In Table 13, these values for the more important points, in order to reconstruct the curve are summarized.

Vibrational Random Noise On-Board the MPO Inside the Frequency Band

The dynamical noise on board the MPO, induced by all the instruments present on it, must be kept at 10 % of the ISA precision requirement: $10^{-9} m/s^2/\sqrt{Hz}$ in the range $10^{-4} - 10^{-3}$ Hz, $10^{-8} m/s^2/\sqrt{Hz}$ at 10^{-1} Hz and $3 \times 10^{-9} m/s^2/\sqrt{Hz}$ at 3×10^{-5} Hz.

This kind of noise is mainly due to the HGA movements and to the propellant sloshing, but also the noise introduced by other instrumentation with moving mechanical parts must be considered. In Table 14, these values are summarized.

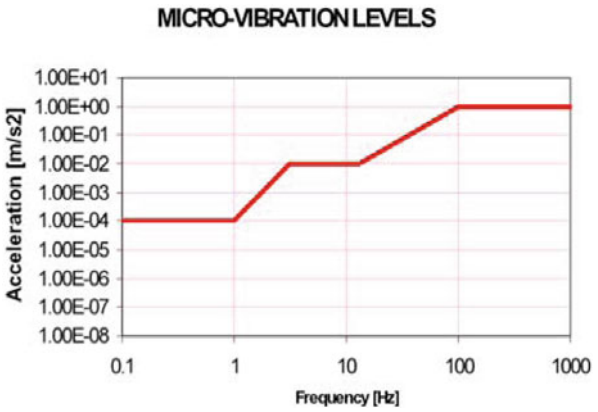


Fig. 60 Out of band total noise

Micro-vibration Random Noise On-Board the MPO, Outside the Frequency Band $10^{-1} - 300$ Hz

The dynamical noise on board the MPO must be below the ISA required precision in the frequency band, and reduced at high frequencies so to avoid aliasing problems during the acquisition phase, the saturation of the preamplifier and the saturation of the A/D converter. In Fig. 60, the micro-vibration requirements in the amplitude of peak noise that can be present on the MPO (m/s^2) so to avoid aliasing noise and ISA saturation, at the relevant frequencies are reported.

It is necessary to remember that the indicated values are valid when only one peak of noise is present. When more than one peak is present it is necessary to take in consideration the value at the different frequencies of their envelope.

MPO Thermal Requirements

The thermal variations induced by the MPO spacecraft on the mounting plate, where ISA is installed, will be attenuated by its active control, providing their reduction at a level such that the induced effects are reduced under the required precision inside the frequency band. The main values to take into account to evaluate these effects are:

ISA thermal instability, equal to $5 \times 10^{-7} m/s^2/°C$.

ISA thermal attenuation factor provided by the active control, equal to 700.

In Table 15 the temperature requirements to the MPO at the sidereal and orbital periods and random is reported.

Table 15 MPO temperature variations at the orbital period, sidereal period, and random

Frequency	T (°C)
Mercury revolution period	25 peak-peak
MPO orbital period	4 °C peak-peak
Random noise	$4°C/\sqrt{Hz}$

Table 16 Values for the components of $\vec{R}$ (in mm) used for the evaluations

Axis (mm)	$\overrightarrow{R_0}$	$\Delta\overrightarrow{R_0}$	$\overrightarrow{R_t}(t)$	$\Delta\overrightarrow{R_t}(t)$
X	0	±4	30	±1
Y	0	±5.5	20	±2
Z	0	±11	40	±4

Requirements on ISA Positioning

Remembering the definition of the vector indicating the distance between the instantaneous MPO COM and that of the ISA center of mass:

$$\vec{R} = \vec{R}_0 + \Delta\vec{R}_0 + \vec{R}_t(t) + \Delta\vec{R}_t(t)$$

given in terms of time-independent and time-dependent parts, we can write the following requirements:

a. The accelerometer must be placed with its COM coincident with that of the MPO.
b. The three ISA elements com are aligned along an axis and this must be made coincident with the MPO rotation axis.
c. A position away from the MPO COM is also possible but this implies more stringent requirements on $\delta\omega$ and $\delta\dot{\omega}$.
d. For ISA installed with its com coincident with the MPO COM, the different values for the components of $\vec{R}$ are as reported in Table 16.

Different choices of these parameters are also possible but it is necessary to give more stringent requirement on the MPO attitude control system.

MPO Attitude Control System Requirements

Starting from the values of $\vec{R}_0$ and $\vec{R}_t\,(t)$ reported in Table 16, we computed the maximum allowable noise of MPO attitude and orbit control system (AOCS) due to angular rate and angular acceleration.

In Figs. 61 and 62 are reported the values of $\delta\omega$ and $\delta\dot{\omega}$ as function of the *out of plane* component of the distance, values enabling to reach the required ISA accuracy in the frequency band $10^{-4} \div 10^{-3}$ Hz, when the other two components are fixed ($X_t = 30$ mm, $Y_t = 20$ mm).

In Figs. 63 and 64, the requirements for $\delta\omega$ and $\delta\dot{\omega}$ spectral density in the whole measurement band ($3 \cdot 10^{-5} \div 10^{-1}$ Hz) in the condition of ISA com coincident with the MPO COM and having fixed all the components of the $\vec{R}$ vector are reported. The green line represents ISA values, while the red one is the expected AOCS spectral density noise derived in the hypothesis of white noise for the attitude noise.

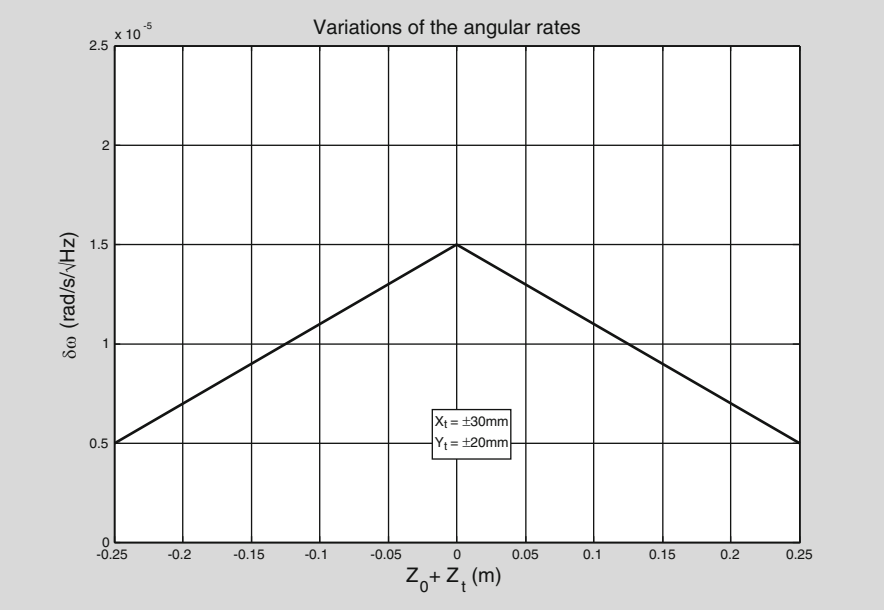


Fig. 61 Values of $\delta\omega$ as function of the out of plane component of the distance $\vec{R} = \vec{R}_0 + \Delta\vec{R}_0 + \vec{R}_t\,(t) + \Delta\vec{R}_t\,(t)$ enabling to reach the required ISA accuracy in the frequency band $10^{-4} \div 10^{-3}$ Hz, when the other two components are fixed ($X_t = 30$ mm, $Y_t = 20$ mm)

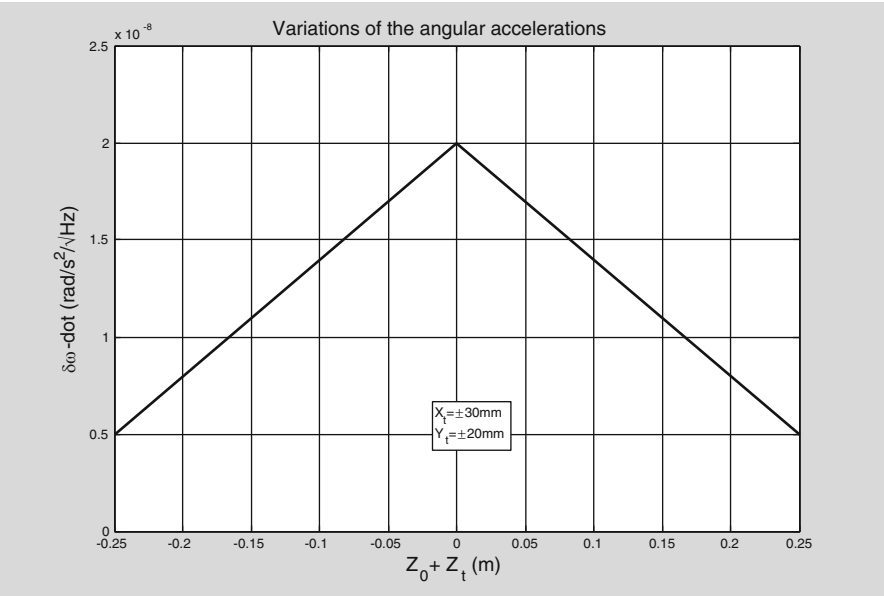


Fig. 62 Values of $\delta\dot{\omega}$ as function of the out of plane component of the distance $\vec{R} = \vec{R}_0 + \Delta\vec{R}_0 + \vec{R}_t\,(t) + \Delta\vec{R}_t\,(t)$ enabling to reach the required ISA accuracy in the frequency band $10^{-4} \div 10^{-3}$ Hz, when the other two components are fixed ($X_t = 30$ mm, $Y_t = 20$ mm)

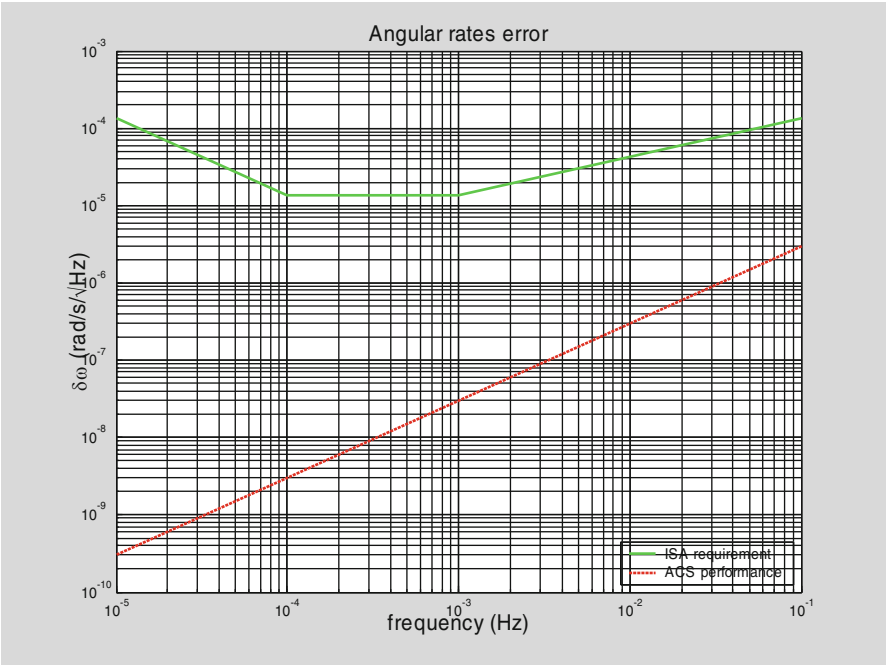


Fig. 63 Values of $\delta\dot{\omega}$ spectral density as function of frequency, having fixed all the components of the $\vec{R}$ vector: $\vec{R}_0(t) = (0;0;0)$ and $\vec{R}_t\,(t) = (\pm 30 \pm 20 \pm 40)$ mm

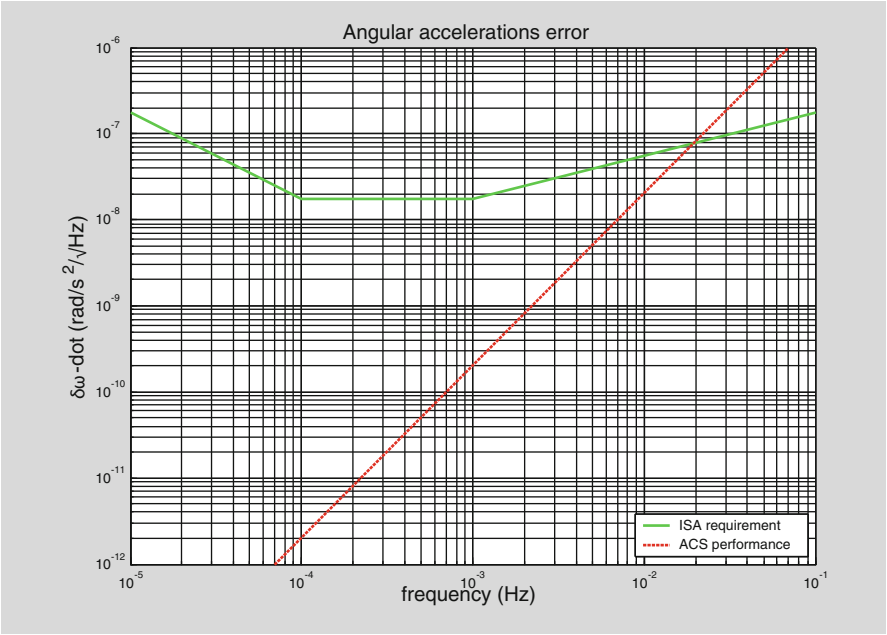


Fig. 64 Values of $\delta\omega$ spectral density as function of frequency, having fixed all the components of the $\vec{R}$ vector: $\vec{R}_0(t) = (0;0;0)$ and $\vec{R}_t(t) = (\pm 30 \pm 20 \pm 40)$ mm

Table 17 Values for $\delta\omega$ and $\delta\dot{\omega}$ necessary to guarantee the ISA accuracy in the frequency band $10^{-4} - 10^{-3}$ Hz

$\delta\omega$[rad/s/$\sqrt{\text{Hz}}$]	1.5×10^{-5}
$\delta\dot{\omega}$[rad/s^2/$\sqrt{\text{Hz}}$]	1.8×10^{-8}

In Table 17, the values for $\delta\omega$ and $\delta\dot{\omega}$ necessary to guarantee the ISA accuracy in the frequency band $10^{-4} - 10^{-3}$ Hz are summarized.

ISA to Star Tracker and MPO Alignment

Because ISA's measurements must be referred to an inertial frame, it is necessary to control the alignments of the ISA sensing axes with respect to the star tracker so to identify the inertial axes. Possible misalignments between these axes give an error directly proportional to the amplitude to be measured. In presence of a signal with amplitude equal to the maximum signal allowed by the dynamic, equal to 3000×10^{-8} m/s^2, only a misalignment less than 1/3000 rad guarantees the precision of 10^{-8} m/s^2 as required.

Data Rate Requirement

The ISA data rate requirement is strongly connected to the MPO environmental noise, in the condition of noise as previously indicated, in order to avoid aliasing noise it is necessary to acquire data at 300 Hz and then send the data to Earth at 1.5 KBs in nominal mode.

References

1. Fuligni F, Iafolla V. "Galileo and the Principle of Equivalence" STEP Symposium, Pisa, Italy 6–8 April 1993. ESA WPP–115.
2. Fuligni F, Iafolla V, Milyukov V, Nozzoli S. Experimental gravitation and geophysics. Il Nuovo Cimento C. 1997;20(5):637.
3. Eötvös Rv, Pekar V, Feteke E. Beitrage zum Gesetze der Proportionalitat von Tragheit und Gravitat. Annalen der Physik. 1922;68:11–66.
4. Eötvös RV, Pekár D, Fekete E. Beiträge zum Gesetze der Proportionalität von Trägheit und Gravität. Ann Phy. 1922;373:11.
5. Everitt CWF. Gravity, microgravity and the approach to picogravity. Prog Astronaut Aeronaut. 1986;108:89. (AIAA, Washington, DC).
6. Lorenzini EC, Gullahorn GG, Fuligni F. Recent developments in gravity gradiometry from the space-shuttle-borne tethered satellite system. J Appl Phys. 1988;63(1):216–23.
7. Lorenzini EC et al. A method for testing the principle of equivalence with improved accuracy. Smithsonian Astrophysical Observatory Technical Note TP91-001, 1991.
8. Nobili AM. GG report of mission study (phase A2 level). April 2009. http://eotvos.dm.unipi.it/PA2/GGPA2.pdf.

9. Raychaudhuri AK, Benerji S, Banerjee A. General relativity, astrophysics, and cosmology. Berlin: Springer; 1992.
10. Braginsky VB, Manukin AB. Measurement of weak forces in physics experiments. Chicago: The University of Chicago Press; 1974. p. 7.
11. Iafolla V, Lorenzini E, Milykov V, Nozzoli S. GiZero: new facility for gravitational experiments in free fall. Gravitat Cosmol. 1997;3(2):151–60.
12. Bordoni F, Maggi G, Ottaviano A, Pallottino GV. Very-low-noise cryogenic amplifier with silicon JFETs. Rev Sci Instrum. 1981;52(7):1079.
13. Fuligni F, Iafolla V. Measurement of small forces in the physics of gravitation and geophysics. Nuovo Cimento C. 1997;20:619–28.
14. Agnew DC, Berger J. Vertial seismic noise at very low-frequencies. JGR. 1978;83(B11):5420–24.
15. Iafolla V, Nozzoli S, Lorenzini EC, Milyukov V. Methodology and instrumentation for testing the weak equivalence principle in stratospheric free fall. Rev Sci Instrum. 1998;69(12):4146–51.
16. Fuligni F, Iafolla V. Measurement of small forces in the physics of gravitation and geophysics. Il Nuovo Cimento. 1997;20C(5):619–28.
17. Fuligni F, Iafolla V. Fifth force Measurements from a free falling platform at balloon altitudes. Kyoto: Procs. of the Marcel Grossman Conference; 1991.
18. Fuligni F, Iafolla V, Nozzoli S, Cosmo M, Grossi MD et al. Gravity gradiometry from the downward tethered system. Fourth International Conference on Tethers in Space. Washington DC, p. 1273 10–14 April (1995).
19. Giffard RP. Ultimate sensitivity limit of a resonant gravitational wave antenna using a linear motion detector. Phys Rev D. 1976;14:2478.
20. Gillies GT, Ritter RC. Torsion balance, torsion pendulums, and related devices. Rev Sci Instrum. 1993;64:283–309.
21. Galilei G. Dialogues concerning two new sciences. New York: Dover; 1954. (H. Crew and A. De Salvio Traslator).
22. Iafolla V, Milyukov V. GiZero: Error budget analysis, IRVIN Technical note RS 001103100 July. 1997.
23. Iafolla V, Lorenzini E, Milyukov V, Nozzoli S. Gravitation and cosmology. J Rus Gravit Soc. 1997;3:151.
24. Iafolla V, Nozzoli S, Mandiello A. High sensitive accelerometer for fundamental physics in space 2nd Joint Meeting of the International Gravity Commission and the International Geoid Commission Trieste 7–12 settembre 1998.
25. Iafolla V, Fiorenza E, Lefevre C, Nozzoli S, Peron R, Persichini M, Reale A, Santoli F, Lorenzini EC, Shapiro II, Ashenberg J, Bombardelli C, Glashow S. TEPEE/GReAT (general relativity accuracy test in an Einstein elevator): ready to start. Nuovo Cim C. 2008;31:497.
26. Iafolla V, Nozzoli S. Italian Spring Accelerometer (ISA): high sensitive accelerometer for "BepiColombo" ESA CORNERSTONE. Planetary and space. Science. 2001;49:1609.
27. Lorenzini EC, Shapiro II, Fuligni F, Iafolla V, Cosmo ML, et al. Test of the weak-equivalence principle in an Einstein elevator. Il Nuovo Cimento. 1994;109B:1195–209.
28. Lorenzini EC, Shapiro II, Fuligni F, Iafolla V, Cosmo ML, Grossi MD, Cheimets PN, Zielinski JB. Test of the weak equivalence principle in an einstein elevator. Il Nuovo Cimento. 1994;109B(12):1195–209.
29. Lorenzini EC, Iafolla V, Orlando P, Grossi MD. Testing the equivalence principle with a free-fall experiment from balloon. 31st Cospar Scientific Assembly. Birmingham: Elsevier Science; 1996 14–21 July.
30. Lavoie FJ, editor. Materials engineering: materials selector 1987. New York: Penton Publication; 1986. p. 72.
31. Iafolla V, Milyukov V, Nozzoli S. Observation of tidal tilts in Gran Sasso, l'Aquila Italy. EGS Conference, Nice 25–29 April 2000.
32. Shapiro II, Lorenzini EC, Ashenberg J, Bombardelli C, Cheimets PN, Iafolla V, Lucchesi DM, Nozzoli S, et al. Testing the principle of equivalence in an einstein elevator. Int J Mod Phys D. 2007;16:2227.

33. Schiff LI. On experimental tests of the general theory of relativity. Am J Phys. 1960;28:340.
34. Touboul P, Métris G, Lebat V, Robert A. The MICROSCOPE experiment, ready for the in-orbit test of the equivalence principle. Class Quantum Grav. 2012;29:184010.
35. Thorne KS, Lee DL, Lightman AP. Foundations for a theory of gravitation theories. Phys Rev D. 1973;7:3563.
36. Will CM. The Confrontation between general relativity and experiment. Living Rev Relativ. 2014;17:4. http://www.livingreviews.org/lrr-2014–4. Accessed 28 Nov 2014.
37. Overduin J, Everitt F, Worden P, Mester J. 2012. STEP and fundamental physics. Class. Quantum Grav. 29, 184012. doi:10.1088/0264 9381/29/18/184012, arXiv:1401.4784.

Fundamental Physics with the LAGEOS Satellites

Roberto Peron

Abstract The space(time) around Earth is a good environment in order to perform tests of gravitational theories. According to Einstein's view of gravitational phenomena, it is curved, mainly by the Earth mass–energy content. This (relatively) quiet dynamical environment enables a good reconstruction of a satellite (test mass) orbit, provided high-quality tracking data are available. This is the case of the LAGEOS satellites, built and launched mainly for geodetic and geodynamical purposes, but equally good for fundamental physics studies.
In this chapter a review will be presented of these studies, focusing on data, models and analysis strategies. Several recent and less recent results will be presented. General relativity once more appears as a very precise and effective theory for the gravitational phenomena.

1 Introduction

In dealing with gravitational phenomena, sometimes the geocentric point of view can be useful and productive. Looking at the hierarchy of masses in the universe (Earth, Solar System, Galaxy, ...) that this particular point of view gives us, the Earth mass–energy content—for example expressed as the Schwarzschild radius $2GM_{\oplus}/c^2 \sim 1$ cm—is the first one giving rise to interesting effects detectable with current technologies. By "interesting effects" here we mean something going beyond the physics described by the Newtonian framework. This is—as far as we know—the realm of Einstein's general relativity theory. After almost a century since its development and after many decades of intense experimental effort, it appears there is still room for further inquiry; this space is motivated by the appearance

R. Peron (✉)
Istituto di Astrofisica e Planetologia Spaziali (IAPS), Istituto Nazionale di Astrofisica (INAF), Via Fosso del Cavaliere 100, Roma, Italy
e-mail: roberto.peron@iaps.inaf.it

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_4

of more and more precise experimental techniques—as well as modeling tools—and by the push from never–ceasing attempts aimed at disputing the theory or its foundations[1].

Among the ways to test gravitational dynamics one of the simplest is to follow (*track*) the motion of an object orbiting in the gravitational field produced by another, bigger one (the *primary*). The orbiting object should be as close as possible to a point mass, in order not to perturb in a significant way the gravitational field of the primary; it should be what is called a *test mass*. A suitable modelization (analytical or numerical) of this system gives a prediction for the resulting orbit which can be compared with experimental tracking data. Such a scheme is rather general and could be applied to a variety of experimental situations. We describe here a particular such situation, given by the availability around the Earth of objects (satellites) specifically designed to be, as much as possible, close to the ideal concept of a test mass: the LAGEOS satellites [1]. These, as well as similar ones, have been designed, built, and launched for geodetic and geodynamical purposes. The study of their orbital motion, indeed, helps to characterize the fine details of the Earth gravitational field (and therefore of its structure and composition) and to establish and maintain a global reference frame with applications that range from astronomy to navigation (see e.g., [2, 3]).

The LAGEOSs are target for laser pulses sent from ground stations, used to calculate their instantaneous distance (*range*); the outstanding precision of this tracking technique, named satellite laser ranging (SLR), allows a precise determination of their orbits. This can be done with dedicated procedures and a fine modeling of their dynamics, which is by itself an interesting topic. Along the years, the availability to the scientific community of the ranging data allowed a variety of studies. Many of them, as said above, are related to geodesy and geophysics. At the same time, however, it is possible to exploit the same data to perform fundamental physics tests, by comparing the (measured and reconstructed) orbit with the ones predicted by several, competing, gravitational theories. This very simple objective requires a number of steps to be performed, which will be described in the following. It has to be stressed that, in this quest, to better data better models must follow. This is especially true since the sought for signals often lie several orders of magnitude below the "competitive" signals.

2 Gravitational Physics Opportunities

As mentioned above, along the years the LAGEOS satellites turned out to be very good targets to be tracked. They materialize very finely (though not exactly) the ideal concept of a *test mass*, which has to satisfy the following requirements [4]:

[1] For a review of this very general issue see the chapter *Gravity: Newtonian, post-Newtonian and General Relativistic* by Clifford Will in this book and [5].

- No electric charge
- Gravitational bindig energy negligible with respect to rest mass–energy
- Angular momentum negligible
- Sufficiently small to neglect tidal effects

An ideal test mass follows a purely gravitational orbit (a *geodesic* in metric theories) and is therefore an appropriate tool to study gravitational phenomena.

2.1 Relativistic Effects on Test Masses Around Earth

General relativity, in its *weak-field* and *slow-motion* limit[2], provides an effective description of the gravitational phenomena around the Earth. A formulation of the relevant equations of motion in a geocentric non-inertial reference system (nonrotating with respect to the barycentric one) is given in [6], from which we quote the relevant terms. The analyses here described are consistent with this formulation.

A test mass orbiting around the Earth is subject in its motion to three main relativistic effects. The biggest contribution comes from the gravitoelectric curvature of space-time induced by the Earth mass–energy:

$$\mathbf{a}_{\rm Schw} = \frac{Gm_{\rm E}}{c^2 r^3}\left[\left(\frac{4Gm_E}{r} - v^2\right)\mathbf{r} + 4(\mathbf{v}\cdot\mathbf{r})\mathbf{v}\right]. \tag{1}$$

This is called *Einstein* or *Schwarzschild precession* [7,8]. The satellite, in its motion around the Earth, follows the revolution of the latter in the space-time curved by the Sun mass–energy; this (via parallel transport of the normal to the satellite orbit) induces the *de Sitter* or *geodetic precession* [9]:

$$\mathbf{a}_{\rm dS} = 2\boldsymbol{\Omega}\times\mathbf{v} \qquad \boldsymbol{\Omega} \approx -\frac{3}{2}(\mathbf{V}_{\rm E} - \mathbf{V}_{\rm S})\times\frac{GM_{\rm S}\mathbf{X}_{\rm ES}}{c^2 R_{\rm ES}^3}. \tag{2}$$

In general relativity, unlike Newtonian physics, mass–energy currents also cause effects, named *gravitomagnetic* (see [4]). In particular, Earth's intrinsic angular momentum curves space-time and induces a further effect on the satellite orbit, called *Lense–Thirring effect* [10, 11] (also termed *dragging of inertial frames* in a more general setting):

$$\mathbf{a}_{\rm LT} = \frac{2Gm_{\rm E}}{c^2 r^3}\left[\frac{3}{r^2}(\mathbf{r}\times\mathbf{v})(\mathbf{r}\cdot\mathbf{J}) + \mathbf{v}\times\mathbf{J}\right]. \tag{3}$$

[2] The weak-field condition considers the space-time curvature so small that the metric can be written as $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ (Minkowski metric plus a "small" perturbation, $|h_{\mu\nu}| \ll 1$). The slow-motion condition requires $v \ll c$. Given the relative smallness of the masses at play, as well as that of their speed when compared with that of light, this approximation of the theory is sufficient for the purpose.

In the notation we follow [6]. In particular, c is the speed of light, G the Newtonian gravitational constant, $m_{\rm E}$ and J are the Earth mass and angular momentum, $\mathbf{r}$ and $\mathbf{v}$ the test mass position and velocity in the geocentric frame, $M_{\rm S}$ is the Sun mass, $\mathbf{V}_{\rm E}$ and $\mathbf{V}_{\rm S}$ are the Earth and Sun geocentric positions, $\mathbf{X}_{\rm ES}$ is the geocentric Earth–Sun vector, with distance $R_{\rm ES}$. We notice that, while the effects described by Eqs. 1 and 3 depend only on the Earth mass–energy and angular momentum, the geodetic precession of Eq. 2 involves also the Sun (relativistic three-body problem)[3].

Using the methods of celestial mechanics (in particular first-order perturbation theory), the secular[4] effects of relativistic corrections in the satellite Keplerian elements can be evaluated (see e.g., [12]). It turns out that the Schwarzschild term is mainly effective on the argument of perigee

$$\dot{\omega}_{\rm Schw} = \frac{3(GM_{\oplus})^{3/2}}{c^2 a^{5/2}(1-e^2)}, \tag{4}$$

the de Sitter one on the longitude of the ascending node

$$\dot{\Omega}_{\rm dS} = |\boldsymbol{\Omega}| \cos\varepsilon \tag{5}$$

(with ε obliquity of the ecliptic) and the Lense–Thirring one on both node

$$\dot{\Omega}_{\rm LT} = \frac{2GJ_{\oplus}}{c^2 a^3 (1-e^2)^{3/2}} \tag{6}$$

and perigee

$$\dot{\omega}_{\rm LT} = \frac{-6GJ_{\oplus}}{c^2 a^3 (1-e^2)^{3/2}} \cos I. \tag{7}$$

Numerical values can be found in Table 1. The dominant effect is the gravitoelectric precession, analogous to Mercury perihelion precession in the field of the Sun, with smaller contributions from geodetic and gravitomagnetic effects. It turns out that the argument of perigee is a good observable to see in action the Schwarzschild and Lense–Thirring secular precessions, having a secular part for both , while the longitude of the node is a good observable for the Lense–Thirring and de Sitter secular precessions.

Therefore, following GR, the secular behavior for the nodal rate is given by

$$\Delta\dot{\Omega}_{\rm rel} = \Delta\dot{\Omega}_{\rm dS} + \Delta\dot{\Omega}_{\rm LT}, \tag{8}$$

while the secular behavior for the pericenter rate is given by

$$\Delta\dot{\omega}_{\rm rel} = \Delta\dot{\omega}_{\rm Schw} + \Delta\dot{\omega}_{\rm LT}. \tag{9}$$

Given its higher orbital eccentricity (~ 0.014 compared to ~ 0.004) LAGEOS II is in fact more suited than LAGEOS for performing such a measurement in the case of pericenter.

[3] This effect appears due to the chosen geocentric non-inertial reference system.

[4] In first-order perturbation theory, two kinds of behavior for a given element can arise. The first is a term $\propto \sin t$ or $\cos t$; this is called *periodic*. The second is a term $\propto t$ (or higher powers); this is called *secular,* since it tends to accumulate over time.

Table 1 Rate (mas/year) and orbital shift (over 14 days) of the different types of secular relativistic precession on LAGEOS and LAGEOS II longitude of ascending node and argument of pericenter (1 mas/year = 1 milli-arc-second per year)

	Precession	Rate (mas/year)	Shift (m)
LAGEOS	$\Delta\dot{\omega}^{\mathrm{Schw}}$	3278.77	7.49
	$\Delta\dot{\Omega}^{\mathrm{LT}}$	32.65	7.46×10^{-2}
	$\Delta\dot{\omega}^{\mathrm{LT}}$	32.00	7.31×10^{-2}
LAGEOS II	$\Delta\dot{\omega}^{\mathrm{Schw}}$	3351.95	7.60
	$\Delta\dot{\Omega}^{\mathrm{LT}}$	31.48	7.14×10^{-2}
	$\Delta\dot{\omega}^{\mathrm{LT}}$	−57.00	-1.29×10^{-1}

Are the expected values compatible with the uncertainty associated to tracking data? An estimate of the orbital shift due to each effect can be obtained for nearly circular orbits by $\Delta x|_{14d} \simeq a\Delta\alpha|_{14d}$; here a is the semimajor axis of the orbit and $\Delta\alpha$ is the precession (on node or perigee) integrated over the 14 days estimation period. The values can be seen in the fourth column of Table 1: given a typical SLR normal point (see Sect. 2.4) precision of $\simeq$ 1 mm, we can notice that the Schwarzschild signal is well above the noise, while the gravitomagnetic and geodetic ones are barely above it[5].

2.1.1 Testing Inverse-Square Law

Tests for the inverse-square law behavior of gravitation, in the Newtonian limit, form an important issue. On one side, they are useful to better characterize gravitation itself, especially in the short- and intermediate-range. On another side, possible violations of this behavior could be related to new interactions between bodies acting at macroscopic distances (New Long Range Interaction—NLRI). In addition, these NLRIs may be thought of as the residual of a cosmological primordial scalar field related with the inflationary stage (dilaton scenario) [13].

Usually, this supplementary interaction is described by means of a Yukawa-like potential with strength α and range λ and transmitted by a field of very small mass $\mu = \hbar/\lambda c$:

$$V_{\mathrm{Yuk}} = -\alpha\frac{G_\infty M_\oplus}{r}e^{-r/\lambda}, \quad \alpha = \frac{1}{G_\infty}\left(\frac{K_\oplus}{M_\oplus}\frac{K_{\mathrm{s}}}{m_{\mathrm{s}}}\right). \tag{10}$$

Here G_∞ represents the Newtonian gravitational constant, $M_\oplus$ and m_{s} are the mass of the primary body (the Earth) and the satellite, r is their separation, $\hbar$ the reduced Planck constant. The strength α depends both on the mass–energy content of the

[5] In any case, the secular (systematic) character of the relativistic signal causes it to appear above the noise upon integration in a sufficiently longer time.

sources and their coupling strengths, $K_\oplus$ and K_s respectively. The suggestion in the 1980s of a possible "fifth force" [14] boosted further research on this supplementary interaction (see also [15, 16] for reviews and [17] for recent results).

An adequate observable in order to test for such non-Newtonian behavior is the pericenter of a binary system. A perturbative analysis of pericenter shift has been performed in [18, 19]. By differentiating the potential, we obtain the radial acceleration between the Earth and the satellite as produced by the long-range interaction:

$$\mathbf{R}_{\rm Yuk} = -\alpha \frac{G_\infty M_\oplus}{r^2}\left(1+\frac{r}{\lambda}\right)e^{-r/\lambda}\hat{\mathbf{r}}. \tag{11}$$

We can write the Gauss disturbing equation for the pericenter rate:

$$\Delta\dot{\omega}_{\rm Yuk} = -\frac{\sqrt{1-e^2}}{ena}\mathcal{R}_{\rm Yuk}\cos f, \tag{12}$$

where f represents the satellite true anomaly. Averaging Eq. (12) over a fast variable, such as the mean anomaly, we obtain the corresponding secular effect produced by the hypothesized Yukawa interaction. The maximum secular effect is given by:

$$< \Delta\dot{\omega}_{Yuk} >_{2\pi} \simeq 8.29\cdot 10^{11}\alpha \quad \text{[mas/year]} \tag{13}$$

and it corresponds to the peak value at a range $\lambda = 6082$ km, very close to one Earth radius.

2.2 Measurement Concept

Among the various techniques used to track satellites, SLR is one of the most precise [20]. It uses the propagation of a collimated laser pulse to measure the instantaneous distance between a station on Earth and a satellite. At the ground station, a definite laser pulse is generated and—through a tel escope— sent towards the satellite, where it is reflected back in the same direction by optical elements called *cube corner retroreflectors (CCR);* it then comes back to the same station, focused by the telescope and detected by a proper sensing device (see Fig. 1). By precisely measuring the start and stop times of the pulses with dedicated instruments, it is then possible to recover the instantaneous station–satellite distance (*range*):

$$\Delta s = \frac{c\Delta t}{2}. \tag{14}$$

This is of course the basic concept of the measurement. In practice, things are made more complex from having to take into account various phenomena, from the propagation of the pulse in the atmosphere to instrumental biases due to (among other things) laser stability, detector, timing device. A hint into the complexities of each

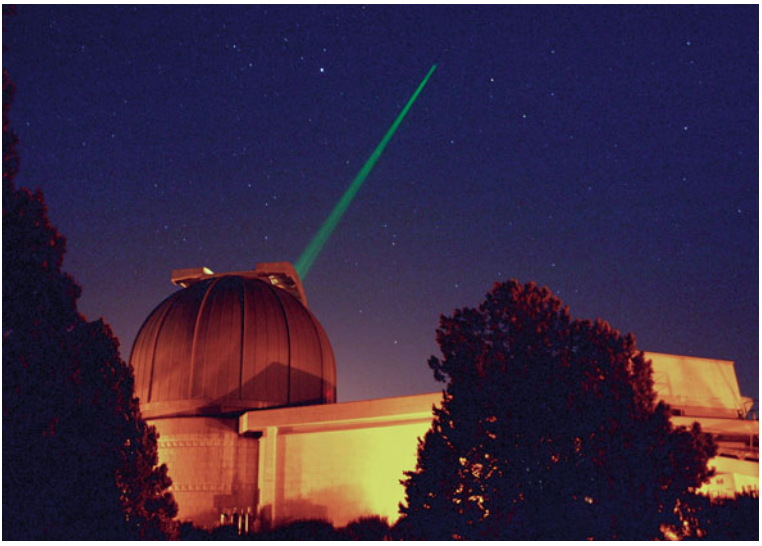


Fig. 1 Matera laser station during a pass. (Photo by Franco Ambrico. Courtesy Giuseppe Bianco, ASI-CGS)

single measurement can be found in [21]. Laser range observations from the various stations on the globe are collected by the International Laser Ranging Service (ILRS) [20] and are publicly available.

Presently, the two LAGEOS satellites are among the best tracked ones through the SLR technique. LAGEOS, launched by NASA (1976), and LAGEOS II, launched by NASA/ASI (1992), have been designed spherical in shape, with high density and small area-to-mass ratio in order to minimize the effects of the nongravitational perturbations [22]. Their radius is just 30 cm and their mass about 407 kg[6]. Their aluminum surface is covered with 426 CCRs.

In our analyses, a multi-arc technique has been employed [23]. The time period considered in the data analysis has been divided into shorter periods, called *arcs*, 15 days long. The arc length has been chosen as a compromise between the need of small not modeled effects, not accumulating too much (arc not too long) and the need to have a sufficient number of data and the accumulated unmodeled secular relativistic effects being larger than the corresponding error in the SLR measurements (arc not too short). For each arc the tracking data are reduced, resulting in an estimate of the state vector (position and velocity) at the beginning of the arc and selected parameters for the dynamics. A very precise orbit is therefore obtained for each arc, which can be expressed in terms of Keplerian elements. The arcs have a

[6] LAGEOS has an almost circular orbit, with an eccentricity $e_I \simeq 0.004$, a semimajor axis $a_I \simeq 12270$ km, and an inclination over the Earth's equator $i_I \simeq 109.8°$. The LAGEOS II corresponding elements are: $e_{II} \simeq 0.014$, $a_{II} \simeq 12162$ km and $i_{II} \simeq 52.66°$.

1 day overlap; calculating the difference in elements at the middle of this overlap provides time series of *residuals* which contain information on the part of dynamics which has not been modeled (or has been mismodeled). The fundamental observable being the range, strictly also the residuals, in their meaning of "observed minus calculated," are range. The elements difference method used in these analyses retains the concept for the various Keplerian elements, as shown in [24]. Analysis of the residuals time series allows recovering *a posteriori* the signature of effects which have not been modeled, as it was purposely done for the relativistic part.

2.3 Analysis Strategy

The tracking data (two-way range, in our case) contain the information associated with the satellite dynamics, as well as with the measurement procedure and the observational "constraints" (i.e., station positions, reference frames). This information has to be extracted in some way from the data. The problem is not trivial, considering the relative magnitudes of the effects involved (see Table 2). A direct comparison between the normal point (NP) precision (~ 1 mm) and the average size of orbit shift due to the relativistic effects shows that these effects could be recovered once the satellite dynamics has been properly modeled (for a description of the models employed see Sect. 3). The recovery of the information could be done with a least-squares procedure, in which data are fit to a model by a proper estimation of a set of selected parameters.

For the analyses here described, the NASA/GSFC software GEODYN II [25,26] has been used. This software is dedicated to satellite orbit determination and prediction, geodetic parameters estimation, tracking instruments calibration, and many other applications in the field of space geodesy.

The orbit determination is based on the numerical integration of the equations of motion of the satellite using the Cowell's method. The integration procedure of the Cowell's method is a predictor–corrector one, with a fixed time step, and this method has the great advantage of the ease of its implementation for any given number of perturbing bodies or nongravitational disturbing effects.

The result for a given perturbing acceleration $\mathbf{a_p}$ is obtained by solving the equations of motion (in terms of the satellite position $\mathbf{r}$) given its initial state vector (position and velocity):

$$\ddot{\mathbf{r}} = -\frac{GM_{\oplus}}{r^2}\hat{\mathbf{r}} + \mathbf{a_p}, \tag{15}$$

where the effects of each perturbation are considered as corrections to the *ideal* two-body Keplerian solution. The equations of motion for the satellite are integrated in an inertial reference frame, which for GEODYN is the mean equinox and equator of J2000.

Table 2 Magnitude of the main disturbing effects on the LAGEOS II spacecraft. (Adapted from [22])

Effect	Estimate	Magnitude (ms^{-2})
Earth's monopole	$\frac{GM_\oplus}{r^2}$	2.69
Earth's oblateness	$3\frac{GM_\oplus}{r^2}\left(\frac{R_\oplus}{r}\right)^2 \bar{C}_{20}$	-1.1×10^{-3}
Low-order geopotential harmonics	$3\frac{GM_\oplus}{r^2}\left(\frac{R_\oplus}{r}\right)^2 \bar{C}_{22}$	5.4×10^{-6}
High-order geopotential harmonics	$19\frac{GM_\oplus}{r^2}\left(\frac{R_\oplus}{r}\right)^{18} \bar{C}_{18,18}$	1.4×10^{-12}
Moon perturbation	$2\frac{GM_\leftmoon}{r_\leftmoon^3} r$	2.2×10^{-12}
Sun perturbation	$2\frac{GM_\odot}{r_\odot^3} r$	9.6×10^{-13}
General relativistic correction	$\frac{GM_\oplus}{r^2}\frac{GM_\oplus}{c^2}\frac{1}{r}$	9.8×10^{-10}
Atmospheric drag	$\frac{1}{2}C_D\frac{A}{M}\rho V^2$	3.4×10^{-12}
Solar radiation pressure	$C_R\frac{A}{M}\frac{\Phi_\odot}{c}$	3.2×10^{-9}
Albedo radiation pressure	$C_R\frac{A}{M}\frac{\Phi_\odot}{c}A_\oplus\left(\frac{R_\oplus}{r}\right)^2$	3.5×10^{-10}
Thermal emission	$\frac{4}{9}\frac{A}{M}\frac{\Phi_\odot}{c}\alpha\frac{\Delta T}{T_0}$	2.8×10^{-11}
Dynamic solid tide	$3k_2\frac{GM_\leftmoon}{r_\leftmoon}\left(\frac{R_\oplus}{r_\leftmoon}\right)^2\frac{R_\oplus^3}{r^4}$	3.7×10^{-6}
Dynamic ocean tide	~ 0.1 of the dynamic solid tide	3.7×10^{-7}

The orbit determination employs then the least-squares solution of the range residuals:

$$O_i - M_i = -\sum_j \frac{\partial M_i}{\partial P_j} dP_j + dO_i, \tag{16}$$

where O_i are the range observations, M_i their modeled values, dP_j the corrections to the vector $\mathbf{P}$ of parameters to be estimated, and dO_i are the errors associated with each observation. Concerning these errors, the dO_i account for both the contribution from the noise in the observations, as well as for the incompleteness of the mathematical model included in the orbit determination software. The least-squares algorithm seeks to minimize the residuals $O_i - M_i$ by adjusting at the same time the state vector at the epoch of arc and the parameters selected for estimation.

A basic choice of the analysis has been to use the residuals in order to recover the relativistic effects. By construction, they provide a measure of the discrepancy between experimental data and models; by purposely not including relativity into the modeling set, the residuals time series is expected to contain signatures of relativity

itself. The basic observable being distance, the residuals are strictly speaking on station–satellite distances. Being interested in effects related to individuals orbital elements, the method outlined in [24] has been employed, as said before, in order to obtain derived residual time series for the various elements. This is the method that has been used in the relativistic precessions measurements performed so far [27–34].

The strategy employed here could be considered as "minimal" or "conservative" in the following sense. The precise modeling of the orbits requires complex models, which depend on thousands of parameters (see Sect. 3). We underline that, while in general geodetic and geophysical problems often the majority of model parameters are estimated, in the analyses it has been chosen to estimate only few of them, namely those most directly related to the particular orbit of the satellites; the other parameters were selected as *consider parameters*, that is, ones which are already known with sufficient accuracy from other sources.

This approach is in line with the strategy of recovering the sought-for signal from the residuals, and considerably simplifies the mathematical structure of the problem being solved, moreover strongly lowering the chance of estimation biases. In particular, it has been chosen not to include the so-called *empirical accelerations* in the set of models fitting the SLR data. These can bias the estimation procedure and corrupt, in particular, the argument of perigee residuals [35].

2.4 LAGEOS Range Data Sets

The ranges so obained are called *Fullrate* ones, and are the basic product of SLR observations. In the 1980s it has been introduced a more compact format, called *NP*, which is the one commonly used. A NP is basically an "average" of the Fullrate observations over a defined time period (bin)[7]. In the formation of NP for bin i, the observation O_i nearest to the midpoint of the bin is located, and a *fit residual* FR_i (a residual from which systematic trends in the predictions have been removed) is calculated. The NP is then calculated as:

$$NP_i = O_i - FR_i + \overline{FR}_i, \tag{17}$$

where $\overline{FR}_i$ denotes the mean value of FR_i. The NP so calculated is characterized by the fact that its random error is reduced to that of the mean of the bin. More details can be found in [36].

The precision of the measurements is mainly related with the pulse width, which is usually $\approx 1 \cdot 10^{-10}$ s down to $3 \cdot 10^{-11}$ s for the best laser ranging stations. In the case of the two LAGEOS satellites, the NPs are characterized by a root mean square (RMS) down to a few mm, that corresponds to an accuracy in the orbit reconstruction at a few cm level, when using the best dynamical models. In the preliminary analyses a fit of the orbit of the satellites at a 1–2 cm (RMS) level in range has been achieved[8].

[7] For the LAGEOSs the bin size amounts to 120 s.

[8] In order to obtain a good reference orbit for the considered period, it has been performed a preliminary data reduction of laser range data with a modeling setup slightly different from the

3 Models

The procedures for determining the satellite orbit at a level comparable with the quality of tracking data require models not only for satellite dynamics, but also for measurement procedure and reference frame transformations.

The dynamics of LAGEOS satellites, seen at the level enabled by the accuracy of SLR data, is rather complex. Several gravitational and nongravitational effects are at work; estimates of their magnitude are provided in Table 2 (see [22, 37]).

The models included in GEODYN are devoted to describe not only the satellite dynamics, but also the measurement procedure and the reference frame transformations. These models include: (i) the geopotential (both in its static and dynamic part), (ii) lunisolar and planetary perturbations, (iii) solar radiation pressure and Earth's albedo, (iv) Rubincam and Yarkovsky-Schach effects (which need the satellite spin-axis orientation in order to be modeled), (v) drag effects, (vi) SLR stations coordinates, (vii) ocean loading, (viii) Earth orientation parameters and, (ix) measurement procedure. Usually, the models implemented in the code also include the general relativistic corrections in the so-called parameterized post-Newtonian (PPN) formalism [38–41]. In the analyses we performed, in order to solve for the relativistic secular precessions, we did not include such corrections in our setup.

The particular models used for the analyses here described are listed in Table 3. For the relevant part, the conventions established by the International Earth Rotation and Reference Systems Service (IERS)[9], which constitute the general framework for reference systems related issues and measurement models, have been followed as much as possible. The reference version has been IERS Conventions (2003) [42].

3.1 Gravitational Perturbations

The deviations of the Earth's gravitational field from the point mass one, due to the inhomogeneous mass density distribution inside the Earth, are by far the most important source of perturbations in the orbits of LAGEOS satellites. It is customary in geodesy and geophysics to represent the gravitational potential by expanding it

nominal one of Table 3. In particular, the radiation coefficient C_R has been estimated, together with corrections to polar motion and length of day. Moreover, empirical acceleration components in the three Gauss directions have been added. The combined estimate led to the reported 1–2 cm level in range. Of course, this slight overestimation comes at the price of some aliasing: Therefore, the results of the analysis could not be directly used for the relativity signal recovery.

[9] IERS is the international organization in charge of maintaining the reference frames used in astronomy, geodesy, and geophysics.

Table 3 Modeling setup as included in a typical analysis of LAGEOS satellites range data

Model for	Model type	Reference
Geopotential (static)	EIGEN-GRACE02S, EGM96	[43, 44]
Geopotential (time-varying, tides)	Ray GOT99.2	[45]
Geopotential (time-varying, nontidal)	IERS Conventions (2003)	[42]
Third body	JPL DE-403	[46]
Relativistic corrections[a]	PPN	[6]
Direct solar radiation pressure	Cannonball	[25]
Earth albedo	Knocke–Rubincam	[47]
Station positions	ITRF 2000	[48, 49]
Ocean loading	Schernek and GOT99.2 tides	[25, 45]
Earth rotation parameters	IERS EOP C04	[50]

[a] In fact, as explained in the text, these corrections have not been included in the modelization setup used in the analysis

in spherical harmonics (real basis):

$$U(\mathbf{r}) = \frac{GM_\oplus}{r}\left[1 + \sum_{l=1}^{\infty}\left(\frac{R_\oplus}{r}\right)^l \sum_{m=0}^{l} \bar{P}_{lm}(\sin\theta)\left(\bar{C}_{lm}\cos(m\phi) + \bar{S}_{lm}\sin(m\phi)\right)\right], \tag{18}$$

see, for example, [37, 51, 52]. Here r, θ, and ϕ represent the polar coordinates of the point at which the potential U is evaluated, $\bar{P}_{lm}$ are the associated Legendre functions, $M_\oplus$ the Earth mass, and $R_\oplus$ the Earth mean equatorial radius. The normalized coefficients $\bar{C}_{lm}$ and $\bar{S}_{lm}$, with l called *degree* and m *order*, are functions of the mass density distribution, and completely characterize the gravitational potential outside the distribution itself. In practice, the series is truncated at some finite $l_{\max}$: the model is then sensitive to inhomogeneities at the scale of $\pi R_\oplus / l_{\max}$. The lower degree harmonics are related to the choice of the reference frame in which the potential itself is expressed. Of paramount importance are the so-called *zonal* harmonics, that is, the ones with $m = 0$; they represent the part of the potential with rotational symmetry, and play an important role in the error budget of the measurements. It has to be noted that some care must be put in dealing with the permanent tide. In GEODYN, a "tide-free" geopotential is modeled, that is, one in which both the permanent part and the related deformation of the Moon and Sun tidal perturbations have been removed.

The Earth gravitational field, also seen in an "Earth fixed" frame, is not static: It varies in time due to a series of phenomena, from tides to mass transport in the Earth/atmosphere system at various scales. The tidal deformations of the Earth—both solid and ocean—and its atmosphere, are of primary interest for the measurement because of their combined periodic variations in the gravitational

attraction of the planet on the satellite [53–60]. In particular, solid tides account for about 90 % of the total response to the Moon and the Sun tidal disturbing potential.

A convenient way to describe these deformations is through the so-called *Love numbers* ($k^f_{2,m} \simeq 0.30$, where f represents the frequency of the tidal wave), which measure the ratio between the response of the real Earth and the theoretical response of a perfect fluid sphere, and are determined with very high accuracy because of their long-term effects on geodetic satellites, as in the case of the two LAGEOS [61,62].

In particular, in the case of solid tides, the degree $l = 2$ terms, that is, those due to the quadrupole tidal potential, are the most important to be considered. The resultant quadrupole tidal potential is the sum of terms like the following ones, where the summation has to be extended over all possible frequencies f (complex basis):

$$U^f_T(\mathbf{r}) = -\frac{GM}{5r_{\rm E}}\left(\frac{R_\oplus}{r_{\rm E}}\right)^2 \sum_m k^f_{2,m}\left[\left(\frac{R_\oplus}{d}\right)^3 Y^\star_{2,m}(\hat{\mathbf{d}})\right]^f Y_{2,m}(\hat{\mathbf{r}}_{\rm E}). \qquad (19)$$

In this expression, M represents the mass of the perturbing body (the Moon or the Sun), d its distance with respect to the Earth, and $r_{\rm E}$ the distance between a given mass element of the Earth and its center of mass.

Tidal effects are important to be considered in geophysics and geodynamics because they influence the satellite orbits under three aspects:

1. Kinematic effect: because they produce periodic pulsations of the Earth and, as a consequence, of the on-ground tracking stations
2. Dynamic effect: because they cause a time variation of the geopotential, that affects the satellites orbit
3. Reference system effect: because they perturb the Earth rotation, thus affecting the reference systems used in the orbit computation

Ocean tides are difficult to model because of the greater complexity of the involved phenomena. Indeed, even if ocean tides account for about 10 % only of the total response to the external potentials, their uncertainties are a factor of 10 larger than those of solid tides.

The effect of third-body perturbations has been modeled as well, using the well-established solar system ephemerides by JPL, DE-403 [46]. As discussed in Sect. 2.1, the relativistic corrections are consistent with the formulation of [6]. In line with the chosen strategy of recovering the relativistic effects *a posteriori* in the residuals time series, in fact these corrections have not been included in the setup.

3.2 Nongravitational Perturbations

An important part of the satellites dynamics is represented by the effects caused by nongravitational forces. These, of various origin, are caused by the interaction of the satellite body with the near-Earth radiation and particle environment. Such

forces are typically surface ones and depend, in a complex way, on the physical properties of the satellite, as well as on its attitude. Even for very simple satellites as the LAGEOS (spherical in shape, very dense, and passive), these effects are relevant and, especially, very difficult to model. A wide literature is available on the subject, see for example, [22, 63–65].

The biggest contribution is given by the push of radiation on the satellite surface (*radiation pressure*), in particular direct visible radiation from the Sun; also reflected visible radiation from the Earth (*albedo*) and infrared radiation emitted from the Earth surface are important. They depend on the way this radiation is reflected–diffused–absorbed by the satellite surface, and therefore on the optical properties of this surface, primarily the CCRs.

The most important nongravitational effect is the direct solar radiation pressure. The resultant acceleration, for a body of spherical shape, can be modeled as:

$$\mathbf{a}_{\odot} = -C_{\mathrm{R}} \frac{A}{m} \frac{\Phi_{\odot}}{c} \left(\frac{< D_{\odot} >}{D_{\odot}} \right)^2 \hat{\mathbf{s}}, \tag{20}$$

where A is the cross-sectional area of the satellite, m its mass, $\Phi_{\odot}$ the solar radiation flux at 1 AU, c the speed of light, and C_{R} (called *radiation coefficient*) summarizes the optical properties of the satellite surface. The last squared term is due to the modulation coming from the eccentricity of the Earth orbit around the Sun ($D_{\odot}$ represents the Earth–Sun distance and $< D_{\odot} >$ its average value, that is, 1 AU) and $\hat{\mathbf{s}}$ is the Sun unit vector direction. Equation (20) corresponds to the so-called *cannonball* model for the direct solar radiation pressure from the Sun, which represents the simplest model for a spherical in shape satellite. The model expressed by Eq. 20 is rather good for the LAGEOS satellites, provided an estimate is done of the C_{R} parameter. Evidences have been provided that LAGEOS II optproperties could have been changed since the launch time [66].

More subtle perturbing effects are due to the so-called *thermal forces*; these are caused by an inhomogeneous temperature distribution of the body (due to its finite thermal inertia), resulting in a thrust force due to emitted radiation. In particular, thermal forces depend on the satellite spin vector, giving different contributions on the orbit as a function of the spin orientation and rate. In the case of a rapidly spinning satellite, a latitudinal distribution for the differential temperature across the satellite surface can be assumed, and the consequent (net) recoil force will be directed along the spin-axis direction, away from the "colder" pole [64, 67–71]. Conversely, if the assumption of a (comparatively) fast rotational motion of the spin is avoided, a longitudinal temperature distribution arises. As a consequence, the perturbing force due to the thermal thrust effect will tilt from the spin-axis direction giving rise to additional "equatorial" components and more complicated equations to be solved [72].

As soon as the rotational period is less than the thermal inertia, the diurnal thermal asymmetry is negligible and the rapid-spin case approximation could be applied, as in the model developed by [67] or in the more complex model

of [70]. Otherwise, the model developed by [72] must be used. Therefore, we have a *seasonal-like Yarkovsky–Schach effect* in the former case of a rapidly spinning satellite, and a *diurnal-like Yarkovsky–Schach effect* when the the fast rotation approximation is no more valid.

The *Yarkovsky–Rubincam effect* [73–75], or *Earth–Yarkovsky effect*, is related with the infrared radiation emitted by the Earth's surface. The effect can be explained in the following way: LAGEOS satellites CCRs have a significant thermal inertia and since this thermal response time is shorter than the satellite orbital period and larger than the spin period (nominal value of about 1 s at launch epoch), the perturbation causes a temperature asymmetry between the hemisphere facing the planet and the one opposite to it. We refer to [64] for further details.

In order to model—as accurately as possible—the perturbing thermal thrust effects, and especially the Yarkovsky–Schach effect, a detailed description of the evolution of the spin axis is crucial. Several authors have focused on this problem and tried to explain the evolution of the LAGEOS satellites spin-axis, in either an analytical [76–78] or a more empirical approach [79].

The most satisfactory approach comes from [76], in which the problem of the rotation of the former LAGEOS satellite was theoretically developed. Subsequently, an extension and improvement of the model appeared in [80], in which solutions were also obtained for LAGEOS II. A refinement of the model described in [80] is contained in [81]. Finally, based on the results of these works, LOSSAM (*LAGEOS Spin Axis Model*) has been created to bring the problem of the rotation of both satellites a significant step further (see [82, 83], and [65] for the final model). LOSSAM combines the corrected mathematical theory with (past) observations of the rate and orientation of the angular velocity of the two LAGEOS satellites. The independent observations form a crucial element of this study.

However, in particular, because of the intrinsic difficulties in the modeling of LAGEOS II spin-axis vector—especially because of the slow down of its rate [84, 85] and the lack of more recent observations of its rate and orientation [65] —and also because of the passage from the fast rotation regime to the slow one during the time span of the analysis [72, 86], a different approach has been followed. Indeed, this perturbation was not included in the dynamical models but, as explained in [32], its unmodeled effects on the pericenter of LAGEOS II were *a posteriori* removed through the final least-squares best fit.

3.3 Empirical Accelerations

A modelization piece that is often used in precise orbit determination is given by the so-called *empirical accelerations*. These are general acceleration terms added to the equations of motion, and are aimed at modeling and/or absorbing small otherwise unknown effects which may be relevant for the dynamics. They are usually decomposed in the three Gauss directions $\hat{r}$, $\hat{t}$, $\hat{w}$ (radial, transverse, and out-of-plane), in

the form

$$A(t) = A_0 + A_1 \sin M + A_2 \cos M, \tag{21}$$

where M represents the satellite mean anomaly,aiming thus at modeling constant or once-per-revolution accelerations. This orbit modeling tool is useful when long wavelengths orbit errors, including secular disturbing effects, need to be removed, as well as for long-period resonances and also nongravitational perturbations that are not included in the software dynamical model. Experience shows that, while they are useful to improve the fit quality, they can easily bias the estimation of other quantities (this is known as over-parameterization of the problem). We highlight once more that, in order to avoid the orbit corruption, in particular of the satellite argument of pericenter, their use has been avoided during the data reduction.

4 Data Reduction

In the analyses described here, more than 10 years of LAGEOS and LAGEOS II laser tracking data, provided by ILRS [20], have been reduced using the GEODYN II software [25]. The selected period has been divided into 15-days arcs. For each of them, the data reduction provides an estimate of the initial conditions (state vector) together with estimates of selected parameters. The models employed (see Table 3) enabled a very good fit of the data, as can be seen in the statistics. In particular, in Fig. 2, the post-fit weighted RMS and a histogram of the residuals in range are shown. These plots are related to runs dedicated to analyse the LAGEOS II perigee behavior (see [33]). The average RMS is somewhat higher than the "ideal" level that could be expected based on data quality; this is due to the fact that in this analysis no relativistic effects were inserted in the modelization set, thereby lowering the overall accuracy. In this way, however, the residuals contain useful information, which is indeed related to relativity itself. This can be seen in the histogram of the residuals in range; their distribution appears close to but is not exactly Gaussian, indicating that some information is still present in the residuals themselves. The same reasoning applies also to the analysis reported in [31]; in that case, however, only the gravitomagnetic contribution was taken out.

5 Zonal Harmonics Related Uncertainties and ombination Formula for Lense–Thirring Measurements

Detailed error budget calculations show the importance of the zonal harmonics uncertainties in the overall effectiveness of the analysis procedure in extracting the relativistic signals. This is especially true in the case of Lense–Thirring measurements employing the nodal residuals. In particular, the quadrupole coefficient $\bar{C}_{20}$

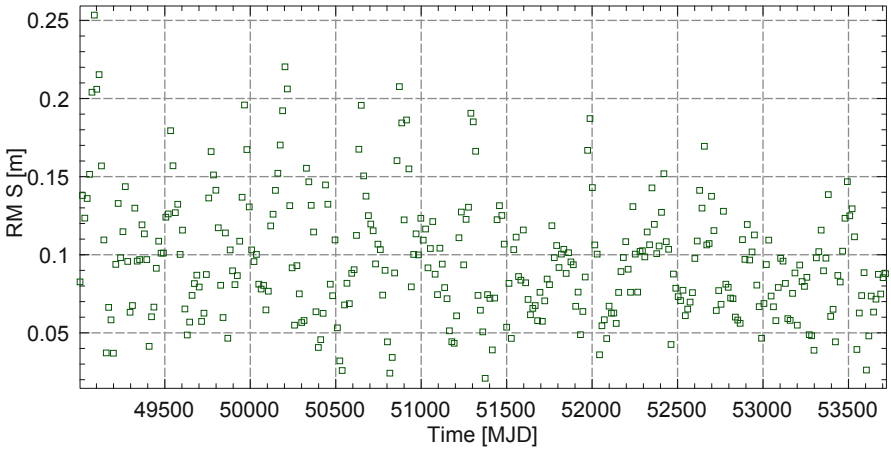


Fig. 2 LAGEOS II post-fit weighted RMS *(left)* and residuals in range *(right)* computed for each of the arcs in which it has been divided the analysed time period

has been found to be the major source of uncertainty. Its secular effect on the nodal longitude is given by

$$\dot{\Omega}_{\rm class} = \frac{3\sqrt{5}}{2} n \left(\frac{R}{a}\right)^2 \frac{\cos I}{(1-e^2)^2} \bar{C}_{20}, \tag{22}$$

see for example, [87, 88][10]. Therefore, the orbit of the satellite is subject to a classical precession whose value is much higher than the relativistic (gravitomagnetic) one to be measured (see Table 4)[11]. Of course, what really matters is the unknown part of this precession due to the uncertainty in the accepted value of $\bar{C}_{20}$; its value based on two geopotential models is shown in Table 4. This precession actually hides the relativistic one and must be handled in some way. To this aim, in [89], a procedure based on the combination of orbital residuals from more than one satellite, to get rid of this masking precession; see also [90].

The simplest case is that of two satellites, along with their two nodal longitudes. Let us consider a single arc of orbit determination; the procedure allows obtaining a residual for each orbital element, in particular for the nodal longitude: $\delta\dot{\Omega}$[12]. If the modelization setup is accurate enough, this residual—that is, the difference between the calculated and observed value—is mainly a function of two quantities, the classical quadrupole precession and the relativistic (Lense–Thirring) one, so that the

[10] We use here the normalized coefficients $\bar{C}_{l0}$ instead of the non-normalized C_{l0} or the J_l. We remember that $J_2 = -C_{20} = -\sqrt{5}\bar{C}_{20}$.

[11] We consider only the quadrupolar part of this classical precession.

[12] In [24] it is shown that the residuals obtained with their method are in fact *rates*.

Table 4 Values (in mas year^{-1}) of the nodal precession for LAGEOS and LAGEOS II orbits due to relativistic and classical gravitational effects

Effect	LAGEOS	LAGEOS II
Lense–Thirring	32.65	31.48
$\bar{C}_{20}$ (EGM96)	2.702×10^{10}	-4.982×10^{10}
$\delta\bar{C}_{20}$ (EGM96)	-3.240×10^{3}	5.975×10^{3}
$\delta\bar{C}_{20}$ (EIGEN-GRACE02S)	-2.960×10^{3}	5.458×10^{3}

following equation can be reasonably[13] considered to hold:

$$\delta\dot{\Omega} = \delta\dot{\Omega}_{\text{class}} + \delta\dot{\Omega}_{\text{rel}}. \tag{23}$$

Regarding the two terms on the right-hand side of the previous equation, these can be expressed as follows. Following equation (22), the classical precession can be written as $\dot{\Omega}_{\text{class}} = N_{20}\bar{C}_{20}$, with N_{20} function of Earth equatorial radius and satellite orbit. Upon the assumption that the biggest uncertainty comes from $\bar{C}_{20}$, one can write

$$\delta\dot{\Omega}_{\text{class}} = \frac{\partial\dot{\Omega}_{\text{class}}}{\partial\bar{C}_{20}}\delta\bar{C}_{20} = N_{20}\delta\bar{C}_{20}. \tag{24}$$

In the same way, the relativistic precession can be written as $\dot{\Omega}_{\text{rel}} = \dot{\Omega}_{\text{LT}}\mu$; here $\dot{\Omega}_{\text{LT}}$ comes from Eq. (6), while μ is an empirical parameter[14] measuring the actual value of the relativistic effect ($\mu = 0$ in Newtonian physics, $\mu = 1$ in general relativity). In writing $\dot{\Omega}_{\text{rel}}$ this way, we in fact parameterize the general relativistic prediction with μ and assume that all the error is contained in this empirical parameter[15]; so:

$$\delta\dot{\Omega}_{\text{rel}} = \frac{\partial\dot{\Omega}_{\text{rel}}}{\partial\mu}\delta\mu = \dot{\Omega}_{\text{LT}}\delta\mu. \tag{25}$$

The total uncertainty from Eq. (23) can therefore be written as

$$\delta\dot{\Omega} = N_{20}\delta\bar{C}_{20} + \dot{\Omega}_{\text{LT}}\delta\mu. \tag{26}$$

We see that each residual can be expressed as a function of two uncertainties, $\delta\bar{C}_{20}$ and $\delta\mu$. As such, Eq. (26) is not useful. But adding a further observable (i.e., taking

[13] In writing this equation, we neglect higher-degree multipoles and other sources of perturbation for the node. We can safely do that since we assume them sufficiently well-modeled in the orbit determination setup, in such a way that the neglected effects are small. How well this assumption holds is determined by the measurement error budget.

[14] In fact, in the post-Newtonian framework $\mu = (1 + \gamma)/2$, with γ the PPN parameter quantifying how much space curvature is produced by unit rest mass; see [5].

[15] In fact, this equals to admit no a priori knowledge on the amplitude of Lense–Thirring effect.

the nodal residuals of two satellites, as LAGEOS *and* LAGEOS II) one can construct a system of two equations

$$\begin{cases} \delta\dot{\Omega}^{\mathrm{I}} = N_{20}^{\mathrm{I}}\delta\bar{C}_{20} + \dot{\Omega}_{\mathrm{LT}}^{\mathrm{I}}\delta\mu \\ \delta\dot{\Omega}^{\mathrm{II}} = N_{20}^{\mathrm{II}}\delta\bar{C}_{20} + \dot{\Omega}_{\mathrm{LT}}^{\mathrm{II}}\delta\mu \end{cases} \tag{27}$$

which can be solved to obtain $\delta\mu$:

$$\delta\mu = \frac{N_{20}^{\mathrm{I}}\delta\dot{\Omega}^{\mathrm{II}} - N_{20}^{\mathrm{II}}\delta\dot{\Omega}^{\mathrm{I}}}{N_{20}^{\mathrm{I}}\dot{\Omega}_{\mathrm{LT}}^{\mathrm{II}} - N_{20}^{\mathrm{II}}\dot{\Omega}_{\mathrm{LT}}^{\mathrm{I}}}. \tag{28}$$

This $\delta\mu$, together with $\delta\bar{C}_{20}$ is just the right one to account for the total residuals $\delta\dot{\Omega}^{\mathrm{I}}$ and $\delta\dot{\Omega}^{\mathrm{II}}$.

It is worth emphasizing two things. First, the solution given by Eq. (28) clearly does not contain $\delta\bar{C}_{20}$, and so the expression overcomes the problem of the uncertainty in $\bar{C}_{20}$, both static and time dependent. Second, $\delta\mu$ as expressed by Eq. (28) is related to a *single* arc of orbit determination; so what one obtains is a time series covering the period of analysis. The outcome can be seen in Fig. 3, in which the relativistic parameter $\delta\mu$ from LAGEOS and LAGEOS II combined nodal residuals is plotted as a function of time. Notice the cancellation of the nonseasonal, "anomalous" change in Earth quadrupole (apparent in the nodal longitude residuals, see Fig. 4), as reported in [31,91].

6 A Review of Recent Measurements

The idea of using laser ranged satellites in order to test selected predictions of general relativity theory dates back to the 1970s and 1980s. The test of Schwarzschild precession has been discussed in [8]. The measurement of the Lense–Thirring effect has been suggested by [92] and proposed by [93]. We review here some recent results which come out from precise orbit determinations of LAGEOS and LAGEOS II satellites. These analyses produced residuals time series of the satellites, Keplerian elements in the way discussed in Sect. 2.2. The expected relativistic signals, both in longitude of ascending node and in argument of perigee, were of a secular type, so they should appear as a secular trend upon time integration of the relevant time series.

In [31], such an analysis has been performed on LAGEOS and LAGEOS II tracking data in order to detect the Lense–Thirring precession. Given the limitations due to Earth's gravitational quadrupole uncertainty, the node residuals of both satellites have been combined as discussed in Sect. 5. The corresponding combined time series have been fitted with a secular trend plus a number of periodic terms (in order to account for mismodeling in some perturbations). They report a value of

$$\mu = 0.984 \pm 5 - 10\%, \tag{29}$$

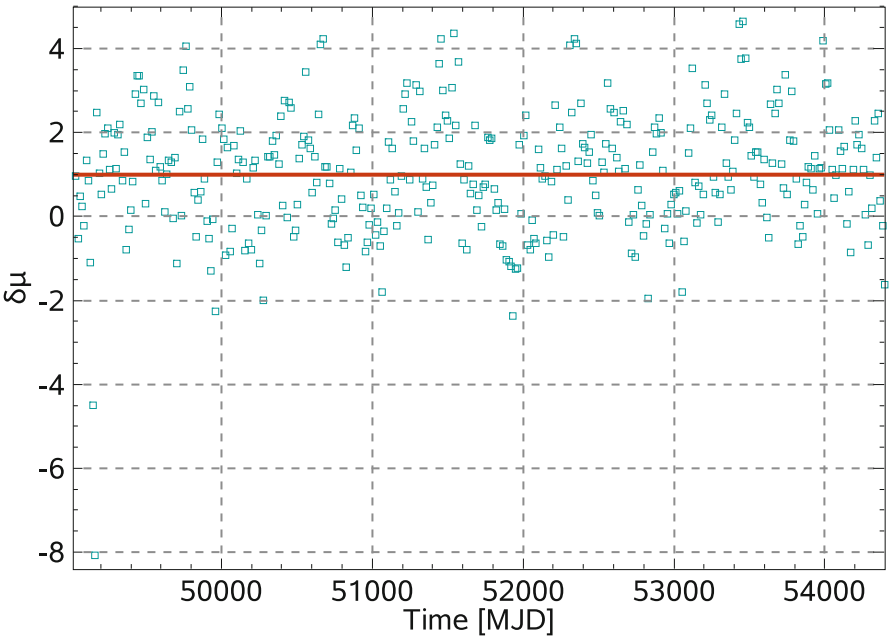


Fig. 3 Relativistic effect from the combined nodal longitude residuals of LAGEOS and LAGEOS II *(squares)*. The average value is 1.056, to be compared with the general relativistic prediction value, 1 *(continuous line)*. MJD stands for "Modified Julian Day;" the considered time period for the analysis starts 1993

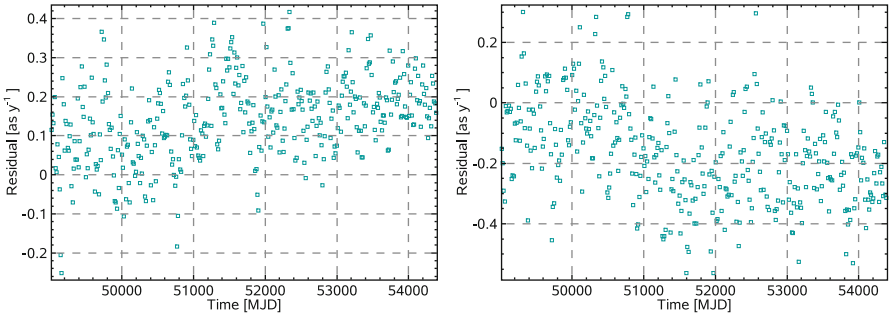


Fig. 4 Nodal longitude residuals, LAGEOS *(left)* and LAGEOS II *(right)*. The average values are 141.3 mas year^{-1} and − 167.3 mas year^{-1} for LAGEOS and LAGEOS II respectively

using the EIGEN-GRACE02S as a geopotential model[16]. See Fig. 5 for the related fit.

[16] The reported value has been obtained fitting the combined residuals with a secular trend plus ten periodic terms.

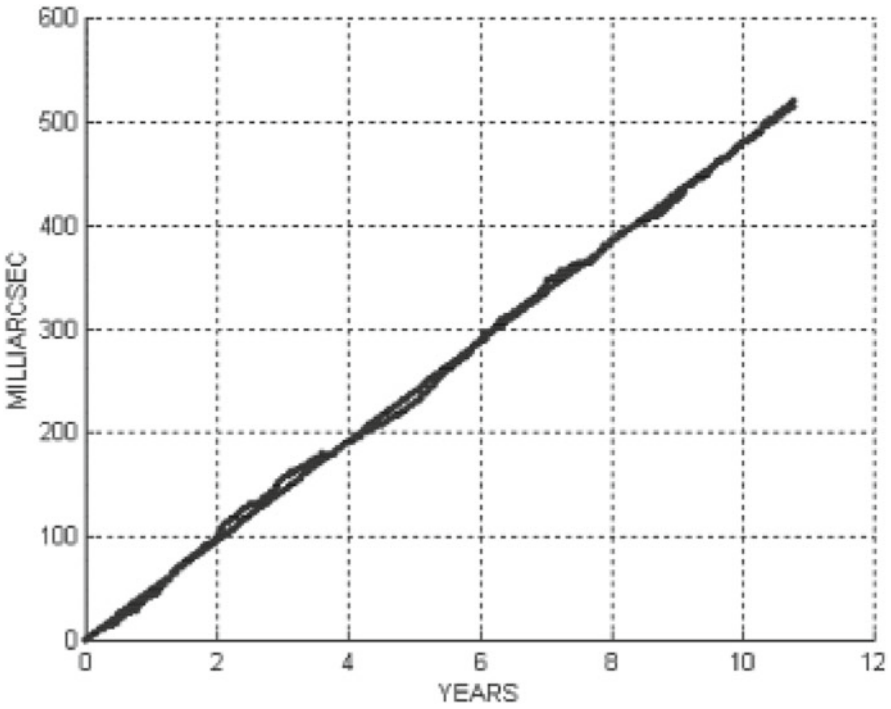


Fig. 5 Figure 14 of [31]. A combination of LAGEOS and LAGEOS II node residuals together with their fit (secular trend plus ten periodic terms) is shown

In [32], a dedicated analysis has been performed, focused on the LAGEOS II perigee behavior. In that case, the residuals being analyzed were directly those of the Keplerian element: the combination was not necessary, since the overall magnitude of the relativistic effects (Schwarzschild plus Lense-Thirring) is much bigger. A fit value $\Delta\dot{\omega}_{\rm meas} = 3306.58$ mas/year This value can be taken as an estimate of the total relativistic perigee precession, given by

$$\Delta\dot{\omega}_{\rm rel} \simeq \varepsilon_{\rm Schw}\Delta\dot{\omega}_{\rm Schw} + \varepsilon_{\rm LT}\Delta\dot{\omega}_{\rm LT}. \tag{30}$$

The slope estimate has small variations depending on the number of periodic effects which are fitted together with the linear trend. The following conservative result for the magnitude of the total relativistic effect has been reported, at the post-Newtonian level:

$$\varepsilon_\omega = 1 + (0.28 \pm 2.14) \times 10^{-3}, \tag{31}$$

where $\varepsilon_\omega = 1$ in general relativity. Since the dominant contribution in Eq. 30 comes from

$$\varepsilon_{\rm Schw} = \frac{2 + 2\gamma - \beta}{3}, \tag{32}$$

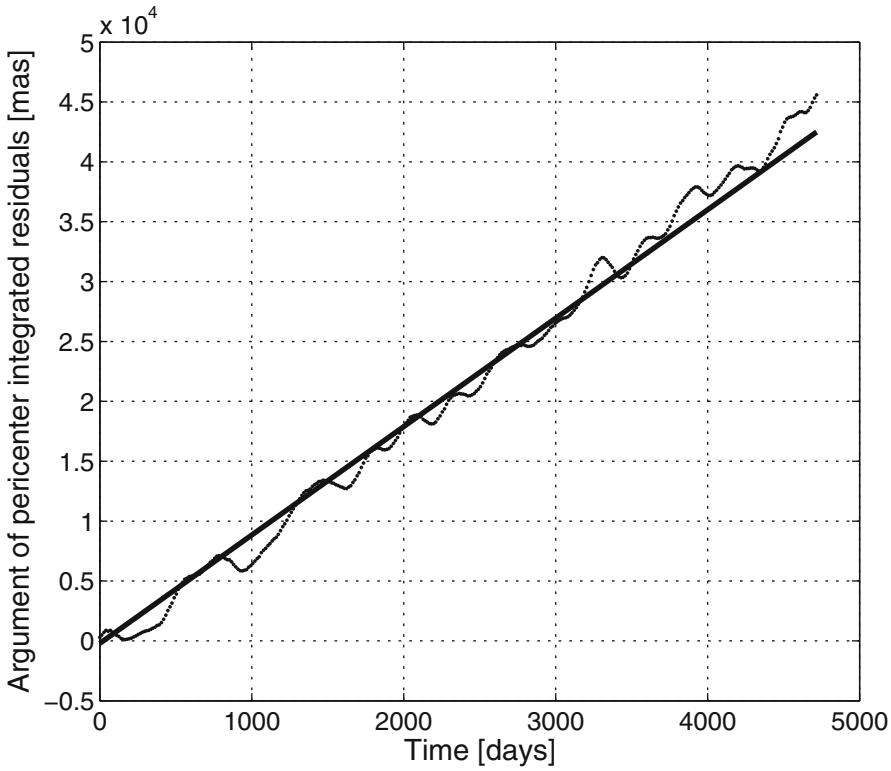


Fig. 6 Figure 1 of [32]. The perigee residuals are shown together with their fit (secular trend plus four periodic terms)

the estimate given by Eq. 31 is mainly a measurement of such a combination of γ and β PPN parameters. A preliminary error budget for the measurement, taking into account the various systematic errors, estimated the error to be at 2% level [18]. A complete error budget has been reported in [34]:

$$\varepsilon_\omega - 1 = \left[-0.12 \cdot 10^{-3} \pm 2.10 \cdot 10^{-3} \right] \pm \left[1.74 \cdot 10^{-2} \right], \tag{33}$$

where in the first square bracket the result and statistical error from the best fit is shown and in the second square bracket, the error budget due to the gravitational and nongravitational systematic sources of error is represented (Fig. 6).

The measured value for the argument of perigee precession can also be used to constrain a non-Newtonian contribution to the satellite dynamics, as discussed in Sect. 2.1.1. Indeed, the absence of such a signal in the residuals time series allows placing a strong constraint to the strength α at $\lambda = a$. In [32], the following upper bound is shown:

$$|\alpha| \simeq |(1.0 \pm 8.9)| \times 10^{-12}, \tag{34}$$

This result has been improved in [34]:

$$|\alpha| \simeq |(0.5 \pm 8.0) \pm 69| \cdot 10^{-12}. \tag{35}$$

These results represent a huge improvement with respect to previous constraints at this scale and are comparable with the lunar laser ranging results.

7 Conclusions

The near-Earth environment is a good place to test the gravitational interaction, at least in the weak-field regime. Nonetheless, the general relativity theory, almost a century after its development by Albert Einstein, has passed so far for all the experimental testsand is continuously challenged by alternative theories motivated by several unsolved problems, both theoretical and experimental. Verifying the equation of motion of a test mass in a given gravitational field provides important series of tests that can indeed be used to verify general relativity and place constraints on alternative theories, thanks also to theoretical instruments such as the PPN formalism. Performing high-precision experiments along this way is nowadays possible thanks to the availability of a class of artificial satellites, the geodetic ones, developed and launched mainly for geodetic and geophysical purposes. These are tracked with a very powerful technique, SLR; the tracking data should contain indeed signatures of the relativistic dynamics.

Extracting the small signals due to the curvature of space-time is not an easy task; even the strongest effects are covered by much bigger signals due to classical (Newtonian) dynamics and to nongravitational sources of "noise." The required techniques of precise orbit determination have been discussed, as applied in particular to the LAGEOS and LAGEOS II data. The quality of data themselves, as well as of the dynamical models, is such that the orbits can be reconstructed at the centimeter level. A subsequent study of the residuals, that is, the quantities representing the difference between what is measured and what is predicted, revealed general relativity in action, especially with peculiar behavior related to the gravitoelectric and gravitomagnetic fields, in the appropriate post-Newtonian approximation. An error budget for the measurements can be developed, with the useful by-product of placing strong constraints on alternative theories of gravitation. Perhaps amazingly, once more, the Einsteinian theory emerges as a very effective description of gravitational phenomena.

Acknowledgements The author would like to thank David M. Lucchesi (Istituto di Astrofisica e Planetologia Spaziali, IAPS-INAF). He acknowledges the ILRS for providing high-quality laser ranging data of the two LAGEOS satellites.

References

1. Cohen SC, King RW, Kolenkiewicz R, Rosen RD, Schutz BE. LAGEOS scientific results. J Geophys Res. 1985;90:9215–438. doi:10.1029/JB090iB11p09215.
2. Xu G. Sciences of geodesy - I; 2010.
3. Xu G. Sciences of geodesy - II. Innovations and future developments; 2013.
4. Ciufolini I, Wheeler JA. Gravitation and inertia. Princeton series in Physics. Princeton: Princeton University Press; 1995.
5. Will CM. The confrontation between general relativity and experiment. Living Rev Relativ. 2006;9:3.
6. Huang C, Ries JC, Tapley BD, Watkins MM. Relativistic effects for near-earth satellite orbit determination. Celest Mech Dyn Astron. 1990;48:167–85. doi:10.1007/BF00049512.
7. Einstein A. Die Grundlage der allgemeinen Relativitätstheorie. Ann Phys. 1916;354:769–822. doi:10.1002/andp.19163540702.
8. Rubincam DP. General relativity and satellite orbits—the motion of a test particle in the Schwarzschild metric. Celest Mech. 1977;15:21–33. doi:10.1007/BF01229045.
9. de Sitter W. On Einstein's theory of gravitation and its astronomical consequences . Second paper. Mon Not R Astron Soc. 1916;77:155–84.
10. Lense J, Thirring H. Über den Einfluß der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie. Phys Z. 1918;19:156.
11. Mashhoon B, Hehl FW, Theiss DS. On the gravitational effects of rotating masses—the Thirring-Lense papers. Gen Rel Grav. 1984;16:711–50. doi:10.1007/BF00762913.
12. Soffel MH. Relativity in astrometry, celestial mechanics and geodesy; 1989. Berlin, Heidelberg, New York: Springer-Verlag
13. Damour T, Piazza F, Veneziano G. Violations of the equivalence principle in a dilaton-runaway scenario. Phys Rev. 2002;D66:046007. doi:10.1103/PhysRevD.66.046007.
14. Fischbach E, Sudarsky D, Szafer A, Talmadge C, Aronson SH. Reanalysis of the Eotvos experiment. Phys Rev Lett. 1986;56:3–6. doi:10.1103/PhysRevLett.56.3.
15. Fischbach E, Gillies GT, Krause DE, Schwan JG, Talmadge C. Non-Newtonian gravity and new weak forces: an index of measurements and theory. Metrologia. 1992;29:213–60. doi:10.1088/0026-1394/29/3/001.
16. Goldhaber AS, Nieto MM. Photon and graviton mass limits. Rev Mod Phys. 2010;82:939–79. doi:10.1103/RevModPhys.82.939.
17. Gundlach JH, Schlamminger S, Wagner T. Laboratory tests of the equivalence principle at the University of Washington. Space Sci Rev. 2009;148:201–16. doi:10.1007/s11214-009-9609-3.
18. Lucchesi DM. LAGEOS II perigee shift and Schwarzschild gravitoelectric field. Phys Lett A. 2003;318:234–40. doi:10.1016/S0375-9601(03)01213-1.
19. Lucchesi DM. The LAGEOS satellites orbit and Yukawa-like interactions. Adv Space Res. 2011;47:1232–37. doi:10.1016/j.asr.2010.11.029.
20. Pearlman MR, Degnan JJ, Bosworth JM. The international laser ranging service. Adv Space Res. 2002;30:135–43.
21. Degnan JJ. Satellite laser ranging: current status and future prospects. IEEE Trans Geosci Remote Sensing. 1985;23:398–413. doi:10.1109/TGRS.1985.289430.
22. Milani A, Nobili AM, Farinella P. Non-gravitational perturbations and satellite geodesy. Bristol: Adam Hilger; 1987.
23. Milani A, Gronchi GF. Theory of orbit determination. Cambridge: Cambridge University Press; 2010.
24. Lucchesi DM, Balmino G. The LAGEOS satellites orbital residuals determination and the Lense Thirring effect measurement. Plan Space Sci. 2006;54:581–93. doi:10.1016/j.pss.2006.03.001.
25. Pavlis DE, et al. GEODYN II operations manual. NASA GSFC; 1998.

26. Putney B, Kolenkiewicz R, Smith D, Dunn P, Torrence MH. Precision orbit determination at the NASA Goddard Space Flight Center. Adv Space Res. 1990;10:197–203. doi:10.1016/0273-1177(90)90350-9.
27. Ciufolini I, Lucchesi D, Vespe F, Mandiello A. Measurement of dragging of inertial frames and gravitomagnetic field using laser-ranged satellites. Nuovo Cim A. 1996;109:575–90. doi:10.1007/BF02731140.
28. Ciufolini I, Lucchesi D, Vespe F, Chieppa F. Measurement of gravitomagnetism. Europhys Lett. 1997;39:359–64. doi:10.1209/epl/i1997-00362-7.
29. Ciufolini I, Pavlis E, Chieppa F, Fernandes-Vieira E, Perez-Mercader J. Test of general relativity and measurement of the Lense-Thirring effect with two Earth satellites. Science. 1998;279:2100–03. doi:10.1126/science.279.5359.2100.
30. Ciufolini I, Pavlis EC. A confirmation of the general relativistic prediction of the Lense-Thirring effect. Nature. 2004;431:958–60. doi:10.1038/nature03007.
31. Ciufolini I, Pavlis EC, Peron R. Determination of frame-dragging using Earth gravity models from CHAMP and GRACE. New Astron. 2006;11:527–50. doi:10.1016/j.newast.2006.02.001.
32. Lucchesi DM, Peron R. Accurate measurement in the field of the Earth of the general-relativistic precession of the LAGEOS II pericenter and new constraints on non-Newtonian gravity. Phys Rev Lett. 2010;105(23):231103. doi:10.1103/PhysRevLett.105.231103.
33. Peron R. Einstein is still right: new tests of gravitational dynamics in the field of the Earth. Nuovo Cim C. 2013;36:125–34. doi:10.1393/ncc/i2013-11493-6.
34. Lucchesi DM, Peron R. LAGEOS II pericenter general relativistic precession (1993–2005): error budget and constraints in gravitational physics. Phys Rev D. 2014;89(8):082002. doi:10.1103/PhysRevD.89.082002.
35. Lucchesi DM. The LAGEOS satellites orbital residuals determination and the way to extract gravitational and non-gravitational unmodeled perturbing effects. Adv Space Res. 2007;39:1559–75. doi:10.1016/j.asr.2007.04.040.
36. Sinclair AT. Data screening and normal point formation—restatement of Herstmonceux Normal Point Recommendation; 1997. http://ilrs.gsfc.nasa.gov/products_formats_procedures/normal_point/np_algo.html. Accessed 28 Sep 2015
37. Montenbruck O, Gill E. Satellite orbits. Models, methods and applications. Berlin: Springer; 2000.
38. Nordtvedt K. Equivalence principle for massive bodies. II. Theory. Phys Rev. 1968;169:1017–25. doi:10.1103/PhysRev.169.1017.
39. Will CM. Theoretical frameworks for testing relativistic gravity. II. Parametrized post-Newtonian hydrodynamics, and the Nordtvedt effect. Astrophys J. 1971;163:611–28. doi:10.1086/150804.
40. Will CM, Nordtvedt K Jr. Conservation laws and preferred frames in relativistic gravity. I. Preferred-frame theories and an extended PPN formalism. Astrophys J. 1972;177:757–74. doi:10.1086/151754.
41. Nordtvedt K Jr., Will CM. Conservation laws and preferred frames in relativistic gravity. II. Experimental evidence to rule out preferred-frame theories of gravity. Astrophys J. 1972;177:775–92. doi:10.1086/151755.
42. McCarthy DD, Petit G. IERS conventions. IERS Technical Note 32, IERS 2004; 2003.
43. Reigber C, Schmidt R, Flechtner F, König R, Meyer U, Neumayer KH, Schwintzer P, Zhu SY. An Earth gravity field model complete to degree and order 150 from GRACE: EIGEN-GRACE02S. J Geodyn. 2005;39:1–10. doi:10.1016/j.jog.2004.07.001.
44. Lemoine FG, Kenyon SC, Factor JK, Trimmer RG, Pavlis NK, Chinn DS, Cox CM, Klosko SM, Luthcke SB, Torrence MH, Wang YM, Williamson RG, Pavlis EC, Rapp RH, Olson TR. The development of the joint nasa gsfc and the national imagery and mapping agency (nima) geopotential model egm96. Technical Paper NASA/TP-1998-206861. NASA GSFC; 1998.
45. Ray RD. A global ocean tide model from TOPEX/POSEIDON altimetry: GOT99.2. Technical Paper NASA/TM-1999-209478. Maryland: Goddard Space Flight Center, Greenbelt; 1999.

46. Standish EM, Newhall XX, Williams JG, Folkner WM. JPL planetary and lunar ephemerides, DE403/LE403. Tech Rep JPL IOM 314.10-127; 1995.
47. Rubincam DP, Knocke P, Taylor VR, Blackwell S. Earth anisotropic reflection and the orbit of LAGEOS. J Geophys Res. 1987;92:11662–8. doi:10.1029/JB092iB11p11662.
48. Altamimi Z, Sillard P, Boucher C. ITRF2000: a new release of the International Terrestrial Reference Frame for earth science applications. J Geophys Res. 2002;107:2214. doi:10.1029/2001JB000561.
49. Boucher C, Altamimi Z, Sillard P, Feissel-Vernier M. The ITRF2000. IERS Technical Note 31, IERS; 2004.
50. International Earth Rotation Service. EOP Combined Series EOP C04. Tech rep. IERS.
51. Bertotti B, Farinella P, Vokrouhlický D. Physics of the solar system. Dynamics and evolution, space physics, and spacetime structure. Astrophysics and Space Science Library. 2003;293. Dordrecht:Kluwer Academic Publishers
52. Hofmann-Wellenhof B, Moritz H. Physical geodesy. Berlin: Springer; 2006.
53. Melchior P. The tides of the planet Earth. New York:Pergamon Press; 1981.
54. Lambeck K. Tidal dissipation in the oceans: astronomical, geophysical and oceanographic consequences. Roy Soc Lon Philos Trans Ser A. 1977;287:545–94. doi:10.1098/rsta.1977.0159.
55. Schwiderski EW. Ocean tides, part i: global ocean tidal equations. Mar Geodes. 1980;3(1–4):161–217. doi:10.1080/01490418009387997. http://www.tandfonline.com/doi/abs/10.1080/01490418009387997.
56. Schwiderski EW. Ocean tides, part ii: a hydrodynamical interpolation model. Mar Geodes. 1980;3(1–4):219–55. doi:10.1080/01490418009387998. http://www.tandfonline.com/doi/abs/10.1080/01490418009387998.
57. Wahr JM. A normal mode expansion for the forced response of a rotating earth. Geophys J R Astron Soc. 1981;64:651–75.
58. Wahr JM. Body tides on an elliptical, rotating, elastic and oceanless earth. Geophys J R Astron Soc. 1981;64:677–703.
59. Wang R. Effect of rotation and ellipticity on Earth tides. Geophys J R Astron Soc. 1994;117:562–5. doi:10.1111/j.1365-246X.1994.tb03953.x.
60. Yoder CF, Williams JG, Parke ME. Tidal variations of Earth rotation. J Geophys Res. 1981;86:881–91. doi:10.1029/JB086iB02p00881.
61. Cheng MK, Shum CK, Tapley BD. Determination of long-term changes in the Earth's gravity field from satellite laser ranging observations. J Geophys Res. 1997;102:22377–90. doi:10.1029/97JB01740.
62. Wu B, Bibo P, Zhu Y, Hsu H. Determination of love numbers using satellite laser ranging. J Geodetic Soc Jpn. 2001;47:174–80.
63. Lucchesi DM. Reassessment of the error modelling of non-gravitational perturbations on LAGEOS II and their impact in the Lense-Thirring determination. Part I. Plan Space Sci. 2001;49:447–63. doi:10.1016/S0032-0633(00)00168-9.
64. Lucchesi DM. Reassessment of the error modelling of non-gravitational perturbations on LAGEOS II and their impact in the Lense-Thirring derivation-Part II. Plan Space Sci. 2002;50:1067–1100. doi:10.1016/S0032-0633(02)00052-1.
65. Andrés de la Fuente JI. Enhanced modelling of LAGEOS non-gravitational perturbations. Ph.D. thesis. Delft University Press; 2007.
66. Lucchesi DM, Ciufolini I, Andrés JI, Pavlis EC, Peron R, Noomen R, Currie DG. LAGEOS II perigee rate and eccentricity vector excitations residuals and the Yarkovsky-Schach effect. Plan Space Sci. 2004;52:699–710. doi:10.1016/j.pss.2004.01.007.
67. Afonso G, Barlier F, Mignard F, Carpino M, Farinella P. Orbital effects of LAGEOS seasons and eclipses. Ann Geophysicae. 1989;7:501–14.
68. Farinella P, Nobili AM, Barlier F, Mignard F. Effects of thermal thrust on the node and inclination of LAGEOS. Astron Astrophys. 1990;234:546–54.
69. Scharroo R, Wakker KF, Ambrosius BAC, Noomen R. On the along-track acceleration of the Lageos satellite. J Geophys Res. 1991;96:729–40. doi:10.1029/90JB02080.

70. Slabinski VJ. A numerical solution for Lageos thermal thrust: the rapid-spin case. Celest Mech Dyn Astron. 1996;66:131–79. doi:10.1007/BF00054962.
71. Rubincam DP, Currie DG, Robbins JW. LAGEOS I once-per-revolution force due to solar heating. J Geophys Res. 1997;102:585–90. doi:10.1029/96JB02851.
72. Farinella P, Vokrouhlický D. Thermal force effects on slowly rotating, spherical artificial satellites-I. Solar Heating Plan Space Sci. 1996;44:1551–61. doi:10.1016/S0032-0633(96)00073-6.
73. Rubincam DP. LAGEOS orbit decay due to infrared radiation from Earth. J Geophys Res. 1987;92:1287–94. doi:10.1029/JB092iB02p01287.
74. Rubincam DP. Yarkovsky thermal drag on LAGEOS. J Geophys Res. 1988;93:13805–10. doi:10.1029/JB093iB11p13805.
75. Rubincam DP. Drag on the Lageos satellite. J Geophys Res. 1990;95:4881–86. doi:10.1029/JB095iB04p04881.
76. Bertotti B, Iess L. The rotation of Lageos. J Geophys Res. 1991;96:2431–40. doi:10.1029/90JB01949.
77. Habib S, Holz DE, Kheyfets A, Matzner RA, Miller WA, Tolman BW. Spin dynamics of the LAGEOS satellite in support of a measurement of the Earth's gravitomagnetism. Phys Rev D. 1994;50:6068–79. doi:10.1103/PhysRevD.50.6068.
78. Williams SE. The Lageos satellite: a comprehensive spin model and analysis. Ph.D. thesis, NCSU PhD Dissertation, pp. i–xii. 2002;1-252.
79. Ries J, Eanes R, Watkins MM. Spin vector influence on LAGEOS ephemeris. Presented at the Second Meeting of IAG Special Study Group 2.130, Baltimore; 1993.
80. Farinella, P., Vokrouhlicky, D., Barlier, F.: The rotation of LAGEOS and its long-term semimajor axis decay: a self-consistent solution. J Geophys Res. 1996;101:17861–72.
81. Vokrouhlický D. Non-gravitational effects and LAGEOS' rotation. Geophys Res Lett. 1996;23:3079–82. doi:10.1029/96GL03025.
82. Andrés JI, Noomen R, Bianco G, Currie DG, Otsubo T. Spin axis behavior of the LAGEOS satellites. J Geophys Res. 2004;109:6403. doi:10.1029/2003JB002692.
83. Andrés JI, Noomen R, Vecellio None S. Numerical simulation of the LAGEOS thermal behavior and thermal accelerations. J Geophys Res. 2006;111:B09406. doi:10.1029/2005JB003928.
84. Kucharski D, Kirchner G, Schillak S, Cristea E. Spin determination of LAGEOS-1 from kHz laser observations. Adv Space Res. 2007;39:1576–81. doi:10.1016/j.asr.2007.02.045.
85. Kucharski D, Kirchner G, Koidl F, Cristea E. 10 years of LAGEOS-1 and 15 years of LAGEOS-2 spin period determination from SLR data. Adv Space Res. 2009;43:1926–30. doi:10.1016/j.asr.2009.01.019.
86. Vokroublický D, Farinella P. Thermal force effects on slowly rotating, spherical artificial satellites-II. Earth infrared heating. Plan Space Sci. 1997;45:419–25. doi:10.1016/S0032-0633(96)00147-X.
87. Kaula WM. Theory of satellite geodesy. Applications of satellites to geodesy. Blaisdell; 1966.
88. Iorio L. The impact of the static part of the Earth's gravity field on some tests of general relativity with satellite laser ranging. Celest Mech Dyn Astron. 2003;86:277–94.
89. Ciufolini I. On a new method to measure the gravitomagnetic field using two orbiting satellites. Nuovo Cim A. 1996;109:1709–20. doi:10.1007/BF02773551.
90. Peron R. On the use of combinations of laser-ranged satellites orbital residuals to test relativistic effects around Earth. Mon Not Roy Astron Soc. 2013;432:2591–5. doi:10.1093/mnras/stt621.
91. Cox CM, Chao BF. Detection of a large-scale mass redistribution in the terrestrial system since 1998. Science. 2002;297:831–3. doi:10.1126/science.1072188.
92. Cugusi L, Proverbio E. Relativistic effects on the motion of Earth's artificial satellites. Astron Astrophys. 1978;69:321–5.
93. Ciufolini I. Measurement of the Lense-Thirring drag on high-altitude, laser-ranged artificial satellites. Phys Rev Lett. 1986;56:278–81. doi:10.1103/PhysRevLett.56.278.

Probing Gravity with Next Generation Lunar Laser Ranging

Manuele Martini and Simone Dell'Agnello

Abstract Lunar and satellite laser ranging (LLR/SLR) are consolidated techniques which provide a precise, and at the same time, cost-effective method to determine the orbits of the Moon and of satellites equipped with laser retroreflectors with respect to the International Celestial Reference System. We describe the precision tests of general relativity and of new theories of gravity that can be performed with second-generation LLR payloads on the surface of the Moon (NASA/ASI MoonLIGHT project), and with SLR/LLR payloads deployed on spacecraft in the Earth–Moon system. A new wave of lunar exploration and lunar science started in 2007–2008 with the launch of three missions (Chang'e by China, Kaguya by Japan, Chandrayaan by India), missions in preparation (LCROSS, LRO, GRAIL/LADEE by NASA) and other proposed missions (like MAGIA in Italy). This research activity will be greatly enhanced by the future robotic deployment of a lunar geophysics network (LGN) on the surface of the Moon. A scientific concept of the latter is the International Lunar Network (ILN, see http://iln.arc.nasa.gov/). The LLR retroreflector payload developed by a US–Italy team described here and under space qualification at the National Laboratories of Frascati (LNF) is the optimum candidate for the LGN, which will be populated in the future by any lunar landing mission.

1 Lunar Laser Ranging

Lunar laser ranging (LLR) is mainly used to conduct high-precision measurements of ranges between laser stations on Earth and a corner cube retroreflector (CCR) array on the lunar surface. Over the years, LLR has benefited from a number of improvements both in observing technology and data modeling, which led to the current accuracy of postfit residuals of $\sim$2 cm.

Nowadays, LLR is a primary technique to study the Earth–Moon system and is very important for gravitational physics, geodesy, and studies of the lunar interior. LLR contributes to the realization of both the terrestrial and selenocentric reference frames. The realization of a dynamical inertial reference frame, in contrast to the

M. Martini (✉) · S. Dell'Agnello
INFN-LNF, via E. Fermi, 40, Frascati, Rome
e-mail: Manuele.Martini@lnf.infn.it

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_5

kinematically realized frame of very long baseline interferometry (VLBI), offers new possibilities for mutual crosschecking and confirmation.

Since 1969, LLR has supplied many tests of general relativity (GR): LLR data have been used to evaluate geodetic precession [1, 2], probe the weak and strong equivalence principle, determine the parametrized post-Newtonian (PPN) parameters, and address the time change of the gravitational constant G and $1/r^2$ deviations of gravitational interactions. LLR has also provided important information on the composition and origin of the Moon through measurement of its rotations and tides. Future lunar missions will expand this broad scientific program.

Initially, the Apollo arrays contributed a negligible portion of the LLR error budget. Today, the ranging accuracy of ground stations has improved by more than two orders of magnitude: the new Apache Point Observatory Lunar Laser-ranging Operation (APOLLO) station at Apache Point, USA, is capable of mm-level range measurements; The Matera Laser Ranging Observatory (MLRO), at the Agenzia Spaziale Italiana (ASI) "Centro di Geodesia Spaziale" in Matera, Italy, has restarted LLR operations. Now, owing to lunar librations, the Apollo arrays dominate the LLR error budget, which is a few cm.

2 The MoonLIGHT Program

LLR has for decades provided the best tests of a wide variety of gravitational phenomena, probing the validity of Einstein's theory of GR [3, 4]. The lunar orbit is obviously influenced by the gravitational field of the Earth and the Sun, but is also sensitive to the presence of many other solar system bodies.

In 2006, Istituto Nazionale di Fisica Nucleare (INFN) proposed the Moon Laser Instrumentation for General Relativity High accuracy Tests (MoonLIGHT) technological experiment, which has the goal of reducing the error contribution of LLR measurements by more than two orders of magnitude. In Table 1, the possible improvements in the measurement of gravitational parameters achievable through reaching the ranging precision of 1 mm or even 0.1 mm are reported.

The MoonLIGHT program [5] is the result of a collaboration between two teams: the Lunar Laser Ranging Retroreflector Array for the twenty-first century (LLRRA21) team in the USA, led by Douglas Currie of the University of Maryland (UMD), and the Italian one led by National Institute of Nuclear Physics-National Laboratories of Frascati (INFN-LNF). We are exploring improvements in both the instrumentation and the modeling of the CCR.

To explain the MoonLIGHT experiment, we have to understand the limitations of a multi-CCR array.

The main problem that affects the Apollo arrays consists of the lunar librations in longitude, that result from the eccentricity of the Moon's orbit around Earth [6–8]. Due to this phenomenon, the Apollo arrays are shifted so that one corner of the array is more distant than the opposite corner. Because of the librations tilt, the arrays

Table 1 Narrowing of parameter bounds due to gains in the accuracy of ranging measurements by one or two orders of magnitude (reaching a precision of 1 mm or even 0.1 mm)

Gravitational measurement	First generation LLR precision	Second generation LLR precision	2^{nd} generation LLR precision	Time scale
	($\sim$ cm)	(1 mm)	(0.1 mm)	
WEP	$\left\vert\frac{\Delta a}{a}\right\vert < 1.4 \times 10^{-13}$	10^{-14}	10^{-15}	Few years
SEP	$\vert\eta\vert < 4.4 \times 10^{-4}$	3×10^{-5}	3×10^{-6}	Few years
β	$\vert\beta - 1\vert < 1.1 \times 10^{-4}$	10^{-5}	10^{-6}	Few years
$\frac{\dot{G}}{G}$	$\left\vert\frac{\dot{G}}{G}\right\vert < 9 \times 10^{-13} yr^{-1}$	5×10^{-14}	5×10^{-15}	$\sim$ 5 years
Geodetic precession	6.4×10^{-3}	6.4×10^{-4}	6.4×10^{-5}	Few years
$\frac{1}{r^2}$ deviation	$\vert\alpha\vert < 3 \times 10^{-11}$	10^{-12}	10^{-13}	$\sim$ 10 years

LLR Lunar laser ranging, *WEP* Weak equivalence principle, *SEP* Strong equivalence principle

increase the dimension of the pulse coming back to the Earth (Fig. 1). The broadening of the pulse will be greater proportionally to the array physical dimensions and to the Moon–Earth distance increase. Therefore, for the largest array, from Apollo 15, the enlargement is about 30 cm, and for the Apollo 11 and Apollo 14 arrays, it is about 15 cm. In agreement with this relationship, the pulse enlargement corresponds to a flight time increase:

- $\pm$ 0.5 ns for Apollo 15;
- $\pm$ 0.25 ns for Apollo 11 and Apollo 14.

In order to solve this problem, LNF in collaboration with UMD, indicated a new design of lunar CCR, named the second generation LLR, whose performance is unaffected not only by lunar librations but also by the motion of the regolith due to its large thermal cycle. The idea that we propose is to move from a multi-CCR array to a series of single, larger CCRs, each with 10 cm of front face diameter (Fig. 2, 3, 4).

Instead of having a single pulse spread by the array and the libration effect, we will have single short pulses coming back with the same dimensions as the incoming one (Fig. 1), with a final laser retroreflector array (LRA) ranging accuracy below 10 μm. When the new CCRs are placed on the lunar surface, it will make sense to improve the station capabilities [9–12].

To summarize, in the past, LLR techniques have employed a large laser pulse fired from the Earth station, larger than array dimensions, which dominated the measurement uncertainty. Now there is a moderately-sized laser pulse, but still a large array, so that the measurement uncertainty is dominated by the array; in the future, with MoonLIGHT/LLRRA21, there will be a single CCR unaffected by librations. The measurement uncertainty will then be dominated by the laser pulse, which could be shortened through modern technology (Fig. 5).

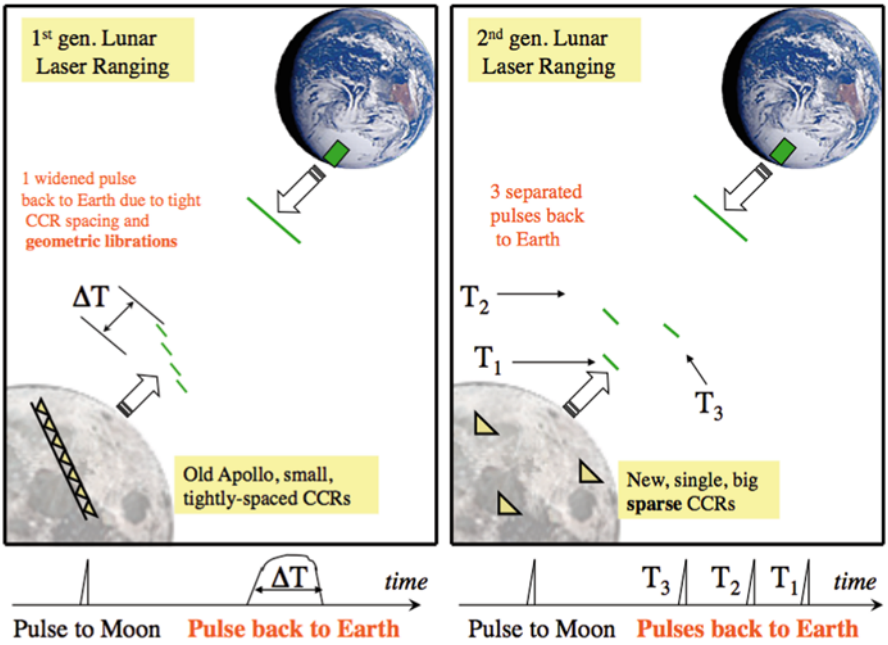


Fig. 1 Comparison between first and second generation laser retroreflector arrays *(LRAs)*. The librations tilt the arrays *(left)*, but the individual large corner cube retroreflectors *(CCRs)* are unaffected *(right)*. So, by using the Moon Laser Instrumentation for General Relativity High accuracy Tests *(MoonLIGHT)* payloads, we receive single short pulses instead of one broadened pulse

Fig. 2 Picture of Apollo, first generation and MoonLIGHT/LLRRA21, second generation CCR

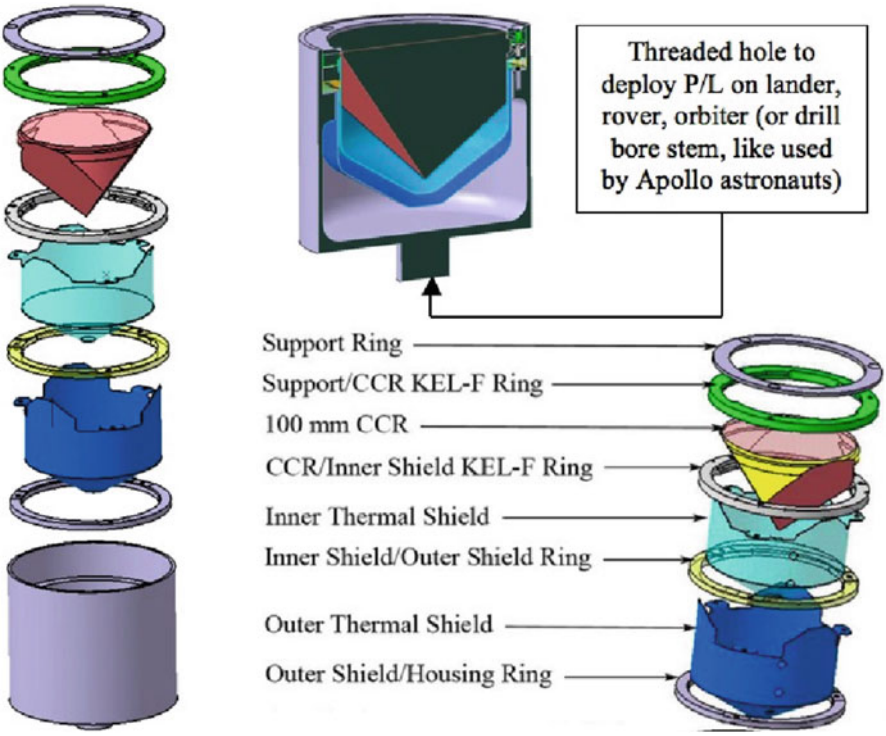


Fig. 3 Drawings of the MoonLIGHT CCR with its internal mounting rings, thermal shields, and aluminum housing

Fig. 4 Inner conformal shield of the MoonLIGHT CCR

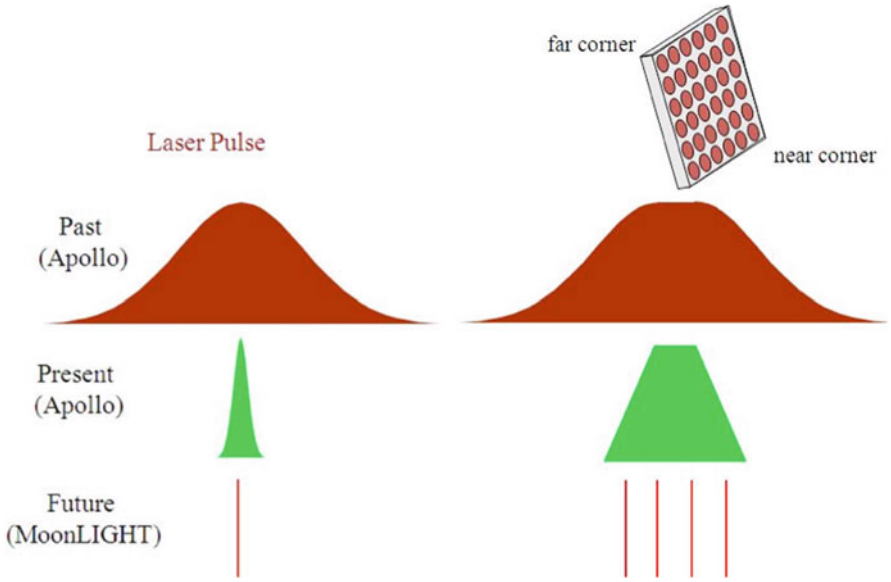


Fig. 5 The figure shows which contribution (between fired and retroreflected laser pulse size) dominates the measurement uncertainty; the *top horizontal stripe* shows the situation in the past; the *middle stripe* shows the current one and the *bottom* one the future possible situation with the MoonLIGHT/LLRRA21 CCR

Opportunities for deployment of the MoonLIGHT/LLRRA21 payload will come from the participation of INFN-LNF and UMD, through their national agency programs, to international missions to develop a lunar geophysical network (LGN) [13].

In order to improve the ranging measurement, we have to investigate the technical and fabrication challenges of MoonLIGHT/LLRRA21, through thermal/optical simulations and vacuum chamber tests performed at the INFN-LNF Satellite/lunar/GNSS laser ranging and altimetry Characterization Facility (SCF). Beyond the simulations, we have performed thermal and optical vacuum chamber tests to further validate the design issues.

2.1 SCF_Lab

In 2004, INFN started to build the SCF_Lab in Frascati (Fig. 6. The main purpose of this apparatus is the thermal and optical characterization of CCR arrays in simulated space conditions.

In Fig. 6, the SCF apparatus is shown. It is a steel cryostat 2 m in length by 1 m in diameter. The inner copper shield is painted with Aeroglaze Z306 black paint (0.95 emissivity and low out-gassing properties) and is kept at T = 77 K with liquid nitrogen. When the SCF is cold ($\sim$ 80 K), the vacuum is typically in the 10^{-6} mbar

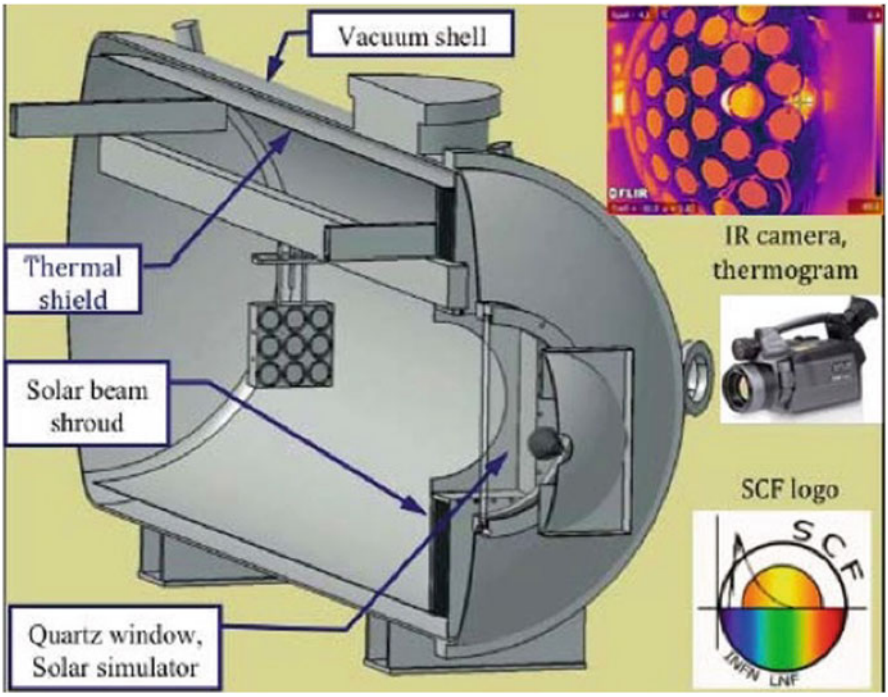


Fig. 6 Drawing of the Satellite/lunar/GNSS laser ranging and altimetry Characterization Facility *(SCF)* cryostat

range. Two distinct positioning systems at the top of the cryostat hold and move the prototype in front of the Earth infrared simulator, the solar simulator (SS), the infrared camera, and the laser, all located outside the SCF.

The experimental apparatus is described in great detail in [14, 15].

2.2 SCF Test of the MoonLIGHT CCR

The SCF-Test [14] is a new test procedure to characterize thermal and optical behavior of laser retroreflectors in space for industrial and scientific applications. We perform an SCF-Test on the MoonLIGHT CCR to evaluate the thermal and optical performance in space environment, (Fig. 7).

The temperature of the housing has been controlled with resistive tape heaters. For thermal measurements, we use both an infrared (IR) camera and temperature probes, which give real time measurements. The IR camera, through a Germanium window, can give thermograms of all the components of the CCR and its housing.

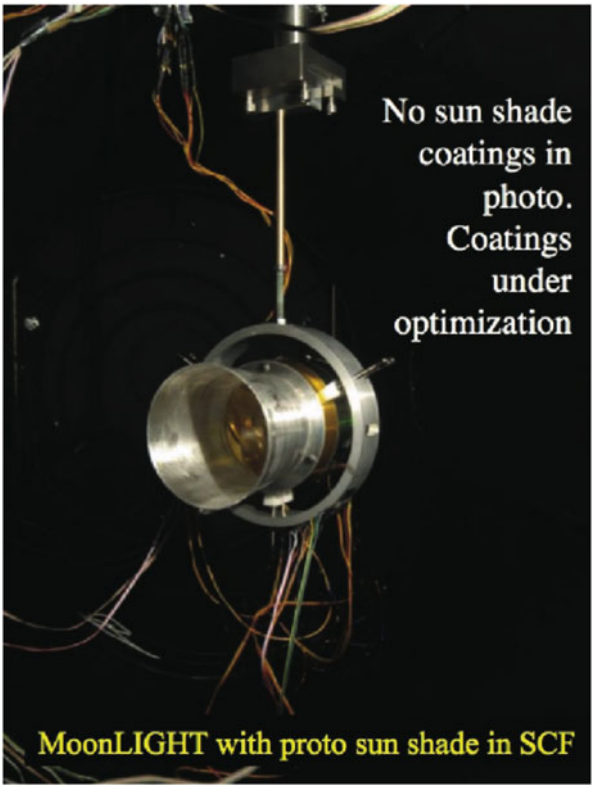


Fig. 7 MoonLIGHT CCR inside the Satellite/lunar/GNSS laser ranging and altimetry Characterization Facility *(SCF)*

Instead, to measure the thermal gradient on the CCR surface, we glued three calibrated temperature sensors (silicon diode) along one of the back faces of the CCR, at a separation of 35 mm.

In particular, we look at the temperature from the front face to the tip, studying how the optical response (the far field diffraction pattern, FFDP) of the CCR changes during the different thermal phases. Various configurations and designs of the CCR and the housing have been tested in the SCF facility, with the solar simulator, the temperature data recording with the infrared camera, and the measurement of the FFDP. At the present time, we have a preliminary result indicating that the CCR has a satisfactory performance.

The orientation of the CCR inside the housing is such that one physical edge was parallel to the axis of the SCFs rotation positioning system. We simulated an illumination of the Sun at lower elevations, so the CCR was rotated 30° clockwise and 30° counterclockwise with respect to the SS. We report a total internal reflection breakthrough in one direction, but not in the other. Figures 8 and 9 show the temperature variation of the housing and the intensity variation of the FFDP.

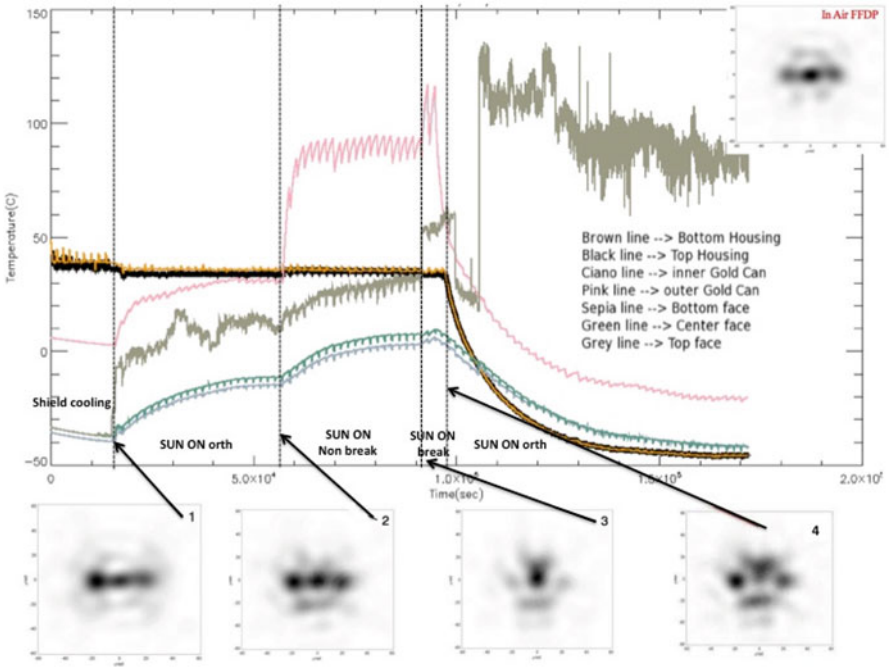


Fig. 8 Temperature variation of the housing and relative far field diffraction pattern *(FFDP)*

Looking at Fig. 9, we conclude that the intensity decreases during the phase when the CCR is not orthogonal to the SS, in particular when the Sun enters the housing cavity during the breakthrough phase. This effect is due to a strong increase of the "Tip-Face" thermal gradient during the test (Fig. 10). When the temperature of the housing is allowed to vary, the intensity increases because the "Tip-Face" gradient is reduced.

3 Analysis of LLR Data

3.1 Planetary Ephemeris Program (PEP)

In order to analyze LLR data we used the PEP software, developed by the CfA (I. Shapiro et al.) starting from 1970s.

PEP was designed not only to generate ephemerides of the planets and the Moon, but also to compare the model with observations [16, 17]. One of the early uses of this software was the measurement of the geodetic precession of the Moon [1]. PEP can handle several observation types (LLR, radar ranging and doppler, optical positional measurements, transponder measurements, pulsar timing data, and so on).

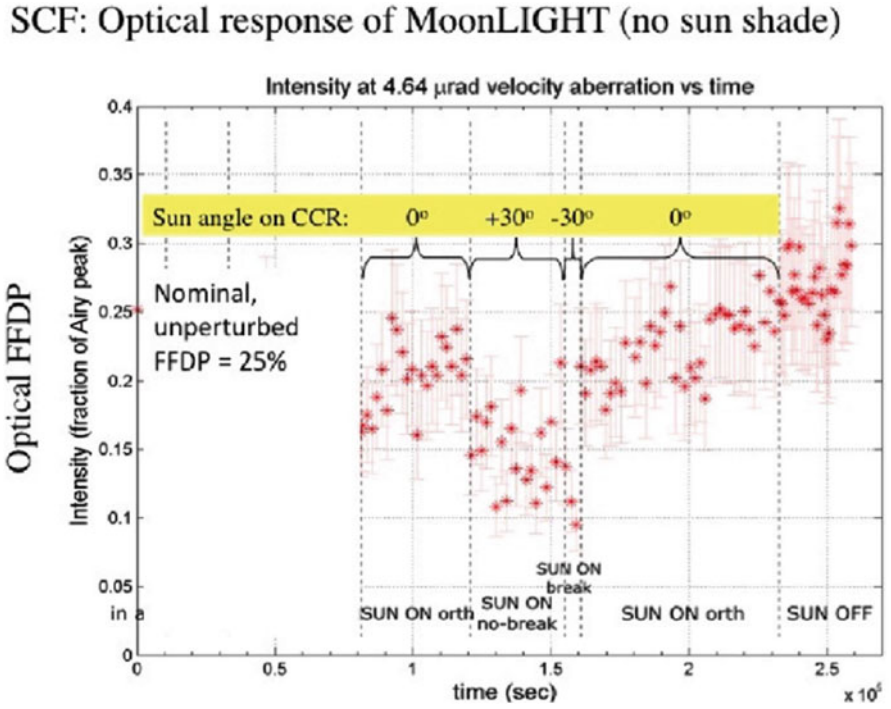


Fig. 9 Intensity variation of the FFDP during the tests

Since we care primarily about LLR observations, we describe the components of a range prediction, though most of the steps along the way are common to other observables as well.

In particular, we are primarily concerned with LLR observations. For this purpose, the software is able to calculate the residuals of the distances between observed LLR data and computed data, derived from the expectations of GR and of terrestrial and lunar geodesy. We have performed a very preliminary analysis of LLR data from three stations: McDonald Observatory in Texas (USA), Grasse in France, and APOLLO [18, 19] in New Mexico (USA). The latter station has been providing the best quality data since 2006. On March 25, 2010, the Matera Laser Ranging Observatory in Italy (MLRO, led by G. Bianco) recorded LLR echoes from the array of Apollo 15.

The histograms in Fig. 11 show photon-by-photon data and are used to form a single LLR "normal point" of the Apollo 15 array taken by the APOLLO station (led by T. W. Murphy) on November 19, 2007. A normal point contains a lot of information, for example, date of observation, atmospheric conditions, as well as time of flight, data quality, and CCR arrays. The APOLLO instrumental accuracy (in terms of laser, detector, timing electronics, and so on) shown by the fiducial returns in Fig. 11 is given by a root mean square contribution of 120 ps (18 mm).

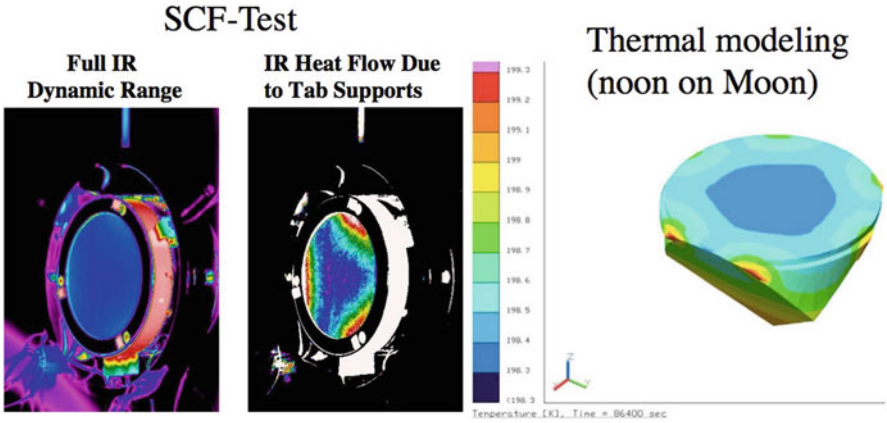


Fig. 10 CCR temperature distribution at noon (orthogonal illumination) with conformal thermal shield

From a comparison between the middle and the last plot we can see how the tilt in the arrays affects the accuracy of the ranging measurements. The model parameter estimates are refined by minimizing the residual differences, in a weighted least-squares sense, between observations (O) and model predictions (C, stands for "Computed"), O-C. "Observed" is the round-trip time of flight. "Computed" is modeled by the PEP software. PEP software has allowed placing constraints on departures from standard physics. For example, it has been used to place limits on the PPN parameters β and γ, the geodetic precession, and the variation of the gravitational constant, $\frac{\dot{G}}{G}$. The equations of motion in a space-time with torsion [20, 21] can be included in PEP and constrained with all LLR data, including the newest APOLLO data (at the present time, the published constraints on space-time torsion are calculated using LLR data from other stations).

3.2 Determination of the Geodetic Precession

With PEP, we are able to measure the possible relative deviation of geodetic precession from the GR value (deviation from zero) that is expressed with K_{GP} parameter (Fig. 12).

Here we show our first determination of the K_{GP} parameter. We have used all the data available to us from Apollo CCR arrays (Apollo 11, Apollo 14, and Apollo 15). The results are reported in two tables, one until 2003, with data acquired by the old ILRS stations (Table 2) and one with data from 2007 to 2009 acquired by the new APOLLO station (Table 4). Results described in the tables are obtained by fixing $\frac{\dot{G}}{G} = 0$ and $\beta = \gamma = 1$, that is, their value in GR.

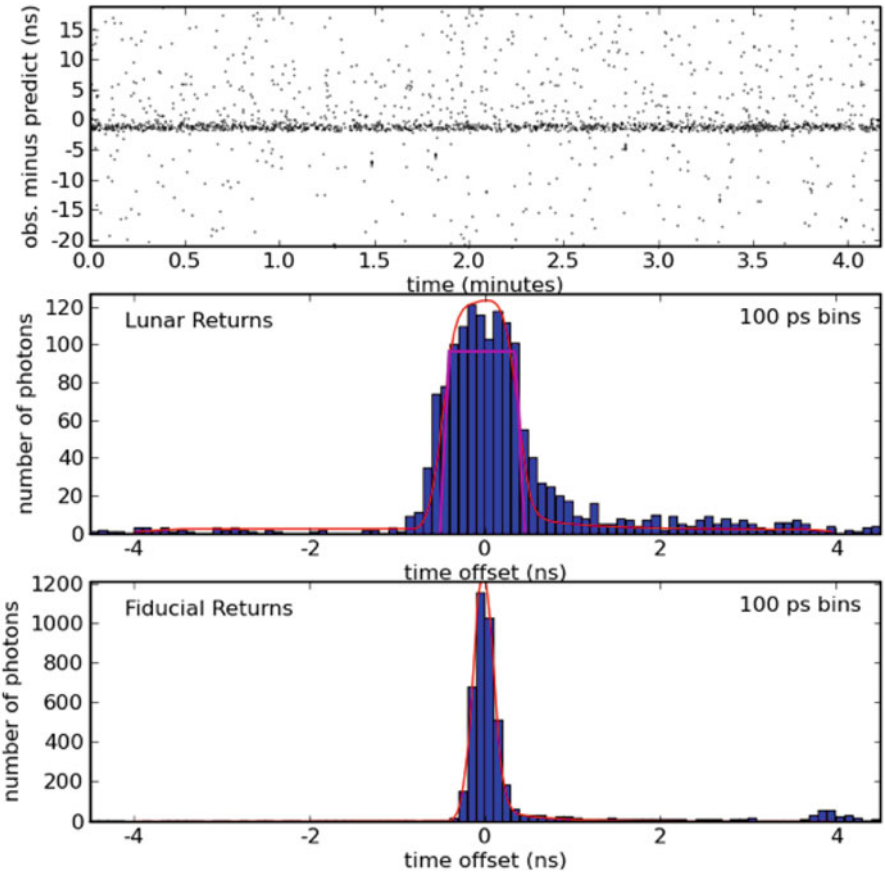


Fig. 11 Example run of Apollo 15. In the plot, the *top panel* shows a 40 ns window of observed round trip time minus the predicted range. Background noise and detector dark current appear as scattered dots, while the lunar return is in the middle. The *middle panel* shows a histogram of the lunar returns, while the *bottom panel* shows the local "fiducial" CCR return, fitted by the red Gaussian. The lunar return is additionally spread by the tilted reflector array modeled by the superimposed magenta trapezoidal shape

Using APOLLO LLR data, the estimated value is consistent with the value obtained using old stations (see Table 3).
The nominal errors returned by the fit are significantly smaller than the value of K_{GP}, and smaller than the best published values. Therefore, we want to use the data to understand and estimate independently the size of the error budget.

We have also performed a fit using every single old station (CERGA: Centre d'Etudes et de Recherches Godynamiques et Astronomiques, MLR2: McDonald Laser Ranging Station, TEXL: Texas Laser Ranging , MAUI: Hawaii Laser Ranging Station). The results are shown in Table 4:

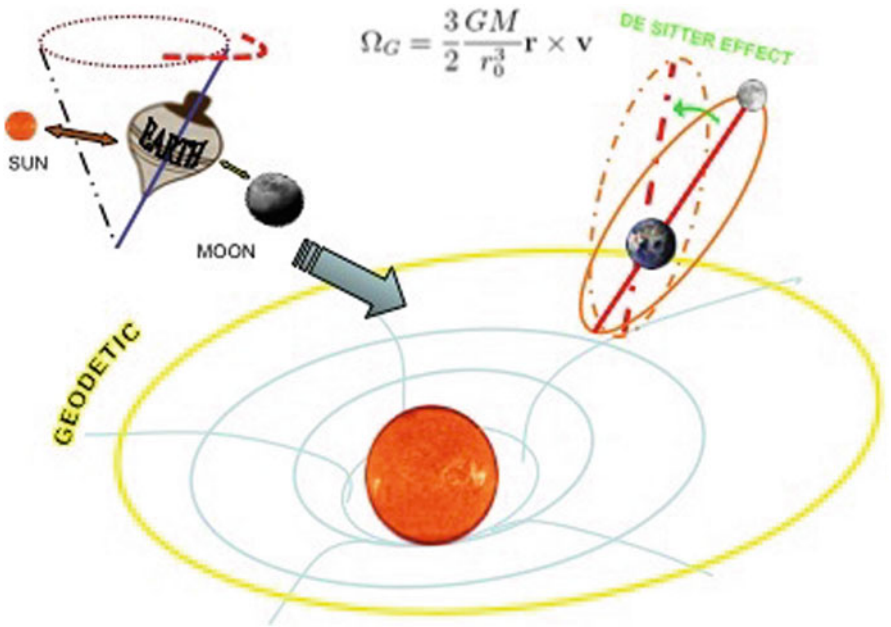


Fig. 12 K_{GP} is the relative deviation of geodetic precession from the general relativity *(GR)* value

Table 2 Estimates of geodetic precession, K_{GP}, with the data set from MLRS, MLR2, and CERGA stations

Parameter	GR initial value	Final value
K_{GP}	0	0.009

GR General relativity, *CERGA* Centre d'Etudes et de Recherches Godynamiques et Astronomiques, *MLR2* McDonald Laser Ranging Station, *MLRS* McDonald Laser Ranging Station

Table 3 Estimates of geodetic precession, K_{GP}, with the new data set from APOLLO station

Parameter	GR initial value	Final value
K_{GP}	0	−0.0096

APOLLO Apache Point Observatory Lunar Laser-ranging Operation

These preliminary measurements are to be compared with the best result published by the Jet Propulsion Laboratory (JPL) ($K_{GP} = (-1.9 \pm 6.4) \times 10^{-3}$), obtained using a completely different software package, developed over the past 40 years. On the contrary, after the original 2 % K_{GP} measurement by CfA in 1988, the use of PEP for LLR has been resumed only since a few years, and it is still undergoing the necessary modernization and optimization.

Table 4 Estimates of geodetic precession, K_{GP} using every single old station

Station	K_{GP}
CERGA	−0.016
MAUI	0.0060
MLR2	0.0095
TEXL	−0.044

CERGA Centre d'Etudes et de Recherches Godynamiques et Astronomiques, *MLR2* McDonald Laser Ranging Station, *TEXL* Texas Laser Ranging, *MAUI* Hawaii Laser Ranging Station

4 Determination of β

Using LLR data is also possible to measure PPN parameter β that is of great interest. In order to do this, we start from the equivalence principle (EP).

There are two different kinds of EP: weak equivalence principle (WEP that concerns nongravitational contribution to mass) and strong equivalence principle (SEP that extends the WEP to include gravitational self-energy of a body).

In GR, SEP is exact and a possible violation can be expressed by:

$$\left[\frac{M_G}{M_I}\right]_{SEP} = 1 + \eta\frac{U}{Mc^2}, \tag{1}$$

where $\frac{M_G}{M_I}$ is the ratio between gravitational and inertial mass, U is the gravitational energy, and η is a constant expressed by:

$$\eta = 4\beta - \gamma - 3. \tag{2}$$

Considering the Earth–Moon–Sun system, we have:

$$\left[\left(\frac{M_G}{M_I}\right)_e - \left(\frac{M_G}{M_I}\right)_m\right]_{SEP} = \left[\frac{U_e}{M_e c^2} - \frac{U_m}{M_m c^2}\right]\eta = -4.45 \times 10^{-10}\eta, \tag{3}$$

where e is relative to the Earth and m is relative to the Moon.

Fitting the data using LLR measurements:

$$\left[\left(\frac{M_G}{M_I}\right)_e - \left(\frac{M_G}{M_I}\right)_m\right]_{EP} = (-1.0 \pm 1.4) \times 10^{-13}. \tag{4}$$

The combination of laboratory experiments of EP [22] with LLR data gives:

$$\left[\left(\frac{M_G}{M_I}\right)_e - \left(\frac{M_G}{M_I}\right)_m\right]_{SEP} = (-2.0 \pm 2.0) \times 10^{-13}. \tag{5}$$

Considering the previous equation and the most accurate result of γ [23], the value of the β-parameter is:

$$\beta - 1 = (1.2 \pm 1.1) \times 10^{-4}. \tag{6}$$

Our value of the β-parameter calculated using the PEP software is:

$$\beta = (6.4 \pm 6.4) \times 10^{-4}, \tag{7}$$

which is consistent with the value obtained by [4].

5 Conclusions and Future Prospects

The analysis of the existing LLR data with PEP is making good progress, thanks to the important collaboration with CfA, as shown with the preliminary measurement of the geodetic precession (de Sitter effect) with an accuracy level of 1 %.

In the future, we are going to enhance our knowledge of data and software in order to better estimate the K_{GP} uncertainty and other GR parameters. A possible way to improve the precision of LLR measurements is to improve the intercalibration among the stations, that is, to range not only to the Moon but also to satellites around the Earth and primarily to laser geodynamics satellites (LAGEOS).

Before the end of the decade, a robotic mission on the lunar surface will deploy new scientific payloads, which include MoonLIGHT-type laser retroreflectors, and thus extend the LLR reach for new physics in three ways: (i) using a significantly improved second-generation retroreflector design; (ii) increasing by a factor of about 2 the geometric lever arm of LLR with missions to the lunar poles or limbs; (iii) combining LLR payloads with radio/microwave transponders (at least two) for same-beam microwave interferometry (SBI) capable of additional accurate measurements of lunar rotations and librations.

In particular, the single, large, fused-silica retroreflector design developed by UMD and INFN-LNF will improve by a factor of 100 or more the performance of current Apollo arrays, removing in this way the dominant contribution to the LLR error budget. Such a contribution is of the order of 2 cm. It is due to the multi-retroreflector structure of the arrays together with the librations and rotations of the Moon with respect to the Earth. The functionality of this specific new design, which inherits and is evolved from the successful Apollo 11, 14, and 15 experience, is being validated by thermal–vacuum–optical testing in space conditions accurately simulated in the laboratory at the INFN-LNF SCF_Lab."

Acknowledgements We thank Ryan Heller, Griffin Hosseinzadeh (DOE/INFN Summer Exchange Program for 2012), and Chiara Mondaini for the English revision of this chapter.

References

1. Shapiro II, Reasenberg RD, Chandler JF, Babcock RW. Measurement of the de Sitter precession of the Moon: a relativistic three-body effect. Phys Rev Lett. 1988;61:2643–6.
2. Bertotti B, Ciufolini I, Bender PL. New test of general relativity: measurement of de Sitter geodetic precession rate for lunar perigee. Phys Rev Lett. 1987;58:1062–5.

3. Williams JG, Newhall XX, Dickey Jean O. Lunar moments, tides, orientation, and coordinate frames. Planet Space Sci. 1996;44:1077–80.
4. Williams JG, Turyshev SG, Boggs DH, Ratcliff JT. Lunar laser ranging science: gravitational physics and lunar interior and geodesy. Adv Space Res. 2006;37(1):67–71.
5. Dell'Agnello S., Currie D., Delle Monache G, et al. MoonLIGHT: a lunar laser ranging retroreflector array for the 21st century. Paper # GLEX-2012.02.1.7x12545 in Proceedings of the Global Lunar Exploration Conference; May 2012; Washington, USA.
6. Alley CO, Chang RF, Currie DG, Mullendore J, Poultney SK, Rayner JD, Silverberg EC, Steggerda CA, Plotkin HH, Williams W, Warner B, Richardson H, Bopp B. Apollo 11 laser ranging retroreflector: initial measurements from the McDonald observatory. Science. 1970;167(3917):368–70.
7. Chang RF, Alley CO, Currie DG, Faller JE. Optical properties of the Apollo laser ranging retroreflector arrays. Space Res XII. 1971;1:247– 59.
8. Bender PL, Currie DG, Dicke RH, Eckhardt DH, Faller JE, Kaula WM, Mulholland JD, Plotkin HH, Poultney SK, Silverberg EC, Wilkinson DT, Williams JG, Alley CO. The lunar laser ranging experiment. Science. 1973;182(4109):229–38.
9. Currie D, Dell'Agnello S, Delle Monache G. A lunar laser ranging retroreflector array for the 21st century. Acta Astronaut 2011;68:667–80.
10. Lops C, Martini M. Horizons in Earth science research. Vol. 5. New York: Nova; 2011. (ISBN: 978-1-61209-923-1).
11. Martini M, et al. Probing gravity with second generation Lunar Laser Ranging. Talk for the XIII LNF Spring School in Nuclear, Subnuclear and Astroparticle Physics, Frascati, 12–16 May 2008.
12. Lops C. et al. Probing gravity in Near Earth orbits with LAGEOS and LARES. Talk for the XIII LNF Spring School in Nuclear, Subnuclear and Astroparticle Physics, Frascati, 12–16 May 2008.
13. A scientific concept for the LGN has been developed by the International Lunar Network (see http://iln.arc.nasa.gov/). See Core Instrument and Communications Working Group Final Reports.
14. Dell'Agnello S, et al. Creation of the new industry-standard space test of laser retroreflectors for the GNSS and LAGEOS. Adv Space Res. 2011;47:822–42. http://ilrs.gsfc.nasa.gov/docs/ETRUSCO_PaperPublished_JASR.pdf.
15. Dell'Agnello S. ETRUSCO-2: An ASI-INFN project of technological development and SCF-Test of GNSS LASER retroreflector arrays. 3rd International Colloquium on Scientific and Fundamental Aspects of the Galileo Programme; August 2011; Copenhagen, Denmark. http://ilrs.gsfc.nasa.gov/satellite_missions/list_of_satellites/ga01_reflector.html.
16. Chandler JF, Reasenberg RD, Shapiro II. On recent developments in theoretical and experimental general relativity, gravitation, and relativistic field theories. In: Jantzen RT, Mac Keiser G, Ruffini R, editors. Proceedings of the Seventh Marcel Grossman Meeting; 24–30 July 1994; p. 1501; Stanford, USA; 1996.
17. Battat JBR, Chandler JF, Stubbs CW. Testing for Lorentz violation: constraints on standard-model-extension parameters via lunar laser ranging. Phys Rev Lett. 2007;99:241103, arXiv:0710.0702.
18. Battat JBR, Murphy TW, et al. The Apache Point Observatory Lunar Laser-ranging Operation (APOLLO): two years of millimeter-precision measurements of the Earth-Moon range. Publications of The Astronomical Society of the Pacific, 2009;121:29–40.
19. Murphy Jr. TW, Adelberger EG, Strasburg JD, Stubbs CW. APOLLO: Multiplexed Lunar Laser Ranging. 13th international workshop on laser ranging; October 2002; Washington DC, USA. http://physics.ucsd.edu/∼tmurphy/apollo/doc/multiplex.pdf.
20. March R, Bellettini G, Tauraso R, Dell'Agnello S. Constraining spacetime torsion with the Moon and Mercury. Phys Rev D. 2011;83:104008.
21. March R, Bellettini G, Tauraso R, Dell'Agnello S. Constraining spacetime torsion with LAGEOS. Gen Relativ Gravit. 2011;43(11):3099–126.
22. Baessler S, Adelberger EG et al. Improved test of the equivalence principle for gravitational self-energy. Phys Rev Lett. 1999;83:3585.
23. Bertotti B, Iess L, Tortora P. A test of general relativity using radio links with the Cassini spacecraft. Nature. 2003;425:374–6.

Space-based Tests of Relativistic Gravitation

Vyacheslav G. Turyshev

Abstract Since its initial publication, Einstein's general theory of relativity had been tested to a very high precision and presently is considered to be the standard theory of gravitation, especially when the phenomena in astrophysics, cosmology, and fundamental physics are concerned. As such, this theory has many practically important applications including spacecraft navigation, relativistic geodesy, time transfer, etc. Here we discuss the foundations of general relativity, present its current empirical status, and highlight the need for the new generation of high-accuracy tests. We present some space-based gravitational experiments and discuss anticipated advances in our understanding of the fundamental laws of nature.

1 Introduction

With a successful explanation of the puzzle of the perihelion advance of the Mercury's orbit calculated by Einstein himself in 1916 [1, 2] and confirmation of the general relativistic gravitational deflection of light reported by the Eddington's expedition in 1919, the Einstein's general theory of relativity has been tested and verified with ever-increasing precision [3]. Presently, after nearly a century-long and very successful reign, the theory accounts for all the empirical data gathered within our solar system and also received from the millisecond binary pulsars [4–6].

The true renaissance in the tests of relativistic gravitation began in early 1970s with advances in microwave tracking of interplanetary spacecraft, high-precision astrometric observations with very long baseline interferometry (VLBI), and the beginning of lunar laser ranging (LLR) operations that followed the Apollo missions to the Moon. In particular, analysis of ranging data provided by the Viking lander on Mars became the first use of then new interplanetary tracking technologies to test Einstein's theory of gravitation. Thus, the value of the Eddington's space curvature parameter γ (whose value is 1 in general relativity, see [7] for details) was obtained at the level of 1.000 ± 0.002 [8]. Later on, tracking of interplanetary spacecraft and

V. G. Turyshev (✉)
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109-0899, USA
e-mail: turyshev@jpl.nasa.gov

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_6

series of planetary radar observations reached the accuracy of $\sim 0.15\,\%$ [9, 10] in testing relativistic gravity via dynamics in the solar system.

At the same time, the geodetic observations with VLBI have reached the accuracy below 100 μas (micro-arcseconds of arc), yielding a number of very precise measurements of the parameter γ. It is well known that bending of starlight by gravity was predicted by Einstein when he published his theory in 1916. It was suggested that the gravity of a massive object curves the nearby space, thereby altering the path of light passing near the object. The phenomenon was first observed by Eddington during a solar eclipse in 1919 [3]. Over the years, multiple independent analyses of various VLBI data have yielded a consistent stream of improvements in measuring the gravitational deflection of light and reaching the present-day accuracy of $\gamma = 0.9998 \pm 0.0003$ [11], resulting in the accuracy of $\sim 0.03\,\%$ in the tests of gravity via astrometric VLBI observations.

LLR, the longest continuously running space experiment to date, has provided series of important constraints on the Eddington parameter β (a measure of nonlinearity of superposition of gravitational fields; equal to 1 in general relativity, see [12] for details). Thus, using the Cassini value for the parameterized post-Newtonian (PPN) parameter γ, the analysis of LLR data in 2004 [13] constrained the combination of parameters $4\beta - \gamma - 3 = (4.0 \pm 4.3) \times 10^{-4}$, leading to the current best accuracy of $\sim 0.011\,\%$ in testing the Eddington's parameter β via precision measurements of the lunar orbit.

Finally, microwave tracking of the Cassini spacecraft on its approach to Saturn in 2002 has provided the most precise measurement of the relativistic gravity parameter γ measuring it with an accuracy of $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$, thereby reaching the current best accuracy of $\sim 0.002\,\%$ achieved in the solar system tests of fundamental gravitation [6, 14]. Figure 1 shows the progress in gravitational tests for the past several decades.

To date, the general theory of relativity is also in agreement with the data collected in studying binary and double pulsars. By measuring relativistic corrections to the Keplerian description of the orbital motion, the recent analysis of the data collected from the double pulsar system, PSR J0737-3039A/B, found agreement with general relativity with an uncertainty of $\sim 0.05\,\%$ [15], which is the most precise pulsar test of gravity yet obtained.

As a result, both in the limit of weak gravitational field (such as in our solar system) and with the stronger fields (present in the vicinity of the binary pulsars) the foundations of the general theory of relativity have been tested extremely well with the theory surviving every test. It is remarkable that nearly 100 years after its discovery, Einstein's theory of gravitation has survived every test [4]. Such a successful longevity makes general relativity the de facto "standard" theory of gravitation for all practical purposes involving spacecraft navigation and astrometry, astronomy, astrophysics, cosmology, and fundamental physics [5].

However, despite a very impressive track record, there are many important reasons to continue with high-precision tests of relativistic gravitation, especially those conducted in the solar system. In fact, our solar system is a very important laboratory to conduct gravitational experiments. The solar system gravitational field is

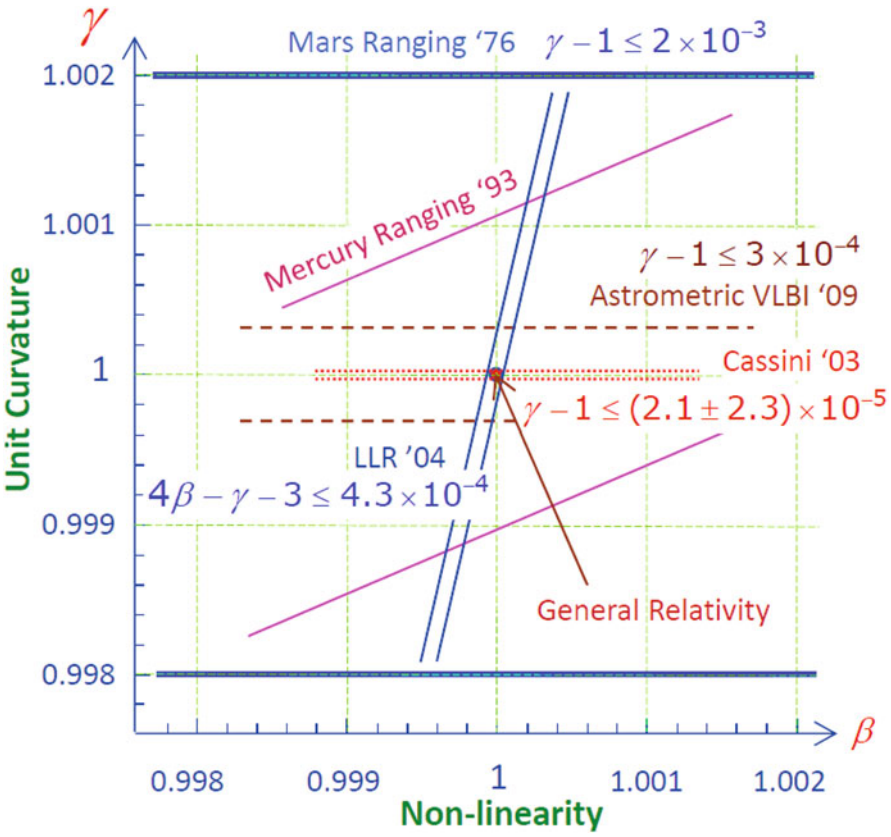


Fig. 1 The progress in measuring the Eddington's parameters γ and β. General theory of relativity survived every test, yielding $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [14] and $\beta - 1 = (1.2 \pm 1.1) \times 10^{-4}$ [13]

determined by the dynamical distribution of matter in the solar system and by that outside the solar system itself. Different relativistic theories of gravity can make different predictions about the properties of the solar system gravitational field with some being within the reach of modern technologies. Hence, precise investigations of the orbits of the celestial bodies, trajectories of spacecraft, and that of electromagnetic signals can be used to test modern theories of gravitation. This is especially important since the highly precise measurements may reveal the presence of physics beyond the framework of general relativity, which currently is highly anticipated.

This chapter discusses a number of recently proposed gravitational experiments in space and anticipated advances in our understanding of the fundamental laws of nature. This short review is organized as follows. Section 2 discusses the foundations of general theory of relativity and reviews the recent results of the tests of its foundations. In Sect. 3 we present the PPN formalism—a phenomenological framework developed to facilitate experimental research in relativistic gravity. In Sect. 4

we review the recent progress in the tests of the general relativity and discuss several proposed space missions, focusing only on the most representative and viable concepts. We conclude in Sect. 5.

2 General Theory of Relativity

Fundamentally, the general theory of relativity is a tensor-field theory of gravitation with universal coupling to the particles and fields of the Standard Model. It describes gravity as a universal deformation h_{mn} of the flat space-time Minkowski metric,[1] γ_{mn}:

$$g_{mn}(x^k) = \gamma_{mn} + h_{mn}(x^k). \tag{1}$$

On a classical level [1, 2], the theory is introduced by the action describing the propagation and self-interaction of the gravitational field [16–19]

$$\mathcal{S}_{\rm G}[g_{mn}] = \frac{c^4}{16\pi G}\int d^4x\sqrt{-g}R, \tag{2}$$

where G is Newton's universal gravitational constant, $g = \det g_{mn}$, and R is the Ricci scalar. Applying the variational principle with respect to g_{mn} to the total action that contains the gravitational field and the fields of matter from the Standard Model, one obtains the well-known Einstein's field equations of general relativity,

$$R_{mn} - \frac{1}{2}g_{mn}R + \Lambda g_{mn} = \frac{8\pi G}{c^4}T_{mn}, \tag{3}$$

where $T_{mn} = g_{mk}g_{nl}T^{kl}$ with $T^{mn} = 2/\sqrt{-g}\;\delta\mathcal{L}_{\rm SM}/\delta g_{mn}$ being the (symmetric) energy–momentum tensor of the matter as described by the Standard Model with the Lagrangian density $\mathcal{L}_{\rm SM}$. With the value for the vacuum energy density $\rho_{\rm vac} \approx (2.3 \times 10^{-3}{\rm eV})^4$, as measured by recent cosmological observations [20, 21], the cosmological constant $\Lambda = 8\pi G\rho_{\rm vac}/c^4$ is too small to be observed by solar system experiments, but may be important for greater scales.

The theory is invariant under arbitrary coordinate transformations: $x'^m = f^m(x^n)$. To solve the field Eq. (3), one needs to fix this coordinate gauge freedom. For example, the "harmonic gauge" (which is the analog of the Lorentz gauge, $\partial_m A^m = 0$, in electromagnetism) corresponds to imposing the condition $\partial_n(\sqrt{-g}g^{mn}) = 0$.

Einstein's equations (3) of general theory of relativity connect the geometry of a four-dimensional Riemannian manifold representing space-time with the energy-momentum contained in that space-time. In classical mechanics, phenomena that

[1] The notational conventions employed here are as follows: Letters from the second half of the Latin alphabet, $m, n, \ldots = 0\ldots3$ denote space-time indices. Greek letters $\alpha, \beta, \ldots = 1\ldots3$ denote spatial indices. The flat Minkowski space-time metic is $\gamma_{mn} = {\rm diag}(+1, -1, -1, -1)$. We rely on the Einstein summation convention with indices being lowered or raised using γ_{mn}.

are ascribed to the action of the force of gravity (such as free-fall, orbital motion, and spacecraft trajectories) correspond to inertial motion within a curved geometry of space-time in general relativity.

Equation (3) represents a set of nonliner hyperbolic-type differential equations of the second order with respect to the metric tensor of the Riemannian space-time. This nonlinearity makes finding a solution to this set of equations in the general case to be a complicated problem to which no analytical solution is known and a full numerical treatment is needed. Depending on a particular situation, one usually introduces relevant small parameters and develops an iterative solution. Quite often, when studying a problem in the weak gravitational field and slow motion approximation, one uses the ratio v/c of the velocity of motion of the bodies v with respect to a barycentric coordinate reference frame to the speed of light c.

For the bodies of the solar system this ratio is a small parameter and typically is of the order $v/c \approx 10^{-4}$. The gravitational field in the solar system is also quite weak (compared to that of, say, a neutron star or a black hole) and, as the virial theorem suggests, the potential energy in the solar system, represented by the Newtonian potential U, is of the same order as the kinetic energy, leading to the following relationship $U/c^2 \approx (G/c^2)(m/r) \sim v^2/c^2 \approx 10^{-8}$, where G is the universal gravitational constant, m is a mass monopole of a body's mass distribution, and r is its radius vector with respect to the barycentric coordinate reference frame.

The metric tensor for a system of N point-like gravitational sources [18, 22, 23] that satisfy Eq. (3) may be written as

$$
\begin{aligned}
g_{00} &= 1 - \frac{2}{c^2}\sum_{j\neq i}\frac{\mu_j}{r_{ij}} + \frac{2}{c^4}\left\{\left[\sum_{j\neq i}\frac{\mu_j}{r_{ij}}\right]^2 - \frac{3}{2}\sum_{j\neq i}\frac{\mu_j \dot{r}_j^2}{r_{ij}}\right. \\
&\left. + \sum_{j\neq i}\frac{\mu_j}{r_{ij}}\sum_{k\neq j}\frac{\mu_k}{r_{jk}} - \frac{1}{2}\sum_{j\neq i}\mu_j \frac{\partial^2 r_{ij}}{\partial t^2}\right\} + \mathcal{O}(c^{-5}), \\
g_{0\alpha} &= \frac{4}{c^3}\sum_{j\neq i}\frac{\mu_j \dot{\mathbf{r}}_j^{\alpha}}{r_{ij}} + \mathcal{O}(c^{-5}), \quad g_{\alpha\beta} = -\delta_{\alpha\beta}\left(1 + \frac{2}{c^2}\sum_{j\neq i}\frac{\mu_j}{r_{ij}}\right) + \mathcal{O}(c^{-5}),
\end{aligned}
\tag{4}
$$

where the indices j and k refer to the N bodies and where k includes body i whose motion is being investigated. μ_j is the gravitational constant for body j given as $\mu_j = Gm_j$, where G is the universal Newtonian gravitational constant and m_j is the isolated rest mass of a body j. In addition, the vector $\mathbf{r}_i$ is the barycentric radius vector of this body, the vector $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ is the vector directed from body i to body j, $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$, and the vector $\mathbf{n}_{ij} = \mathbf{r}_{ij}/r_{ij}$ is the unit vector along this direction.

The point-mass Newtonian and relativistic perturbative accelerations in the solar system's barycentric reference frame have the form [16, 18, 24]:

$$\ddot{\mathbf{r}}_i = \sum_{j\neq i} \frac{\mu_j(\mathbf{r}_j - \mathbf{r}_i)}{r_{ij}^3}\Bigg\{1 - \frac{4}{c^2}\sum_{l\neq i}\frac{\mu_l}{r_{il}} - \frac{1}{c^2}\sum_{k\neq j}\frac{\mu_k}{r_{jk}} + \big(\frac{\dot{r}_i}{c}\big)^2 + 2\big(\frac{\dot{r}_j}{c}\big)^2 - \frac{4}{c^2}\dot{\mathbf{r}}_i\dot{\mathbf{r}}_j$$
$$- \frac{3}{2c^2}\left[\frac{(\mathbf{r}_i - \mathbf{r}_j)\dot{\mathbf{r}}_j}{r_{ij}}\right]^2 + \frac{1}{2c^2}(\mathbf{r}_j - \mathbf{r}_i)\ddot{\mathbf{r}}_j\Bigg\}$$
$$+ \frac{1}{c^2}\sum_{j\neq i}\frac{\mu_j}{r_{ij}^3}\Big\{\Big[\mathbf{r}_i - \mathbf{r}_j\Big]\cdot\Big[4\dot{\mathbf{r}}_i - 3\dot{\mathbf{r}}_j\Big]\Big\}(\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j) + \frac{7}{2c^2}\sum_{j\neq i}\frac{\mu_j\ddot{\mathbf{r}}_j}{r_{ij}} + \mathcal{O}(c^{-4}). \tag{5}$$

To describe the propagation of electromagnetic signals between any two points in space, the following light-time equation is derived from the metric tensor Eq. (4)

$$t_2 - t_1 = \frac{r_{12}}{c} + 2\sum_i \frac{\mu_i}{c^3}\ln\left[\frac{r_1^i + r_2^i + r_{12}^i + \frac{2\mu_i}{c^2}}{r_1^i + r_2^i - r_{12}^i + \frac{2\mu_i}{c^2}}\right] + \mathcal{O}(c^{-5}), \tag{6}$$

where t_1 refers to the signal transmission time and t_2 refers to the reception time. $r_{1,2}$ are the barycentric positions of the transmitter and receiver and r_{12} is their spatial separation. The terms proportional to μ_i^2 are important only in the proximity of the Sun and are negligible for all other bodies in the solar system [22]. General relativistic equations of motion form the "backbone" of the entire suite of models used for interplanetary spacecraft navigation [22, 25].

3 Metric Theories of Gravity and PPN Formalism

Metric theories hold a special position among other alternative theories of gravity [3, 7]. This is because independently of many different principles in their foundations, the gravitational field in these theories affects the matter directly through the metric tensor g_{mn}, which is determined from the field equations of this particular theory. As a result, this tensor expresses the properties of a particular gravitational theory and carries information about the gravitational field of the bodies.

In many alternative theories of gravity, the gravitational coupling strength depends on a certain field. In scalar–tensor theories, this is a scalar field φ. The scalar–tensor theories of gravity are some of the most established and well-studied alternative theories of gravity that exist in the literature. They are often used as the prototypical way in which deviations from general relativity are modeled, and are of particular interest as the relatively simple structure of their field equations allows exact analytic solutions to be found in a number of physically interesting situations.

Scalar–tensor theories arise naturally as the dimensionally reduced effective theories of higher dimensional theories, such as Kaluza–Klein and string models. They are also often used as simple ways to self-consistently model possible variations in

Newton's constant, G. The Brans–Dicke theory [26] is the best-known alternative scalar–tensor theory of gravity with action being

$$S = \frac{c^3}{16\pi G}\int d^4x\sqrt{-g}\Big[\varphi R - \frac{2\omega}{\varphi}\partial_m\varphi\partial^m\varphi\Big] + \sum_i q_i(\varphi)\mathcal{L}_i, \tag{7}$$

where $q_i(\varphi)$ are coupling functions, $\mathcal{L}_i$ is the Lagrangian density of the matter fields of the Standard Model, and ω is an arbitrary coupling constant.

Note that in the Brans–Dicke theory (Eq. 7), the kinetic energy term of the field φ is noncanonical and has a dimension of energy squared. In the context of this theory [26], one can operationally introduce Mach's principle which states that the inertia of bodies is due to their interaction with the matter distribution in the universe. Indeed, in this theory the gravitational coupling is proportional to φ^{-1}, which depends on the energy–momentum tensor of matter through the field equations.

The presence of an unknown dimensionless parameter ω in the action Eq. (7) yields non-general relativistic values for the two PPN parameters, $\beta = 1$, $\gamma = (1+\omega)/(2+\omega)$, a feature of the theory that may lead to observational deviations from general relativity. The stringent observational bound resulting from the 2003 experiment with the Cassini spacecraft requires that $|\omega| \gtrsim 40,000$ [14]. There exist additional alternative theories that provide guidance for gravitational experiments.

3.1 PPN-Renormalized Extension of General Relativity

Using Eddington's phenomenological parameterization of the gravitational metric tensor as a starting point, a method called the PPN formalism has been developed (see details in [7]). The PPN formalism represents the gravity tensor's potentials for slow moving bodies and weak interbody gravity, and is valid for a broad class of metric theories, including general relativity as a unique case. The several parameters in the PPN metric expansion vary from theory to theory and are individually associated with various symmetries and invariance properties of the underlying theory.

Note that in the complete PPN framework, a particular metric theory of gravity in the PPN formalism with a specific coordinate gauge is fully characterized by means of ten PPN parameters [7]. Thus, in addition to the parameters γ and β, there are eight other parameters, not included in Eq. (4). The formalism uniquely prescribes the values of these parameters for the particular theory under study. Gravity experiments can be analyzed in terms of the PPN metric, and an ensemble of experiments determine the unique value for these parameters and hence the metric field itself.

In the special case, when only two PPN parameters (γ, β) are considered, these parameters have a clear physical meaning. The parameter γ represents the measure of the curvature of the space-time created by a unit rest mass; parameter β represents a measure of the nonlinearity of the law of superposition of the gravitational fields in the theory of gravity. General relativity, when analyzed in standard PPN gauge,

gives $\gamma = \beta = 1$, and the other eight parameters vanish; the theory is thus embedded in a two-dimensional space of theories.

Given the phenomenological success of general relativity, it is convenient to use this theory to describe experiments. At the same time, any possible deviation from general relativity would appear as a small perturbation to this general relativistic background. Such perturbations are proportional to renormalized PPN parameters (i.e., $\bar{\gamma} \equiv \gamma - 1$, $\bar{\beta} \equiv \beta - 1$, etc.), which are zero in general relativity, but which may have nonzero values for some gravitational theories. In terms of the metric tensor, this PPN-perturbative procedure may be conceptually presented as

$$g_{mn} = g_{mn}^{\rm GR} + \delta g_{mn}^{\rm PPN}, \tag{8}$$

where metric $g_{mn}^{\rm GR}$ is derived from Eq. (4) by taking the general relativistic values of the PPN parameters and where $\delta g_{mn}^{\rm PPN}$ is the PPN metric perturbation. The PPN-renormalized metric perturbation $\delta g_{mn}^{\rm PPN}$ for a system of N point-like gravitational sources in four dimensions may be given as

$$\begin{aligned}
\delta g_{00}^{\rm PPN} &= -\frac{2\bar{\gamma}}{c^4}\sum_{j\neq i}\frac{\mu_j \dot{r}_j^2}{r_{ij}} + \frac{2\bar{\beta}}{c^4}\left(\Big[\sum_{j\neq i}\frac{\mu_j}{r_{ij}}\Big]^2 + 2\sum_{j\neq i}\frac{\mu_j}{r_{ij}}\sum_{k\neq j}\frac{\mu_k}{r_{jk}}\right) + \mathcal{O}(c^{-5}),\\
\delta g_{0\alpha}^{\rm PPN} &= \frac{2\bar{\gamma}}{c^3}\sum_{j\neq i}\frac{\mu_j \dot{\mathbf{r}}_j^{\alpha}}{r_{ij}} + \mathcal{O}(c^{-5}), \quad \delta g_{\alpha\beta}^{\rm PPN} = -\delta_{\alpha\beta}\frac{2\bar{\gamma}}{c^2}\sum_{j\neq i}\frac{\mu_j}{r_{ij}} + \mathcal{O}(c^{-5}).
\end{aligned} \tag{9}$$

Thanks to the smallness of the current values for the PPN parameters $\bar{\gamma}$ and $\bar{\beta}$, the PPN metric perturbation $\delta g_{mn}^{\rm PPN}$ in (9) represents a very small deformation from the general relativistic background $g_{mn}^{\rm GR}$. The expressions (9) embody the "spirit" of many gravitational tests that assume that general relativity provides the correct description of the experimental situation at the first order and search for small non-Einsteinian deviations from that background.

The equations of motion (5) may also be presented in the PPN-renormalized form with explicit dependence on the PPN-perturbative acceleration terms:

$$\ddot{\mathbf{r}}_i = \ddot{\mathbf{r}}_i^{\rm GR} + \delta\ddot{\mathbf{r}}_i^{\rm PPN}, \tag{10}$$

where $\ddot{\mathbf{r}}_i^{\rm GR}$ are the equations of motion from Eq. (5) with the values of the PPN parameters γ and β set to their general relativistic values of unity. Then the PPN-perturbative acceleration term $\delta\ddot{\mathbf{r}}_i^{\rm PPN}$ is given as

$$\begin{aligned}
\delta\ddot{\mathbf{r}}_i^{\rm PPN} = \sum_{j\neq i}\frac{\mu_j(\mathbf{r}_j - \mathbf{r}_i)}{r_{ij}^3}\Bigg\{&\left(\left[\frac{m_G}{m_I}\right]_i - 1\right) + \frac{\dot{G}}{G}\cdot(t - t_0)\\
&- \frac{2(\bar{\beta}+\bar{\gamma})}{c^2}\sum_{l\neq i}\frac{\mu_l}{r_{il}} - \frac{2\bar{\beta}}{c^2}\sum_{k\neq j}\frac{\mu_k}{r_{jk}} + \frac{\bar{\gamma}}{c^2}(\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j)^2\Bigg\}\\
&+ \frac{2\bar{\gamma}}{c^2}\sum_{j\neq i}\frac{\mu_j}{r_{ij}^3}\Big\{(\mathbf{r}_i - \mathbf{r}_j)\cdot(\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j)\Big\}(\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j)
\end{aligned}$$

$$+\frac{2\bar{\gamma}}{c^2}\sum_{j\neq i}\frac{\mu_j\ddot{\mathbf{r}}_j}{r_{ij}}+\mathcal{O}(c^{-4}). \tag{11}$$

Equation (11) provides a useful framework for gravitational tests conducted in the solar system. Thus, besides the terms with PPN-renormalized parameters $\bar{\gamma}$ and $\bar{\beta}$, it also contains the expression $([m_G/m_I]_i - 1)$ that signifies a possible inequality between the gravitational and inertial masses of the same body; this contribution is needed to facilitate the investigation of a possible violation of the equivalence principle (EP; see Sect. 4.1.2). In addition, Eq. (11) also includes the parameter $\dot{G}/G$, which is needed to investigate a possible temporal variation in the gravitational constant (see Sect. 4.2), during the period starting from the initial moment t_0. Note that in this framework, general relativity corresponds to $\delta\ddot{\mathbf{r}}_i^{\mathrm{PPN}} \equiv 0$.

Finally, we present the similar expressions for the light-time equation, Eq. (6). This equation that can be written as $t_2 - t_1 = \Delta t_{12} = \Delta t_{12}^{\mathrm{GR}} + \delta\Delta t_{12}^{\mathrm{PPN}}$ with the PPN perturbation is given as below

$$\delta\Delta t_{12}^{\mathrm{PPN}} = \frac{\bar{\gamma}}{c^3}\sum_i \mu_i \ln\left[\frac{r_1^i + r_2^i + r_{12}^i}{r_1^i + r_2^i - r_{12}^i}\right] + \mathcal{O}(c^{-5}), \tag{12}$$

where, again, general relativity is characterized by $\delta\Delta t_{12}^{\mathrm{PPN}} = 0$.

The PPN expansion serves as a useful framework to test relativistic gravitation in the context of the gravitational experiments. The main properties of the PPN metric tensor given by Eq. (4) are well established and are widely used in modern astronomical practice [7]. For practical purposes one uses the metric to derive the equations of motion for gravitating bodies and light, namely Eqs. (10)–(12). These general relativistic equations are then used to produce numerical codes for the purposes of construction of the solar system's ephemerides, spacecraft orbit determination [22], and analysis of the gravitational experiments in the solar system [7, 23].

Equations (10)–(12) are used to focus the science objectives of gravitational experiments and also to interpret their results. So far, the general theory of relativity survived every test [5], yielding the ever-improving values for the set of the PPN parameters (γ, β), namely $\bar{\gamma} = (2.1 \pm 2.3) \times 10^{-5}$ using the data from the Cassini spacecraft taken during solar conjunction experiment [14] and $\bar{\beta} = (1.2 \pm 1.1) \times 10^{-4}$, which resulted from the analysis of the multiyear segments of LLR data [13].

4 Search for New Physics Beyond General Relativity

The fundamental physical laws of nature, as we know them today, are described by the Standard Model of particles and fields and the general theory of relativity [5, 27]. The Standard Model specifies the families of fermions (i.e., leptons and quarks) and their interactions by vector fields that transmit the strong, electromagnetic, and weak

forces. General relativity is a tensor-field theory of gravity with universal coupling to the particles and fields of the Standard Model.

However, despite the beauty and simplicity of general relativity and the success of the Standard Model, our present understanding of the fundamental laws of physics has several shortcomings. From a theoretical standpoint, the challenges appear from the strong gravitational field regime and include appearance of space-time singularities and breakdown of classical approach in describing the physics of very strong gravitational fields. It is believed that a way out of the difficulties above may be through gravity quantization. However, despite encouraging progress in string theory, the search for a realistic theory of quantum gravity remains a challenge. These continued difficulties with gravity quantization indicate that, perhaps, the pure tensor gravity of general relativity needs alteration. In addition, the recent progress in observational cosmology has subjected the general theory of relativity to increased scrutiny by suggesting a non-Einsteinian scenario of the universe's evolution. It is now believed that new physics is needed to resolve these issues.

The theoretical models of new physics that can solve the problems described above typically involve new interactions, some of which may manifest themselves as violations of the EP, variation of fundamental constants, modification of the gravitational inverse-square law at short distances, breaking of the Lorenz symmetry, or large-scale gravitational phenomena [5, 23, 28, 29]. Each of these manifestations offers an opportunity for space-based experiment and, hopefully, could lead to a major discovery.

Gravitation is the weakest of the four fundamental particle interactions in physics. Experiments designed to test gravitation fall into five categories [3]: those which attempt to prove the invariance of conventional constants; those designed to confirm the predictions of the special theory of relativity; tests of the EP involving the question of how various forms of energy contribute to mass; those attempting to determine the mathematical relationship between the curvature of space-time and mass energy; and those to detect gravitational radiation [5].

Below we discuss space-based gravitational experiments that aim to test the EP, search for variability of the fundamental constants, tests of the gravitational inverse-square law, and tests of alternative and modified gravity theories.

4.1 Tests of the Equivalence Principle

Since the early days of general relativity, there was a need to develop a framework to test the theory against other possible theories of gravity compatible with special relativity. To that extent Einstein's equivalence principle (EEP) was suggested. The EEP states that the result of a local nongravitational experiment in an inertial frame of reference is independent of the velocity or location in the universe of the experiment. This is a kind of Copernican extension of Einstein's original formulation, which requires that suitable frames of reference all over the universe behave identically. It is an extension of the postulates of special relativity in that it requires

that dimensionless physical values such as the fine-structure constant and electron-to-proton mass ratio be constant. From the theoretical standpoint, the EEP is at the foundation of general relativity; therefore, testing the principle is very important.

One can test the validity of both the EP and the field equations that determine the geometric structure created by a mass distribution. There are two different "flavors" of the EP, the weak and the strong forms (weak equivalence principle, WEP and strong equivalence principle, SEP, respectively), which are being tested in various experiments performed with laboratory test masses and with bodies of astronomical size [7, 12].

4.1.1 The Weak Equivalence Principle

The *weak form of the EP* (the WEP) assumes freely falling bodies are bound by non-gravitational forces only and also strong and electroweak interactions obey the EP. In this case the relevant test-body differences are their fractional nuclear-binding differences, their neutron-to-proton ratios, their atomic charges, etc. Furthermore, the equality of gravitational and inertial masses implies that different neutral massive test bodies have the same free-fall acceleration in an external gravitational field, and therefore in freely falling inertial frames, the external gravitational field appears only in the form of a tidal interaction.

General relativity and other metric theories of gravity assume that the WEP is exact. However, many extensions of the Standard Model that contain new macroscopic-range quantum fields predict quantum exchange forces that generically violate the WEP because they couple to generalized "charges" rather than to mass/energy as does gravity [30–32].

In a laboratory, precise tests of the EP can be made by comparing the free-fall accelerations, a_1 and a_2, of different test bodies. When the bodies are at the same distance from the source of the gravity, the expression for the EP takes the form

$$\frac{\Delta a}{a} = \frac{2(a_1 - a_2)}{a_1 + a_2} = \left[\frac{m_G}{m_I}\right]_1 - \left[\frac{m_G}{m_I}\right]_2 = \Delta\left[\frac{m_G}{m_I}\right], \tag{13}$$

where m_G and m_I are the gravitational and inertial masses of each body, respectively. The sensitivity of the EP test is determined by the precision of the differential acceleration measurement divided by the degree to which the test bodies differ (e.g., composition).

Various experiments have been performed to measure the ratios of gravitational to inertial masses of bodies. Recent experiments on bodies of laboratory dimensions have verified the WEP to a fractional precision of $\lesssim 1.4 \times 10^{-13}$ [33]. The accuracy of these experiments is high enough to confirm that the strong, weak, and electromagnetic interactions each contribute equally to the passive gravitational and inertial masses of the laboratory bodies.

Currently, the most accurate results in testing the WEP have been reported by ground-based laboratories [34]. The most recent result [33, 35] for the fractional differential acceleration between beryllium and titanium test bodies gave $\Delta a/a = (1.0 \pm 1.4) \times 10^{-13}$. A review of the most recent laboratory tests of gravity can be found in Ref. [36]. Significant improvements in the tests of the EP are expected from dedicated space-based experiments [5].

The composition independence of acceleration rates of various masses toward the Earth can be tested in space-based laboratories to a precision of many additional orders of magnitude, down to levels at which some models of the unified theory of quantum gravity, matter, and energy suggest a possible violation of the EP [30–32]. In some scalar–tensor theories, the strength of EP violations and the magnitude of the fifth force mediated by the scalar can be drastically larger in space than on the ground [37], providing further justification for space deployment. Importantly, many of these theories predict observable violations of the EP at various levels of accuracy ranging from 10^{-13} to 10^{-16}. Therefore, even a confirmation of no EP violation will be exceptionally valuable, as it will place useful constraints on the range of possibilities in the development of a unified physical theory.

Compared with Earth-based laboratories, experiments in space can benefit from a range of conditions, including free fall and significantly reduced contributions due to seismic, thermal, and other nongravitational noise [5]. As a result, many experiments have been proposed to test the EP in space.

The Micro-Satellite à traînée Compensée pour l'Observation du Principe d'Equivalence (MicroSCOPE) mission is a room-temperature EP experiment in space that utilizes electrostatic differential accelerometers [38]. The mission is currently under development and is scheduled for launch in 2016. The design goal is to achieve a differential acceleration accuracy of 10^{-15}.

The Galileo Galilei (GG) mission [39] is an Italian space experiment proposed to test the EP at room temperature with an accuracy of 1 part in 10^{17}. The key instrument of GG is a differential accelerometer made of weakly coupled coaxial, concentric test cylinders that spin rapidly around the symmetry axis and are sensitive in the plane perpendicular to it. GG is included in the National Aerospace Plan of the Italian Space Agency (ASI) for implementation in the near future.

The Satellite Test of Equivalence Principle (STEP) mission [40] is a proposed test of the EP to be conducted from a free-falling platform in space provided by a drag-free spacecraft orbiting the Earth. STEP will test the composition independence of gravitational acceleration for cryogenically controlled test masses by searching for a violation of the EP with a fractional acceleration accuracy of 1 part in 10^{18}. As such, this ambitious experiment will be able to test very precisely for the presence of any new nonmetric, long-range physical interactions.

The evidence of the validity of the WEP for laboratory bodies is incomplete for astronomical body scales. The experiments searching for WEP violations are conducted in laboratory environments that utilize test masses with negligible amounts of gravitational self-energy; therefore, a large-scale experiment is needed to test the postulated equality of gravitational self-energy contributions to the inertial and passive gravitational masses of the bodies. Once the self-gravity of the test bodies is

nonnegligible (which is currently true only for bodies of astronomical sizes), the corresponding experiment tests the ultimate version of the EP—the SEP.

4.1.2 The Strong Equivalence Principle

In its *strong form* (SEP) the EP is extended to cover the gravitational properties resulting from gravitational energy itself. It is an assumption about the way gravity acts on itself, that is, about the nonlinear property of gravitation. Although general relativity assumes that the SEP is exact, alternate metric theories of gravity—such as those involving scalar fields—typically violate the SEP. For the SEP case, the relevant test-body differences are the fractional contributions to their masses by gravitational self-energy. Because of the extreme weakness of gravity, SEP test bodies must have astronomical size.

The SEP states that the results of any local experiment, gravitational or not, in an inertial frame of reference are independent of where and when in the universe it is conducted. This is the only form of the EP that applies to self-gravitating objects (such as stars), which have substantial internal gravitational interactions. It requires that the gravitational constant be the same everywhere in the universe and is incompatible with a fifth force. It is much more restrictive than the EEP. General relativity is the only known theory of gravity compatible with this form of the EP.

Nordtvedt [41] suggested several solar system experiments for testing the SEP. In fact, the PPN formalism [7] describes the motion of celestial bodies in a theoretical framework common to a wide class of metric theories of gravity. To facilitate investigation of a possible violation of the SEP, Eq. (11) allows for a possible inequality of the gravitational and inertial masses, given by the parameter $[m_G/m_I]_i$, which in the PPN formalism is expressed as

$$\left[\frac{m_G}{m_I}\right]_{\text{SEP}} = 1 + \eta\left(\frac{E}{mc^2}\right), \tag{14}$$

where m is the mass of a body, E is the body's (negative) gravitational self-energy, mc^2 is its total mass energy, and η is a dimensionless constant for SEP violation. Any SEP violation is quantified by the parameter η: in fully conservative, Lorentz-invariant theories of gravity [4, 7], the SEP parameter is related to the PPN parameters by $\eta = 4\beta - \gamma - 3 \equiv 4\bar{\beta} - \bar{\gamma}$ (in general relativity, $\eta = 0$, see [4, 12]).

The quantity E is the body's i gravitational self-energy ($E < 0$), which is given by

$$\left[\frac{E}{mc^2}\right]_i = -\frac{G}{2m_i c^2}\int_i d^3x \rho_i U_i = -\frac{G}{2m_i c^2}\int_i d^3x d^3x' \frac{\rho_i(\mathbf{r})\rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{15}$$

For a uniform density sphere with a radius R, $E/mc^2 = -3Gm/5Rc^2 = -0.3v_E^2/c^2$, where v_E is the escape velocity. Accurate evaluation for solar system bodies requires numerical integration of the expression (15). Evaluating the standard solar model [42] results in $(E/mc^2)_\odot \sim -3.52 \times 10^{-6}$. Because gravitational

self-energy is proportional to m_i^2 and also because of the extreme weakness of gravity, the typical values for the ratio (E/mc^2) are $\sim\ 10^{-25}$ for bodies of laboratory size. Therefore, the experimental accuracy of 1 part in 10^{13} [33], which is so useful for the WEP, is not sufficient to test how gravitational self-energy contributes to the inertial and gravitational masses of small bodies. To test the SEP, one must consider planet-sized extended bodies, where the ratio in Eq. (15) is considerably higher.

Currently, the Earth–Moon–Sun system provides the best solar system arena for testing the SEP. LLR experiments involve reflecting laser beams off retroreflector arrays placed on the Moon by the Apollo astronauts and by two unmanned Soviet landers [12,13]. Solutions using LLR data give $(-0.8\pm1.3)\times10^{-13}$ for any possible inequality in the ratios of the gravitational and inertial masses for the Earth and Moon. This result, in combination with laboratory experiments on the WEP, yields an SEP test of $([m_G/m_I]_i - 1) = (-1.8 \pm 1.9) \times 10^{-13}$ that corresponds to the value of the SEP violation parameter of $\eta = (4.0 \pm 4.3) \times 10^{-4}$. In addition, using the recent Cassini result for the PPN parameter γ, PPN parameter β is determined at the level of $\bar{\beta} = (1.2 \pm 1.1) \times 10^{-4}$ (see [13,43] for details).

With the Apache Point Observatory Lunar Laser-ranging Operations (APOLLO) facility [44,45], LLR goes through a significant progress. APOLLO's 1-mm range precision will translate into order-of-magnitude accuracy improvements in tests of the WEP and the SEP (leading to accuracy at levels of $\Delta a/a \lesssim 1 \times 10^{-14}$ and $\eta \lesssim 2 \times 10^{-5}$ respectively), in the search for variability of Newton's gravitational constant (see Sect. 4.2), and in the test of the gravitational inverse-square law (see Sect. 4.3) on scales of the Earth–Moon distance (anticipated accuracy of 3×10^{-11}) [45].

One possible improvement would be an upgrade to the corner-cube retroreflector arrays currently on the moon [3]. There is a proposal to develop and deploy, on the lunar surface, new retroreflector instrument which is based on a single, hollow corner cube with a large aperture [46]. The new instrument will be able to reach an Earth–Moon range precision of 1 mm in a single pulse while being subject to significant thermal variations present on the lunar surface, and will have low mass to allow robotic deployment.

The next step in this direction is interplanetary laser ranging [47] to, for example, a lander on Phobos. Technology is available to conduct such measurements with a few-picosecond timing precision, which could translate into millimeter-class accuracies ranging between the Earth and Mars. The resulting Phobos laser ranging (PLR) experiment [48] could (a) test the strong form of the EP with accuracy at the 2×10^{-6} level, (b) measure the PPN parameter γ (see Sect. 4.4) with accuracy below the 10^{-6} level, and (c) test the gravitational inverse-square law at $\sim$2-AU distances with an accuracy of 1×10^{-14}, thereby greatly improving the accuracy of the current tests [47]. PLR could also advance research in several areas of science including remote-sensing geodesic and geophysical studies of Mars and Phobos.

Furthermore, with the recently demonstrated capabilities of reliable laser links over large distances in space [48], there is a strong possibility of improving the accuracy of gravity experiments with precision laser ranging over interplanetary

scales [48]. With PLR, the best venue for gravitational physics will be expanded to interplanetary distances, representing an upgrade in both the scale and the precision of this promising technique.

The experiments described above are examples of the rich opportunities offered by the fundamental physics community to explore the validity of the EP. These experiments could potentially offer a major improvement up to five orders of magnitude over the accuracy of the current tests of the EP. Such experiments would dramatically enhance the range of validity for one of the most important physical principles or they could lead to a spectacular discovery.

4.2 Search for Variability of the Gravitational Constant

Dirac's nearly 80-year-old idea that cosmological evolution may lead to variation of fundamental physical constants has been recently revisited. This was motivated by the advent of models unifying the forces of nature based on the symmetry properties of possible extra dimensions, such as the Kaluza–Klein-inspired theories, Brans–Dicke theory, and supersymmetry models. Alternative theories of gravity [4] and theories of modified gravity [3] include cosmologically evolving scalar fields that lead to variability of the fundamental constants. It has been hypothesized that a variation of the cosmological-scale factor with epoch could lead to temporal or spatial variation of the constants, especially the gravitational constant, G.

A possible variation of Newton's gravitational constant G could be related to the expansion of the universe depending on the cosmological model considered. Variability in G can be tested in space with a much greater precision than on Earth [13, 49]. For example, a decreasing gravitational constant, G, coupled with angular momentum conservation is expected to increase a planet's semimajor axis, a, as $\dot{a}/a = -\dot{G}/G$. The corresponding change in orbital phase grows quadratically with time, providing for strong sensitivity to the effect of $\dot{G}$.

Currently, experiments relying on lunar and planetary ranging measurements are the best means of searching for very small spatial or temporal gradients in the values of G [12, 13]. Thus, analysis of LLR data performed in 2004 strongly limited such variations and constrained a local scale (~ 1 AU) expansion of the solar system as $\dot{G}/G = (5 \pm 6) \times 10^{-13}$ year^{-1}, including the expansion resulting from cosmological effects [50]. Interestingly, the achieved accuracy in $\dot{G}/G$ implies that, if this rate is representative of our cosmic history, then G has changed by less than 1 % over the 13.4 Gyr age of the universe.

The ever-extending temporal span of the LLR data set and the increase in the accuracy of lunar ranging (i.e., APOLLO) could lead to improvements in the search for variability of Newton's gravitational constant; an accuracy at the level of $\dot{G}/G \sim 1 \times 10^{-14}$ year^{-1} is feasible with LLR [47]. High-accuracy timing measurements of binary and double pulsars could also provide a good test of the

variability of the gravitational constant [6, 15, 51]. A preliminary analysis of the accuracy achievable with PLR indicates that $\dot{G}/G$ could be determined with an accuracy at the level $\dot{G}/G \sim 3 \times 10^{-15}$ year^{-1}, limited by the asteroids and the lifetime of the experiment.

4.3 Testing the Gravitational Inverse-Square Law

Many modern theories of gravity, including the string, supersymmetry, and brane-world theories, suggest that new physical interactions will appear at short ranges. This may happen because at submillimeter distances new dimensions can exist, thereby changing the gravitational inverse-square law [52]. Similar forces that act at short distances are predicted in supersymmetric theories with weak-scale compactifications [53], in some theories with very low energy supersymmetry breaking [54], and also in theories of very low quantum gravity scale [55, 56]. These multiple predictions provide strong motivation for experiments that would test for possible deviations from Newton's gravitational inverse-square law at very short distances, notably on millimeter-to-micrometer ranges.

Most of the investigations to date were devoted to study the behavior of gravity at short distances. However, it is possible that small deviations from the inverse-square law occur at much larger distances. In fact, there is a possibility that non-compact extra dimensions could produce such deviations at astronomical distances [57], motivating new tests.

By far the most stringent constraints on a possible violation of the inverse-square law to date come from very precise measurements of the lunar orbit about the Earth [58]. Even though the lunar orbit has a mean radius of 384,000 km, the models agree with the data at the level of 2 mm. As a result, analysis of the LLR data tests the gravitational inverse-square law to 1.5×10^{-11} of the gravitational field strength at the scale of the Earth–Moon distance [45].

Additionally, the emerging interplanetary laser ranging could provide the conditions needed to improve the tests of the inverse-square law on interplanetary scales [47]. Thus, PLR could reach the accuracy of 1×10^{-14} at ~ 2 AU distances, thereby improving the current tests by several orders of magnitude.

We note that with the successful resolution of the puzzling Pioneer anomaly [59], there are other intriguing puzzles in the solar system dynamics still awaiting a proper explanation, notably the so-called fly-by anomaly [60] that occurred during Earth gravity assists performed by several interplanetary spacecraft.

4.4 Tests of Alternative and Modified Gravity Theories in Space

The challenge posed by the unexpected discovery of the accelerated expansion of the universe makes it very important to explore every option to explain and probe

the underlying physics. Theoretical efforts in this area offer a rich spectrum of new ideas, some discussed below, that can be tested by experiment.

Motivated by the dark-energy and dark-matter problems, long-distance gravity modification is one of the radical proposals that has recently gained attention [61]. Theories that modify gravity at cosmological distances exhibit a strong coupling of extra graviton polarizations [62]. This effect plays an important role in this class of theories in allowing them to agree with solar system constraints. In particular, the "brane-induced gravity" model [63] provides a new and interesting way of modifying gravity at large distances to produce an accelerated expansion of the universe, without the need for a nonvanishing cosmological constant [61,64]. One of the peculiarities of this model is by means of recovering the usual gravitational interaction at small (i.e., non-cosmological) distances, motivating precision tests of gravity on solar system scales [65].

The Eddington parameter γ is perhaps the most fundamental PPN parameter [4, 7], in that $\frac{1}{2}\bar{\gamma}$ is a measure, for example, of the fractional strength of the scalar–gravity interaction in scalar–tensor theories of gravity [66]. Currently, the most precise value for this parameter, $\bar{\gamma} = (2.1 \pm 2.3) \times 10^{-5}$, was obtained using radiometric tracking data received from the Cassini spacecraft [14]. This accuracy approaches the region where multiple tensor–scalar gravity models, consistent with the recent cosmological observations [21], predict a lower bound for the present value of this parameter at the level of $\bar{\gamma} \sim 10^{-6} \div 10^{-7}$ [30–32]. Therefore, improving the measurement of this parameter would provide crucial information to separate scalar–tensor theories of gravity from general relativity, probe possible ways for gravity quantization, and test modern theories of cosmological evolution.

As far as optical observations are concerned, future astrometric missions such as Gaia will push the accuracy to the level of a few microarcseconds, and the expected accuracy in the determination of γ will be 10^{-6} to 5×10^{-7} [67]. We note that this result will be obtained simply as a by-product of its astrometric campaign.

Interplanetary laser ranging could lead to a significant improvement in the accuracy of the parameter γ. Thus, precision ranging between the Earth and a lander on Phobos during solar conjunctions may offer a suitable opportunity (i.e., PLR). If the lander were equipped with a laser transponder capable of reaching a precision of 1 mm, a measurement of γ with accuracy of 2×10^{-7} would be possible [48]. In addition, a Mercury lander equipped with a laser ranging transponder would be interesting as it would probe a stronger gravity regime while providing measurements that will not be affected by the dynamical noise from the asteroids. If the technical challenges related to landing a spacecraft in polar regions of Mercury and relevant thermal design issues are resolved (see [48]), this would be a valuable test.

The Gravitational Time Delay Mission (GTDM) [68] proposes to use laser ranging between two drag-free spacecraft (with spurious acceleration levels below 1.3×10^{-13} m/s^2 Hz$^{-1/2}$ at 0.4 μHz) to accurately measure the Shapiro time delay for laser beams passing near the Sun. One spacecraft will be kept at the L1 Lagrange point of the Earth–Sun system, the other one will be placed on a 3:2 Earth-resonant, LATOR-type orbit (see below and [69] for details). A high-stability frequency standard ($\delta f/f \lesssim 1 \times 10^{-13}$ Hz$^{-1/2}$ at 0.4 μHz) located on the L1 spacecraft will

permit accurate measurement of the time delay. If requirements on the performance of the disturbance compensation system, the timing-transfer process, and the high-accuracy orbit determination are successfully addressed, then determination of the time delay of interplanetary signals to a 0.5-ps precision in terms of the instantaneous clock frequency could lead to an accuracy of 2×10^{-8} in measuring γ.

The Astrodynamical Space Test of Relativity using Optical Devices (ASTROD) [70, 71] is a mission concept with three spacecraft, one near the L1/L2 point, one with an inner solar orbit, and one with an outer solar orbit, ranging coherently with one another using lasers to test relativistic gravity, measure the solar system, and detect gravitational waves. ASTROD will place two spacecraft in separate solar orbits carrying a payload of a proof mass, two telescopes, two 1–2 W lasers, a clock, and a drag-free system, together with a similar L1/L2 spacecraft. The three spacecraft range coherently with one another using lasers to map the solar system gravity, test relativistic gravity, and detect gravitational waves. The three spacecraft, advanced drag-free systems, and mature laser interferometric ranging will be used and the resolution is sub-wavelength. The accuracy of measuring γ will depend on the stability of the lasers and clocks. Preliminary analysis of the missions' performance indicates that an uncertainty of 1×10^{-9} on the determination of γ is achievable.

The Beyond Einstein Advanced Coherent Optical Network (BEACON) [72] is an experiment designed to reach a sensitivity of 1 part in 10^9 in measuring the PPN parameter γ. The mission will place four small spacecraft in 80,000-km circular orbits around the Earth with all spacecraft in the same plane. Each spacecraft will be equipped with three sets of identical laser ranging transceivers, which will send laser metrology beams between the spacecraft to form a flexible light-trapezoid formation. In Euclidean geometry this system is redundant. By measuring only five of the six distances one can compute the sixth. To enable its primary science objective, BEACON will precisely measure and monitor all six inter-spacecraft distances within the trapezoid using transceivers capable of reaching an accuracy of ~ 0.1 nm in measuring these distances. The resulting geometric redundancy is the key element that enables BEACON's superior sensitivity in measuring a departure from Euclidean geometry. In the Earth's vicinity, this departure is primarily due to the curvature of the relativistic space-time. It amounts to ~ 10 cm for laser beams just grazing the surface of the Earth, then falls off inversely proportional to the impact parameter. The simultaneous analysis of the resulting time series of these distance measurements will allow BEACON to measure the curvature of the space-time around the Earth with an accuracy of better than 1 part in 10^9 [72].

The Laser Astrometric Test of Relativity (LATOR) [69] proposes to measure the parameter γ with an accuracy of 1 part in 10^9, which is a factor of 30,000 beyond the best currently available, Cassini's 2003 result [14]. The key element of LATOR is a geometric redundancy provided by the long-baseline optical interferometry and interplanetary laser ranging. By using a combination of independent time series of gravitational deflection of light in immediate proximity to the Sun, along with measurements of the Shapiro time delay on interplanetary scales (to a precision better than 0.01 prad and 3 mm, respectively), LATOR will significantly improve our

knowledge of relativistic gravity and cosmology. LATOR's primary measurement, the precise observation of the non-Euclidean geometry of a light triangle that surrounds the Sun, pushes the search for cosmologically relevant scalar–tensor theories of gravity to unprecedented accuracy by looking for a remnant scalar field in today's solar system. LATOR could lead to very robust advances in the tests of fundamental physics. It could discover a violation or extension of general relativity or reveal the presence of an additional long-range interaction.

As we see, there is an impressive list of modern experiments that are designed to test fundamental gravitation to a remarkable precision. The science is clearly important [3] and the needed technology is ready with some of its components already being space-qualified [5]. It is likely that some of these missions could be realized in the near future; if that happens, we will have an important opportunity either to verify the theory in a different setting or to make a profound discovery.

5 Discussion and Outlook

For many years since its discovery in 1916, Einstein's general theory of relativity has been considered as a purely mathematical structure rather than as a physical theory. This was due partly to an insufficient recognition of the deep physical significance of the nongeometrical nature of the right-hand side of the Einstein's equations [18], and partly to the lack of experimental or observational challenges to the theory.

The situation has changed drastically in the past 40 years. From the conceptual point of view, it has been realized through the work of many people that the elegant, geometrical nature of the left-hand side of Einstein's equations followed, as a necessary consequence, from the physical postulate that the source of gravity be the energy–momentum tensor. From the empirical point of view, starting in the 1960s, the implementation of high-precision gravity tests, and the discovery of quasars and pulsars motivated one to tackle, in detail, the deep physical implications of general relativity. As we know it today, the theory has successfully passed all the challenges.

Although general theory of relativity had been successfully and extensively tested in various configurations, we, nevertheless, should continue testing this fundamental physical theory to the highest precision available. This is simply because this theory is one of the essential pillars of the framework of modern physics. Second, some very crucial qualitative features of this theory have not yet been verified; in particular, the direct detection of gravitational waves. Also, accurate tests in the strong-gravity regime are still missing.

Presently, there is well-supported expectation of new physics. In particular, the remarkable recent progress in observational cosmology has subjected the general theory of relativity to increased scrutiny by suggesting a non-Einsteinian scenario of the universe's evolution. From a theoretical standpoint, the challenge is even stronger—if gravity is to be quantized, general relativity must be modified! Furthermore, recent advances in the scalar–tensor extensions of gravity, and brane-world gravitational models, along with efforts to modify gravity on large scales, predict

the existence of very small deviations from general relativity on various scales, including on the spacecraft-accessible distances in the solar system. These theoretical advances motivate searches for deviations from Einstein's theory, at the level of three to five orders of magnitude below the level currently tested by experiment.

Space itself, deep-space probes, and orbiting space platforms are all components of a new laboratory for gravitation experiments and their use offers exciting prospects for the future. As far as modern tests of gravitation are concerned, space deployment offers access to unique conditions that are beneficial to gravitational experiments and are not readily available on Earth, but are important for many highly accurate investigations exploring the limits of modern physics. There are ongoing efforts to develop a highly accurate framework to benefit the new generation of the gravitational tests [25, 73, 74] that would replace the PPN formalism in the case of highly accurate measurements [75]. The work is nearing completion, its results will soon be available and are expected to be useful for many areas related to spacecraft navigation, precision metrology, and time transfer.

We presented the current empirical status of the general theory of relativity based on the results of solar system experiments and discussed a number of recently proposed advanced tests of this fundamental theory. Concluding, we emphasize that physics is an empirical science, so is gravitation. The road map for gravitation is clearly empirical. As the precision of the fundamental tests will be increased by orders of magnitude, we will be in a position to explore deeper into the origin of gravitation and, thus, into the evolution and the fate of the universe. The current and coming generations of space-based experiments are holding such promises.

Acknowledgements The work described here was carried out at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, under a contract with the National Aeronautics and Space Administration.

References

1. Einstein A. Die Feldgleichungen der Gravitation (The field equations of gravitation). Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin; 1915. p. 844–7.
2. Einstein A. Die Grundlage der allgemeinen Relativitätstheorie. Ann Physik. 1916;49:146.
3. Turyshev SG. Experimental tests of general relativity: Recent progress and future directions. Usp Fiz Nauk. 2009;179:3–34.
4. Will CM. The confrontation between general relativity and experiment. Living Rev Relativ. 2006;9(3).
5. Turyshev SG, Israelsson UE, Shao M, Yu N, Kusenko A, Wright EL, Everitt CWF, Kasevich M, Lipa JA, Mester JC, et al. Space-based research in fundamental physics and quantum technologies. Int J Mod Phys D. 2007;16(12a):1879–1925.
6. Possenti A, Burghay B. 35 years of testing relativistic gravity: where do we go from here? In: Peron R, Gorini V, Moschella U, editors. Astrophysics, clocks and fundamental constants. Berlin: Springer; 2013.
7. Will CM. Theory and experiment in gravitational physics. Cambridge: Cambridge University Press; 1993.

8. Reasenberg RD, Shapiro II, MacNeil PE, Goldstein RB, Breidenthal JC, Brenkle JP, et al. Viking relativity experiment—verification of signal retardation by solar gravity. Astrophys J Lett. 1979;234:L219–21.
9. Anderson JD, Lau EL, Turyshev SG, Williams JG, Nieto MM. Recent results for solar-system tests of general relativity. BAAS. 2002;34:833.
10. Pitjeva EV. High–precision ephemerides of planets—EPM and determination of some astronomical constants. Sol Syst Res. 2005;39(3):176–86.
11. Fomalont E, Kopeikin S, Lanyi G, Benson J. Progress in measurements of the gravitational bending of radio Waves Using the VLBA. ApJ. 2009;699(2):1395–1402.
12. Williams JG, Turyshev SG, Boggs DH. Lunar laser ranging tests of the equivalence principle with the earth and moon. Int J Mod Phys D. 2009;18:1129–75.
13. Williams JG, Turyshev SG, Boggs DH. Progress in lunar laser ranging tests of relativistic gravity. Phys Rev Lett. 2004;93:261101.
14. Bertotti B, Iess L, Tortora P. A test of general relativity using radio links with the Cassini spacecraft. Nature. 2003;425:374.
15. Kramer M, Stairs IH, Manchester RN, McLaughlin MA, Lyne AG, Ferdman RD, Burgay M, Lorimer DR, et al. Tests of general relativity from timing the double pulsar. Science. 2006;314:97–102.
16. Fock VA. The theory of space, time and gravitation. Moscow: Fizmatgiz; 1959. (English translation (1959), Pergamon, Oxford).
17. Landau LD, Lifshitz EM. The classical theory of fields. 7th edn. Moscow: Nauka; 1988.
18. Misner CW, Thorne KS, Wheeler JA. Gravitation. W. H. Freeman and Co.; 1973.
19. Yao WM, et al. Review of particle physics. J Phys G. 2006;33:205–9.
20. Weinberg S. The cosmological constant problem. Rev Mod Phys. 1989;61(1):1–23.
21. Spergel DN, Bean R, Doré O, Nolta MR, Bennett CL, Hinshaw G, Jarosik N, Komatsu E, Page L, Peiris HV, Verde L, et al. Wilkinson Microwave Anisotropy Probe (WMAP) three year results: implications for cosmology. Astrophys J Suppl. 2007;170:377.
22. Moyer TD. Formulation for observed and computed values of deep space network data types for navigation. JPL deep-space communications and navigation series. Wiley: Hoboken; 2003.
23. Turyshev SG. Experimental tests of general relativity. Ann Rev Nucl Part Sci. 2008;58:207.
24. Einstein A, Infeld L, Hoffmann B. The gravitational equations and the problem of motion. Ann Math. 1938;39:65.
25. Turyshev SG, Toth VT, Sazhin MV. General relativistic observables of the GRAIL mission. Phys Rev D. 2013;87:024020.
26. Brans C, Dicke RH. Mach's principle and a relativistic theory of gravitation. Phys Rev. 1961;124(3):925.
27. Bertolami O, Paramos J, Turyshev SG. In: Dittus H, Lämmerzahl C, Turyshev S, editors. Lasers, clocks and drag-free control: exploration of relativistic gravity in space. Berlin: Springer; 2007. p. 27–67.
28. Liberati S, Mattingly D. Lorentz breaking effective field theory models for matter and gravity: theory and observational constraints. In: Peron R, Gorini V, Moschella U, editors. Gravity: where do we stand? Berlin: Springer; 2016, p. 367 (Chap. 11, this volume).
29. Braccini S, Fidecaro F. The detection of gravitational waves. In: Peron R, Gorini V, Moschella U, editors. Gravity: where do we stand? Berlin: Springer; 2016, p. 235 (Chap. 7, this volume).
30. Damour T, Nordtvedt K. General relativity as a cosmological attractor of tensor-scalar theories. Phys Rev Lett. 1993;70:2217.
31. Damour T, Polyakov AM. String theory and gravity. Gen Rel Grav. 1994;26:1171.
32. Damour T, Piazza F, Veneziano G. Runaway dilaton and equivalence principle violations. Phys Rev Lett. 2002;89:081601.
33. Adelberger E. New tests of Einstein's equivalence principle and Newton's inverse-square law. Class Quant Grav. 2001;18:2397–2405.
34. Baeßler S, Heckel BR, Adelberger EG, Gundlach JH, Schmidt U, Swanson HE. Improved test of the equivalence principle for gravitational self-energy. Phys Rev Lett. 1999;83(18):3585.

35. Schlamminger S, Choi KY, Wagner TA, Gundlach JH, Adelberger EG. Test of the equivalence principle using a rotating torsion balance. Phys Rev Lett. 2008;100:041101.
36. Gundlach JH, Schlamminger S, Spitzer CD, Choi KY, Woodahl BA, Coy JJ, Fischbach E. Laboratory test of Newton's second law for small accelerations. Phys Rev Lett. 2007;98(15):150801.
37. Mota DF, Barrow JD. Local and global variations of the fine-structure constant. Mon Not Roy Astron Soc. 2004;349:291.
38. Touboul P, Rodrigues M. The MICROSCOPE space mission. Class Quant Grav. 2001;18:2487–98.
39. Nobili A, Shao M, Pegna R, Zavattini G, Turyshev S, et al. 'Galileo Galilei' (GG): space test of the weak equivalence principle to 10^{-17} and laboratory demonstrations. Class Quant Grav. 2012;29:184011.
40. Mester J, Torii R, Worden P, Lockerbie N, Vitale S, Everitt CWF. The STEP mission: principles and baseline design. Class Quant Grav. 2001;18:2475–86.
41. Nordtvedt K. Equivalence principle for massive bodies. I. Phenomenology. Phys Rev. 1968;169:1014–6.
42. Ulrich RK. The influence of partial ionization and scattering states on the solar interior structure. Astrophys J. 1982;258:404.
43. Williams JG, Turyshev SG, Boggs D. Lunar laser ranging tests of the equivalence principle. Class Quant Grav. 2012;29:184004.
44. Murphy TW, Michelson EL, Orin AE, Adelberger EG, Hoyle CD, Swanson HE, Stubbs CW, Battat JE. APOLLO: next-generation lunar laser ranging. Int J Mod Phys D. 2007;16:2127.
45. Williams JG, Turyshev SG, Murphy TW. Improving LLR tests of gravitational theory. Int J Mod Phys D. 2004;13:567–82.
46. Turyshev SG, Williams JG, Folkner WM, Gutt GM, et al. Corner-cube retro-reflector instrument for advanced lunar laser ranging. Exp Astron. 2012. doi:10.1007/s10686-012-9324-z, arXiv:1210.7857 [physics.ins-det].
47. Turyshev SG, Williams JG. Space-based tests of gravity with laser ranging. Int J Mod Phys D. 2007;16:2165–79.
48. Turyshev SG, Farr W, Folkner WM, Girerd AR, et al. Advancing tests of relativistic gravity via laser ranging to phobos. Exp Astron. 2010;28:209–49.
49. Uzan JP. The fundamental constants and their variation: observational status and theoretical motivations. Rev Mod Phys. 2003;75:403.
50. Williams JG, Turyshev SG, Boggs DH. Williams et al. Reply (to the comment by Dumin on Progress in Lunar Laser Ranging Tests of Relativistic Gravity). Phys Rev Lett. 2007;98:059002.
51. Nordtvedt K. Space-time variation of physical constants and the equivalence principle. Int J Mod Phys A. 2002;17:2711–5.
52. Adelberger EG, Heckel BR, Nelson AE. Tests of the gravitational inverse-square law. Ann Rev Nucl Part Sci. 2003;53:77–121.
53. Antoniadis I, Dimopoulos S, Dvali GR. Millimeter range forces in superstring theories with weak-scale compactification. Nucl Phys B. 1998;516:70–82.
54. Dimopoulos S, Giudice GF. Macroscopic forces from supersymmetry. Phys Lett B. 1996;379:105–14.
55. Sundrum R. Towards an effective particle-string resolution of the cosmological constant problem. JHEP. 1999;07:001.
56. Dvali G, Gabadadze G, Kolanovic M, Nitti F. Scales of gravity. Phys Rev D. 2002;65:024031.
57. Dvali G, Gruzinov A, Zaldarriaga M. The accelerated universe and the Moon. Phys Rev D. 2003;68:024012.
58. Peron R. Fundamental physics with the LAGEOS satellites. In: Peron R, Gorini V, Moschella U, editors. Gravity: where do we stand? Berlin: Springer; 2016, p.167 (Chap. 4, this volume).
59. Turyshev SG, Toth VT, Kinsella G, Lee SC, et al. Support for the thermal origin of the Pioneer anomaly. Phys Rev Lett. 2012;108:241101.
60. Turyshev SG, Toth VT. The Puzzle of the flyby anomaly. Space Sci Rev. 2010;148:169–74.

61. Deffayet C, Dvali GR, Gabadadze G. Accelerated universe from gravity leaking to extra dimensions. Phys Rev D. 2002;65:044023.
62. Dvali G. Predictive power of strong coupling in theories with large distance modified gravity. New J Phys. 2006;8:326.
63. Dvali G, Gabadadze G, Porrati M. 4D gravity on a brane in 5D Minkowski space. Phys Lett B. 2000;485:208–14.
64. Deffayet C. Cosmology on a brane in Minkowski bulk. Phys Lett B. 2001;502:199–208.
65. Magueijo J, Bekenstein J. Testing strong MOND behavior in the solar system. Int J Mod Phys D. 2007;16:2035–53.
66. Damour T, Nordtvedt K. Tensor–scalar cosmological models and their relaxation toward general relativity. Phys Rev D. 1993;48:3436–50.
67. Turyshev SG. Relativistic gravitational deflection of light and its impact on the modeling accuracy for the Space Interferometry Mission. Astron Lett. 2009;35:215–34.
68. Ashby N, Bender P. Measurement of the shapiro time delay between drag-free spacecraft. In: Dittus H, Lämmerzahl C, Turyshev SG, editors. Lasers, clocks, and drag-free: technologies for future exploration in space and tests of gravity. Berlin: Springer; 2006. p. 219.
69. Turyshev SG, Shao M. Laser astrometric test of relativity: science, technology, and mission design. Int J Mod Phys D. 2007;16:2191–203.
70. Ni WT, Shiomi S, Liao AC. ASTROD, ASTROD I and their gravitational-wave sensitivities. Class Quant Grav. 2004;21(5):S641.
71. Appourchaux T, Burston R, Chen Y, Cruise M, Dittus H, et al. Astrodynamical Space Test of Relativity using Optical Devices I (ASTROD I)—A class-M fundamental physics mission proposal for Cosmic Vision 2015–2025. Exp Astron. 2009;23(2):491–527.
72. Turyshev SG, Lane B, Shao M, Girerd AR. A search for new physics with the BEACON mission. Int J Mod Phys D. 2009;18:1025–38.
73. Turyshev SG, Minazzoli OL, Toth VT. Accelerating relativistic reference frames in Minkowski space-time. J Math Phys. 2012;53(1):032501.
74. Turyshev SG, Toth VT. New perturbative method for solving the gravitational N-body problem in general relativity. 2013. arXiv:1304.8122 [gr-qc].
75. Turyshev SG, Sazhin MV, Toth VT. General relativistic laser interferometric observables of the GRACE-Follow-On mission. Phys Rev D. 2014;D89:105029. doi:10.1103/PhysRevD.89.105029.

Part II
Astrophysics

The Detection of Gravitational Waves

Stefano Braccini and Francesco Fidecaro

Abstract The detection of gravitational waves is challenging researchers since half a century. The relative precision required, 10^{-21}, is difficult to imagine, this is 10^{-5} the diameter of a proton over several kilometres, using masses of tens of kilograms, or picometres over millions of kilometres. A theoretical description of gravitational radiation and its effects on matter, all consequence of the general theory of relativity, is given. Then the astrophysical phenomena that are candidates of gravitational wave emission are discussed, considering also amplitudes and rates. The binary neutron star system PSR1913+16, which provided the first evidence for energy loss by gravitational radiation in 1975, is briefly discussed. Then comes a description of the experimental developments, starting with ground-based interferometers, their working principles and their most important sources of noise. The earth-wide network that is being built describes how these instruments will be used in the observation era. Several other detection techniques, such as space interferometry, pulsar timing arrays and resonant detectors, covering different bands of the gravitational wave frequency spectrum complete these lectures.

1 Foreword

Detecting gravitational waves (GW) is challenging human ingenuity since several decades. Stefano Braccini had taken it up putting into it his energies and passion. These lectures were given by Stefano in beautiful Villa Olmo, where he transmitted all his enthusiasm for the construction of knowledge through scientific discovery. These lectures describe where we stand now, with Stefano's conviction that the next steps in instrumentation will record the first signals transmitted gravitationally. Through the work going into building and commissioning the new instruments and achieving GW detection, we wish to honour his memory.

F. Fidecaro (✉)
"Enrico Fermi" Physics Dept., University of Pisa and INFN, L.go Bruno Pontecorvo 3, 56127 Pisa, Italy
e-mail: francesco.fidecaro@unipi.it

S. Braccini
INFN-Sezione di Pisa, L.go Bruno Pontecorvo 3, 56127 Pisa, Italy

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_7

2 Gravitational Waves: What Are They?

The general theory of relativity [1–3] has proven to be highly successful in describing gravitation as a geometric phenomenon involving the structure of space-time [4]. The Einstein equivalence principle suggests that experimentalists can measure curvature using uncharged test masses or photons that follow geodesic lines. Light deflection, radar echo delay and Mercury precession measure the amount of curvature generated by the Sun, finding agreement of general relativity with experiment at the ‰ level. Here the gravitational field is static, while general relativity allows curvature to vary with time and to propagate at the speed of light. These travelling variations of curvature are called GW and are necessary to signal at a distance where gravity changes. These are generated by the motion of massive objects or, more generally said, by changes of the stress–energy tensor [5, 6]. Curvature is usually very small, but its effects could be detected in the laboratory by Cavendish. Curvature changes are even smaller and one has to turn to astrophysical processes with solar mass compact objects to generate sizeable GW.

How GW arise from general relativity and which are the most promising sources is covered in the first part. Then various approaches to achieve GW detection are discussed with attention to experimental challenges encountered.

2.1 Einstein's Equations

The theory Einstein published in 1916 was the result of a long work that required facing many conceptual problems. Differential geometry concepts had to be acquired, among which the Riemann tensor $R^{\alpha}{}_{\beta\mu\nu}$ that characterizes curvature. This tensor describes the change of a vector when parallelly transported along a closed infinitesimal path; it has 20 independent components. Derived from it by contraction over two indices is the Ricci tensor,

$$R_{\alpha\beta} = R^{\mu}_{\alpha\mu\beta} \tag{1}$$

that carries most of the curvature information. The Einstein tensor is then obtained by making $R^{\alpha\beta}$ traceless

$$G^{\alpha\beta} = R^{\alpha\beta} - \frac{1}{2} g^{\alpha\beta} R^{\mu}{}_{\mu}, \tag{2}$$

to be part of the left-hand side of Einstein's equations:

$$G_{\alpha\beta} + g_{\alpha\beta}\Lambda = \frac{8\pi G}{c^4} T_{\alpha\beta}. \tag{3}$$

$T^{\alpha\beta}$ is the stress–energy tensor, and Λ is called the cosmological constant.

These equations are highly nonlinear because of the presence of the metric in the equations, metric that is in turn obtained from the solution for the Einstein tensor of these equations. Curvature from gravitation generates additional curvature, that is gravitation.

2.2 Weak-Field Approximation

Flat space-time is described by the Minkowski metric

$$\begin{aligned} \mathrm{d}s^2 &= c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \\ &= \eta_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu. \end{aligned} \tag{4}$$

Tiny deviations from flat space-time can be described by setting

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{5}$$

where $h_{\mu\nu} \ll 1$ in some coordinate system. A very useful feature of $h_{\mu\nu}$ can be noted imagining to apply a "background Lorentz transformation"

$$x^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\beta} x^\beta, \tag{6}$$

where $\Lambda^{\bar{\alpha}}{}_{\beta}$ coincides with a Lorentz transformation of special relativity, that is with constant coefficients. The metric tensor transforms as

$$\begin{aligned} g_{\bar{\alpha}\bar{\beta}} &= \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}}\, g_{\mu\nu} \\ &= \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}}\, \eta_{\mu\nu} + \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}}\, h_{\mu\nu}. \end{aligned} \tag{7}$$

But $\eta_{\mu\nu}$ is invariant under Lorentz transformations, so

$$g_{\bar{\alpha}\bar{\beta}} = \eta_{\bar{\alpha}\bar{\beta}} + h_{\bar{\alpha}\bar{\beta}} \tag{8}$$

having defined

$$h_{\bar{\alpha}\bar{\beta}} = \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}} h_{\mu\nu}. \tag{9}$$

One can think in terms of a flat space-time with a Lorentz "tensor" defined on it. Other fields depending on $g_{\mu\nu}$, like $R_{\mu\nu\alpha\beta}$ appear as deviations from their flat space-time form:

$$R_{\alpha\beta\mu\nu} = \frac{1}{2}\left[h_{\alpha\nu,\beta\mu} + h_{\beta\mu,\alpha\nu} + h_{\alpha\mu,\beta\nu} - h_{\beta\nu,\alpha\mu}\right]. \tag{10}$$

Under gauge transformations

$$x^{\alpha'} = x^\alpha + \xi^\alpha(x^\beta), \quad |\xi^\alpha, \beta| \ll 1 \tag{11}$$

the Riemann tensor is invariant under gauge transformations, neglecting second-order terms. Gauge transformations identify classes of equivalent coordinates of curved space-time.

Following the idea that $h_{\mu\nu}$ is a special relativity tensor on a flat background one can set up some rules for index manipulation, making use of the following definitions:

$$h^{\mu}{}_{\beta} := \eta^{\mu\alpha} h_{\alpha\beta} \tag{12}$$

$$h^{\mu\nu} := \eta^{\nu\beta} h^{\mu}{}_{\beta} \tag{13}$$

$$h := h^{\alpha}{}_{\alpha} \tag{14}$$

$$\bar{h}^{\alpha\beta} := h^{\alpha\beta} - \frac{1}{2}\eta^{\alpha\beta}h. \tag{15}$$

In this framework, the Einstein tensor becomes

$$G_{\alpha\beta} = -\frac{1}{2}\left[\bar{h}_{\alpha\beta,\mu}{}^{,\mu} + \eta_{\alpha\beta}\bar{h}_{\mu\nu}{}^{,\mu\nu} - \bar{h}_{\alpha\mu,\beta}{}^{,\mu} - \bar{h}_{\beta\mu,\alpha}{}^{,\mu} + O(h^2_{\alpha\beta})\right]. \tag{16}$$

The form of $G_{\alpha\beta}$ can be simplified performing a gauge transformation such that

$$h^{\mu\nu}{}_{,\nu} = 0 \tag{17}$$

as is done in electromagnetism. With this choice of gauge, the weak-field Einstein equations simplify to

$$\bar{h}^{\mu\nu,\alpha}{}_{,\alpha} = \Box\bar{h}^{\mu\nu} = -\frac{16\pi G}{c^4}T^{\mu\nu}. \tag{18}$$

This is a wave equation with the stress–energy tensor $T^{\mu\nu}$ as a source term.

Consider the static case of a point-like source with mass M at rest and spherical symmetry. The equation with nonzero right-hand side is

$$\triangle\bar{h}^{00} = -\frac{16\pi GM}{c^2}\delta(\mathbf{r}), \tag{19}$$

where

$$\triangle := \partial^i\partial_i \tag{20}$$

is the Laplacian operator. Comparing Eq. 19 with the equation

$$\triangle\phi = 4\pi GM \tag{21}$$

that determines the gravitational potential of Newtonian gravitation, one sees that

$$\bar{h}^{00} = -4\phi/c^2, \tag{22}$$

while $\bar{h}^{0i} = \bar{h}^{i0} = \bar{h}^{ij} = 0$. Going back to $h_{\mu\nu}$

$$h^{00} = -2\phi/c^2, h^{ii} = 2\phi/c^2, \tag{23}$$

which allows to write the metric, shown here for the spherically symmetric case

$$\mathrm{d}s^2 = (1+2\phi)c^2\mathrm{d}t^2 - (1-2\phi)\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2). \tag{24}$$

By using the geodesic equation, one obtains Newton's equation of motion as classical limit of general relativity

$$x^i{}_{;00} = -\phi^{,i}. \tag{25}$$

These equations show mass playing the role of charge in electromagnetism. The comparison can be pursued further by considering masses in motion. There will be a "current" of matter and one expects "gravitomagnetism". The Lense–Thirring effect [7, 8] recently observed [9, 10] gravitomagnetism, which is generated by Earth's rotation, dragging space-time around it.

2.3 Plane Waves

In vacuum, Eq. 18 admits as a possible solution for plane waves. The following form will be assumed:

$$\bar{h}^{\alpha\beta} = A^{\alpha\beta} \exp(ik_\mu x^\mu), \tag{26}$$

so that

$$\bar{h}^{\alpha\beta}{}_{,\mu} = ik_\mu \bar{h}^{\alpha\beta}. \tag{27}$$

The wave equation becomes

$$\bar{h}^{\alpha\beta,\mu}{}_{,\mu} = \eta^{\mu\nu}\bar{h}^{\alpha\beta}{}_{,\mu\nu} = -\eta^{\mu\nu}k_\mu k_\nu \bar{h}^{\alpha\beta} = 0, \tag{28}$$

from which one has

$$\eta^{\mu\nu}k_\mu k_\nu = k^\nu k_\nu = \mathrm{c}^2\omega^2 - \mathbf{k}\cdot\mathbf{k} = 0. \tag{29}$$

In other words, planes of constant gravitational wave phase move at the speed of light. The group velocity v_g

$$v_g = \frac{\partial\omega}{\partial k} \tag{30}$$

is also equal to c, without any dispersion.

A further condition derived from the gauge condition of Eq. 17 that was imposed to simplify the Einstein equations implies

$$A^{\alpha\beta}k_\beta = 0, \tag{31}$$

that is $A^{\alpha\beta}$ is orthogonal to k_β.

2.4 TT Gauge and Solution

As mentioned before, the weak-field Einstein equations are simplified by applying a gauge transformation whose generator $\xi_\alpha(x^\mu)$ has to satisfy a wave equation

$$\Box\xi_\alpha = 0. \tag{32}$$

It is a good idea to select a solution that has the same wave number k_μ as the gravitational plane wave

$$\xi_\alpha = B_\alpha \exp(ik_\mu x^\mu). \tag{33}$$

The change in $\bar{h}^{\alpha\beta}$ under a gauge transformation is such that

$$\bar{h}_{\alpha\beta} \to \bar{h}_{\alpha\beta} - \xi_{\alpha,\beta} - \xi_{\beta,\alpha} + \eta_{\mu\nu}\xi^\alpha{}_{,\alpha}, \tag{34}$$

so that $A_{\alpha\beta}$ transforms as

$$A_{\alpha\beta} \to A_{\alpha\beta} - iB_\alpha k_\beta - iB_\beta k_\alpha + i\eta_{\alpha\beta}B^\mu k_\mu. \tag{35}$$

The freedom given by B_α can be used to impose

$$A^\alpha{}_\alpha = 0, \quad A_{\alpha\beta}U^\beta = 0, \tag{36}$$

where U^β is some fixed four-velocity. Applying the first constraint to the transformed $A_{\alpha\beta}$

$$\begin{aligned} A^\alpha{}_\alpha - 2iB^\alpha k_\alpha + i\eta^\alpha{}_\alpha B^\mu k_\mu = \\ A^\alpha{}_\alpha + 2iB^\alpha k_\alpha = 0, \end{aligned} \tag{37}$$

while for the second constraint

$$\begin{aligned} A_{\alpha\beta}U^\beta - iB_\alpha k_\beta U^\beta - iB_\beta k_\alpha U^\beta + i\eta_{\alpha\beta}U^\beta B^\mu k_\mu = \\ A_{\alpha\beta}U^\beta - iB_\alpha k_\beta U^\beta - iB_\beta k_\alpha U^\beta + iU_\beta B^\mu k_\mu = 0. \end{aligned} \tag{38}$$

If one chooses a reference frame such that U^β becomes the time basis vector, or the four-velocity of a particle in its rest frame, then in that frame $A_{\alpha 0} = 0$. By further choosing the orientation of axes so that k_μ represents a wave traveling in the z-direction, then $A_{\alpha z} = 0$. One obtains in this coordinate system called "transverse-traceless (TT)"

$$\bar{h}^\alpha{}_\alpha = 0 \tag{39}$$

and

$$\bar{h}^{TT}_{\alpha\beta} = h^{TT}_{\alpha\beta}. \tag{40}$$

Using the trace condition

$$A^{TT_{\alpha\beta}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & A_+ & A_\times & 0 \\ 0 & A_\times & -A_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{41}$$

which contains only two independent quantities, A_+ and $A_\times$. These describe a plane wave with wave number k_μ.

2.5 Polarization of Gravitational Waves

The above result can be interpreted in terms of polarization by studying how to detect the passage of a gravitational wave. First, consider a particle moving with four-velocity U^α along a geodesic determined by a flat metric deformed by a gravitational wave. The geodesic equation requires

$$\frac{\mathrm{d}}{\mathrm{d}\tau}U^\alpha + \Gamma^\alpha{}_{\mu\nu}U^\mu U^\nu = 0, \tag{42}$$

where $\Gamma^{\alpha}{}_{\mu\nu}$ is the Cristoffel symbols used in covariant derivation. For a particle initially at rest, the initial value of its acceleration is

$$\frac{\mathrm{d}}{\mathrm{d}\tau}U^{\alpha} + \Gamma^{\alpha}{}_{00} = -\frac{1}{2}\eta^{\alpha\beta}(h_{\beta 0,0} + h_{0\beta,0} - h_{00,\beta}) = 0, \tag{43}$$

that is the particle, in that reference frame, has constant space coordinates or, in other words, stays at rest.

To detect the passage of a gravitational wave, one needs to measure changes in curvature. The simplest conceptual (and practical) case is to use two masses in free fall and initially at rest to mark two geodesics. Variations of invariant distance as function of the proper time of one of the masses measure curvature variations. This measurement can be performed by sending a light ray from one mass to the other and return measuring the round trip time. Consider a gravitational wave with $h_{+} \neq 0$ and $h_{\times} = 0$ and two masses with coordinates (0,0,0) and (L,0,0), that is on the x-axis. The metric element is

$$\mathrm{d}s^2 = \mathrm{d}(\mathrm{c}t)^2 - (1 + h_{+})\mathrm{d}x^2 - (1 - h_{+})\mathrm{d}y^2 - \mathrm{d}z^2. \tag{44}$$

For a light beam that propagates along the x-axis, one has $\mathrm{d}s^2 = 0$ or

$$\mathrm{d}(\mathrm{c}t)^2 = (1 + h_{+}(t))\mathrm{d}x^2, \tag{45}$$

where $h_{+}(t)$ is computed at constant z and does not depend on x or y for a plane wave. One has

$$\mathrm{d}(\mathrm{c}t)^2 = (1 + h_{+}(t - x/\mathrm{c}))\mathrm{d}x^2 \tag{46}$$

$$\mathrm{d}(\mathrm{c}t)^2 = (1 + h_{+}(t + x/\mathrm{c}))\mathrm{d}x^2 \tag{47}$$

for light travelling toward increasing or decreasing x, respectively. By following the light ray in space and time, the round trip time is computed by

$$\int_{t_0}^{t_1} \mathrm{d}t = \int_{0}^{L} \sqrt{1 + h_{+}(t_0 + x/\mathrm{c})}\mathrm{d}x + \int_{L}^{0} \sqrt{1 + h_{+}(t_0 + (L - x)/\mathrm{c})}\mathrm{d}x, \tag{48}$$

which definitely changes if h_{+} varies with time.

A similar result is obtained for a light ray travelling along the y-direction. As above,

$$\mathrm{d}(\mathrm{c}t)^2 = (1 - h_{+}(t - y/\mathrm{c}))\mathrm{d}y^2 \tag{49}$$

$$\mathrm{d}(\mathrm{c}t)^2 = (1 - h_{+}(t + y/\mathrm{c}))\mathrm{d}y^2, \tag{50}$$

and

$$\int_{t_0}^{t_1} \mathrm{d}t = \int_{0}^{L} \sqrt{1 - h_{+}(t_0 + y/\mathrm{c})}\mathrm{d}y + \int_{L}^{0} \sqrt{1 - h_{+}(t_0 + (L - y)/\mathrm{c})}\mathrm{d}y. \tag{51}$$

The variation of the round trip time has now opposite sign. If the light goes along -x the effect again will be the one observed along +x. This is the effect of the so-called + polarization

Along the same lines, one can show that if $h_\times \neq 0$ and $h_+ = 0$, the maximum effect is obtained with light travelling along $x = y$ and $x = -y$, and a similar pattern is obtained. This is the effect of the x polarization , identical to the + one, but for a $\pi/4$ rotation.

A different point of view can be adopted by considering the local inertial frame where one of the masses is at rest. For small relative velocities and a small separation between the two masses, we can work in a Newtonian reference frame where the proper distance between the masses is obtained from the components of the vector ξ^α connecting them. In such a frame, the trajectory of the second mass is given by the equation of geodesic deviation:

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\xi^\alpha = R^\alpha{}_{\mu\nu\beta}U^\mu U^\nu \xi^\beta, \tag{52}$$

where U^μ is the four-velocity of the second particle. At rest, $U^\mu = (1, 0, 0, 0)$ and since $R^\alpha{}_{\mu\nu\beta}$ is already of the order of $h_{\mu\nu}$ changes in U^μ will be of higher order. For $\xi^\alpha = (0, \varepsilon, 0, 0)$, the equation of geodesic deviation becomes:

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\xi^\alpha = \frac{\partial^2}{\partial t^2}\xi^\alpha = R^\alpha{}_{00x}\varepsilon = -R^\alpha{}_{0x0}\varepsilon. \tag{53}$$

The Riemann tensor is gauge invariant, ξ^α has initially as non-zero component a proper length, so the above equation is gauge invariant.

For a wave traveling in the $z-$direction and masses along the $x-$axis, one has

$$\frac{\partial^2}{\partial t^2}\xi^x = \frac{1}{2}\varepsilon\frac{\partial^2}{\partial t^2}h^{TT}_{xx} \tag{54}$$

$$\frac{\partial^2}{\partial t^2}\xi^y = \frac{1}{2}\varepsilon\frac{\partial^2}{\partial t^2}h^{TT}_{xy}. \tag{55}$$

Similar expressions are obtained for masses along the y-axis.

This acceleration can also be seen as the result of a force acting on the second mass, called Riemann force. Being proportional to separation, these forces are called "tidal" forces, thinking of the elongation observed on the Earth surface when it deforms due to the tide. This is best shown in Fig. 1 where the displacement of masses initially on a circle is shown for the + and x polarizations.

Two different points of view, that translate into different coordinate systems, have been shown. In the *TT* approach, masses do not change their coordinates and the measurement rod (the light) stretches. In the other case, the rod is rigid and masses move. Other coordinate choices are possible, but the result of the measurement of the round-trip time is invariant.

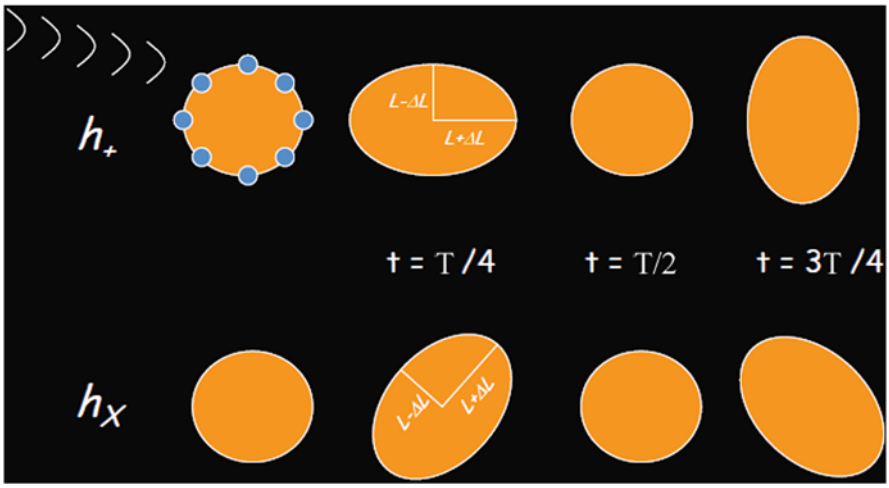


Fig. 1 Displacement of free masses for the + and the x wave polarizations

Coming back to the *TT* gauge, consider a wave described by $h\cos(\Omega t+\phi)$. The integral along the light path gives

$$t_1 - t_0 = \frac{2L}{c}\left[1 - \frac{h}{2}\frac{\sin\frac{\Omega L}{c}}{\frac{\Omega L}{c}}\cos\left(\Omega\left(t_1 - \frac{L}{c}\right)+\phi\right)\right].$$

Considering the Fourier components of a generic waveform, if $\Omega L/c \ll 1$ then the measurement of the round-trip time follows the shape of the incoming gravitational wave. To increase the gain, one can consider N round trips; in that case, L is replaced by NL. Note that if $\Omega L/c = k\pi$, that is if $2L$ is a multiple of the gravitational wavelength, the response vanishes; the change accumulated during, say, the positive half wave gets cancelled during the negative one.

It is anticipated here that the most promising astrophysical sources have $h \sim 10^{-23}$. With $L = 0.75$ km for ground-based detectors, the displacement in a rigid frame is

$$\delta x = L\frac{h}{2} = 1.5 \div 2 \times 10^{-20}\,\mathrm{m} \tag{56}$$

or $1/10^{10}$ of the size of hydrogen atom.

The measurement of the round-trip time or, for interferometers, the comparison between the round trip times for two different paths is the technique adopted for most of the gravitational wave spectrum.

3 The Generation of Gravitational Waves

The weak-field equations have been discussed until now without the source term for gravitational wave emission. As indicated before, the complete equation is

$$\Box \bar{h}^{\mu\nu} = -\frac{16\pi G}{c^4} T^{\mu\nu} \tag{57}$$

where the coupling constant is $G/c^4 = 8.23 \times 10^{-45}\ \mathrm{kg}^{-1}\ \mathrm{m}^{-1}\ \mathrm{s}^{-2}$.

The outgoing solution of this equation is

$$\bar{h}_{\mu\nu}(t, \mathbf{x}) = 4\frac{G}{c^4} \int \frac{T_{\mu\nu}(t - R/c, \mathbf{z})}{R} \mathrm{d}^3 z, \quad R = ||\mathbf{x} - \mathbf{z}||, \tag{58}$$

where the integrand is to be computed at the "retarded time", that is on the past light cone of $(t, \mathbf{x})$. Computing the solution far away from the source with respect to the source size, $R \gg ||\mathbf{x}||$, and with slowly moving sources then

$$\bar{h}_{\mu\nu}(t, \mathbf{x}) \approx \frac{4G}{c^4 r} \int T_{\mu\nu}(t - R/c, \mathbf{z}) \mathrm{d}^3 z. \tag{59}$$

Now, the conservation law $T^{\mu\nu}{}_{,\nu} = 0$ is also written in integral form as

$$\int T_{0\mu} \mathrm{d}^3 z = \text{const.} \tag{60}$$

implying that $\bar{h}_{0\mu}$ cannot contain a propagating field.

From the conservation law $T^{\mu\nu}{}_{,\nu} = 0$, it can be shown that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \int T^{00} x^l x^m \mathrm{d}^3 x = 2 \int T^{lm} \mathrm{d}^3 x \tag{61}$$

or

$$\bar{h}_{jk}(t, \mathbf{x}) = -\frac{2}{r} I_{jk,00}(t - r/c), \tag{62}$$

that is the second derivative of the mass quadrupole moment at retarded time

$$I_{jk}(t - r/c) = \int x_i x_j \mathrm{d}m. \tag{63}$$

Expressing this in the TT gauge

$$\begin{aligned} \bar{h}^{TT}_{xx}(t, \mathbf{x}) &= \frac{G}{c^4 r} \left[\text{I}\!\text{I}_{xx,00}(t - r/c) - \text{I}\!\text{I}_{yy,00}(t - r/c) \right] \\ \bar{h}^{TT}_{xy}(t, \mathbf{x}) &= \frac{G}{c^4 r} \text{I}\!\text{I}_{xy,00}(t - r/c), \end{aligned} \tag{64}$$

where $\text{I}\!\text{I}_{ij,00}(t - r/c)$ is the TT part of the quadrupole moment of the source.

The presence of the factor G/c^4 implies that the amplitude of the wave emitted by whatever device could be conceived in the laboratory is so small that one has to consider the mass of astrophysical objects in violent motion, as shown below.

The luminosity, or emitted power of a source, can be computed from the energy flux of a gravitational wave (see e.g. Schutz [3]. The expression

$$L = \frac{c^5}{5G}\langle \dddot{\mathcal{I}}_{ij} \dddot{\mathcal{I}}^{ij} \rangle \tag{65}$$

is valid, however, for slow moving bodies and weak fields. This expression, through the square of $\dddot{\mathcal{I}}_{ij}$, is proportional to the average $(v/c)^6$. Higher luminosity is obtained by considering rapid and coherent motion of large masses. The factor $c^5/G \simeq 3.6 \times 10^{52}$ W sets the scale of luminosity for gravitational wave sources. This is a huge number that however does not translate in a large wave amplitude.

To compute the amplitude at the detector, one has to consider propagation, with energy flow decreasing as $1/R^2$. Consider a source emitting in 1 s an amount of energy corresponding to one solar mass and size one Schwarzschild radius $R_S \sim 3$ km situated at a distance $R = 100$ Mpc (approximately 300 million light years or 3×10^{24} m). At the source, gravity will be strong, $h \sim O(1)$, but on the Earth the gravitational wave amplitude is reduced by a factor of the order of $(R_S/R)^2 \sim 10^{-42}$.

This still gives a high energy flux, related to amplitude by

$$F = \frac{\omega^2 c^3}{32\pi G}\langle \bar{h}_{ij}^{TT} \bar{h}^{TTij} \rangle. \tag{66}$$

The amplitude estimate is

$$\bar{h}_{ij}^{TT} \sim \left(\frac{32\pi G}{\omega^2 c^3} F\right)^{1/2}, \tag{67}$$

where, leaving aside the ω^{-2} behaviour, the conversion factor from the square root of flux to amplitude is

$$\left(\frac{32\pi G}{c^3}\right)^{1/2} \sim 10^{-16}.$$

A small deformation of vacuum stores an enormous amount of energy: it can be said that vacuum is quite rigid.

One can now estimate the order of magnitude of the gravitational wave amplitude in various cases. A ground-based source could be a bar rotating with angular speed ω around an axis perpendicular to the bar itself. For a thin homogenous bar, the quadrupole moment would oscillate with amplitude

$$I_{xx}^0 = 2\int_0^{L/2} x^2 \mathrm{d}m = 2\int_0^{L/2} x^2 \mathrm{d}m = M\frac{L^2}{12} = \rho S \frac{L^3}{12}, \tag{68}$$

and the second derivative would be

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} I^0_{xx} = -\omega^2 \rho S \frac{L^3}{12}. \tag{69}$$

On the other hand, the bar has to sustain the stress, that is going to be maximum at its centre

$$\sigma = \frac{F}{S} = \frac{1}{S} \int_0^{L/2} \omega^2 x \mathrm{d}m = \frac{\omega^2 M L^2}{8S} = \frac{\omega^2 \rho L^3}{8} \tag{70}$$

that has to be less than the fracture modulus which is of the order of 10 GPa for iron.

Assume a monster of 1000 t and 100 m length rotating at 10 rad/s. The luminosity is going to be

$$L \sim \frac{G}{c^5} M^2 L^4 \omega^6 \sim 10^{-26}\,\mathrm{W}, \tag{71}$$

which is a very tiny power.

On the other hand, consider two neutron stars in their last phase of inspiral, that is a system of $2.8 M_\odot$ of size $r \sim 20$ km. For such a system in a regime of strong gravity, the Schwarzschild radius is going to be of the order of the size of the neutron star:

$$L \sim \frac{G}{c^5} M^2 L^4 \omega^6 \sim \frac{G}{c^5} \frac{c^4 R_S}{G^2} r^4 \frac{\beta^6 c^6}{r^6} \sim \frac{c^5}{G} \left(\frac{R_S}{r} \right)^2 \beta^6 \sim 10^{43}\,\mathrm{W} \tag{72}$$

or 10^{17} Suns, but for a fraction of a second.

Another example of interest is a rotating neutron star (radius $r \sim 10$ km) with moment of inertia

$$I \sim \frac{2}{5} 1.4 M_\odot R^2 \sim 10^{38}\,\mathrm{kg\,m^2}, \tag{73}$$

rotation frequency 10^3 rad/s and ellipticity $\varepsilon = (I_{xx} - I_{yy})/I_{zz} \sim 10^{-7}$. The luminosity is

$$L \sim \frac{G}{c^5} M^2 L^4 \omega^6 \sim \frac{G}{c^5} (\varepsilon I_{zz})^2 (\omega)^6 \sim 3 \times 10^{28}\,\mathrm{W}. \tag{74}$$

If this neutron star is at 10 kpc (3×10^{19} m) away, the flux on Earth is $3 \times 10^{-12}\,\mathrm{W/m^2}$, which corresponds to an amplitude for h_{ij}^{TT} of the order of 10^{-26}.

Consider now a binary system made of two neutron stars ($m = 1.4 M_\odot$) that are practically touching themselves, that is with an orbit radius r_0 of 10 km. This would be the very end of the life of a binary system that lost almost all its energy. Making the crude approximation of considering the two masses as point-like with Newtonian dynamics, the orbital frequency is given by

$$\omega^2 = Gm/r_0^3 \tag{75}$$

or $\omega \sim 2 \times 10^3$ rad/s.

The TT quadrupole moment is given by

$$\begin{aligned} I_{xx} &= mr_0^2 \cos 2\omega t + C \\ I_{yy} &= -mr_0^2 \sin 2\omega t + C' \\ I_{xy} &= mr_0^2 \sin 2\omega t. \end{aligned} \tag{76}$$

For this system, the order of magnitude of the amplitude of the radiation, observed at distance R, is then

$$h \sim \frac{G}{Rc^5} mr_0^2 \omega^2 = \frac{G^2 m^2}{c^5 R r_0} \sim \frac{1}{c^5} \phi \phi_0, \tag{77}$$

where $\phi = Gm/R$ is the potential of the system at the observation point and $\phi_0 = Gm/r_0$ is the typical potential in the system.

4 GW Sources

The possibility of detecting GW, studying their properties, and analyzing the astrophysics message they carry depends on many factors. Detectors have to give a response that can stand out of the instrument noise, and an effort started several decades ago has gone into the design of more and more sensitive instruments. This determines the amplitude of detectable waves. But sufficient sources at the right distance have to be available. Strong but unlikely events may never occur during the lifetime of an instrument, while events that are feeble due to distance may be more frequent as they come from a broader portion of the universe. Today, predictions for amplitudes and populations for the various sources still span several orders of magnitude.

The sources being considered are binary systems in their final inspiral and merging phase, supernova core collapse, rapidly rotating neutron stars and stochastic background. The first two are transient, the other two are permanent. Considering detection, another distinction has to be made; knowing the waveform allows to significantly increase the signal-to-noise ratio (SNR) in the detector. This is the case for binary systems and rotating neutron stars. Core collapse cannot really be modelled; stochastic background results in additional noise in a detector, possibly with different statistical properties.

4.1 Inspiralling Binary Systems

Binary systems of compact astrophysical objects are the simplest systems that are able to emit GW. As shown above, waveforms can be computed at first order using Newtonian orbits, having care of taking into account secular variations of energy and

Table 1 Predicted observing rates per year for second generation interferometers Advanced Ligo-Advanced VIRGO (see [11] for details. Lo, Re, Hi and Max refer to the posterior probability density function *(PDF)* when available. R_{Re} refers to the PDF peak, R_{Lo} and R_{Hi} to the 95 % pessimistic and optimistic confidence intervals. R_{Max} is based on very basic limits set by other astrophysical knowledge

System	R_{Lo} yr^{-1}	R_{Re} yr^{-1}	R_{Hi} yr^{-1}	R_{Max} yr^{-1}
Binary neutron stars	0.4	40	400	1000
Neutron star-black hole	0.2	10	300	
Binary black holes	0.4	20	1000	

angular momentum. Much more can be learnt when the system becomes more relativistic and higher order terms in $(v/c)^2$ need to be taken into account. Ground-based experiments, sensitive in the band $10 \div 10^4$ Hz, will be able to observe only the very last orbits of systems of $1 \div 100$ solar masses, their merging and ring down. Whether in these extreme conditions general relativity will need modifications or will successfully compare with observation is a fundamental question for our understanding of gravitation. In addition, in the case of neutron stars the observed behaviour will allow to discriminate between various equations of state for nuclear matter.

Much less certain is the rate of observation of these events. One has to predict the probability to build a binary system made of two compact objects such as two neutron stars, two black holes or a black hole-neutron star system. Theoretical predictions are uncertain, being largely model dependent. Only a handful of neutron star binary systems has been identified and is thus available to tune models. However, lower limits are now within the reach of ground-based interferometers that are being upgraded now, as shown in Table 1 (computed in [11]) for the expected performance of the second generation interferometers.

The table indicates that in the next years, when the second generation detectors will be operating we should observe gravitational wave signals. Not observing them would have major implications on our understanding of physical laws.

These events may be accompanied by electromagnetic emission, in particular gamma-ray bursts. The simultaneous observation of GW and electromagnetic radiation over the whole spectrum would allow to study these violent phenomena in great detail. In particular, GW will allow to probe the inner matter of the system, opaque to other radiation.

4.2 Supernova Core Collapse

At the end of a star life, when the pressure originated by the nuclear reactions is not able to sustain gravitational force anymore, one has the collapse of the star core, resulting then in a supernova (SN) explosion. At the end of the process, a neutron star may have formed, if the parent star mass is in the range $8 \div 25 M_{\odot}$. The collapse

has a duration of the order of 1 ms and involves large quantities of mass. This is very promising to make a core collapse an interesting gravitational wave source; if the collapse is not axisymmetric, there should be a large quadrupole emission. The short duration of the pulse makes it accessible to the ground-based detectors, as discussed later, and indeed supernovae are a primary target for gravitational wave detection by acoustic detectors resonating around 1 kHz.

Knowledge of these sources has evolved since the first experiments both for rate estimates and for emission amplitude. It is estimated that in the Milky Way there will be $2 \div 3$ SN explosions every century. This rate becomes high if one is able to listen to tens or hundreds of galaxies, extending the sensitivity range (or horizon) well outside our own galaxy. For the amplitude, estimates were based on the amount of energy going into gravitational radiation. Early estimates assumed that $10^{-2} M_\odot$ solar masses would be converted, but the development of general relativistic magnetohydrodynamic codes drastically reduced that amount. Nowadays, estimates are of $10^{-7} - 10^{-8} M_\odot$ [12]; so, the perspective of detecting SN signals is dim even for second-generation detectors.

The waveform from a core collapse is unpredictable; matter motion toward the star centre is prone to instability and will depend strongly on initial conditions. Detection of these signals against background noise can be achieved primarily by estimating the energy content of the wave. On the other hand, the recorded signal will carry unique information about the last milliseconds of life of the star.

4.3 Periodic Sources

Estimates on star population in our galaxy indicate that there are of the order of 10^9 neutron stars. These neutron stars have a mass that is typically 1.4 $M_\odot$ as higher masses would lead to the formation of a black hole. Their radius is of the order of 10 km; the density is 3×10^{17} kg/m^3. They come from progenitor stars and due to angular momentum conservation when they were created, or by successive matter accretion, they can spin quite fast. They have a strong magnetic dipole, with a field at the surface of 10^8 T and turn out to be powerful electromagnetic pulse emitters. Pulsars were discovered in radioastronomical signals and were immediately interpreted as coming from rotating neutron stars. As of today, more than 2000 pulsars have been discovered, from tens of seconds to millisecond period [13]. As discussed above, rapidly rotating neutron stars are gravitational wave emitters. The amplitude of the wave at the Earth is given by

$$h = 3 \times 10^{-27} \left(\frac{10\,\text{kpc}}{r} \right) \left(\frac{I}{10^{38}\text{kg m}^2} \right) \left(\frac{f}{200\,\text{Hz}} \right)^2 \left(\frac{\varepsilon}{10^{-6}} \right), \qquad (78)$$

where I is the actual moment of inertia and r is the distance of the source.

It is anticipated that this amplitude is nevertheless very small. However, one can profit from the periodicity of the signal and use the detector as a lock-in amplifier, accumulating the signal period after period, while averaging the noise down

as $T^{-1/2}$, where T is the observation time. In this way with $T \sim 1$ year, the signal-to-noise ratio can be optimally enhanced by a factor of 5×10^3.

With an observation time of 1 year, the frequency resolution $\delta f = 1/T = 3 \times 10^{-7}$ Hz but the signal observed will not be that stable, due to the Doppler effect coming from the Earth orbiting around the Sun at 30 km/s $= 10^{-4}\beta$c. For a source emitting GW at frequency f the frequency span Δf will be up to $\Delta f = \pm 10^4 f \gg \delta f$, making the lock-in scheme ineffective. Data analysis becomes much more complex as frequency follows a different pattern for every different location in the sky, so one needs to set up in data analysis programs something like 10^{16} different lock-in devices, for 10^{10} different frequency bins up to 1 kHz and of the order of 10^6 different patches in the sky. This is not manageable with current computers so other methods have been developed, attempting to compromise between the computing time needed and a lower detection efficiency.

The detection of GW from these sources will provide direct information on the neutron star structure and the nuclear matter equation of state. The distribution of these pulsars will give a completely new view of the galaxy, and it may be that there is a pulsar nearby the solar system that just does not illuminate Earth with its lighthouse beam and would be discovered only by the ripples in space-time it generates.

4.4 Stochastic Background

Gravitational wave detectors are sensitive to waves coming from almost any direction. In presence of several waves, these detectors are going to measure a linear combination of the resulting spacetime deformations. Waves from all directions, frequencies and polarizations form a stochastic background, whose origin can be astrophysical, cumulating many sources at different red-shifts, or cosmological, coming from the early moments of the universe. The microwave background gives a picture of the universe at an age of 380,000 years, when the temperature has dropped sufficiently so that nuclei and electrons can recombine to form neutral atoms that are transparent to electromagnetic radiation. For GW, due to their very low interaction with matter, the universe becomes transparent much earlier and information about that epoch is kept in the stochastic background.

Many cosmology-related measurements have been performed, resulting in a highly coherent picture for the standard model of the Big Bang. These measurements provide constraints on the energy density of the gravitational wave stochastic background

$$\rho_{GW} = \frac{c^2}{32\pi G}\langle \dot{h}_{ij}\dot{h}^{ij}\rangle = 1.34 \times 10^{-25}\langle \dot{h}_{ij}\dot{h}^{ij}\rangle, \tag{79}$$

related to the energy flow of Eq. 66. The main limit is given by Big Bang Nucleosynthesis and leads to an upper limit [14]:

$$\Omega_{GW}(f) = \frac{1}{\rho_c}\frac{\mathrm{d}\rho_{GW}}{\mathrm{d}\ln f} < 8.6 \times 10^6, \tag{80}$$

flat in frequency, and

$$\rho_c = \frac{3c^2 H_0^2}{8\pi G} \approx 1.6 \times 10^{20} H_0^2 = 5.49 \times 10^{-20}\,\mathrm{J/m^3}, \tag{81}$$

with $H_0 = 73.8 \pm 2.4\,\mathrm{km/sMpc}$ [15]. An additional limit at low frequency comes from the observation of cosmic microwave background and matter inhomogeneities:

$$\Omega_{GW} h^2 < 10^{-14}, \quad f < 10^{-15} \div 10^{-10}\,\mathrm{Hz}. \tag{82}$$

The energy density of the gravitational wave stochastic background is related to the amplitude of GW by [16]

$$\langle \dot{h}_{ij} \dot{h}^{ij} \rangle < 10^{-26}. \tag{83}$$

In the standard scenario, the expected amplitude seems to be out of reach for ground-based detectors. However, if the cosmological gravitational wave background has an appropriate frequency dependence, the limits set by cosmology might not apply. This could be the case if some new physics were present at the Big Bang.

Detecting a stochastic background amounts to detect an excess noise in a detector. Unless one has precise control of the instrumental noise, this task cannot be carried out but by observing correlation between two different instruments. The task of detecting a stochastic background signal is very difficult indeed, but scientific interest for it is beyond doubt.

5 PSR 1913+16

Having spent some time in describing the various sources of GW, it is appropriate to discuss the first evidence for gravitational wave emission [13]. The pulsar PSR 1913+16 was discovered by Hulse and Taylor in 1974 using the Arecibo radiotelescope. The pulsar has a rotation frequency of about 17 Hz but it was immediately apparent that this frequency was fluctuating in a periodic way (see Fig. 2. This allowed to deduce that the pulsar was a member of a tight binary system with an orbital period of 7.75 h as there was no other way to explain the size of the effect (1.3 ‰) over such a time scale.

It can be seen that the orbit is highly eccentric and the ratio of speeds at periastron and apastron allows to compute the eccentricity in Newtonian approximation:

$$\frac{v_p}{v_a} = \frac{1+e}{1-e}, \quad e = \frac{v_p - v_a}{v_p + v_a} \simeq 0.6. \tag{84}$$

The periastron precession is given by

$$\dot{\varphi} = \frac{6\pi G M'}{a(1-e^2)\tau c^2}, \tag{85}$$

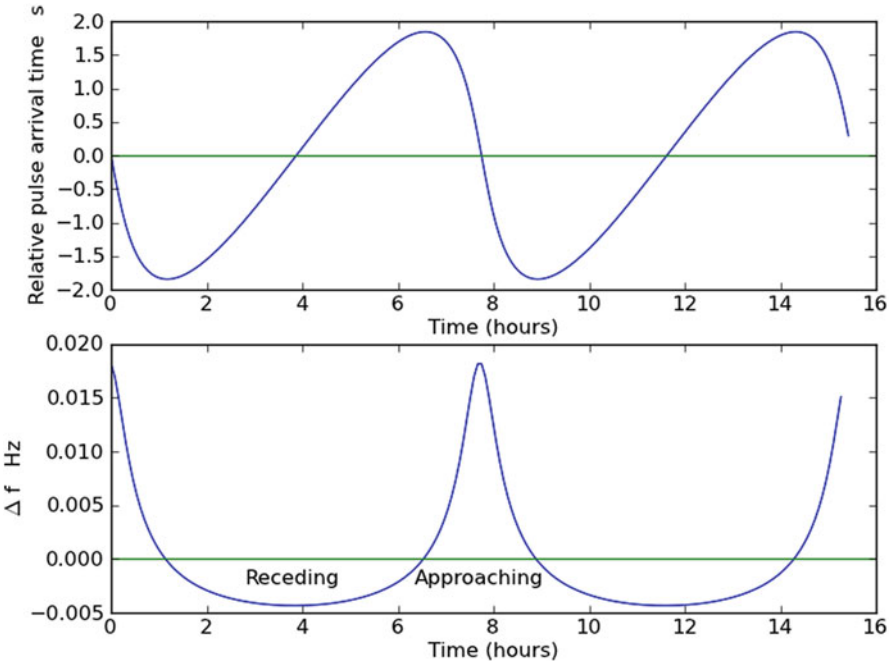


Fig. 2 *Top:* PSR1913+16 pulse arrival time relative to centre of mass [17]. *Bottom:* Doppler shift of the pulsar frequency. The 7.75 h period is explained most naturally by the presence of a companion in a tight binary orbit

where M' is the mass of the pulsar companion and τ is the orbital period. For PSR 1913+16, this precession is of about 4°year^{-1}, orders of magnitude larger than the Mercury's.

From the same data, one can deduce the size of the orbit along the line of sight which is about 4 light-seconds, $\simeq 1.2 \times 10^9$ m, much less than the diameter of the Sun. The distance between the two stars varies by a factor of 4, enough to be able to observe a gravitational redshift for the pulsar period P in the field of the other star:

$$P = P_0 \left(1 - 2\frac{GM'}{rc^2}\right)^{-1/2}. \tag{86}$$

It is also possible to observe the time dilation of special relativity, as the pulsar speed is high and varies sufficiently to detect the related period variations:

$$P = P_0 \left(1 - 2\frac{GM'}{rc^2}\right)^{-1/2} \left(1 - \frac{v^2}{c^2}\right)^{-1/2}. \tag{87}$$

All this makes PSR1913+16 a unique laboratory for special and general relativity, moreover at a distance of about 20,000 light-years.

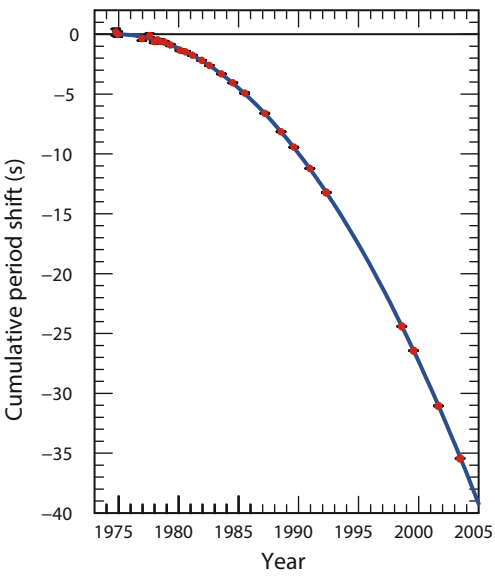


Fig. 3 Cumulative shift of periastron time for PSR 1913+16 (data from [18])

However, the most important result was the observation that the orbital period of the system, which should stay constant in Newtonian mechanics, decreases in a sizeable way:

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = -2.422(6) \times 10^{-12}, \tag{88}$$

as shown in Fig. 3. The decrease comes from the energy decrease of the system due to gravitational wave emission. The parameters related to the system PSR 1913+16 are listed in Table 2.

The hypothesis of gravitational radiation as origin of the energy loss is clearly compatible with these observations; once all effects are included, general relativity is in agreement with the observations, the combination of theoretical uncertainties and experimental errors being at the 2 ‰ level. This is a clear evidence for gravitational radiation; however, until now no gravitational wave signal has ever been detected.

6 Gravitational Wave Detection

The discussion on sources of GW has a first conclusion: detecting them is not going to be an easy task. The expected changes in the metric range from 10^{-12} for stochastic background with timescale of million of years down to 10^{-21} for inspiral events within some tens of Mpc. These very small values have to be compared on one side with how precisely test masses follow a geodesic and on the other to the precision in distance measurement. Experimentalists know that the best techniques are needed

Table 2 Parameters of PSR 1913+16. Data from [19] and [18]

Parameter	Symbol (units)	Value
Pulsar parameters		
Right Ascension	α	$19^h\,15^m\,28.^s0018(15)$
Declination	δ	$+16\,°06'27.''4043(3)$
Distance	kpc	7.13
Pulsar period	P_p (ms)	59.029997929613(7)
Period derivative	$\dot{P}_p$	$8.62713(8) \times 10^{-18}$
Binary system parameters		
Projected semimajor axis	$a_p \sin\iota$ (light-s)	2.3417592(19)
Eccentricity	e	0.6171308(4)
Orbital period	P_b (day)	0.322997462736(7)
Longitude of periastron	$\omega_0(\,°)$	226.57528(6)
Julian date of periastron	T_0 (MJD)	46443.99588319(3)
Post-Keplerian parameters		
Mean rate of periastron advance	$\langle\dot{\omega}\rangle$ ($\,°\,\mathrm{yr}^{-1}$)	4.226621(11)
Redshift/time dilation	γ' (ms)	4.295(2)
Orbital period derivative	$\dot{P}_b$ ($\times 10^{-12}$)	−2.422(6)

and will have to be improved: There is a strong drive toward improving that precision realizing that the number of sources increases with the cube of the distance that can be reached.

Interferometry is a high precision technique; optical paths can be very effectively compared by having two beams interfere destructively, performing a null experiment. The natural measurement unit is the wavelength, typically of the order of 10^{-6} m. Still more than 12 orders of magnitude need to be gained but interferometry seems well suited for gravitational wave detection. It has been shown that distances in perpendicular directions vary with opposite signs (see 41), so that a Michelson–Morley type interferometer (Fig. 4) is an excellent device for gravitational wave detection. A first experimental advantage of the use of two-arm interferometry is that for equal arms the measurement is insensitive to wavelength changes. Although exact symmetry cannot be achieved, nevertheless the requirements on the light source can be relaxed with respect to the above 10^{-21} of relative precision. Calling ϕ_N and ϕ_W the phases of the beams (N is for North and W for West), k and λ the wave number and wavelength, respectively, one has

$$\begin{aligned}\phi_N - \phi_W = 2k(L_N - L_W) &= \frac{4\pi}{\lambda}\left(L_N\left(1+\frac{h_+}{2}\right) - L_W\left(1-\frac{h_+}{2}\right)\right)\\ &= \frac{4\pi}{\lambda}\left(\Delta L + h_+\frac{L_N+L_W}{2}\right)\end{aligned} \tag{89}$$

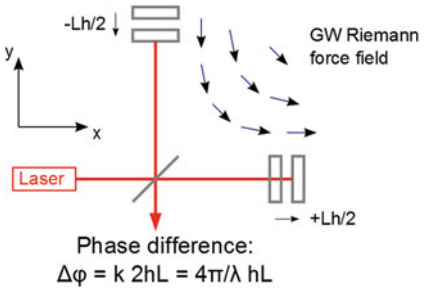


Fig. 4 Michelson interferometer principle

where $\Delta L = L_N - L_W$ is the difference in arm length and h_+ is the amplitude of the gravitational wave. The error on phase is given by

$$\sigma_\phi = 4\pi \frac{\Delta L}{L} \frac{\sigma_\lambda}{\lambda}. \tag{90}$$

Geometrically, $\Delta L/L$ can be as low as 10^{-6}, depending on how well mirror survey is performed, but if L is increased by means of a Fabry–Perot cavity the effective cavity asymmetry is determined by the difference in transmission coefficients of the mirrors. A few ‰ asymmetry has been achieved in present generation interferometers and this allows to reduce significantly the requirements on the frequency noise of the light source.

7 Fundamental Noises in the Interferometers

The sort of noise discussed above, although at the hearth of the measurement, is not fundamental at the present level, in the sense that more symmetric arms could be achieved, and frequency fluctuations can be controlled at the required level by appropriate feedback. The first sort of fundamental noise that will be discussed is related to the measurement process, which involves, at some point, counting photons. The inherent statistical nature of this process limits the achievable precision.

The actual position of the masses used in the measurement is modified by local forces. Most of them can be reduced by an appropriate experimental set-up but there are sources of noise that cannot be easily shielded. Masses are in a thermal bath, and the average energy associated to any degree of freedom is on average kT, where k is the Boltzmann constant and T is the absolute temperature. The actual kinetic energy fluctuates, resulting in random motion. The actual effect on the measurement represents the main limit to the performance of today's interferometers. The masses are also in a gravitational field, that is not constant. In fact, fluctuating gravity gradients modify mass separation making them undistinguishable from GW.

For completeness, as this is becoming relevant in second generation interferometers, one has to mention the effect of the measurement process on the masses.

Fluctuations in light intensity introduce time-dependent thermal effects resulting in mass deformation and therefore mirror surface motion. The same fluctuations move the mass as a whole limiting the measurement capability. Fluctuations from the ultimate quantum nature of light enter the measurement process too, a macroscopic manifestation of the Heisenberg' principle. This can be evaded by changing the quantum nature of the light used, measuring better the phase while tolerating larger amplitude fluctuations, but this intriguing subject is beyond the scope of this chapter.

7.1 Measurement Noise

Consider a Michelson interferometer with a light source of power P_{in} and detected power P_{out} and assume a phase difference between the two arms containing an offset $\pi + \alpha$ and the gravitational wave signal ϕ_{GW}

$$\Delta\varphi = \pi + \alpha + \phi_{GW}, \tag{91}$$

where $\alpha = 0$ means destructive interference toward the detection port. The detected power is

$$P_{out} = P_{in} \sin^2\left(\frac{\alpha + \phi_{GW}}{2}\right), \tag{92}$$

which becomes, for $\phi_{GW} \ll 1$

$$P_{out} = P_{in}\left(\sin^2\frac{\alpha}{2} + \frac{1}{2}\phi_{GW}\sin\alpha\right). \tag{93}$$

The interferometer response is:

$$\begin{aligned}\frac{\mathrm{d}P_{out}}{\mathrm{d}\phi_{GW}} &= P_{in}\sin\left(\frac{\alpha + \phi_{GW}}{2}\right)\cos\left(\frac{\alpha + \phi_{GW}}{2}\right)\\ &= \sin(\alpha + \phi_{GW}),\end{aligned} \tag{94}$$

which is maximum for $\alpha = \pi/2$. This corresponds to $P_{out} = P_{\bar{out}} = P_{in}/2$ with $P_{\bar{out}}$ being the power back to the light source. The response of an instrument to a small variation of the physical quantity to be measured is traditionally called *sensitivity*. In the field of GW, however, this term is used in a different way, as discussed below. $\frac{\mathrm{d}P_{out}}{\mathrm{d}\phi_{GW}}$ can be called instead GW transfer function or GW gain, following electronic engineering terminology.

However, P_{out} is affected with fluctuations and this will limit the precision of the measurement. If on average N photon counts per second are expected, and these counts have Poisson statistics, as is the case for usual and laser light, the counting

process is affected by "shot noise" and the variance of the counting process is N. The fluctuation in the measurement of P_{out} is

$$\sigma_{P_{out}} = \sqrt{\eta N} = \sqrt{\frac{\eta P_{in}}{h\nu}\left(\sin^2\frac{\alpha}{2} + \frac{1}{2}\phi_{GW}\sin\alpha\right)}, \tag{95}$$

η being the detection quantum efficiency and $h\nu$ the energy of the single photon. In absence of the signal,

$$\sigma_{P_{out}} = \sqrt{\frac{\eta P_{in}}{h\nu}}\left|\sin\frac{\alpha}{2}\right|. \tag{96}$$

The SNR for a 1-s observation time (or a 1-Hz frequency band) is

$$SNR = \sqrt{\frac{\eta P_{in}}{2h\nu}}\left|\cos\frac{\alpha}{2}\right|\phi_{GW}, \tag{97}$$

maximum for $\alpha = 0$. Gravitational wave interferometric detectors work near the "dark fringe" (although not exactly at $\alpha = 0$ to avoid a quadratic response).

The amplitude of h that generates a signal equal to the noise level (SNR=1) is conventionally called "sensitivity" in the gravitational wave community and from now this meaning will be adopted. The phase sensitivity for a 1-s measurement is

$$\tilde{\phi}_{GW}^{min} = \sqrt{2\frac{h\nu}{\eta P_{in}}}. \tag{98}$$

If the measurement time is T, the average count varies accordingly and the phase sensitivity becomes

$$\phi_{GW}^{min} = \sqrt{2\frac{h\nu}{\eta P_{in}T}}. \tag{99}$$

Note that $\tilde{\phi}_{GW}^{min}$ and ϕ_{GW}^{min} have different units. ϕ_{GW}^{min} is in radians and $\tilde{\phi}_{GW}^{min}$ is in radians $s^{1/2}$ or radians/$\sqrt{\text{Hz}}$. What has been measured is ϕ_{GW}^{min}, which has physical units. However, to be independent of its measurement time, which may depend on the particular gravitational wave source, $\tilde{\phi}_{GW}^{min}$ is usually mentioned. This is a *(linear) spectral density* in the sense that the square represents the variance of the physical quantity downstream a 1-Hz wide bandpass filter. Discussion of power spectra for random signals can be found in books on stochastic processes [20] or on telecommunications. For our purposes, linear spectral densities (or for short spectra), represent the standard deviation of a given quantity in a 1-Hz bin. Contributions from independent noise sources must be added quadratically. Noise over a wider frequency band is also obtained adding quadratically the various frequency bins. One has to be careful specifying whether the spectrum is to be considered for both negative and positive frequency, or only positive. As both sides of the spectrum sum up, the two-sided spectrum is a factor $\sqrt{2}$ lower that the one-sided one.

Table 3 Sensitivity in h for a table top and a km class interferometer

	L	h_{min} ($\mathrm{Hz}^{-1/2}$)
Table top	1 m	1.6×10^{-17}
VIRGO	3 km	5.3×10^{-21}

As a typical example, consider an infrared 10-W light source ($\lambda = 1064$ nm and an ideal photon detector $\eta = 1$). One obtains

$$\phi_{GW}^{min} \sim \sqrt{\frac{2h\nu}{\eta P_{in}}} = 2.0 \times 10^{-10} \mathrm{rad}\,\mathrm{Hz}^{-1/2}. \tag{100}$$

For a gravitational wave with + polarization impinging perpendicularly to the interferometer

$$\phi_{GW} = \frac{2\pi}{\lambda} 2L \frac{h}{2} \times 2, \tag{101}$$

so that

$$h^{min} = \frac{\lambda}{4\pi L}\sqrt{\frac{2h\nu}{\eta P_{in}}} = \frac{1}{4\pi L}\sqrt{\frac{2hc\lambda}{\eta P_{in}}}, \tag{102}$$

which is a spectral density in $\mathrm{Hz}^{-1/2}$. In Table 3, two cases of interest are presented, a table top and the 3-km arm VIRGO interferometer.

One sees that for a kilometre-class interferometer the sensitivity is respectable but not yet very useful, given the amplitude of the expected signals. Equation 102 shows which improvements are possible; an increase in L, although this has consequences on the maximum detectable frequency, an increase in P_{in} and a decrease in λ. All these parameters hit practical limitations; L is not easily increased on the Earth for ground-cost reasons, increasing P_{in} requires a serious technological effort to stay within other very stringent requirements for the light source, like frequency stability and beam quality, while decreasing λ significantly would mean operating in the near UV, where light sources and optical components do not have the same level of quality that is available at 1064 nm.

Remaining in the near infrared, L can be increased by using multiple reflections in order to increase the optical path difference. The simplest way is to have a beam bounce on different places of a test mass. Light diffusion by mirror imperfections and the natural broadening of the beam require each reflection spot to be well separated from the other otherwise there is crosstalk between the various segments of the optical path. This quickly sets a limit to this approach; either very large transversal dimension for the test masses are adopted or one needs more than one test mass at each end of the arm. A different approach is to use a Fabry–Perot resonant cavity, made of two mirrors between which light goes back and forth. At the input, a semi-reflective mirror allows light to enter the cavity that is closed by a reflective mirror. By properly adjusting the distance between the mirrors, the light can be stored with little light being reflected. As a consequence, length variations are detected through the light reflected by the cavity.

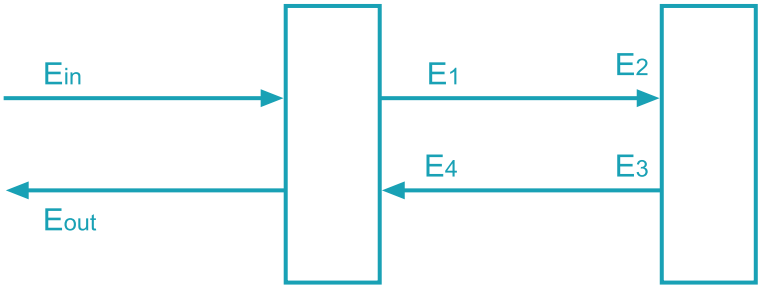


Fig. 5 Schematic Fabry–Perot cavity

Consider plane waves with incoming and reflected electric fields E_{in} e E_{out}. The fields inside the cavity at both ends and in both directions are E_1, E_2, E_3 and E_4 (see Fig. 5).

The boundary conditions are as follows:

$$\begin{aligned} E_1 &= t_1 E_{in} + j r_1 E_4 \\ E_2 &= \exp[jkl] E_1 \\ E_3 &= j r_2 E_2 \\ E_4 &= \exp[jkl] E_3 \\ E_{out} &= j r_1 E_{in} + t_1 E_4 \end{aligned}$$

with a reflection phase difference equal to $\pi/2$ (dielectrics).

The result is

$$E_1 = \frac{t_1}{1 + r_1 r_2 \exp[2jkl]} E_{in} \tag{103}$$

that has a resonance condition for $2kL = \pi$ giving for the field inside the cavity

$$E_1 = \frac{t_1}{1 - r_1 r_2} E_{in}. \tag{104}$$

The beam intensity at resonance is

$$P_1 = \frac{\varepsilon_0 c}{2} |E_1|^2 = \left(\frac{t_1}{1 - r_1 r_2} \right)^2 P_{in}, \tag{105}$$

and more in general

$$P_1 = \left(\frac{t_1}{1 - r_1 r_2} \right)^2 \frac{P_{in}}{1 + \frac{4\mathcal{F}^2}{\pi^2} \sin(kx)^2 [x/2]}, \tag{106}$$

where x is the deviation from the resonance length L and $\mathcal{F}$ is the cavity finesse:

$$\mathcal{F} = \frac{\pi \sqrt{r_1 r_2}}{1 - r_1 r_2}. \tag{107}$$

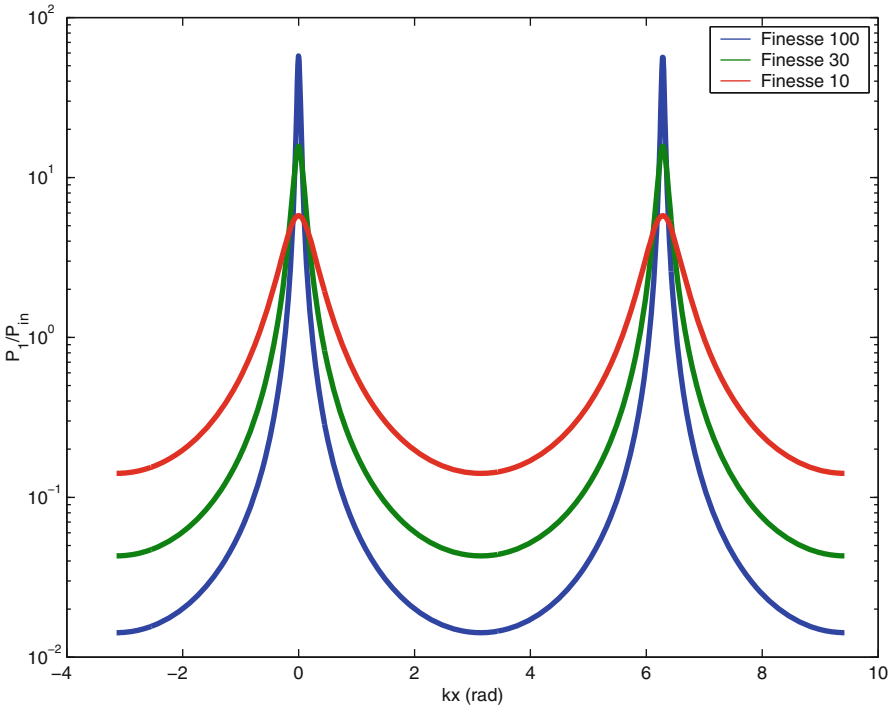


Fig. 6 Beam intensity as function of detuning

It is important to note that the cavity resonates at many different frequencies as shown in Fig. 6:

$$\nu_n = \left(n + \frac{1}{2}\right)\frac{c}{2L}. \tag{108}$$

$2L/c$ is called the free spectral range (FSR). For a 3-km cavity, FSR $\cong$ 50 kHz.

In analogy with resonating LC circuits, one can define as figure of merit the field inside the cavity divided by the incident field. The reflected field E_{out} is given by

$$E_{out} = jr_1E_{in} + jr_2t_1\exp[j2kL]E_1 = j\mathcal{R}E_{in} \tag{109}$$

with

$$\mathcal{R} = \frac{r_1 + (1-p_1)r_2\exp[j2kL]}{1 + r_1r_2\exp[j2kL]}. \tag{110}$$

p_1 represents the cavity absorption. There will be a phase variation corresponding to a length variation:

$$\frac{d\Phi}{dL} = \frac{8\mathcal{F}}{\lambda}. \tag{111}$$

However, as in the case of the Michelson interferometer working point, this does not correspond to the highest sensitivity because being at resonance implies that little

light is reflected. More detailed computations show that a good working point for the cavity leads to a reduction by a factor of 2 for the sensitivity.

The phase difference between the two arm cavities in presence of a gravitational wave is

$$h^{min} = \frac{\lambda}{8\mathcal{F}L} 2\sqrt{\frac{h\nu}{\eta P_{in}}}\sqrt{1 + \frac{4\mathcal{F}^2}{\pi^2}\sin^2\left(\frac{\Omega L}{c}\right)} \tag{112}$$

to be compared with the simple Michelson case

$$h^{min} = \frac{\lambda}{2\pi L}\sqrt{\frac{h\nu}{\eta P_{in}}}. \tag{113}$$

At low frequency ($\Omega \ll c/L$), one obtains a sensitivity gain of

$$\frac{2\mathcal{F}}{\pi}, \tag{114}$$

which can also be read as the effective length increase of the arms. For $\mathcal{F} = 150$, the gain is 92.4, which brings the sensitivity level of the VIRGO interferometer to $5.7 \times 10^{-23}\,\mathrm{Hz}^{-1/2}$.

While the physical arm length difference can be controlled at 10^{-5} relative level of 10 μm, the difficulty of having exactly the same finesse in the two arms introduces a much more severe asymmetry. Since the effective length is proportional to $L\mathcal{F}$, Eq. 90 becomes a sum in quadrature

$$\sigma_\phi = \left(\frac{\Delta L}{L} \oplus \frac{\Delta \mathcal{F}}{\mathcal{F}}\right), \tag{115}$$

making the instrument much more sensitive to frequency fluctuations; current interferometers have achieved $\Delta\mathcal{F}/\mathcal{F} \simeq 0.5\%$, still one order of magnitude worse than $\Delta L/L$.

It should be mentioned here that keeping these cavities optically resonant requires controlling the mirror relative positions with a sophisticated feedback system, with error signals that have to be extracted directly from the circulating light. The Pound–Drever–Hall technique [21,22] is often used but its discussion would be out of the scope of this work (see [23] for a pedagogical introduction).

Another technique based on resonant cavities is used to increase the sensitivity of gravitational wave interferometers. By inserting a semi-reflective mirror between the light source and the beam splitter, one can create a large complex optical cavity. This mirror behaves as the input mirror of a Fabry–Perot cavity with the E_4 field of the above discussion being determined by the reflected light from the optical system. Since the interferometer is operated on the dark fringe, in principle all power is coming back to the mirror and can be recycled into the interferometer. The incident power on the beam splitter is multiplied by a recycling factor that will depend on the losses in the interferometer. A recycling factor of the order $C = 30$ was achieved with VIRGO.

The resulting sensitivity, including the filtering behavior of the Fabry–Perot cavities, is

$$h^{min} = \frac{\lambda}{4\mathcal{F}L}\sqrt{\frac{h\nu}{\eta C P_{in}}}\sqrt{1+\frac{4\mathcal{F}^2}{\pi^2}\sin^2\left(\frac{\Omega L}{c}\right)}, \tag{116}$$

which in the low-frequency limit results in

$$h^{min} = \frac{\lambda}{4\mathcal{F}L}\sqrt{\frac{h\nu}{\eta C P_{in}}} = 0.8\times 10^{24}\,\mathrm{Hz}^{-1/2}. \tag{117}$$

The introduction of "power recycling", the availability of more powerful light sources and the increase in finesse for advanced detectors results in beams of several MW circulating in the cavities. At this level, even with mirrors of several tens of kilograms, the fluctuations in radiation pressure introduce a spurious motion, limiting the measurement capability.

For a beam of power P, the radiation force on a fully reflective mirror is $F = 2P/\mathrm{c} = 2\bar{N}h\nu/\mathrm{c}$, where $\bar{N}$ is the average number of photons per second. The force fluctuations are then

$$\delta_F = 2\sqrt{N}h\nu/\mathrm{c} = 2\sqrt{\frac{P}{h\nu}}\frac{1}{\mathrm{c}}, \tag{118}$$

the mirror acceleration fluctuations per unit time will be

$$\delta_a = 2\sqrt{\frac{P}{h\nu}}\frac{1}{m\mathrm{c}} \tag{119}$$

resulting in a position fluctuation

$$\delta_x = 2\sqrt{\frac{P}{h\nu}}\frac{1}{m\Omega^2\mathrm{c}}, \tag{120}$$

Ω being the angular frequency at which the measurement takes place.

On the other hand, the precision of the measurement by the cavity is

$$\sigma_x = \frac{2\pi}{\lambda}\sqrt{\frac{P}{h\nu}}\frac{1}{m\Omega^2\mathrm{c}}. \tag{121}$$

The overall precision is the quadratic sum of the two combinations and has a minimum that is frequency dependent. This is the "Standard Quantum Limit" (SQL), the macroscopic manifestation of Heisenberg's uncertainty principle anticipated before.

This is a noise of fundamental origin. As mentioned above, there are developments that allow to better measure the phase while tolerating a larger amplitude fluctuation, still verifying the relation

$$\sigma_\phi\sigma_N \geq \frac{1}{2} \tag{122}$$

that originates from the quantum nature of light.

A factor of 10 has been achieved in the laboratory [24] while the applicability of these techniques to kilometre-size interferometric detectors was demonstrated, although with much smaller performance [25]. Even a factor of 4 would otherwise require increasing the laser power by a factor of 16!

7.2 Seismic and Newtonian Noise

As discussed above, gravitational wave detection requires to observe the relative motion of masses that mark geodesic lines in space-time. This requires that motion due to local perturbations is reduced below something like 10^{-19} m. This very small value is not really appreciated until one learns that, in 1 s, ground motion is of the order of 10^{-7} m. Fortunately, this displacement spectrum decreases with frequency; a typical parametrization of "seismic noise" is given by

$$\tilde{x}(f) = 10^{-7} \frac{1\,\mathrm{Hz}}{f^2}\,\mathrm{mHz}^{-1/2}. \tag{123}$$

Devices for seismic attenuation have to be introduced that should also filter all sort of mechanical vibrations present in a laboratory. Here, the solution adopted for the VIRGO interferometer is presented.

The VIRGO "Superattenuator" is based on the behaviour of a pendulum in presence of a displacement of the suspension point. For small angles, one obtains for the oscillation amplitude

$$|x(\omega)| = x_0 \frac{\omega_0}{\sqrt{(\omega^2 - \omega_0^2)^2 + \omega^2 \Gamma^2}}, \tag{124}$$

where x_0 is the oscillation amplitude of the suspension point, ω_0 is the pendulum angular frequency, and Γ measures dissipation, assuming viscous damping. One sees that for $\omega \gg \omega_0$ the response decreases as $(\omega_0/\omega)^{-2}$, decoupling the mass from the suspension point. By using a cascade of several pendula, sufficient attenuation can be achieved already at a frequency of a few hertz.

Experimentally, it turns out to be extremely difficult to achieve orders of magnitude of attenuation in one degree of freedom without acting on the others. The pendulum can attenuate in the horizontal direction, and rotation degrees of freedom can have quite a low resonant frequency if the pendulum body has large moments of inertia. The vertical degree of freedom is a challenge; passive attenuation is achieved using a mass upon which acts a restoring force that has to be as gentle as possible. This would lead to a very soft spring that would have to elongate in an unpractical way. The chosen solution is to use a strong supporting spring in parallel with an unstable device (with a negative elastic constant) acting over a small position interval. The two "elastic constants" are finely tuned to achieve locally a resonant frequency below 1 Hz. In VIRGO, this is achieved using magnets in a repulsive configuration, to obtain a force in a small interval:

$$F(y) = + ky, \tag{125}$$

where y is the deviation from the unstable equilibrium position.

A device based on the same principle is introduced to suspend the full pendulum chain. An inverted pendulum, in our case, is a set of three beams connected to the ground by means of elastic joints supports a platform to which the superattenuator

is hooked. The beams would fall under the weight of the pendula but are prevented to do so by the elastic joints that generate a restoring momentum. By tuning the weight an oscillation frequency of 30 mHz has been achieved, introducing additional attenuation in the horizontal direction. The resulting restoring force is therefore very low.

This allows to make use of the suspension to compensate for another effect, the elongation of the Earth's crust due to tides, which amounts up to 60 μm with spring tides. The force needed to displace 1 t is low and therefore less noisy. It is applied upstream the attenuation chain so that the tide compensation at the level of the mirror can be performed without introducing additional noise. The dynamic range of the device thus spans more than 12° of freedom.

It was possible to measure the attenuation of this system only by making full use of VIRGO sensitivity and stability, averaging for hours the output of the interferometer with a force applied at the top of the pendulum chain. The results [26] indicate that seismic noise is filtered at a frequency as low as 4 Hz, well below the VIRGO observation band of 10 Hz ÷ 10 kHz.

Local gravity fluctuations, which generate a gravity gradient, mimic a gravitational wave signal. These fluctuations come from the motion of local masses and generate what is called "Newtonian noise". Some effects can be avoided by keeping moving foreign bodies away from the test masses. However, the motion of large masses can act at a distance. One example is atmospheric density variation, that occurs fortunately on a time scale below the second. More important are surface waves, that are a manifestation of seismic activity. A test mass placed above the ground will be attracted more by the crest of the wave than by the trough, and the resulting force will change with the wave passage. The force is proportional to the seismic amplitude and using the seismic noise spectrum of Eq. 123 and taking into account the transfer function from force to acceleration one deduces an ω^{-4} frequency dependence, meaning that this noise becomes important when attempting detection at low frequency. This can be reduced by going underground, where mass is distributed all around the test mass. In that case, a density variation at the passage of the wave still generates a force on the test mass, but the resulting amplitude is much lower. However, for a detector above ground the effect is relevant up to several hertz.

7.3 Thermal Noise

The test masses are at a finite temperature and have thus some thermal energy. Although being macroscopic objects with a mass of several tens of kilogram, the amplitude of the motion along the light beam direction is sufficiently large to limit seriously the performance of current interferometers. The thermal pendulum motion limits sensitivity in the tens of hertz band. At medium and high frequency (100 ÷ 10000 Hz), the vibration modes of the mirrors introduce fluctuation of the

reflecting surface with respect to the centre of mass of the mirror, spoiling the measurement of the separation of the geodesics. Noise comes from thermodynamic fluctuations that disturb an otherwise predictable motion; for example, the pendulum oscillation. Thermodynamic fluctuations have the effect of erasing the memory of the initial conditions of the system. The rate of thermodynamic fluctuations is related to the rate of dissipation of the system. The dissipation mechanism for energy flowing toward the thermal bath is the same through which the thermal bath introduces energy fluctuations in the mechanical system. This is summarized by the fluctuation-dissipation theorem, that relates the power spectrum of velocity of the system to the strength of the dissipation mechanism.

Consider an harmonic oscillator with the equation of motion

$$m\ddot{x} + \Gamma\dot{x} + kx = F(t). \tag{126}$$

If a force is applied on the mass, going into the frequency domain the transfer function from force to velocity is

$$Y(\omega) = \frac{v(\omega)}{F(\omega} = \frac{1}{Z(\omega)} = \frac{\omega}{m}\frac{-\omega_0^2 + \omega^2 - i\omega\gamma}{(-\omega_0^2 + \omega^2)^2 + \gamma^2\omega^2}, \tag{127}$$

where $\gamma = \Gamma/m$ and $\omega_0^2 = k/m$.

The fluctuation dissipation theorem states that the velocity power spectrum is proportional to the absolute temperature T and to the real part of the admittance $Y(\omega) = 1/Z(\omega)$ of the system

$$S_v(\omega) = \Re(Y(\omega))4k_BT, \tag{128}$$

where k_B is the Boltzmann constant and $S_v(\omega)$ is the one-sided power spectrum that sums both positive and negative ω contributions.

The thermal noise for the position of this oscillator is

$$\tilde{x}(\omega) = \sqrt{\frac{1}{\omega^2}S_v(\omega)} = \sqrt{\frac{4k_bT\gamma}{m((-\omega_0^2 + \omega^2)^2 + \gamma^2\omega^2)}}. \tag{129}$$

Thermal noise can be greatly reduced outside the resonance region by reducing as much as possible γ. This amounts to design very low dissipation mechanical systems with a high-merit figure, $Q = \omega_0/\gamma$. By use of suitable materials and configurations, Q of $10^7 \div 10^8$ has been achieved.

The presence of a pendulum to suspend the mirror is a great advantage. Gravitation does not dissipate (but for negligible gravitational wave emission...), so that dissipation occurs only where the pendulum wire flexes, and little energy is available for dissipation in that flexure. High Q pendula have been achieved using low-dissipation silica fibers. Mirrors themselves are made of silica and have a very high Q, but these mirrors have to be coated to create reflective layers, and as of today these layers introduce a dissipation that determines thermal noise between 100 and 10,000 Hz. Intense research is going on to increase the Q factor of interferometer mirrors.

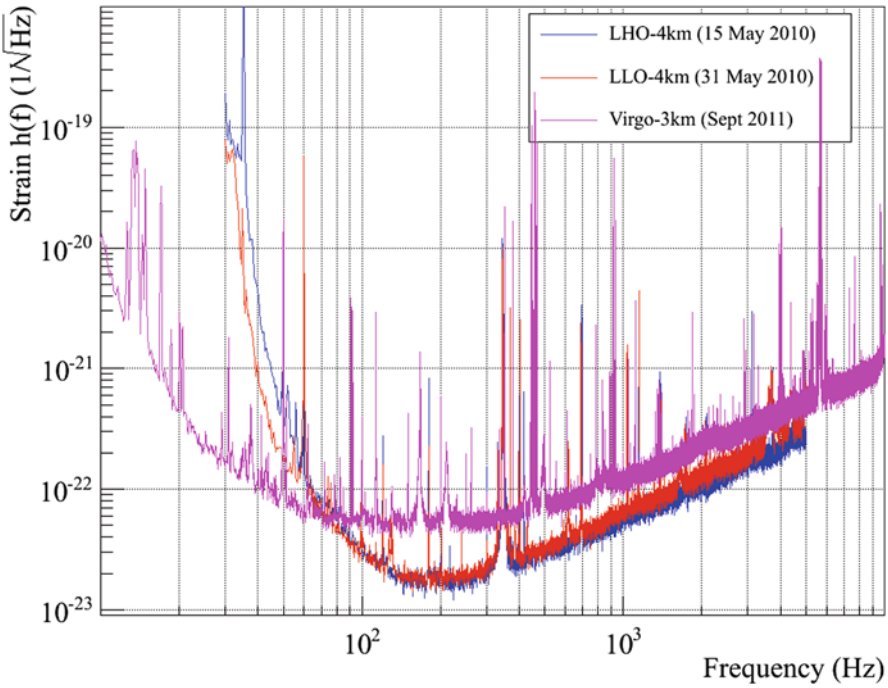


Fig. 7 Sensitivity of the LIGO and VIRGO interferometers

7.4 Typical Interferometer Sensitivity

The resulting sensitivity of a typical interferometer is shown in Fig. 7.

Detailed studies of the various sources of noise, fundamental and technical, indicate that the observed spectrum is essentially understood, giving confidence that upgrades currently being done rely on a solid base.

This noise curve is used to compute an "horizon", that is the distance at which a binary neutron star system can be detected, assuming an amplitude that is averaged over all possible orientations of the system and the detector. To ensure a false alarm rate much less than one event per year a signal-to-noise ratio larger than 8 is required. That curve indicates an horizon of almost 12 Mpc. Advanced detectors now under construction have been designed to reach an horizon of $130 \div 200$ Mpc, for the 3-km VIRGO and the 4-km LIGO, respectively.

8 Ground-Based Interferometer Network

The signal recorded by a single interferometer does not allow to measure all the parameters of the incoming gravitational wave at a given time. One has to determine the incoming direction and the amplitude of the two polarizations. In contrast with

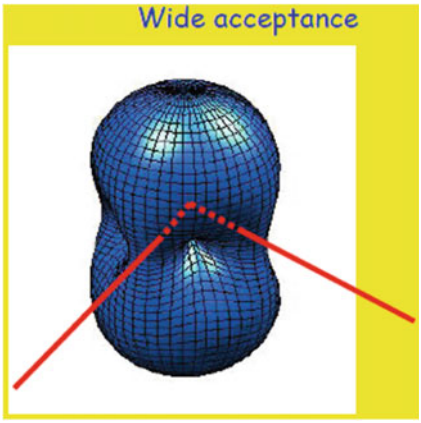


Fig. 8 Response of an interferometer to a nonpolarized gravitational wave as function of direction

a telescope, the interferometer is sensitive to waves coming from any direction, except for those traveling along a symmetry axis or along one arm for the + and x polarizations, respectively. The antenna pattern for nonpolarized waves is shown in Fig. 8.

To reconstruct the incoming direction of a gravitational wave, several interferometers have to detect it so that the time of arrival at the various sites can be recorded, with an expected precision of the order of 1 ms. For two interferometers separated by an Earth radius the resulting angular precision on the plane defined by the wave vector and the segment connecting the two interferometers is

$$\sigma_\theta \sim \frac{c\sigma_t}{R_\oplus} \simeq 5 \times 10^{-2}\,\text{rad}, \tag{130}$$

which is about 3°. Detailed studies have shown that with three interferometers an acceptable coverage can be achieved, with some directions that are less precisely measured. In the optimal case, error boxes of 3×3 degrees are expected. Adding a fourth interferometer improves significantly precision and coverage [27].

Instruments that are sensitive to any direction in the sky can record sources from any direction. As opposed to the electromagnetic telescopes, they are not blind to waves that have to cross the Earth. A network of interferometers, while recording the transit of a gravitational wave, can send alerts to instruments that need to be pointed, like telescopes and detectors on satellites. In fact in violent events, that are anyway triggered by gravity, a gravitational wave signal should be the first component to be emitted (but for possible precursors). Other kinds of more interacting radiation (neutrinos and photons from the whole electromagnetic spectrum) will be emitted once the outgoing energy flow reaches a region transparent to radiation.

Triggered on gravitational wave candidate events, a permanent watch for violent transient events can be organized. Being able to record what is happening at the very hearth of violent events is a unique opportunity for astrophysics, apart from allowing to study gravitational radiation by itself. However, the resources needed are very large, both for the construction and for the commissioning, operation and

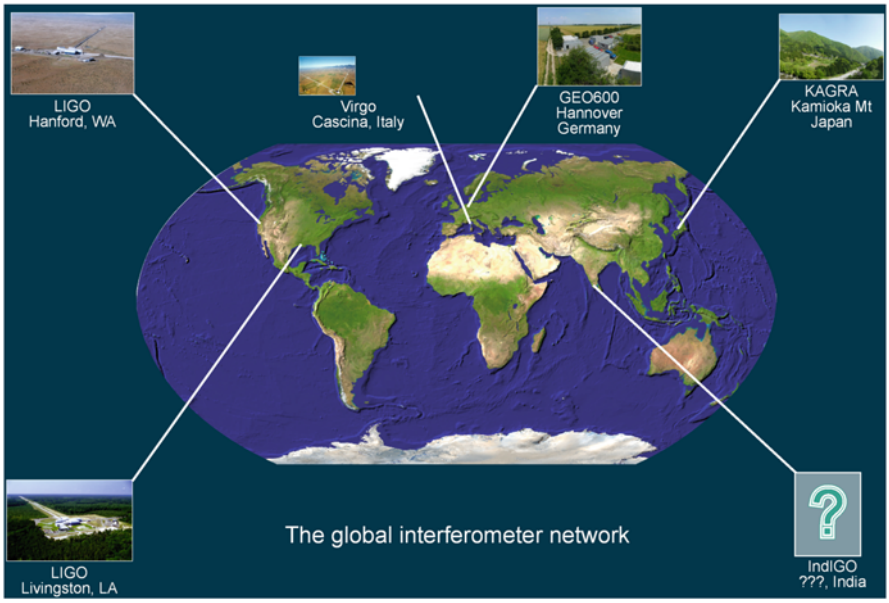


Fig. 9 The gravitational wave network

data analysis of the instruments. This has led to establish a network of interferometers that share their data and are able to perform a low latency analysis to alert astronomical observatories. Currently, the network is composed of the two LIGO interferometers located in Louisiana and Washington, VIRGO in Italy and GEO600 in Germany. In Japan, the construction of KAGRA has just started, while discussions are at different stages for an interferometer in Australia and in India. With a full network complete observation will be possible, maximizing the live time of the network while allowing for maintenance of the instruments. Agreements between the various collaborations are such that all members of the network will participate in producing scientific results. The situation is summarized in Fig. 9 showing the gravitational wave network.

Current plans are for advanced LIGO and VIRGO detectors to start operation in 2015. By 2020 the full network might be operating at a respectable sensitivity, recording tens of events per year.

9 Space Detectors

The detection of GW by ground-based detectors is limited at low frequency by gravity gradient fluctuations caused by local masses. This can be avoided by having masses in orbit in space, in perfect free fall, far away from local gravity fluctuations. This idea has been elaborated imagining to locate an interferometer in space,

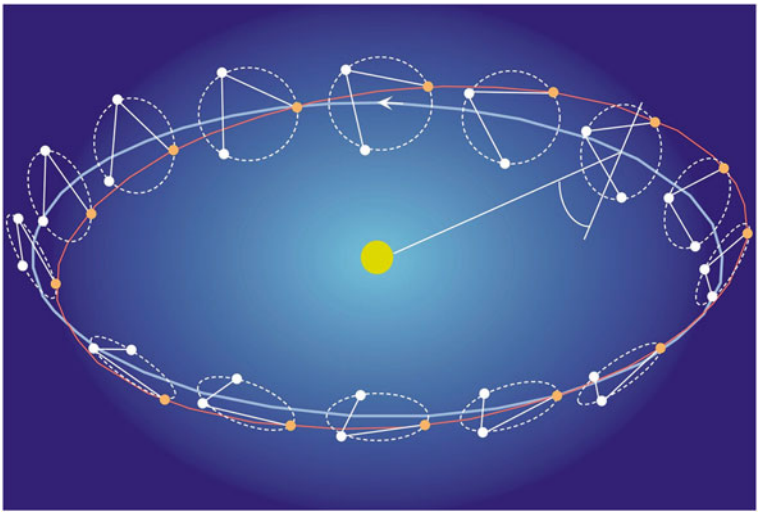


Fig. 10 The orbits of the three eLISA spacecrafts

orbiting around the Earth or around the Sun. The arm length could be from several hundred thousand kilometres up to a couple of astronomical units, spanning a frequency band from 1 Hz down to 0.1 mHz. In the galaxy, there are many galactic binary systems that emit gravitational wave in this band with a comfortable amplitude. But most interesting would be the possibility to record the inspiral and coalescence of massive (million solar masses) black holes. These events, coming from extremely powerful sources, could be detected at cosmological distance with high signal-to-noise ratio providing precise signals to be compared against gravity theories in a stronger field.

The largest effort toward achieving a space gravitational observatory was devoted jointly by NASA and ESA to laser interferometer space antenna (LISA). This has now evolved into *evolved* LISA (eLISA), recently inserted by ESA in their long-term planning for the L3 [28] launch slot, in 2034. The detector test masses will be housed in three spacecraft in heliocentric orbit lagging behind the Earth, staying 1 million km from one another in a triangular constellation, as shown in Fig. 10. Light will be received at the end of the arm and will drive a local source acting as a transponder, otherwise the returning beam would be too faint. Phases of arrival will be compared as in a usual interferometer but fringes will not be stable as eLISA arms will be affected by length changes.

By staying sufficiently far from the Earth, the residual acceleration noise on the test masses is at the level of 3×10^{-15} m s^{-2} below 3 mHz. On the high-frequency side, the distance measurement, including shot noise, is at the level of 12 pm Hz$^{-1/2}$ or in strain 10^{-21} Hz$^{-1/2}$.

Space operation, in addition to many technical constraints, exposes the masses to radiation pressure. This is enough to disturb significantly the motion of test masses

in free fall that have to measure a differential acceleration of the order of 10^{15} m s^{-2}. The spacecraft must act as a shield against these external agents, creating a "drag-free" environment. This is achieved by having the spacecraft, whose motion is influenced by the solar wind, track its position relative to the shielded test mass that ideally is following a geodesic to within 2 nm $Hz^{-1/2}$. Corrections to the spacecraft motion are applied by means of microthrusters. The nominal test mass position inside the spacecraft is such that forces exerted by the spacecraft itself perturb negligibly its free-fall motion.

Due to the extremely ambitious goals and the high cost of the space mission, a test mission called LISA Pathfinder, to demonstrate in particular the concept of drag-free satellite and interferometry at very low frequency, is expected to be launched in 2015.

When eLISA will be launched unique information shall be collected from the band 0.1 mHz $\div$1 Hz. However 10^{-4} Hz might not be low enough to detect the cosmological background, possibly more accessible at much lower frequencies. This has led to an effort based on the use of pulsar timing arrays like EPTA [29], NANOGrav [30] and PPTA [31], to attempt detecting stochastic background profiting of the extreme regularity of pulse emission shown by several millisecond pulsar. The one-way timing of pulses emitted with high regularity allows to detect metric variations. In this case, the arm length is of the order of 10 kpc and the frequency band is around 10^{-8} Hz, limited by the observation time. Such a stochastic background can be detected by observing correlations in the fluctuations of arrival times of pulses from different sources. This is the largest "interferometer" in space that has been imagined.

Finally, an effort to detect traces of the interaction of primordial gravitational wave and electromagnetic background is in progress. Here, the primary detector is the whole early universe; the polarization distribution of the microwave background is predicted to have been influenced by the quadrupolar nature of gravitational background. The latest data on polarization fluctuations were collected by the BICEP2 instrument [32] and by the Planck mission [33]. Very recently, the BICEP2 Collaboration published the detection of polarization patterns (B-mode) that can be attributed to primordial GW [34]. However, due to the presence of foreground contamination further analysis is needed to establish with confidence the result. The analysis of Planck data showed that the BICEP2 observations could as well be originated by foreground contamination.

10 Bar Detectors

Detection of GW by means of a resonant detector was first attempted by Joseph Weber. The principle is that a gravitational wave transit results in forces being applied to masses separated by some distance ξ, as shown in Eq. 54.

For a simple detector made of two masses connected by a spring, the equation of motion for the mass separation is

$$\xi_{,00} + 2\gamma\xi_{,0} + \omega_0^2\xi = 0, \tag{131}$$

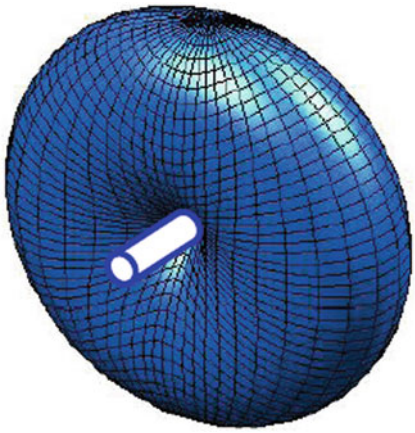


Fig. 11 The antenna pattern for a resonant bar

where $\omega_0^2 = k/2m$ and $\gamma = \Gamma/m$. Then one has to find the proper length of the spring $l(t)$, to be compared with the rest length l_0:

$$l(t) = (x_2(t) - x_1(t))\left[1 + \frac{1}{2}h_{xx}^{TT}(t)\right] + O(|h_{\mu\nu}|^2). \tag{132}$$

To first order in $h_{\mu\nu}$ one has

$$\xi(t) = x_2 - x_1 - l_0 + \frac{1}{2}(x_2 - x_1)h_{xx}^{TT} + O(|h_{\mu\nu}|^2), \tag{133}$$

then the equation of motion for the separation becomes

$$\xi_{,00} + 2\gamma\xi_{,0} + \omega_0^2\xi = \frac{1}{2}l_0 h_{xx}^{TT} \tag{134}$$

to first order in $h_{\mu\nu}$.

For a resonant bar, one has a continuum of masses and springs and the equation of motion for the end of the bar becomes [35]

$$\xi_{,00} + 2\gamma\xi_{,0} + \omega_0^2\xi = \frac{1}{2\pi^2}l_0 h_{xx}^{TT} \tag{135}$$

The antenna pattern is shown in Fig. 11.

These first attempts to detect GW stimulated the ingenuity of many other researchers toward attempting gravitational wave detection. After intense efforts, various groups were able to operate cryogenic resonant bars, with significantly reducing the thermal noise. A typical sensitivity is shown in Fig. 12.

As can be seen, the bar is most sensitive in a frequency band of approximately 100 Hz around 900 Hz. This is well suited to detect a supernova core collapse that has a timescale of the order of 1 ms and also try to catch some millisecond pulsar. In the first case, an international effort by five bars was set up, the International

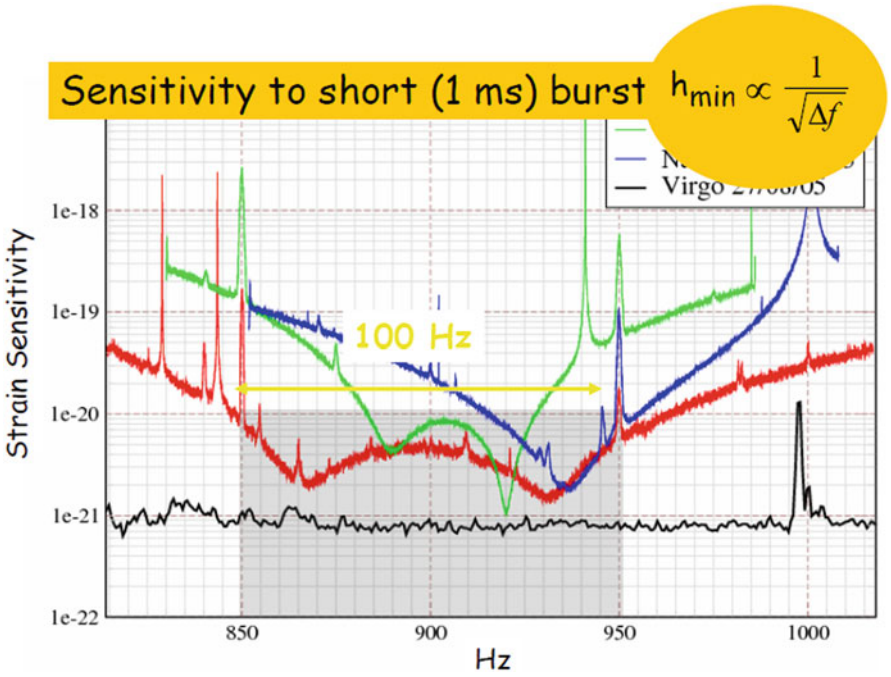


Fig. 12 The sensitivity achieved by the Auriga resonant bar

Gravitational Event collaboration (IGEC), composed by ALLEGRO [36], AURIGA [37], EXPLORER and NAUTILUS [38] and NIOBE [39]. The full configuration ran from 1997 until 2000, setting an upper limit to the rate of gravitational wave bursts as a function of their amplitude [40].

Further studies have shown the interest of having a spherical resonant detector, whose quadrupolar vibration modes would be excited by a gravitational wave allowing reconstruction by a single detector of the wave incoming direction and its polarization. Two such detectors have been built: MiniGRAIL [41] and Schenberg [42]. However, as of today, it appears that the advantage given by the arm length of interferometric detectors cannot easily be matched by resonant detectors, unless one considers gravitational wave frequencies much above the kilohertz, a band that seems of no astrophysical interest, but that it contains a cosmological background is not excluded at the moment.

11 Summary

The existence of GW appears to have a strong support. A theory that has seen many verifications as of today predicts them, and a solid astronomical evidence has been accumulating for more than 30 years. The understanding of the universe content

has also made many steps forward since the first attempts to detect gravitational radiation. Rates are still uncertain by two or three orders of magnitude, but the expected performance of second generation interferometers is such that a first detection may happen soon after begin of operation. Detectors themselves have evolved from prototypes to reliable instruments that are suitable to continuously listen to the universe. In the meantime, detector improvement is continuing, achieving what was only described in visionary theoretical papers, like manipulating vacuum on the kilometre scale to improve photon shot noise. This progress will be very rewarding: Every factor in noise decrease will expand the volume of universe by the cube of that factor, and a handful of events will turn into hundreds starting from gravitational wave astronomy.

12 Acknowledgements

I tried to follow the trail of Stefano Braccini's lectures, knowing well that out of the work of many decades a selection was made. I would like to thank Roberto Peron for his patient assistance. Stefano was very busy within the gravitational wave community and had the opportunity to exchange many thoughts with many of its members. Therefore, I would like to thank Stefano and all the members for their support. Stefano had a passion for his work and for science and for taking difficult challenges. His efforts were essential to reach the nominal performance of the VIRGO detector, helping building confidence in the possibility of detecting GW. The gravitational wave community is missing him.

References

1. Einstein A. Die grundlage der allgemeinen relativitätstheorie. Ann Phys. 1916;49:769–822.
2. Misner CW, Thorne KS, Wheeler JS. Gravitation. San Francisco:WH Freeman; 1973.
3. Schutz BF. A first course in general relativity. 2nd edn. Cambridge: Cambridge University Press; 2009.
4. Will CM. Theory and experiment in gravitational physics. Revised edn. Cambridge: Cambridge University Press; 1993.
5. Einstein A. Näherungsweise integration der feldgleichungen der gravitation. Preuss. Akad. Wiss. Berlin Sitzber: p. 688. 1916.
6. Peters PC, Mathews J. Gravitational radiation from point masses in a Keplerian orbit. Phys Rev. 1963;131:435–40. doi:10.1103/PhysRev.131.435.
7. Lense J, Thirring H. Uber den einfluss der eigenrotation der zentralkorper auf die bewegung der planeten und monde nach der einsteinschen gravitationstheorie. Phys Z. 1918;19:156–63. Translation in Gen Relativ Gravit 1984;16:727–41.
8. Peron R. Fundamental Physics with the LAGEOS Satellites. In: Peron R, Gorini V, Moschella U, editors. Gravity: where do we stand? Berlin: Springer; 2016, p. 167 (Chap.4 this volume).
9. Ciufolini I, Pavlis EC. A confirmation of the general relativistic prediction of the Lense–Thirring effect. Nature. 2004;431(7011):958–60. doi:dx.doi.org/10.1038/nature03007.

10. Everitt CWF, DeBra DB, Parkinson BW, Turneaure JP, Conklin JW, Heifetz MI, Keiser GM, Silbergleit AS, Holmes T, Kolodziejczak J, Al-Meshari M, Mester JC, Muhlfelder B, Solomonik VG, Stahl K, Worden PW, Bencze W, Buchman S, Clarke B, Al-Jadaan A, Al-Jibreen H, Li J, Lipa JA, Lockhart JM, Al-Suwaidan B, Taber M, Wang S. Gravity probe b: Final results of a space experiment to test general relativity. Phys Rev Lett. 2011;106:221101. doi:10.1103/PhysRevLett.106.221101.
11. Abadie J, Abbott BP, Abbott R, Abernathy M, Accadia T, Acerneseac F, Adams C, Adhikari R, Ajith P, Allen B, Allen G, Ceron EA, Amin RS, Anderson SB, Anderson WG, Antonuccia F, Aoudiaa S, Arain MA, Araya M, Aronsson M, Arun KG, Aso Y, Aston S, Astonea P, Atkinson DE, Aufmuth P, Aulbert C, Babak S, Baker P, Ballardin G, Ballmer S, Barker D, Barnum S, Baroneac F, Barr B, Barriga P, Barsotti L, Barsuglia M, Barton MA, Bartos I, Bassiri R, Bastarrika M, Bauchrowitz J, Bauera TS, Behnke B, Beker MG, Benacquista M, Bertolini A, Betzwieser J, Beveridge N, Beyersdorf PT, Bigottaab S, Bilenko IA, Billingsley G, Birch J, Birindellia S, Biswas R, Bitossi M, Bizouard MA, Black E, Blackburn JK, Blackburn L, Blair D, Bland B, Bloma M, Blomberg A, Boccara C, Bock O, Bodiya TP, Bondarescu R, Bondu F, Bonelli L, Bork R, Born M, Bose S, Bosi L, Boyle M, Braccini S, Bradaschia C, Brady PR, Braginsky VB, Brau JE, Breyer J, Bridges DO, Brillet A, Brinkmann M, Brisson V, Britzger M, Brooks AF, Brown DA, Budzynski R, Bulik T, Bulten HJ, Buonanno A, Burguet-Castell J, Burmeister O, Buskulic D, Byer RL, Cadonati L, Cagnoli G, Calloni E, Camp JB, Campagna E, Campsie P, Cannizzo J, Cannon KC, Canuel B, Cao J, Capano C, Carbognani F, Caride S, Caudill S, Cavagli M, Cavalier F, Cavalieri R, Cella G, Cepeda C, Cesarini E, Chalermsongsak T, Chalkley E, Charlton P, Mottin EC, Chelkowski S, Chen Y, Chincarini A, Christensen N, Chua SSY, Chung CTY, Clark D, Clark J, Clayton JH, Cleva F, Coccia E, Colacino CN, Colas J, Colla A, Colombini M, Conte R, Cook D, Corbitt TR, Corda C, Cornish N, Corsi A, Costa CA, Coulon JP, Coward D, Coyne DC, Creighton JDE, Creighton TD, Cruise AM, Culter RM, Cumming A, Cunningham L, Cuoco E, Dahl K, Danilishin SL, Dannenberg R, D'Antonio S, Danzmann K, Dari A, Das K, Dattilo V, Daudert B, Davier M, Davies G, Davis A, Daw EJ, Day R, Dayanga T, De Rosa R, Debra D, Degallaix J, Del Prete M, Dergachev V, DeRosa R, DeSalvo R, Devanka P, Dhurandhar S, Di Fiore L, Di Lieto A, Di Palma I, Emilio MDP, Di Virgilio A, Diaz M, Dietz A, Donovan F, Dooley KL, Doomes EE, Dorsher S, Douglas ESD, Dragocd M, Drever RWP, Driggers JC, Dueck J, Dumas JC, Eberle T, Edgar M, Edwards M, Effler A, Ehrens P, Engel R, Etzel T, Evans M, Evans T, Fafone V, Fairhurst S, Fan Y, Farr BF, Fazi D, Fehrmann H, Feldbaum D, Ferrante I, Fidecaro F, Finn LS, Fiori I, Flaminio R, Flanigan M, Flasch K, Foley S, Forrest C, Forsi E, Fotopoulos N, Fournier JD, Franc J, Frasca S, Frasconi F, Frede M, Frei M, Frei Z, Freise A, Frey R, Fricke TT, Friedrich D, Fritschel P, Frolov VV, Fulda P, Fyffe M, Gammaitoni L, Garofoli JA, Garufiab F, Gemme G, Genin E, Gennai A, Gholami I, Ghosh S, Giaime JA, Giampanis S, Giardina KD, Giazotto A, Gill C, Goetz E, Goggin LM, Gonzalez G, Gorodetsky ML, Gossler S, Gouaty R, Graef C, Granata M, Grant A, Gras S, Gray C, Greenhalgh RJS, Gretarsson AM, Greveriea C, Grosso R, Grote H, Grunewald S, Guidi GM, Gustafson EK, Gustafson R, Hage B, Hall P, Hallam JM, Hammer D, Hammond G, Hanks J, Hanna C, Hanson J, Harms J, Harry GM, Harry IW, Harstad ED, Haughian K, Hayama K, Heefner J, Heitmann H, Hello P, Heng IS, Heptonstall A, Hewitson M, Hild S, Hirose E, Hoak D, Hodge KA, Holt K, Hosken DJ, Hough J, Howell E, Hoyland D, Huet D, Hughey B, Husa S, Huttner SH, Huynh-Dinh T, Ingram DR, Inta R, Isogai T, Ivanov A, Jaranowski P, Johnson WW, Jones DI, Jones G, Jones R, Ju L, Kalmus P, Kalogera V, Kandhasamy S, Kanner J, Katsavounidis E, Kawabe K, Kawamura S, Kawazoe F, Kells W, Keppel DG, Khalaidovski A, Khalili FY, Khazanov EA, Kim C, Kim H, King PJ, Kinzel DL, Kissel JS, Klimenko S, Kondrashov V, Kopparapu R, Koranda S, Kowalska I, Kozak D, Krause T, Kringel V, Krishnamurthy S, Krishnan B, Krolak A, Kuehn G, Kullman J, Kumar R, Kwee P, Landry M, Lang M, Lantz B, Lastzka N, Lazzarini A, Leaci P, Leong J, Leonor I, Leroy N, Letendre N, Li J, Li TGF, Lin H, Lindquist PE, Lockerbie NA, Lodhia D, Lorenzinia M, Lorietteb V, Lormand M, Losurdo G, Lu P, Luan J, Lubinski M, Lucianetti A, Lueck H, Lundgren A, Machenschalk B, MacInnis M, Mackowski JM, Mageswaran M,

Mailand K, Majoranaa E, Mak C, Man N, Mandel. Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors. Classical Quant Grav 2010;27(17):173001. doi:10.1088/0264-9381/27/17/173001.
12. Ott CD. Probing the core-collapse supernova mechanism with gravitational waves. Classical Quant Grav 2009;26(20):204015.
13. Possenti A, Burgay M. The Role of binary pulsars in testing gravity theories In: Peron R, Gorini V, Moschella U, editors. Gravity: where do we stand? Berlin: Springer; 2016, p. 279 (Chap.8, this volume).
14. Smith TL, Pierpaoli E, Kamionkowski M. New cosmic microwave background constraint to primordial gravitational waves. Phys Rev Lett. 2006;97:021301. doi:10.1103/PhysRevLett.97.021301.
15. Riess AG, Macri L, Casertano S, Lampeitl H, Ferguson HC, Filippenko AV, Jha SW, Li W, Chornock R. A 3% solution: determination of the Hubble constant with the Hubble space telescope and wide field camera 3. Astrophys J. 2011;730(2):119.
16. Allen B, Romano JD. Detecting a stochastic background of gravitational radiation: signal processing strategies and sensitivities. Phys Rev D. 1999;59:102001. doi:10.1103/PhysRevD.59.102001.
17. Wex N. A timing formula for main-sequence star binary pulsars. MNRAS. 1998;298(1):67–77. doi:10.1046/j.1365-8711.1998.01570.x.
18. Weisberg JM, Nice DJ, Taylor JH. Timing measurements of the relativistic binary pulsar psr b1913+16. Astrophys J. 2010;722(2):1030.
19. Heasarc pulsar catalogue. http://heasarc.gsfc.nasa.gov/W3Browse/radio-catalog/pulsar.html.
20. Papoulis A, Pillai SU. Probability, random variables and stochastic processes. New York: McGraw-Hill; 2001.
21. Pound RV. Electronic frequency stabilization of microwave oscillators. Rev Sci Instrum. 1946;17(11):490–505. doi:10.1063/1.1770414.
22. Drever RWP, Hall JL, Kowalski FV, Hough J, Ford GM, Munley AJ, Ward H. Laser phase and frequency stabilization using an optical resonator. Appl Phys B. 1983;31:97–105.
23. Black ED. An introduction to Pound–Drever–Hall laser frequency stabilization. Am J Phys. 2001;69(1):79–87. doi:10.1119/1.1286663.
24. Vahlbruch H, Mehmet M, Chelkowski S, Hage B, Franzen A, Lastzka N, Goßler S, Danzmann K, Schnabel R. Observation of squeezed light with 10-db quantum-noise reduction. Phys Rev Lett. 2008;100:033602. doi:10.1103/PhysRevLett.100.033602.
25. Abadie J et al. A gravitational wave observatory operating beyond the quantum shot-noise limit. Nat Phys. 2011;7:962–5. doi:10.1038/nphys2083.
26. Acernese F et al. Measurements of superattenuator seismic isolation by virgo interferometer. Astropart Phys. 2010;33:182–9.
27. Fairhurst S. Source localization with an advanced gravitational wave detector network. Classical Quant Grav 2011;28(10):105021.
28. Amaro-Seoane P, e.a. The gravitational universe. 2013. arXiv:1305.5720 [astro-ph.CO].
29. van Haasteren R, Levin Y, Janssen GH, Lazaridis K, Kramer M, Stappers BW, Desvignes G, Purver MB, Lyne AG, Ferdman RD, Jessner A, Cognard I, Theureau G, D'Amico N, Possenti A, Burgay M, Corongiu A, Hessels JWT, Smits R, Verbiest JPW. Placing limits on the stochastic gravitational-wave background using European Pulsar Timing Array data. MNRAS. 2011;414:3117–28. doi:10.1111/j.1365-2966.2011.18613.x.
30. Demorest, PB, Ferdman RD, Gonzalez ME, Nice D, Ransom S, Stairs IH, Arzoumanian Z, Brazier A, Burke-Spolaor S, Chamberlin SJ, Cordes JM, Ellis J, Finn LS, Freire P, Giampanis S, Jenet F, Kaspi VM, Lazio J, Lommen AN, McLaughlin M, Palliyaguru N, Perrodin D, Shannon RM, Siemens X, Stinebring D, Swiggum J, Zhu WW. Limits on the stochastic gravitational wave background from the North American nanohertz observatory for gravitational waves. Astrophys J. 2013;762(2):94.
31. Manchester RN, Hobbs G, Bailes M, Coles WA, van Straten W, Keith MJ, Shannon RM, Bhat NDR, Brown A, Burke-Spolaor SG, Champion DJ, Chaudhary A, Edwards RT, Hampson G, Hotan AW, Jameson A, Jenet FA, Kesteven MJ, Khoo J, Kocz J, Maciesiak K, Oslowski

S, Ravi V, Reynolds JR, Sarkissian JM, Verbiest JPW, Wen ZL, Wilson WE, Yardley D, Yan WM, You XP. The Parkes pulsar timing array project. PASA - Publications of the Astronomical Society of Australia. 2013;30. doi:10.1017/pasa.2012.017.
32. BICEP2 Collaboration. Bicep2 ii: Experiment and three-year data set. 2014. arXiv:1403.4302 [astro-ph.CO].
33. Lamarre JM, Puget JL, Bouchet F, Ade PAR, Benoit A, Bernard JP, Bock J, De Bernardis P, Charra J, Couchot F, Delabrouille J, Efstathiou G, Giard M, Guyot G, Lange A, Maffei B, Murphy A, Pajot F, Piat M, Ristorcelli I, Santos D, Sudiwala R, Sygnet JF, Torre JP, Yurchenko V, Yvon D. The Planck High Frequency Instrument, a third generation CMB experiment, and a full sky submillimeter survey. New Astro Rev. 2003;47:1017–24. doi:10.1016/j.newar.2003.09.006.
34. Ade PAR, Aikin RW, Barkats D, Benton SJ, Bischoff CA, Bock JJ, Brevik JA, Buder I, Bullock E, Dowell CD, Duband L, Filippini JP, Fliescher S, Golwala SR, Halpern M, Hasselfield M, Hildebrandt SR, Hilton GC, Hristov VV, Irwin KD, Karkare KS, Kaufman JP, Keating BG, Kernasovskiy SA, Kovac JM, Kuo CL, Leitch EM, Lueker M, Mason P, Netterfield CB, Nguyen HT, O'Brient R, Ogburn RW, Orlando A, Pryke C, Reintsema CD, Richter S, Schwarz R, Sheehy CD, Staniszewski ZK, Sudiwala RV, Teply GP, Tolan JE, Turner AD, Vieregg AG, Wong CL, Yoon KW. Detection of *b*-mode polarization at degree angular scales by bicep2. Phys Rev Lett. 2014;112:241101. doi:10.1103/PhysRevLett.112.241101.
35. McHugh MP, Johnson WW, Hamilton WO, Hanson J, Heng IS, McNeese D, Miller P, Nettles D, Weaver J, Zhang P. Calibration of the allegro resonant detector. Classical Quant Grav 2005;22(18):S965.
36. Mauceli E, Geng ZK, Hamilton WO, Johnson WW, Merkowitz S, Morse A, Price B, Solomonson N. The allegro gravitational wave detector: data acquisition and analysis. Phys Rev D. 1996;54:1264–75. doi:10.1103/PhysRevD.54.1264.
37. Vinante A the AURIGA Collaboration. Present performance and future upgrades of the auriga capacitive readout. Classical Quant Grav. 2006;23(8):S103.
38. Astone P, Ballantini R, Babusci D, Bassan M, Carelli P, Cavallari G, Cavanna F, Chincarini A, Coccia E, Cosmelli C, D'Antonio S, Dubath F, Fafone V, Foffa S, Gemme G, Giordano G, Maggiore M, Marini A, Minenkov Y, Modena I, Modestino G, Moleti A, Murtas GP, Pai A, Palamara O, Pallottino GV, Parodi R, Mortari GP, Pizzella G, Quintieri L, Rocchi A, Ronga F, Sturani R, Terenzi R, Torrioli G, Vaccarone R, Vandoni G, Visco M. Status report on the explorer and nautilus detectors and the present science run. Classical Quant Grav 2006;23(8):S57.
39. Blair DG, Ivanov EN, Tobar ME, Turner PJ, van Kann F, Heng IS. High sensitivity gravitational wave antenna with parametric transducer readout. Phys Rev Lett. 1995;74:1908–11. doi:10.1103/PhysRevLett.74.1908.
40. Astone P, Babusci D, Baggio L, Bassan M, Blair DG, Bonaldi M, Bonifazi P, Busby D, Carelli P, Cerdonio M, Coccia E, Conti L, Cosmelli C, D'Antonio S, Fafone V, Falferi P, Fortini P, Frasca S, Giordano G, Hamilton WO, Heng IS, Ivanov EN, Johnson WW, Marini A, Mauceli E, McHugh MP, Mezzena R, Minenkov Y, Modena I, Modestino G, Moleti A, Ortolan A, Pallottino GV, Pizzella G, Prodi GA, Quintieri L, Rocchi A, Rocco E, Ronga F, Salemi F, Santostasi G, Taffarello L, Terenzi R, Tobar ME, Torrioli G, Vedovato G, Vinante A, Visco M, Vitale S, Zendri JP. Methods and results of the IGEC search for burst gravitational waves in the years 1997–2000. Phys Rev D. 2003;68:022001. doi:10.1103/PhysRevD.68.022001.
41. Gottardi L, de Waard A, Usenko O, Frossati G, Podt M, Flokstra J, Bassan M, Fafone V, Minenkov Y, Rocchi A. Sensitivity of the spherical gravitational wave detector minigrail operating at 5 k. Phys Rev D. 2007;76:102005. doi:10.1103/PhysRevD.76.102005.
42. Aguiar OD, Andrade LA, Barroso JJ, Castro PJ, Costa CA, de Souza ST, de Waard A, Fauth AC, Frajuca C, Frossati G, Furtado SR, Gratens X, Maffei TMA, Magales NS Jr, RMM Jr, NFO, Pimentel GL, Remy MA, Tobar ME, Abdalla E, Alves MES, Bessada DFA, Bortoli FS, Brando CSS, Costa KMF, de Araujo HAB, de Araujo JCN, de Gouveia Dal Pino EM, de Paula W, de Rey Neto EC, Evangelista EFD, Lenzi CH, Marranghello GF, Miranda OD, Oliveira SR, Opher R, Pereira ES, Stellati C, Weber J. The Schenberg spherical gravitational wave detector: the first commissioning runs. Classical Quant Grav 2008;25(11):114042.

The Role of Binary Pulsars in Testing Gravity Theories

Andrea Possenti and Marta Burgay

Abstract Radio pulsars are neutron stars (NSs) which emit collimated beams of radio waves, observed as pulses, once per rotation of the NS. A subgroup of the radio pulsars behave as highly stable clocks and monitoring the times of arrival of their radio pulses can provide an accurate determination of their positional, rotational, and orbital parameters, as well as indications on the properties of their space-time environment. In this chapter, we focus on the so-called relativistic binary pulsars, recycled NSs orbiting around a compact companion star. Some of them can be used as unique tools to test general relativity and other gravitational theories. The methodology for exploiting these sources as laboratories for gravity theories is first explained and then some of the most relevant recent results are reviewed.

1 The Many Faces of Pulsar Science

Pulsars are radio sources representing a paradigm for the aims and the procedures of modern astrophysics. In fact, on one side, they are very intriguing celestial objects per se, whereas, on another side, they can be exploited as unique tools for investigating the behavior of nature and constraining its most fundamental laws.

As to the former aspect, the investigation of the radio pulsars (i.e., highly magnetized and rapidly spinning neutron stars (NSs)) is, for example, an important ingredient for understanding the final fate of massive stars, as well as for clarifying the processes occurring in the evolution of binary systems involving at least one compact star. The explanation of the broadband emission of electromagnetic waves (from the radio to the GeV and TeV bands) has been representing a challenging task for generations of experts of electrodynamics, while the expected emission of accelerated particles impacts on the studies of the cosmic rays. Also, the dispersion of the pulsed signal while it goes across the interstellar medium opens the possibility of determining the distribution of free electrons in the Milky Way (e.g., [1]), and the observation of the Faraday rotation of the polarized signals helps in mapping the large-scale structure of the galactic magnetic field (e.g., [2]). Furthermore,

A. Possenti (✉) · M. Burgay
INAF-Osservatorio Astronomico di Cagliari, via della Scienza 5, 09047 Selargius, Cagliari, Italy
e-mail: possenti@oa-cagliari.inaf.it

M. Burgay
e-mail: burgay@oa-cagliari.inaf.it

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_8

when discovered in a globular cluster, pulsars provide a way for studying the cluster's dynamics and potential well, potentially unveiling the presence of nonluminous matter in the form of black holes [3].

Looking at the second aspect, the study of pulsars is of utmost relevance for investigating two of the fundamental "forces" in nature, the nuclear interaction and the gravitational interaction. No terrestrial laboratory can compress a sizable amount of matter to nuclear density; hence, the equation of state for the nuclear matter at ultrahigh density can only be investigated in a NS, where nature provides matter experiencing those extreme conditions. In particular, each equation of state for the nuclear matter predicts a precise value for the radius and for the moment of inertia of a NS, depending on the gravitational mass (below a maximum mass) and the rotational frequency (below a maximum frequency) of the star. The maximum allowed mass and frequency are also dependent on the equation of state. The observation of the very high spin rate [4] and/or high mass of some pulsars [5, 6] has already been used for constraining the equation of state for the nuclear matter. The measurement of moment of inertia is a more difficult task, but it might be attained in the near future for some particularly favorable pulsar (i.e., the double pulsar [7]).

This chapter mostly deals with the application of the pulsars to the investigation of the gravitational theories. We will first briefly summarize the rudiments of pulsar science (Sect. 2), and the evolutionary scenario leading to the formation of the so-called *relativistic binary pulsars*. Then, in Sect. 3 the principles and methodology of the pulsar timing will be reviewed and the most interesting applications presented in Sect. 4. Last Sect. 5 will illustrate where this field of research is heading to.

2 Rudiments for Pulsars' Investigators

Pulsars are celestial objects which only progressively unveil their parameters. The peculiarities and/or the degree of scientific interest of a new source only emerges after having undertaken a patient monitoring of its radio signal, over timescales starting from few weeks and sometimes reaching few decades. One of the basic parameters is the spin period first derivative, $\dot{P}$, a determination of which typically requires almost 1 year of regular observations (see Sect. 3).

Assuming that the emission is due to the rotational energy loss in the form of magneto-dipole radiation (see Eq. 1), its measurement allows us to estimate important physical parameters, such as the dipolar surface magnetic field B_s and the spin-down age of the pulsar. The basic equation is:

$$-I_{NS}\omega\dot{\omega} = \frac{2}{3}\frac{1}{c^3}\omega^4 B_s^2 R_{NS}^6 \sin^2\alpha, \tag{1}$$

where I_{NS} is the moment of inertia of the NS, $\omega = 2\pi/P$ its angular velocity, R_{NS} its radius, and α the angle between the rotation and the magnetic axes. Expressing Eq. 1 in terms of the spin period, assuming $\alpha = 90°$, and using the standard values

for $I_{NS} = 10^{45}$ g cm^2 and $R_{NS} = 10^6$ cm, we can derive an estimate of the surface dipolar magnetic field at the magnetic equator as:

$$B_s = 3.2 \times 10^{19} \sqrt{P\dot{P}} \qquad \text{G} \tag{2}$$

and, after integration of the differential equation in time, we can calculate the so-called spin-down age (or characteristic age) of the pulsar as:

$$\tau_c = \frac{P}{2\dot{P}} \left(1 - \frac{P_0^2}{P^2} \right) \sim \frac{P}{2\dot{P}}, \tag{3}$$

where P_0 is the initial spin period of the pulsar, usually considered negligible with respect to the current one (whence the last approximation in Eq. 3).

According to [13, 14] more than 2000 radio pulsars have been discovered in the field of our Galaxy, in the galactic globular clusters, and in the Magellanic Clouds. The most interesting of which are the so-called *millisecond* or, better saying *recycled* pulsars (see later) located in the bottom-left corner of the $P - \dot{P}$ diagram of Fig. 1, whose short pulse duration and particularly regular pulsations are very helpful to study many physical and astrophysical problems listed in Sect. 1.

2.1 Basics of Evolution

According to the current paradigm, the recycled pulsars, to which the relativistic binaries belong, are believed to be formed in binary systems in which the companion star, during its evolution, transfers matter and angular momentum, via wind or Roche-lobe overflow, onto the NS surface (*recycling model*) [16]. The NS is hence spun-up to periods of the order of few or few tens of milliseconds, depending on the amount of matter accreted, which, in turn, depends on the initial mass of the companion.

The initial mass of the companion star, as well as the orbital separation at the time of formation of the NS, leads to very different evolutionary paths for the binary, whose in-depth description goes beyond the scope of this chapter. For a thorough discussion of (some of) the different possible evolutionary paths of binary pulsars, the reader can rely on the review [17], or on [18] for the case of the evolution of the increasing number of double neutron stars (DNS), as well as [19] for the case of the evolution of binaries in globular clusters.

A sketch in the $P - B_s$ diagram of the two evolutionary paths which are most relevant for the relativistic binaries is displayed in Fig. 2. At birth, the NS spins fast (tens of ms) and has a high magnetic field ($10^{11 \div 13}$ G), hence a high $\dot{P}$. Therefore, the pulsar slows down in a relatively short time scale moving left in the diagram. It is still debated whether a significant spontaneous decay of the surface magnetic field occurs [20] or does not [21] during this stage. When the death-valley (a region surrounding the nominal model-dependent death-line) is crossed, the pulsar switches

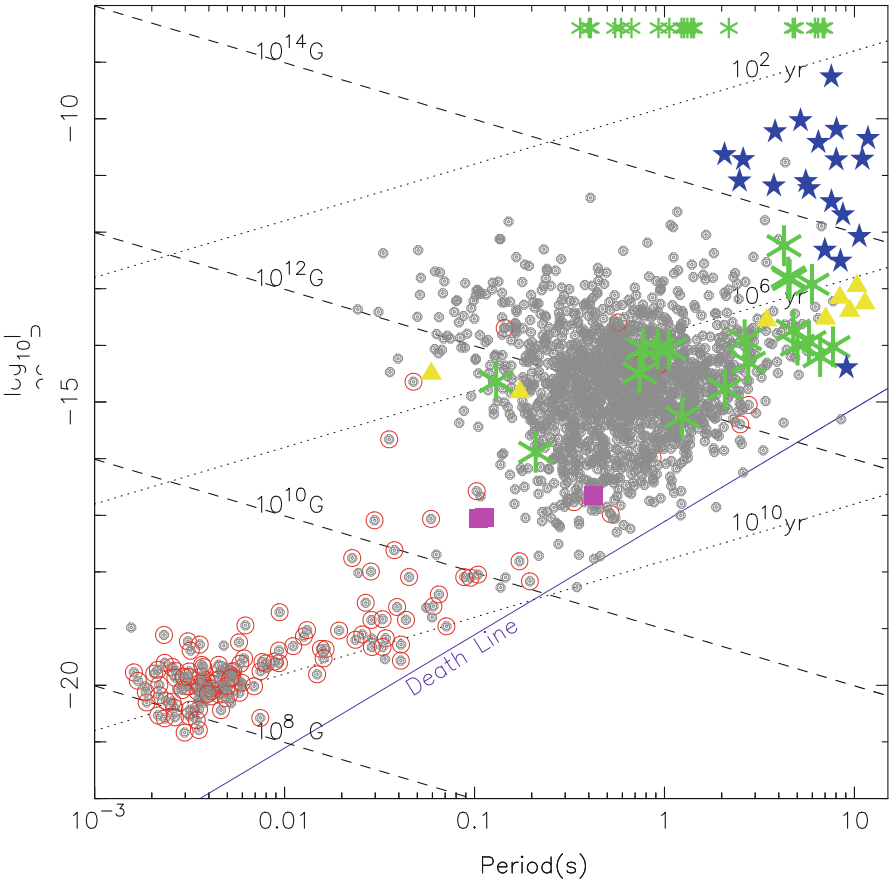


Fig. 1 Period-period derivative diagram. *Grey dots* are galactic field radio pulsars (those surrounded by a *red circle* being in a binary system); *purple squares* are X-ray dim isolated NS (XDINSs; e.g., [8]); *yellow triangles* are central compact objects (CCOs; e.g., [9]); *green asterisks* are rotating radio transients (RRATs; the *top* smaller ones do not have yet a measured $\dot{P}$; e.g., [10]); *blue stars* are magnetars [11]. *Dashed lines* denote equal dipolar magnetic field, calculated as in Eq. 2, while *dotted* ones are equal spin-down age (Eq. 3) lines. The *violet line* is the so-called *death line* (in particular the *death-line C* of [12]) for values of P and $\dot{P}$ below which the mechanism responsible for radio emission is not efficient anymore and the pulsar switches off. We point out that the use of a specific line is only for the sake of simplicity; a death valley [12], across which the pulsars' signal slowly fades out in time, should better replace a single line. (Data taken [13] from [14] and [15])

off. If the NS is isolated, it ends its electromagnetic life in the so-called pulsar graveyard.

If the progenitor of the pulsar companion is a star more massive than 8–10 $M_{\odot}$ and the orbital separation is suitable, a phase of mass transfer from the companion can establish, followed by a phase of common envelope, during which the two

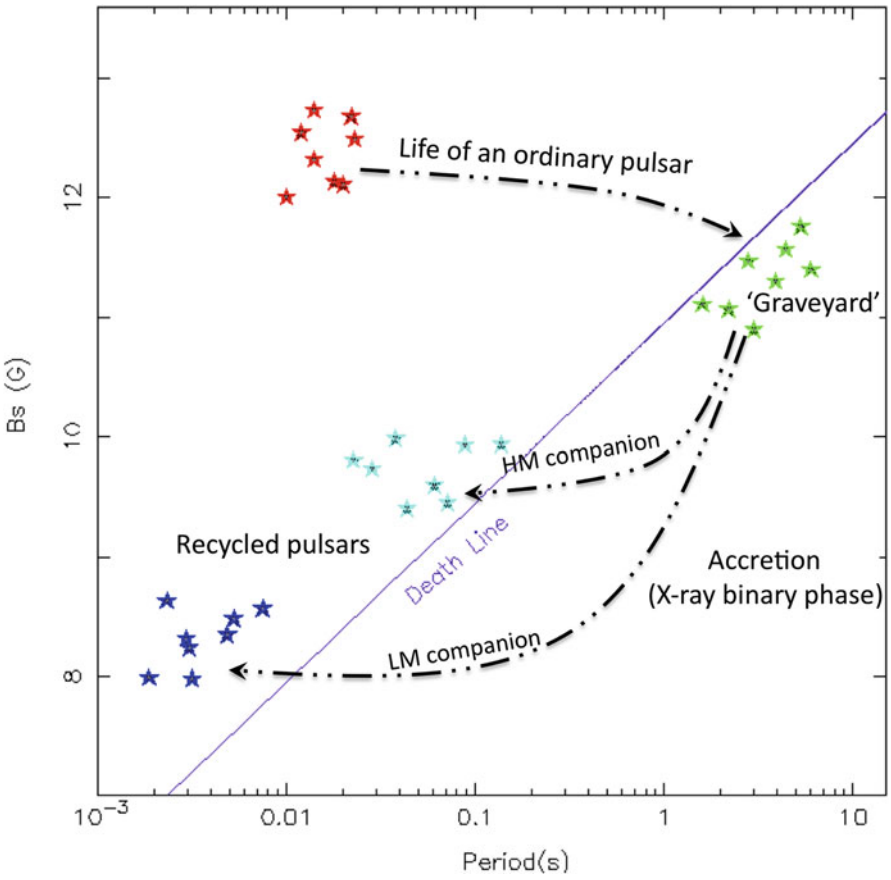


Fig. 2 Spin evolution of a neutron star in the period—magnetic field diagram (analogous to the $P - \dot{P}$ diagram and obtained using Eq. 2)

stars are included in the material lost by the companion and during which the orbit shrinks. Since the evolution of a massive star is relatively fast, the amount of transferred matter, hence of angular momentum, is small. The massive star ends its life in a supernova explosion leaving behind, under favorable circumstances, a binary system containing two NSs, that is, a so-called DNS binary. Due to the effect of the second supernova, the binary acquires a high eccentricity. The first born NS is a so-called *mildly recycled* pulsar, spun-up to tens of milliseconds and typically brought back above the death-line: in fact the short duration of the accretion phase also decreases the surface magnetic field of a couple of orders of magnitude only, thus leaving an NS with $B_s \sim 10^{10}$ G. The second born NS behaves as a young radio pulsar, slowing down rapidly and reaching the death-line much faster than the mildly recycled pulsar. This path is labeled as "HM companion" in Fig. 2.

Among the known pulsars with a high mass companion, ten have a NS companion. Sorted according to right ascension (RA), they are: J0737−3039A [22] and J0737−3039B [23], composing the first (and so far the only) known double pulsar system (this will be described in details in Sect. 4.2.2); J1518+4904 [24]; J1537+1151 (better known as B1534+12 and described in Sect. 4.2.1) [25,26], the only other DNS besides the double pulsar for which all five post-Keplerian parameters have been measured; J1756−2251 [27]; J1811−1736 [28, 29]; J1829+2456 [30]; J1906+0746 [31], whose undetected companion is likely the recycled pulsar in the system; J1915+1606 (better known as B1913+16, see Sect. 4.2.1), the first known binary pulsar whose discoverers were awarded with the Nobel prize for physics in 1993 [32,33] and J2129+1210C in the globular cluster M15 [34,35].

If the progenitor of the pulsar companion has a low mass, below $\sim 8\ M_\odot$, its evolution is slower. When the companion star expands due to its nuclear evolution, it fills its Roche lobe starting to transfer mass through the inner Lagrangian point. The mass-transfer phase lasts up to $\sim 10^8$ years, during which the mass accreted onto the NS surface is able to accelerate the NS up to few milliseconds, to "bury" the NS surface magnetic field down to $\sim 10^8 \div 10^9$ G and to circularize the orbit. Once the accretion stops, the NS is again above the death-line and is able to shine as a *fully recycled* millisecond pulsar. The companion, not having an initial mass high enough to ignite a supernova explosion, typically ends quietly its life as a white dwarf (WD) in an almost circular orbit. This evolutionary path is labeled as "LM companion" in Fig. 2.

The sample of the pulsars with a low mass companion (in most cases, an NS–WD system) is larger than the sample of DNS binaries. In a dozen of these systems, some relativistic effects have been already observed. Two remarkable cases are that of J1614−2230 and J0348+0432. For the former, the determination of both the post-Keplerian parameters related to the Shapiro delay led to determine that the NS has a mass very close to two solar masses ($1.97 \pm 0.04\ M_\odot$ with a companion mass of $0.5\ M_\odot$, [5]). The masses of the two stars were separately measured with high precision ($2.01 \pm 0.04\ M_\odot$ for the pulsar and $0.172 \pm 0.003\ M_\odot$ for the WD [6]) also in the second binary, by combining the determination of the post-Keplerian parameter related to the decay of the orbit of the NS with the phase-resolved optical spectroscopy of its white-dwarf companion. These are the highest NS masses measured with such accuracy to date. That constrains the equation of state for the nuclear matter, effectively ruling out the softest equations of state. In the framework of the study of the gravity theories, another very interesting binary of this class is J1738+0333, which is providing at the moment the most stringent test to the class of the scalar–tensor gravity theories [36,37] (see Sect. 4.2.3).

In effect, the masses of the companion star in NS–WD systems span a large range from $0.02\ M_\odot$ (typically Helium WDs) to $\sim 1\ M_\odot$ (typically carbon–oxygen WDs). As we will show inSect. 4.2.3, the often comparable masses (but very different radius) of the two stars in these systems can make them a better target than the DNS binaries for constraining gravity theories alternative to general relativity.

Finally, we note that the pulsars' zoo has been very recently enriched by a very intriguing object, PSR J0337+1715, which is a a *fully recycled* radio pulsar with two

WD companions in a hierarchical triple system [38]. Besides raising new questions about its formation and evolution [39], it promises to be an excellent laboratory for studying the strong equivalence principle (SEP), as it will be described in Sect. 4.1.

3 Keynotes on Pulsar Timing

When a new pulsar is discovered, the only known parameters are the approximate rotational period P and dispersion measure (DM)[1], as well as the very approximate position of the source in celestial coordinates. The possibility of an accurate determination of P, DM, and position in the sky, as well as the measurement of many other parameters of the targeted pulsar, results from undertaking a procedure dubbed *timing*. Although timing can be applied to any pulsar, the best results are obtained for those pulsars which behave as very stable rotators and whose flux density is large enough for an accurate determination of the "times of arrival" (ToAs) of the radio pulses. This is the case for some of the rapidly spinning recycled pulsars.

Excellent references for a thorough discussion on the pulsar timing procedures are represented, for example, by Chaps. 7 and 8 of [40], as well as Chaps. 4 and 5 of [41] (and Chap. 6 of [42] for an historical view).

3.1 First Step: Determination of the ToAs

On a practical ground, timing a pulsar means performing a campaign of semi-regular observations of the ToAs of some recognizable feature—for example, the peak—in the radio light curve (the so-called pulse profile) of the source. The total duration of the campaign depends on the aims of the experiment and ranges from about 1 year up to decades, when the pulsar is part of a pulsar timing array (PTA, see last section). As to the cadence of repetition of the observations, a dense coverage (up to few observations a day for pulsars in a binary system) is required at the beginning of the timing procedure. At later times it usually ranges from biweekly to bimonthly observations.

The vast majority of radio pulsars are weak radio sources, with their single radio pulses being well below the typical noise resulting from the contribution of the sky and that of the detector system. In view of that, a timing observation typically results in an array of $N_{sub} \times N_{ch}$ integrated pulse profiles, related to N_{sub} subsequent intervals of the observation time (sub-integrations) and to N_{ch} frequency intervals (sub-bands), spanning the total available bandwidth. Each of these profiles is obtained by de-dispersing the signal within each sub-band using the best available

[1] That represents the integrated column density of free electrons along the line of sight to the pulsar

value for the DM of the source and folding the data in each sub-band and each sub-integration at a constant period, corresponding to the expected apparent spin period of the source. Each pulse profile is also carefully time tagged, using an accurate clock which is installed at the radio observatory and is in turn regularly monitored and compared with the time distributed by the global positioning system (GPS).

Each of the pulse profiles is then compared with a very high signal-to-noise (S/N) standard profile, typically obtained from summing in phase a large number of observations of the given pulsar at the given frequency. This comparison produces a set of so-called *topocentric* ToAs, which are calculated by adding at the reference time of any pulse profile, the fraction of spin period by which the pulse profile is shifted with respect to the standard profile. Roughly speaking, the characteristic root mean square (rms) uncertainty in the determination of a topocentric ToA scales as the ratio between the width of the pulse in seconds and the S/N of the pulse profile.

3.2 Second Step: Modeling the ToAs

The aim is to use the pulses from a pulsar as the ticks of a clock, that is, being able to count all the pulses arrived from a reference epoch $t_{\rm ep}$ to the present time t and to predict the times of arrival of all next pulses. With no loss of generality, we can assume that a pulse arrived exactly at $t_{\rm ep}$ and model the rotational evolution of the NS with a power series:

$$N(t) \;=\; \nu_{\rm ep} \times (t - t_{\rm ep}) + \frac{1}{2}\dot{\nu}_{\rm ep} \times (t - t_{\rm ep})^2 + \frac{1}{6}\ddot{\nu}_{\rm ep} \times (t - t_{\rm ep})^3 + ..., \qquad (4)$$

where $\nu_{\rm ep}, \dot{\nu}_{\rm ep}, \ddot{\nu}_{\rm ep}, ...$ are the NS-spin frequency and its first- and second- and higher-order derivatives at the reference epoch $t_{\rm ep}$, whereas $N(t)$ represents the number of rotations occurred from $t_{\rm ep}$ to t. The aim of the timing procedure is to determine $\nu_{\rm ep}, \dot{\nu}_{\rm ep}, \ddot{\nu}_{\rm ep}, ...$ with an accuracy high enough that $N(t_{\rm next})$ will be very close to be an integer, for any future time of occurrence $t_{\rm next}$ of a pulse. In this framework, it is useful to introduce the so-called *timing residuals*, given by $R(t_{\rm i}) = N(t_{\rm i}) - n(t_{\rm i})$, where $n(t_{\rm i})$ is the nearest integer to $N(t_{\rm i})$. We can state to have got a satisfactorily coherent timing solution for a pulsar over a time-span $\Delta t_{\rm span}$, if $R(t_{\rm i}) \ll 1$ for all the observed ToAs $t_{\rm i}$ in the time range from $t_{\rm ep}$ to $t_{\rm ep} + \Delta t_{\rm span}$.

Of course, $R(t_{\rm i}) = R(t_{\rm i};\alpha_1, \alpha_2, ..., \alpha_m)$, where $\alpha_1, \alpha_2, ..., \alpha_m$ are the m parameters of the adopted timing model (i.e., $\nu_{\rm ep}, \dot{\nu}_{\rm ep}, \ddot{\nu}_{\rm ep}$ in the simple 3-parameter case discussed so far) and that naturally provides an operational way for improving a timing solution with a multi-parametric least squares fit, aimed to minimize the expression

$$\chi^2 \;=\; \Sigma_i \left(\frac{R(t_{\rm i};\alpha_1, \alpha_2, ..., \alpha_m)}{\varepsilon_{\rm i}} \right)^2, \qquad (5)$$

where $\varepsilon_{\rm i}$ is the uncertainty on the ith ToA in units of the pulsar spin period and i runs over all the available ToAs in the given time-span.

At the end of the day, one ideally expects to have a set of uncorrelated timing residuals, which should appear randomly scattered (i.e., with no evident trend) about a zero mean value when they are plotted versus the times of their collection t_i. A useful figure of merit for evaluating the quality of the timing solution is the ratio η between the rms of the residuals and the rotational rate of the pulsar, with $\eta \lesssim 0.001$ usually indicating a good solution.

What described above would be the whole story if the topocentric ToAs were collected in an inertial observing frame and when the NS had no binary companion. However, at least the first hypothesis is always wrong and the first step in the timing procedure is to convert the topocentric ToAs to the so-called *barycentric* ToAs. For each topocentric ToA t_i, this implies to calculate the corresponding ToA $t_{i,bary}$, as it were detected at the Solar System Barycenter (SSB, an acceptable approximation of an inertial frame) by a detector operating at an (in principle) infinite frequency. In a formal way, the conversion formula reads like:

$$t_{i,bary} = t_i + t_{\rm clock} - \frac{D}{f^2} + \Delta R_\odot + \Delta E_\odot + \Delta S_\odot. \tag{6}$$

The value of $t_{\rm clock}$ results from the sum of various terms of correction (the so-called *clock correction chain*) and it is added retroactively, using tabulated values published by the *Bureau International des Poids et Mesures* (BIPM): the aim is to convert the time obtained at the telescope to the terrestrial time (TT) realization of the *Temps Atomique International (TAI)* (i.e., TT(TAI)), which differs (since 1971) from the TAI by a constant offset.

The next term in Eq. 6 takes into account the effect of the dispersion of the radio pulses in the interstellar medium (see Sect. 1). In particular,

$$D(t_i) = \frac{e^2}{2\pi m_e c} \int_0^d n_e {\rm d}l = \mathcal{D} \times {\rm DM(t_i)}, \tag{7}$$

where $\mathcal{D} = (4.148808 \pm 0.000003) \times 10^3$ MHz^2 pc^{-1} cm^3 s is the dispersion constant, d the distance to the pulsar, and f is the Doppler-corrected observing frequency.

In general, the fourth term in Eq. 6—also known as Roemer delay—gives the dominant contribution to the barycentric correction. It can be written as

$$\Delta R_\odot = \frac{\mathbf{r}\cdot\mathbf{n}}{c} + \frac{(\mathbf{r}\cdot\mathbf{n})^2 - |r|^2}{2cd}, \tag{8}$$

where $\mathbf{n}$ is the unit vector on the line going from the SSB to the pulsar and $\mathbf{r}$ is the vector connecting the SSB and the Earth. Accounting for this term requires the knowledge of the location and motion of the major bodies in the solar system and of the nonuniform Earth rotation. They are provided by the ephemeris published by the *Jet Propulsion Laboratory* [43] and by the bulletin regularly published by the *International Earth Rotation Service* [44].

Einstein's delay $\Delta E_\odot$ is a combination of the relativistic time delay and the gravitational red-shift due, respectively, to the motion of the Earth and the mass of the other solar system bodies. Its time derivative is given by:

$$\frac{d\Delta E_\odot}{dt} = \sum_k \frac{G m_k}{c^2 d_{k,\oplus}} + \frac{v_\oplus^2}{2c^2} - \text{constant}, \tag{9}$$

where G is the gravitational constant, m_k are the masses of the other solar system bodies, $d_{k,\oplus}$ their distances to the Earth, and $v_\oplus$ the velocity of the Earth with respect to the SSB.

Finally, the Shapiro delay $\Delta S_\odot$ [45] measures the extra time required for an electromagnetic wave to move in the curved gravitational field of a celestial body. In the framework of the pulsar timing, only the Sun and, sometimes, Jupiter, causes observable effects. In the case of the Sun, the formula takes the form:

$$\Delta S_\odot = -\frac{2GM_\odot}{c^3} \ln(1 + \cos\theta), \tag{10}$$

where θ is formed by the vector from the pulsar to the telescope with the vector from the telescope to the third body.

If the pulsar is included in a binary system, the *pulsarcentric* ToAs (i.e., the ToAs expressed in pulsar proper time at the pulsar surface) must also be corrected calculating them at the Pulsar Binary System Barycenter, before translating them to the Solar System Barycenter. This involves the introduction of four additional terms in the Eq. 6:

$$t_{i,bary}^{Bin} = t_{i,bary} + \Delta R_{Bin} + \Delta E_{Bin} + \Delta S_{Bin} + \Delta A_{Bin}. \tag{11}$$

Again, the Roemer term ΔR_{Bin} generates the largest effects: the ToAs anticipate when the pulsar is in front of the companion and are delayed when it is behind it. In a purely Newtonian framework, fitting for the orbital modulations of the ToAs allows one to derive five Keplerian parameters describing the binary system: namely, the orbital period P_b, the eccentricity e, the projection of the semimajor axis along the line of sight $x = a \sin i$, the longitude of the periastron ω (typically calculated with respect to the ascending node), and the epoch of the passage at the periastron T_0. Only in few favorable cases [46], the position angle on the sky of the ascending node (i.e., a sixth Keplerian parameter) can also be measured. From the Keplerian parameters one can derive the *mass function*:

$$f(M) = \frac{(M_c \sin i)^3}{(M_p + M_c)^2} = \frac{4\pi^2 (a \sin i)^3}{G P_b^2}, \tag{12}$$

where M_p is the pulsar mass and M_c is the companion mass. Assuming a value for M_p and an edge-on orbit ($i = 90°$), it is possible to derive a lower limit for M_c.

The other terms in Eq. 11 express the deviations from the predictions of the classical physics expected for a binary pulsar when it experiences strong gravitational fields and/or an high orbital velocity. These effects, and the related measurable

parameters, will be explained in next Sect. 4. It is worth noting that ΔA_{Bin} incorporates the effects of the changing aberration along the orbit and, strictly speaking, it is a classical physics effect. However, since the two related aberration parameters are almost degenerate with some post-Keplerian relativistic parameters (see later), this term is usually also described in the context of the relativistic effects.

In practice, one typically starts with a model having the minimum number of parameters, typically $\nu_{\rm ep}$ and the celestial coordinates of the pulsar. A poor determination of the pulsar spin frequency $\nu_{\rm ep}$ imprints a linear trend in the timing residuals plotted over the usually very short initial time-span. Instead, an error in the spin derivative imprints a parabolic trend in the residuals and the errors in the celestial coordinates leave a sinusoidal signature with a period of 1 year. If the pulsar is in a binary system, many additional trends in the residuals overlap with the ones mentioned above. *Solving* a pulsar means removing all these trends over a yearlong time-span. This is not only the result of a large series of trials and errors, but requires a good deal of experience, perseverance, and educated feeling, as well as a touch of insight, in turn throwing a flash of artistic inspiration on the whole procedure.

We conclude these keynotes on the timing procedure noticing that, unfortunately, not all the, about 2000, radio-emitting NSs are equally good timekeepers: for example, there are many sources exhibiting glitches (i.e., sudden increases in the spin frequency) and/or other timing irregularities. The former have been long thought to be related to some event occurring in the NS interior, although the exact origin of the process is still a matter of debate (e.g., [47–49]). The timing noise has been only recently recognized to be most likely a phenomenon related with instabilities occurring in the magnetosphere of the pulsars [50]. Observations show that the two effects pertain mostly to the youngest and the ordinary pulsars, whose timing residuals display long-term and unpredictable variations [51]. On the contrary, they are virtually absent (or present at a very low level, e.g., [52]) in the population of the recycled pulsars, at least within the present level of precision in the measurements of the ToAs [53].

The considerations above indicate that the recycled pulsars are intrinsically better clocks than the bulk of the ordinary pulsars. Moreover, the accuracy of the measurement of the ToAs roughly scales with the width of the pulses and hence, in turn, with the rotational rate. As a consequence, the fast spin rate of the recycled pulsars allows one to determine their ToAs with an higher accuracy than for the ordinary pulsars. However, at least three additional characteristics are very important in evaluating the quality and the potentialities of a recycled pulsar as a clock: the flux density, the shape of the pulse, and the timing stability.

4 Pulsar Tests of Relativistic Gravity

To first approximation, the amplitude of the deviations of general relativity with respect to Newtonian gravity can be evaluated by comparing the classical gravitational potential energy $E_{\rm gr} \sim -GM^2/R$ of a body of mass M and radius R

with the rest mass energy of the same body, $E_{\rm rm} = Mc^2$ (c being the speed of light in vacuum and G the gravitational constant). For the Earth, the dimensionless ratio $\varepsilon = |E_{\rm gr}/E_{\rm rm}|$ is tiny ($\sim 10^{-10}$) and then most of the tests of general relativity have been carried on in space, involving various bodies of the solar system (e.g., [54]). A detailed description of some of these experiments (the earliest of them, performed by Eddington on 1919, dates more than 90 years ago) is reported in the chapters by Iafolla, Peron, and Turyshev in this book, whereas the chapter by Dell'Agnello mostly deals with future experiments of this class. All these experiments fully exploit a unique feature of general relativity, that is, the absence of tunable parameters in the theory. As a consequence, even a single observation in disagreement with the predictions of the theory would unambiguously determine the falsification of the theory, at least in the physical regime of validity of the given observation. It is well known that to date general relativity passed *cum laude* all these potentially lethal tests (e.g., [55] for a recent review). However, even for the Sun ε is only $\sim 10^{-6}$, implying that all the experiments carried on in the solar system can only explore the so-called *weak-field* limit of gravitational physics.

Despite the observational successes of general relativity, a wealth of alternate gravity theories kept on being proposed. Some of these theories have been already disproved by the observations, whereas many others preserve only a historical and/or academic relevance. However, few of them are still particularly interesting, since they emerged in the context of the studies aimed to build the long sought unified theory of the physical interactions, that is, a theory capable to include electroweak and nuclear interactions in the same framework as the gravitational effects. Nowadays, a major obstacle in this direction is to match the probabilistic approach involved in the treatment of the first two phenomena with the deterministic predictions of general relativity. Following this reasoning, one may wonder whether other theories of gravity may lead to solve the dichotomy, while providing a better description of the behavior of nature in the extreme physical conditions of application of the putative unified theories, for example, in the first fractions of second after the Big Bang (some of these theories are described in depth elsewhere in this book, e.g., in the chapter by Diaferio and Angus, the one by Liberati and Mattingly, as well as the chapter by Antoniadis).

The considerations above emphasize the need of performing tests of general relativity and/or of alternate gravity theories under the so-called *strong-field* limit (corresponding to a value of ε close to unity), which is the regime associated to the notion of *extreme physical conditions* in presence of gravitational effects alone. In particular, on one side it is very important the derive tight constraints to the falsifiability of general relativity in the strong-field regime, where the Einsteinian theory could finally reveal the existence of limits to its application. On another side, it has been shown [56] that there exist alternative theories which could pass all the tests in the weak-field limit, but would be violated as soon as the strong-field regime is approached; that makes the strong-field tests mandatory to properly test those theories.

In this context, two factors—described below—allow some binary recycled pulsars to become superb tools for investigating gravity theories. (I) It holds $\varepsilon \sim 0.2$

at the pulsar surface, which reflects the large gravitational binding energy (i.e., *self-field energy*) of the NS. We here note that for all the known binary pulsar systems, the orbital separation is large with respect to the NS radius and so any binary component moves in the weak gravitational field of the companion. However, in most alternate theories of gravity (but not in general relativity), the orbital motion and the gravitational radiation damping depend on the *self-field energy* of the binary components. Therefore, when ε is large (i.e., of order unity), significant deviations—attributable to strong-field effects—are expected with respect to the orbital motion predicted by general relativity. (II) The radio signal behaves as an accurate and stable clock, which leads to accurately map the rotation and the orbit of the pulsar by means of the timing procedure described in Sect. 3. From the combination of the facts (I) and (II), it results the intriguing possibility of using some binary pulsars as unique laboratories for testing the strong-field regime of the gravity theories. In particular, the best binaries are those for which the properties expounded in (I) and (II) are particularly prominent.

General relativity and all the other in-principle acceptable (i.e., with some chance to be *viable* [57]) alternate gravity theories can be grouped under the very large class of the so-called *metric theories of gravity*. They satisfy the following three assumptions: *[a]* a symmetric metric exists; *[b]* all test bodies follow geodesics of the metric; *[c]* in local freely falling reference frames, all the *non*gravitational laws of physics are those written in the language of special relativity. In other words, in all metric theories, gravitation must be a phenomenon related with the occurrence of a *curved space-time*. Although matter and *non*gravitational fields respond only to the metric, additional fields can occur, giving rise to, for example, tensor/scalar theories, tensor/vectorial theories, and so on... These additional fields prescribe how matter and *non*gravitational fields contribute to create the metric; once determined, the metric alone acts back on the matter. Of course, at variance with the case of general relativity, these theories include a set of tunable parameters, associated to the additional fields. A detailed review of various gravity theories appeared, for example, in the textbook [54], which can be complemented with some recent updates [55, 57], as well as with the chapter by Will in this book.

4.1 Gravity Theory Tests Using the PPN Parameters

A first class of tests of metric theories of gravity involving pulsar observations is based on the parametrized post-Newtonian (PPN) formalism. In this framework, the deviations of any given theory from Newtonian physics are reflected in the values assumed by ten parameters (the so-called PPN parameters, see e.g., [55] and the chapter by Will in this book), each of them connected with a given physical effect, such as the existence of preferred frames, the occurrence of preferred locations, the nonconservation of the momentum, the nonlinearity in the superposition of gravitational effects, the amount of space curvature produced by a unit mass. On the theoretical point of view, this formalism allows one to easily interpret the physical

implications of the various proposed alternative gravity theories in the weak-field limit. At the same time, on the experimental point of view, the comparison between the observed and the predicted values of the PPN parameters provides an observational tool for directly constraining the falsifiability of general relativity and other theories. In effect, the PPN formalism was originally conceived (e.g., [58]) and then fruitfully used for performing a variety of tests in the weak-field environment of the solar system [54, 55, 57]. Its extension to the case of the compact objects (i.e., including strong self-field effects) was elaborated in the 1990s [56] at least for a large class of tensor–multi-scalar gravitational theories [59]. That implied a partial redefinition of the ten original PPN parameters, with the introduction of correction factors which are approximately dependent on the value of the dimensionless ratio ε introduced above. As a matter of fact, these factors can be safely neglected for the solar system tests ($\varepsilon \lesssim 10^{-6}$), but not when dealing with NSs ($\varepsilon \sim 0.2$), which highlights again the fact that the pulsar tests of gravity are complementary to the tests involving bodies of the solar system.

A detailed reading about pulsar experiments using the PPN formalism is given in the excellent online summary [60]. A particularly interesting case is that of the tests of the SEP. That is not only satisfied by general relativity, but it is conjectured to imply general relativity itself (e.g., [54]). It includes both the *weak equivalence principle* (WEP) and the *Einstein equivalence principle* (EEP). WEP states that the trajectory of a free falling test body in a gravitational field is independent of its composition and internal structure (i.e., what is often referred to as the universality of free fall), whereas EEP adds that the outcome of any local nongravitational experiment is independent of (i) when, (ii) where in the universe, and (iii) the velocity of the freely falling reference frame in which it is performed (Lorentz and positional invariance of the nongravitational laws of physics, see the chapter by Liberati and Mattingly in this book for details). SEP extends WEP to self-gravitating bodies (i.e., not only test bodies but also bodies with a significant amount of internal gravitational energy, i.e., having a large value of the parameter ε) and EEP to experiments involving gravitational forces (like for instance the measurement of the position of the axes of the orbit followed by two bodies).

Therefore, a simple test of the SEP is to check for the occurrence of differences in the trajectories of two massive bodies in a gravitational field. In particular, Nordtvedt [61] first proposed to search for the occurrence of a "polarization" in the direction of the Sun of the orbit of the [Earth+Moon] system, caused by the different *self-field energy* of the two bodies (a phenomenon often called *Nordtvedt effect* or *gravitational Stark effect*). Using lunar laser ranging (LLR) experiments, it has been possible to set a strong limit to a linear combination of various PPN parameters in the typical conditions of the solar system. In a similar way, the study of the "polarization" of the [pulsar+WD] binaries in the galactic gravitational potential allows one to set a limit to the violation of the SEP also in presence of strong self-field effects (in principle provided by the NS in the binary). At variance with the case of the [Earth+Moon] system, the geometrical and orbital parameters of a given [pulsar+WD] binary are usually only partially known, which prevents the use of a single binary to perform the test. However, a statistical approach can be adopted:

for example, [62] used 21 [pulsar+WD] binaries with long orbital periods ($P_b > 4$ days) and very low eccentricity ($10^{-6} \lesssim e \lesssim 10^{-3}$) in order to put 95 % confidence upper limit on Δ of 5.6×10^{-3}, where $\Delta = (M_{gr}/M_{in})_{\rm NS} - (M_{gr}/M_{in})_{\rm WD}$, with $(M_{gr}/M_{in})_i$ being the ratio between the gravitational and the inertial mass of the ith body (of course, if SEP holds true, $(M_{gr}/M_{in})_i = 1$ for any ith body and so $\Delta \equiv 0$). More recently, using a extended sample of 27 [pulsar+WD] binaries, [63] improved the constraint to $\Delta < 4.6 \times 10^{-3}$. The possibility for a direct (vs the *statistical* approach presented above) determination of Δ relies on the measurement of the derivative of the eccentricity $\dot{e}$ in pulsar binaries [64]. Although the best limits from this approach [65] are not yet as constraining as those resulting from the statistical analyses above, the prospects are very promising, in view of the continuous improvement in the technology of the pulsar instrumentation and the availability of decadelong data spans [65].

Besides the aforementioned *gravitational Stark effect*, many other phenomena related to violations of the SEP manifest as *polarization* of the orbit or as *precession* of the orbit or of the spin axis of the NS, as well as the occurrence of an additional acceleration applied to the NS or to the centre of mass of the binary, Accurate pulsar timing of one binary (or of an ensemble of suitable binaries) can unveil and measure those effects, thus putting constraints to various PPN parameters. These limits are often much better than those obtained in the solar system: an illustrative case is that of the parameter $\hat{\alpha}_3$ (where $\hat{\alpha}_i$ denotes the strong-field generalization of the PPN parameter α_i). A nonzero value of $\hat{\alpha}_3$ would imply the occurrence of preferred reference frames, as well as the nonconservation of the momentum. By using the data from an ensemble of [pulsar+WD] binaries one can get a limit $\hat{\alpha}_3 < 4 \times 10^{-20}$ [62] (95 % confidence level), which is 13 orders of magnitude better than that derived from the measurements of the perihelion shift of the Earth and of Mercury [54]. Similarly, strong-field limits on the other two PPN parameters ($\hat{\alpha}_1 = -0.4^{+3.7}_{-3.1} \times 10^{-5}$ [66] and $|\hat{\alpha}_2| < 1.6 \times 10^{-9}$ [67], 95 % confidence level) which involve, like $\hat{\alpha}_3$, the existence of preferred frame effects (i.e., a violation of the Lorentz invariance) have been obtained from the timing analysis of the binary pulsar PSR J1738+0333 [36, 37] (in this case, the limit is about one order of magnitude better that than derived for α_1 from lunar laser ranging experiments [68]), as well as exploiting the observations of two isolated millisecond pulsars, PSRs B1937+21 and J1744−1134 (in this case, the constraint is two orders of magnitude stronger than that obtained for $|\alpha_2|$ with the best test performed in a weak field, that is, the alignment of the Sun spin with the total angular momentum of the solar system [69]). It is worth noting that the current best limit on $|\hat{\alpha}_2|$ is not obtained from the study of binary pulsars, but it results from the analysis of the secular stability of the pulse profile of the two aforementioned solitary pulsars. In fact, a nonvanishingly small $\hat{\alpha}_2$ would induce the precession of a pulsar spin (and hence a secular variation in the observed pulse profile) around the pulsar direction of motion with respect to the putative preferred frame. An analogous precessional effect leads to the possibility of using the same pulse profile data of the two millisecond pulsars above for deriving the strongest available limit on the parameter $|\hat{\xi}| < 3.9 \times 10^{-9}$ (95 % confidence level) [70] (the strong-field counterpart of the Whitehead PPN

parameter ξ), that turns out to be six orders of magnitude more constraining than the previous best limit on ξ obtained in the weak gravitational field environment of the superconducting gravimeter experiments [71].

On other cases, the pulsar timing tests are nominally less constraining than the solar system tests; however—as explained above—they explore a different regime of gravity with respect to the solar system tests. For example, an upper limit $\zeta_2 < 4 \times 10^{-5}$ to the nonconservation of the momentum (embodied in the PPN parameter ζ_2) also resulted from the analysis of PSR B1913+16 [72]. Furthermore, we note that the occurrence of preferred positions and times in the universe may naturally lead to variations in the fundamental constants, such as the gravitational constant G. Pulsar timing can put constraints also on this effect, as described in Sect. 4.2.3, when dealing with the case of the binary pulsar PSR J0437−4715. Interestingly, the limit on $|\hat{\xi}|$ can also be converted into an upper limit of 4×10^{-16} on the spatial anisotropy of the gravitational constant [70].

Additional relevant improvements in constraining the SEP are expected from the timing of the first known millisecond pulsar (PSR J0337+1715) in a hierarchical triple system with two other compact stars, both WDs, a recently announced (Jan 2014) milestone discovery [38] in pulsar science. The classical Newtonian effects occurring in this 3-body system (complemented with the inclusion in the model of the special-relativistic transverse Doppler effect) have already led to a precise determination of the masses of the pulsar (1.4378 ± 0.0013 $M_\odot$) and of the two companions (0.19751 ± 0.00015 $M_\odot$ and 0.4101 ± 0.0003 $M_\odot$), as well as to a measurement of the inclinations of the orbits, surprisingly almost coplanar (close to ~ 39.2 deg [38]). The pulsar and the inner WD (orbiting the common center of mass in about 1.6 days) are two bodies having a very different self-field energy ($\varepsilon \sim 0.2$ for the NS, whereas $\varepsilon \sim 10^{-4}$ for the WD, see Sects. 4 and 4.2.3) and both moving in the gravitational field provided by the outer white dwarf (having an orbital period of about 327 days). Since the latter field is 6–7 orders of magnitude larger than the gravitational field due to the Galaxy, any putative SEP violation occurring in the J0337+1715 inner binary would be strongly magnified with respect to the case of all the previously described [pulsar+WD] binaries, thus making the new triple system the best laboratory for investigating the limits of the SEP.

4.2 Gravity Theory Tests Using the PK Parameters

When a pulsar is orbiting another compact object (a second NS or a WD) in a close enough orbit, the very high orbital velocity and the large deformation of the space-time produced by the two massive stars in the system can lead to stronger relativistic effects than those described in Sect. 4.1. This may open the possibility of directly measuring these effects on the times of arrival of the pulses from the pulsar(s) and, in turn, to carry on a direct comparison between the observations and the predictions of the various theories of gravity.

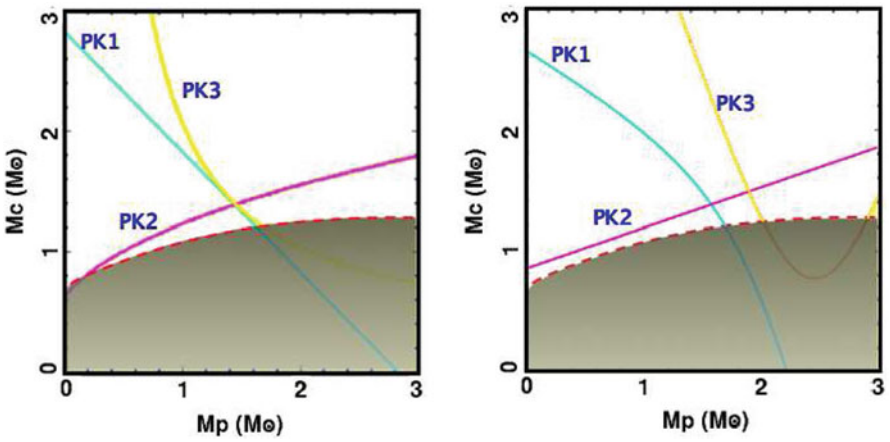


Fig. 3 Mass-vs-mass diagrams of a binary system for which three post-Keplerian parameters have been measured. The two panels refer to two alternate gravity theories. Each color-coded curve in the panels describes the constraint on the masses of the two orbiting bodies resulting from the same observed value of a given PK parameter (and its uncertainty), combined with the adoption of the gravity theory specific to that panel. *Left panel*: Since a common area exists for the three curves, this gravity theory survives the test. *Right panel:* No overlap exists for the three curves, implying the rejection of this gravitational theory

In this context, Damour and Deruelle [73,74] developed a powerful and successful framework, tailored for constraining a very large class of gravitational theories. It relies on the introduction of the so-called *post-Keplerian* (PK) parameters, which satisfy the following two very useful properties: *(i)* the PK parameters are phenomenological quantities, which can be measured according to a well-established operational prescription, independent on the adopted gravity theory; *(ii)* in any specific gravity theory (chosen in a large range of metric theories), the PK parameters can be written as function of the pulsar and companion masses and of the Keplerian parameters of the binary system.

Since the Keplerian parameters are easily determined with a precision much better than any PK parameter, they can be regarded as well known terms in the formulae for the PK parameters, which leaves only two unknown quantities (the pulsar and companion masses) in each formula. As a consequence, the measurement of two PK parameters results in the determination of the masses of the two stars in the system, the values of which are expected to depend on the adopted theory of gravity. On the other hand, the measurement of $N_{\rm PK} > 2$ PK parameters yields $N_{\rm PK} - 2$ independent self-consistency tests for the given theory. The methodology can be illustrated in a simpler way in the mass-vs-mass plot for the two binary components, where each PK parameter (with its uncertainty) is associated to a pair of lines (see Fig. 3). The region in the plot located between a pair of lines is only allowed by the theory under investigation. If an area of overlap between the various pairs of lines exists, that specific gravity theory passes the test (see left panel of Fig. 3). Otherwise (e.g. right panel of Fig. 3) the theory has to be unequivocally rejected.

In the specific case of general relativity, the equations describing the five most used PK parameters assume the form [74–76]:

$$\dot{\omega} = 3\left(\frac{P_b}{2\pi}\right)^{-5/3}(T_\odot M)^{2/3}\,(1-e^2)^{-1} \tag{13}$$

$$\gamma = e\left(\frac{P_b}{2\pi}\right)^{1/3} T_\odot^{2/3}\,M^{-4/3}\,m_2\,(m_1+2m_2) \tag{14}$$

$$\dot{P}_b = -\frac{192\pi}{5}\left(\frac{P_b}{2\pi T_\odot}\right)^{-5/3}\left(1+\frac{73}{24}e^2+\frac{37}{96}e^4\right)(1-e^2)^{-7/2}\,\frac{m_1\,m_2}{M^{1/3}} \tag{15}$$

$$r = T_\odot\, m_2 \tag{16}$$

$$s = x\left(\frac{P_b}{2\pi}\right)^{-2/3} T_\odot^{-1/3}\,M^{2/3}\,m_2^{-1} \equiv \sin i, \tag{17}$$

where m_1 and m_2 are the two star masses, $M = m_1 + m_2$, $x = a\sin i$ and $T_\odot \equiv GM_\odot/c^3 = 4.925490947\,\mu$s. The PK parameter $\dot{\omega}$ is phenomenologically associated with the advance of the periastron, γ is a parameter accounting for gravitational red-shift and time dilation, $\dot{P}_b$ is the orbital damping and, in the framework of general relativity, measures the rate at which the orbital period decreases due to emission of gravitational radiation. Finally, r and $s \equiv \sin i$ represent respectively the *rate* and the *shape* of the so-called Shapiro delay [45], a time delay of the radio signal caused by the space-time deformations around the companion star. As expected, no tunable parameter appears in the set of equations above, whereas one or more theory-dependent tunable parameters are contained in the arrays of equations resulting from adopting alternate gravity theories (see e.g., [54, 57] for a few examples).

4.2.1 The Case of the Double Neutron Star Binaries

The first very famous case of application of this class of tests is represented by the binary pulsar system B1913+16, discovered at Arecibo in 1974 [77]. It is a DNS system whose pulsar is rotating with a 59-ms period while completing a highly eccentric ($e = 0.61$) orbit in about 7.8 h. Three PK parameters have been measured for this system: from Eqs. 13 and 14, for $\dot{\omega}$ and γ respectively, the most precise values ever for the masses of two pulsars have been determined [78], and an accurate prediction for $\dot{P}_b^{\rm GR}$ resulted from inserting these masses in Eq. 15. It turned out that the observed intrinsic value $\dot{P}_b^{\rm int}$ (see later) for the rate of the orbital damping matched spectacularly with $\dot{P}_b^{\rm GR}$ (the agreement, after more than three decades of timing observations, is now at 0.2 % level [78]). On one hand, that provided a spectacular proof of the occurrence of quadrupolar gravitational waves emission from the system, while, on another hand, that also showed that the internal structure (i.e., the *self-gravity*) of the NSs does not affect—at least not at more than the 0.2 % level—the dynamics of the system, which can be described as it were composed

by two point masses. This supports the so-called *effacement* of the interior of the bodies [79], which is a property peculiar to general relativity (in turn arising from the SEP), while not holding in other gravity theories.

It is here important to note that $\dot{P}_b^{\rm int}$ is not directly the value $\dot{P}_b^{\rm ToA}$ resulting from the fitting of the observed ToAs to the timing formula, but it was corrected for various effects which can contaminate $\dot{P}_b^{\rm ToA}$. In fact in general it holds

$$\left(\frac{\dot{P}_b}{P_b}\right)^{\rm ToA} = \left(\frac{\dot{P}_b}{P_b}\right)^{\rm int} + \left(\frac{\dot{P}_b}{P_b}\right)^{\rm z} + \left(\frac{\dot{P}_b}{P_b}\right)^{\rm rot} + \left(\frac{\dot{P}_b}{P_b}\right)^{\rm Shk} + \left(\frac{\dot{P}_b}{P_b}\right)^{\rm other}, \quad (18)$$

where $\dot{P}_b^{\rm int}$ is the intrinsic rate of the orbital decay, while $\dot{P}_b^{\rm z}$ is the contribution to the apparent orbital damping due to the vertical acceleration in the Galactic potential [80] , $\dot{P}_b^{\rm rot}$ is the contribution due to differential rotation in the plane of the Galaxy [80], $(\dot{P}_b/P_b)^{\rm Shk} = [v_\perp^2/(cd)]$ (with d the distance and $v_\perp$ the transverse velocity of the binary system with respect to the SSB) is the Shklovskii effect, arising from the transverse component of the relative velocity of the binary barycenter with respect to the SSB. Finally, $\dot{P}_b^{\rm other}$ summarizes other contributions which have been thoroughly investigated in [80] and which, in most cases, are negligible with respect to the other terms. A relevant exception is for the binary pulsars harbored in a globular cluster, whose gravitational potential well can impart a significant line-of-sight acceleration to the pulsar; in this case, $\dot{P}_b^{\rm other}$ can overcome all the other contributions[2]. In the case of PSR B1913+16, the imperfect knowledge of the shape of the Galactic potential, and in turn of $\dot{P}_b^{\rm z}$ and $\dot{P}_b^{\rm rot}$, emerges now as the limiting factor acting against further large improvements of the accuracy of the test of general relativity.

The impact of the Galactic potential in the orbital period derivative is even more evident for the case of the 38-ms pulsar PSR B1534+12, discovered at Arecibo on 1990 [81] and also hosted in a DNS binary. All the five PK parameters listed in the Eqs. 13→17 have been measured. The three PK parameters $\dot{\omega}$, γ, and s provided the first accurate test (at better than 1 % level) for the nonradiative prediction of general relativity [82] and the range r of the Shapiro delay is also fully compatible with what expected from general relativity. However, the observed value $\dot{P}_b^{\rm ToA}$ is not compatible (at 1σ level) with the expectations of general relativity. This is interpreted as due to the large uncertainty on the correction to $\dot{P}_b^{\rm ToA}$ resulting from the Shklovskii contribution $(\dot{P}_b/P_b)^{\rm Shk} = [v_\perp^2/(cd)]$, in turn arising from the uncertain determination of the pulsar distance d. Assuming that general relativity is correct, one can invert the line of reasoning, obtaining a precise determination of the distance of the binary, which should be located at $d_{\rm GR} = 1.02 \pm 0.05$ pc. This is about a factor 30 % farther away with respect to the distance inferred from the DM of the pulsar $d_{\rm DM} \sim 0.7$ kpc, in agreement with the typical uncertainty on the pulsar distances inferred from the value of their dispersion measure.

[2] This is likely the case for the DNS system J2129+1210C [35] associated with the globular cluster M15.

4.2.2 The Unique Case of the Double Pulsar

The second and unique other binary for which five PK parameters have been determined is the spectacular J0737−3039 system, also known as double pulsar. In April 2003, the discovery at Parkes of this unprecedented system [22, 23] represented a breakthrough in the science of the compact objects and started a new era in the study of relativistic gravity. The discovery occurred during a search for pulsars at high galactic latitude (the so-called *PH-survey* [83]) performed at 1.4 GHz using the multi-beam receiver of the Parkes radio telescope. The discovery plot clearly manifested the occurrence of a strong line-of-sight acceleration (about 100 m s^{-2}) on the pulsar, indicating that it was experiencing a relevant gravitational pull from a massive companion.

Three subsequent $\sim$ 5-h long integrations provided the radial velocity curve suggesting that it was a DNS binary system with an orbital period of only 2.4 h and an eccentricity $e \sim 0.09$. Subsequent observations also showed that both the compact objects were observable as a radio pulsar: PSR J0737−3039A (hereafter labeled as psrA) is a mildly recycled pulsar with a rotational period of about 22 ms, whereas PSR J0737−3039B (hereafter dubbed psrB) is an ordinary pulsar with a spin period of about 2.7 s.

At least four features conjure in order to make this binary an unprecedented laboratory for studying general relativity: *(a)* the large velocity of the stars along their orbit (more than 300 km s^{-1}, i.e., about 1000th of the speed of light in the vacuum); *(b)* the small distance between the two stars (between about 2 and 3 times the Earth–Moon distance); *(c)* the high orbital inclination (above 88°); *(d)* the observability of the radio pulses from both the stars, allowing one to use them as two *clocks* in the binary. In particular, the first two factors lead to significant relativistic effects, while the features *(c)* and *(d)* simplify the detectability of the effects with the timing procedure.

As a result, only few timing observations (spanning no more than about 1 week) were enough to measure the relativistic advance of periastron $\dot{\omega}$ of psrA [22]: about 16.9 $year^{-1}$, significantly larger than ever observed before. Total 6 months of additional ToAs led to determine three further PK parameters, that is, the combined time dilation and gravitational red-shift parameter γ, as well as the range r and the shape s of the Shapiro delay. The observability of both the pulsars made also possible to independently determine the size of the two orbits, which in turn—using the Kepler third law—led to the first direct measurement of the mass ratio R [23] in a DNS system. It is worth noting that the value of R is not affected by the *self-field* effects (at least for a very large class of metric theories of gravity), in contrast with the PK parameters [74]. In terms of the mass-vs-mass plot, that implies that the location of the pair of straight lines—both passing from the origin of the plot—which are associated with R (and with the observational uncertainty on R) is independent of the specific gravity theory.

The fifth PK parameter (i.e., the orbital decay $\dot{P}_b$) was also detected with high significance after another year of data taking. This confirmed that the system shrinks at the pace of about 7 mm day^{-1} and it will end up into a merging of the two NSs in

$\sim$ 85 million years, owing to emission of gravitational waves. This fact, combined with the short distance of this system, makes the double pulsar the dominant binary in the calculations for the merger rate of DNS systems in our galaxy and in the rest of the universe [22]. Despite many uncertainties still affecting the prediction of the absolute rate for those events, the discovery of the double pulsar determined a significant relative increase (up to about an order of magnitude, e.g., [84]) in any model, triggering new hopes for the ongoing ground-based interferometric experiments (e.g., LIGO and VIRGO) aimed at the detection of gravitational waves.

An additional relativistic effect—the *relativistic precession* of the spin axis of psrB[3]—has been more recently measured in the double pulsar system, exploiting the occurrence of short eclipses ($\sim$ 30 s long) of the radio signal from psrA (at least at frequencies smaller than $\sim$ 1.4 GHz) when it transits at the superior conjunction. About 4 years of observations revealed a clear linear evolution in the azimuthal spin axis angle, that is the angle which describes the precessional motion of the spin vector of psrB around the total angular momentum vector. The measured rate $\Omega_B = (4.77^{+0.66}_{-0.65})$ year^{-1} [86] is compatible, at the 13 % level, with the prediction of general relativity [87], $\Omega_B^{\rm GR} = (5.0734 \pm 0.0007)$ year^{-1}.

Although this effect has been somehow observed in other binary pulsars (e.g., see [88] for PSR B1534+12 and [89–91] for PSR B1913+16, and more recently see [92] for PSR J1141−6545), only in the case of the double pulsar it provides a significant test both to general relativity and alternate gravity theories. In fact, when dealing with the class of the Lorentz-invariant gravity theories based on a Lagrangian [76]—general relativity of course belongs to that class—one can write the following general formula for the relativistic spin precession rate of the NS labeled as B in the system:

$$\Omega_B = \left[\frac{x_A x_B}{s^2}\right]\left[\frac{8\pi^3}{(1-e^2)P_b^3}\right]\left[\frac{c^2\sigma_B}{G_{AB}}\right], \tag{19}$$

where x_A, x_B are the projected semimajor axes of the orbits of the two NSs in the system, σ_B is a (theory dependent) *strong-field spin-orbit* coupling constant and G_{AB} is the gravitational constant—also theory dependent!—for the gravity interaction between the two pulsars, whereas the other quantities are the usual Keplerian and PK parameters and constants as defined in previous sections. In order to solve for the rightmost factor, one needs a measurement of the spin precession, a precise determination of s, and separately, of x_A and x_B, which means that the binary must satisfy at least the three following properties: *(i)* a high spin precession rate, *(ii)* a highly inclined orbital plane, and *(iii)* the observability of both the NSs as pulsars. These features simultaneously hold only for the fortunate and so far unique case of the double pulsar, for which, inserting the measured parameters in Eq. 19, one obtains $[\frac{c^2\sigma_B}{G_{AB}}] = 3.38^{+0.49}_{-0.46}$. This provides an unprecedented strong-field test for all

[3] In many papers related with pulsars, this effect is referred to as *geodetic* (or as *De Sitter*) *precession*. For a brief description of the apparently discrepant terminology, see e.g., [85].

the gravity theories belonging to the large aforementioned class: a theory can survive only if it predicts for $[\frac{c^2\sigma_B}{G_{AB}}]$ a value in agreement with the one above. General relativity passes also this test, since in this theory $[\frac{c^2\sigma_B}{G_{AB}}]_{GR} = 2 + \frac{3m_A}{2m_B}$, and from the measured value of R, it results $\frac{3m_A}{2m_B} + 2 = 3.60677 \pm 0.00035$. As it involves the *strong self-field* of a NS, this test of general relativity has a wider nature with respect to the tests of the relativistic spin precession which have been performed and/or are in progress in the weak-field regime of the solar system. In particular, it supports the extension to the rotating bodies of the already introduced concept of *effacement* of gravity in general relativity [79], which had been previously tested for the orbital motion only. Namely the internal structure of the NS does not prevent the star to behave like a spinning test particle in a weak external field[4].

With the measurement of the spin precession rate of psrB—implying the appearance of a related pair of lines, labeled as Ω_B, in the mass-vs-mass diagram of the binary—there are now seven relativistic constraints (i.e., seven pairs of lines in Fig. 4), for the values of the masses of the two NSs in the J0737−3039 system: five from the PK parameters, one from the mass-ratio R, and one arising from Ω_B. On top of those, there are the additional classical constraints derived from the Newtonian mass function of the pulsars. Inspecting the inset in Fig. 4, one can see the existence of a common, although tiny, region of overlap for all the pairs of lines. As a consequence, after less than a decade from the discovery of the system, the observations of the double pulsar binary are already providing $7 - 2 = 5$ successful tests of self-consistency of general relativity in presence of strong-field effects [95]. The most stringent published test to date—and the best so far among all the gravitational experiments involving *strong-field* effects—is for the shape s of the Shapiro delay, which matches with the observations at 0.05 % level [96].

Since there is still a large room for future improvement of the timing solution for the double pulsar[5], a wealth of further intriguing results are expected to flood from keeping on observing this system and from the introduction of more sensitive observational apparatuses. For example, a recent interferometric determination of the parallax of the system [98] indicated that over a decade, $\dot{P}_b$ may be observed with an accuracy of $\gtrsim 0.01\%$, paving the way to stringent tests of the theories that predict dipolar gravitational radiation (contrary to the quadrupolar only contribution resulting from general relativity, see Sect. 4.2.3).

Continuous accumulation of ToAs will also lead to measure new PK parameters (such as the aberration ones), and to derive constraints on the existence of a preferred frame and thus on the violation of the local Lorentz invariance of gravity in the strong internal fields of the NSs [99]. Moreover, the availability of a (decennial) series of high quality ToAs may lead to unveil the signature of the second order post-Newtonian corrections to the orbital periastron advance $\dot{\omega}$ of the system. This would

[4] Note that, contrary to common expectations, no precession of the spin axis of psrA in the double pulsar has been detected so far [93], which may suggest the occurrence of a small misalignment angle between the orbital angular momentum and the spin axis of psrA, likely less than $\sim 15°$ [94].

[5] The recent (but supposedly temporary) disappearance of the radio signal from psrB [97] will not hamper the improvement of the timing accuracy of the solution for psrA.

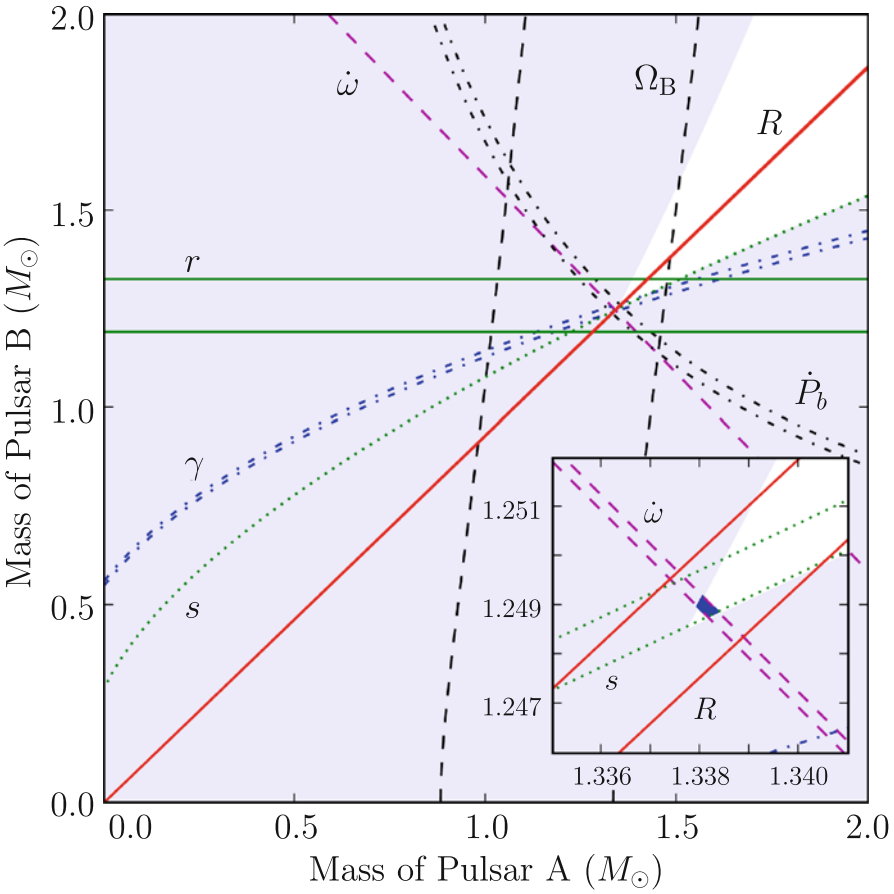


Fig. 4 Mass-vs-mass diagram for the double pulsar system J0737−3039. The *shaded regions* are those that are excluded by the Newtonian mass functions of the two pulsars. Further constraints are shown as pairs of lines enclosing permitted regions as predicted by general relativity and related to the observation of the mass ratio R, the PK parameters, and the precession rate Ω_B of the spin axis of psrB. *Inset* is an enlarged view of the small quadrilateral encompassing the intersection of all these constraints. (Courtesy of René Breton 2009)

be of paramount importance, since these corrections will include the value of the spin angular momentum of psrA, opening the intriguing perspective to measure that and therefore to determine the moment of inertia of psrA [100]. As a matter of fact, this measurement may finally reveal to be at the borderline of the possibilities of the current instrumentation [7], but might be attained when new more powerful radio telescopes will enter in play. The scientific payback from that would be enormous, since the simultaneous determination of the mass and the moment of inertia of a NS would significantly constrain the equation of state for the nuclear matter [101], shedding light on a long standing and still unanswered fundamental question of the nuclear physics.

4.2.3 Other Interesting Cases and Constraints on Tensor–Scalar Theories

As discussed in previous sections there are few binary pulsars with a white dwarf companion for which one or more PK parameters have been measured. In most cases, they are only the shape s and range r of the Shapiro delay, which allow one to infer with good accuracy the mass of the companion, but are not enough for constraining the gravitational theories.

However, at the time of writing these notes, there are at least four notable exceptions, the binary pulsars PSR J0437−4715, PSR J1141−6545, PSR J1738+0333, and PSR J0348+0432 (sorted in chronological order of discovery), the cases of which we briefly illustrate here, since their continuous timing is producing results which are complementary to those of the double pulsar and the other two relativistic DNSs (PSR B1913+16 and PSR B1534+12) which we previously described.

The former was discovered in 1993 during a survey at 400 MHz at the Parkes radio telescope [102]. Two factors make PSR J0437−4715 a primary target for high precision timing: its proximity to the Earth (it is indeed the closest known recycled pulsar, $d = 156.3 \pm 1.3$ pc [103]) and its brightness (flux density of 142 mJy at 1.4 GHz, a factor ~ 14 larger than the second most luminous recycled pulsar). In fact, the closeness to us produce significant geometrical effects, such us *(i)* the so-called *annual-orbital parallax* [104] (i.e., the apparent variation of the inclination angle of the binary system) due to the orbital motion of our planet, and *(ii)* the apparent secular change in the projected semimajor axis of the pulsar orbit due to the pulsar proper motion [105]. Accounting in the timing model for these two apparent phenomena, and exploiting the exceptionally accurate ToAs achievable for this system, it was possible [106] to infer both the inclination angle i and the longitude of the ascending node (a Keplerian parameter which is usually not measurable via pulsar timing) of the system. The derived value of i—thus determined in a completely independent way—was then compared with the expected effects on the ToAs of the Shapiro delay, finding a nice agreement of the data with the prediction of general relativity, for which the PK parameter s is simply equal to $\sin i$. The best pulsar test on the time variation of the Newtonian constant of gravity G is also so far provided by this system: $|\dot{G}/G| < 23 \times 10^{-12}$ year^{-1} (see [107]).

PSR J1141−6545 (discovered on 1999 at Parkes [108]) is a rare example of a binary pulsar, the WD companion[6] of which is older than the NS, due to a reversal in the mass ranking of the two stars in the system—that is—the originally less massive object became the most massive one after the first stages of mass transfer during the binary evolution [110]. Whence, the originally more massive star ended its evolution as a massive (1.02 ± 0.01 M$_\odot$) WD [111], whereas the originally less massive star later exploded as a supernova leaving a relatively young (~ 1.4 Myear [108]) and not recycled (spin period of 394 ms) pulsar with a mass of 1.27 ± 0.01 M$_\odot$ [111], which is now orbiting the WD in an elliptical and short orbit ($e \sim 0.17$, $P_b \sim 4.74$ h) [108]. Despite the relatively long rotational period, the latter two properties

[6] The recent optical detection of the companion supports its being a WD [109]

(as well as the brightness of the source) led to clearly detect three PK parameters (namely $\dot{\omega}$, γ, $\dot{P}_b$), allowing [111] to infer the masses of the two bodies (as reported above) and—after correction for the spurious effects of Eq. 18—to test the radiative prediction (i.e., the damping of the orbit due to gravitational waves emission) of general relativity at better than 10 % level [111]. Also, it has been very recently detected the occurrence of the *relativistic precession* [92] of the spin axis of the pulsar.

Although the aforementioned test of general relativity with PSR J1141−6545 is far from approaching the accuracy of those obtained with the J0737−3039 or the B1913+16 systems, it is more useful than the other tests in constraining a particularly important class of alternate gravity theories, that is, the so-called *tensor–scalar* theories, in which the usual tensor field is complemented with one scalar field ψ. Besides many other additional advantages (see e.g., [59, 112]), these gravity theories are a natural outcome of many unified theories (i.e., superstring, Kaluza-Klein, etc. [59, 112]) invoking a scalar counterpart for the graviton and they often appear as a key ingredient in cosmology for explaining the past or the current phases of accelerated expansion (i.e., inflation, quintessence) of the universe. In particular, the *tensor–mono-scalar* theories can be grouped in a family, parametrized by two constants, α_0 and β_0, which describe the linear and the quadratic coupling between the matter and the additional scalar field, according to the expression $a(\psi) = \alpha_0\psi + 0.5\beta_0\psi^2$, where $a(\psi)$ is the coupling function, and $\exp[2a(\psi)]g_{\mu\nu}$ replaces the role of the ordinary spin-2 metric tensor $g_{\mu\nu}$ of general relativity. Of course, the Einsteinian theory is fully recovered for $\alpha_0 = \beta_0 = 0$, whereas, for example, the Jordan–Fierz–Brans–Dicke gravitational theory [113] is accounted for by assuming $\beta_0 = 0$ and $\alpha_0^2 = 1/(2\omega_{BD} + 3)$, where ω_{BD} is the Brans–Dicke parameter (e.g., [114]). In this context, the experimental tests aim to put constraints in the two-dimensional space of (α_0, β_0), where the origin coincides with the theory of general relativity, the vertical axis with the Brans–Dicke theory, and so on [115]: one such plot is reported in Fig. 5.

The orbital damping rate $\dot{P}_b$ is the most affected PK parameter by the occurrence of a scalar field. In particular the value of $\dot{P}_b$ should be enhanced by the emission of *dipolar* gravitational radiation (due to emission of scalar spin-0 waves) with a power of order $\mathcal{O}(1/c^3)$, superimposed to the quadrupolar component (associated to spin-2 gravitons) of general relativity having a much smaller power of order $\mathcal{O}(1/c^5)$. However, for two perfectly identical NSs in a binary, the dipolar term obviously vanishes and it almost cancels out also for two NSs of similar masses, since their compactness parameters (related to the dimensionless ratio ε) are supposed to be similar. This is not the case for a WD, whose compactness parameter (and hence the coupling of matter with the scalar field) is expected to be much smaller than for a NS. Therefore, the binary pulsars orbiting a WD in clean binary systems (i.e., with no uncontrolled effects affecting $\dot{P}_b$, like tidal interactions or mass loss or magnetic braking...) are primary targets for investigating the occurrence of scalar waves and in turn for constraining the space of the parameters of the *tensor–scalar* theories.

Figure 5 illustrates the important role played by PSR J1141−6545 in this framework. However, another binary pulsar, PSR J1738+0333 (discovered during 2001

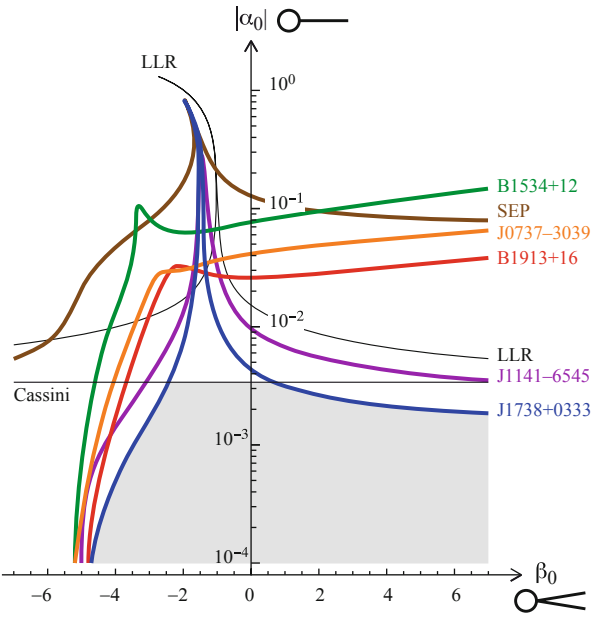


Fig. 5 Experimental constraints on the $|\alpha_0|$ (logarithmic scale) and β_0 (linear scale) constants of the coupling function $a(\psi) = \alpha_0\psi + 0.5\beta_0\psi^2$ of the matter with a scalar field ψ. The point in the diagram at coordinates ($|\alpha_0| \to 0$; $\beta_0 = 0$) corresponds to general relativity. The *vertical axis* corresponds to Brans–Dicke theory, whereas the *horizontal axis* is the locus of the theories which are indistinguishable from general relativity in the weak-field conditions of the solar system experiments. The allowed regions are below the various *solid lines*, related to the observations of various binary pulsar systems (B1534+12, J0737−3039, B1913+16, J1141−6545, J1738+0333) and to the outcomes of several experiments performed in the weak-field limit of the solar system, including the lunar laser ranging (*LLR*) and the Cassini spacecraft's results. The *shaded area* is the one allowed so far by all the tests. It is evident the primary role of PSR J1738+0333 in setting some of the current limits. (Courtesy of Paulo Freire 2012 [37])

at Parkes [116] and orbiting a low-mass WD companion of $\sim 0.18\ M_\odot$ in a ~ 8.5 h orbit of low eccentricity, $e \sim 3 \times 10^{-7}$), recently gained the pole position in the ranking of the most useful binary pulsars in testing (and constraining the parameters of) the tensor–scalar theories [37]. A decadelong timing campaign at Arecibo provided a determination of $\dot{P}_b$, as well as an extremely accurate determination of the proper motion and the parallax of the system, allowing for a precise subtraction of the kinematic contribution to the observed orbital decay. This led to a determination of the intrinsic $\dot{P}_b^{\rm int} = -25.9 \pm 3.2 \times 10^{-15}$, to be compared with the prediction of general relativity $\dot{P}_b^{\rm GR} = -27.7^{+1.5}_{-1.9} \times 10^{-15}$ (the latter prediction relies on the values of the mass of the WD and of the pulsar, obtained from optical photometry and spectroscopy [36]). The nice agreement between the two values introduces a very strong upper limit on *dipolar* gravitational wave emission, which is reflected on the diagram of Fig. 5.

Finally, it is worth noting also the rejection of any model with $\beta_0 \lesssim -5$ which is imposed by J1141−6545, J1738+0333, as well as by B1534+12, J0737−3039, and B1913+16 systems. This is a pure strong-field effect (e.g., [117]), which can be exploited only using the radio pulsars, while the investigation of such low negative values of β_0 is escaping all the solar system experiments.

Additional constraints on deviations from GR due to strong self-field effects derive from the discovery at the Green Bank radio telescope (reported by [118,119] during 2013), and the subsequent radio timing and optical observations, of the J0348+0432 system [6]. It includes a WD and a massive pulsar (2.01 ± 0.04 $M_\odot$) in a 2.46 h almost circular orbit. Given its large mass, PSR J0348+0432 has a gravitational binding energy $\sim$ 60% larger than that of all the other known pulsars with a measured orbital decay (the highest mass of which is $1.46^{+0.06}_{-0.05}$ $M_\odot$ for PSR J1738+0333), and the self-field effects (scaling nonlinearly with the binding energy) are expected to be much more prominent in this binary than in the ones described above. Therefore, the consistency (at the 18 % level at the time being [6]) of the observed orbital decay with the predictions of GR can be used [6] to pose stringent limits to the occurrence of strong-field deviations from the Einsteinian theory in the form of a long-range field[7].

5 Perspectives

As anticipated in the previous sections, a wealth of improved/new tests of gravity theories is expected to emerge from keeping on observing the already known systems, both with the current generation or with new timing instruments, like MeerKAT (in South Africa) and/or the Large European Array for Pulsars (LEAP), which combines the five largest European radio telescopes to generate the equivalent of a $\sim$ 200-m dish [120]. Other relativistic [pulsar + NS] or [pulsar + WD] binaries should also be discovered by the ongoing large search experiments, at Parkes [121], Green Bank [122], Arecibo [123], and Effelsberg radio telescopes, all of them having better sensitivity than all previous surveys. Targeted searches towards gamma-ray point sources [124] have also already demonstrated to be a new promising channel for unveiling recycled pulsars, some of them likely included in relativistic binaries.

Looking on a longer decennial timescale, the perspectives in this field of research will undergo a new gigantic step forward when novel large instruments, like the Square Kilometer Array (SKA) [125] or the Five-hundred meter A-spherical Single-aperture radio Telescope (FAST) [126] will enter in operation. In particular, they will almost certainly lead to the discovery of some [pulsar+black hole] binaries, the

[7] Obviously, short-range fields, not affecting the binary dynamics, cannot be constrained with this kind of tests.

long-dreamt and still unsatisfied objective of all the current pulsar search collaborations. These binaries are expected to be very rare, but, besides the most obvious case of a pulsar orbiting the super-massive black hole in the galactic centre ([127] and references therein), at least some pulsars should be found with a stellar-mass black hole companion (e.g., [128]) and, even more interestingly, some recycled pulsars could be discovered in a globular cluster orbiting an intermediate mass black hole [129].

Timing a natural clock orbiting a spinning black hole will give the unprecedented possibility of a direct measurement not only of the mass of the black hole $M_{\rm BH}$, but also of its angular momentum $S_{\rm BH}$ and its quadrupole moment $Q_{\rm BH}$, thus in turn allowing one to experimentally test [130, 131] the so-called *Cosmic Censorship Conjecture* [132], and the *No-Hair Theorem*. The first states that no-naked singularity (i.e., a singularity not hidden behind an event horizon) exists, while the second affirms that any black hole can be described only on the basis of $M_{\rm BH}$, $S_{\rm BH}$, and its electric charge, the latter not being usually relevant in an astrophysical context. Since the event horizon tends to shrink as a black hole spins up, the *Cosmic Censorship Conjecture* implies that there is a maximum spin for a black hole of a given mass; on the other hand the *No-Hair Theorem* requires that $Q_{\rm BH}$ must be expressed in terms of $M_{\rm BH}$ and $S_{\rm BH}$. Introducing the dimensionless spin and quadrupole parameters χ and q,

$$\chi = \frac{c}{G}\frac{S_{\rm BH}}{M_{\rm BH}^2} \;\; ; \;\; q = \frac{c^4}{G^2}\frac{Q_{\rm BH}}{M_{\rm BH}^3}, \tag{20}$$

one then expects a black hole to satisfy in general relativity the equations

$$\chi \leq 1 \;\; ; \;\; q = -\chi^2, \tag{21}$$

which can be experimentally checked with the timing of a [pulsar+black hole] binary.

Although it is not within the scope of this review, it is worth concluding by briefly recalling the very promising expectations of the so-called *pulsar timing arrays* for tests of gravitational theories [133]. They are based on the consideration of the Earth and a pulsar as a pair of test masses in the space-time metric; the passage of a gravitational wave perturbs the metric and is then expected to leave a signature on the observed timing residuals of the pulsar. The size of the effect on the residuals is very small and is proportional to the amplitude of the characteristic strain (see the chapter by Braccini and Fidecaro in this book for an introduction to the gravitational waves and a description of their basic parameters). The correlation between the data of many Earth-pulsar pairs [134] distributed throughout the sky (thus forming a PTA) is required to remove spurious effects on the timing residuals and lead to a significant detection of gravitational waves. In particular, the PTAs are most sensitive to gravitational waves with periods $\gtrsim$ 1 year (i.e., frequency $\lesssim$ 10 nHz), and therefore probe the nanoHertz gravitational wave sky, thus being nicely complementary to the other experiments aimed at a direct detection of gravitational waves (like LIGO/VIRGO, Advanced LIGO/VIRGO, eLISA and CMB-POL), all operating in other frequency

bands. The most suitable target for the current PTAs [the Australian one (PPTA; [135]), the European one (EPTA, embedding the aforementioned LEAP experiment; [136]), and the US–Canadian one (nanoGrav; [137])], as well as for the international joined effort, labeled IPTA [138], is the stochastic background—or better saying the stocastic *foreground*—of gravitational waves produced by coalescing super-massive black hole binaries in the early stage of assembling of the galaxies (red-shift ~ 1) in the universe. Although no claim for a detection has been announced so far by the various involved groups (the most updated upper limits are reported in [139–141]), the present trend in the timing accuracy of the experiments and the calculation of the cosmological models lend support to the possibility of a clear detection within a few years [139, 142, 143]. Since most metric gravitational theories (e.g., all of those which are Lorentz invariant) unavoidably predict emission of waves from coalescing binaries [144], it would be very surprising (and would determine a revolution in physics or in cosmology) if not only the current instruments, but also SKA and FAST will fail to detect the aforementioned foreground. What seems more likely is that the superb sensitivity of SKA and FAST will not only lead to a simple detection, but will give the chance of using the properties of the observed gravitational waves for constraining the alternative gravity theories (e.g., by studying the mass and the spin of the gravitons [145]), as well as for characterizing the parameters of the massive black hole binary systems (e.g., [146, 147]).

With all these exciting perspectives, pulsar science is more vital than ever and promises to give additional fundamental contributions to our understanding of fundamental physics and relativistic gravity just on the eve of its first half-century.

Acknowledgements AP and MB received support for this review under the program *Progetti per Ricerca di Base 2010* of Regione Autonoma della Sardegna (L.R. 7/2007). AP and MB acknowledge the collaborators at INAF-Osservatorio di Cagliari and the people all around the world who are contributing to make the study of the pulsars an always stimulating and often surprising activity. In particular P. Esposito for having carefully read an early version of this manuscript, giving useful suggestions for making it handier for the neophytes of this field of research.

References

1. Cordes JM, Lazio TJW. NE2001. I. A new model for the galactic distribution of free electrons and its fluctuations. 2002;arXiv:astro-ph/0207156.
2. Han JL, Manchester RN, Lyne AG, Qiao GJ, van Straten W. Pulsar rotation measures and the large-scale structure of the galactic magnetic field. ApJ. 2006;642:868–81.
3. Colpi M, Possenti A, Gualandris A. The case of PSR J1911-5958A in the outskirts of NGC 6752: signature of a black hole binary in the cluster core? ApJ. 2002;570:L85–8.
4. Hessels JWT, et al. A radio pulsar spinning at 716 Hz. Science. 2006;311:1901.
5. Demorest P, Pennucci T, Ransom S, Roberts M, Hessels JWT. Shapiro delay measurement of a 2 solar mass neutron star. Nature. 2010;467:1081.
6. Antoniadis J, et al. A massive pulsar in a compact relativistic binary. Science. 2013;340:1233232.
7. Wex N, Kramer M. Generic gravity tests with the double pulsar. Proceedings of the 12th Marcel Grossmann meeting on general relativity (MG 12), Paris 2010;arXiv:1001.4733.

8. Haberl F. The magnificent seven: magnetic fields and surface temperature distributions. Ap&SS. 2007;308:181.
9. Pavlov GG, Sanwal D, Teter MA. Central compact objects in supernova remnants. In: Camilo F, Gaensler B, editors. Young neutron stars and their environments, IAUS. 2004;218:239.
10. Keane EF, McLaughlin MA. Rotating radio transients. BASI. 2011;39:333.
11. Woods PM, Thompson C. Soft gamma repeaters and anomalous X-ray pulsars: magnetar candidates. In: Lewin W, van der Klis M, editors. Compact stellar X-ray sources. Cambridge Astrophysics Series. Vol. 39. 2006. p. 547–86.
12. Chen K, Ruderman M. Pulsar death lines and death valley. ApJ. 1993;2:264–70.
13. Manchester RN, Hobbs GB, Teoh A, Hobbs M. The Australia telescope national facility pulsar catalogue. AJ. 2005;129:1993.
14. http://www.atnf.csiro.au/research/pulsar/psrcat/. Accessed 30 Sep 2015.
15. http://www.physics.mcgill.ca/∼pulsar/magnetar/main.html. Accessed 30 Sep 2015.
16. Alpar MA, Cheng AF, Ruderman MA, Shaham J. A new class of radio pulsars. Nature. 1982;300:728–30.
17. Verbunt F. Origin and evolution of X-ray binaries and binary radio pulsars. Ann Rev Astron Astrophys. 1993;31:93.
18. Podsiadlowski P. Electron-capture supernovae and accretion-induced collapse of ONeMg white dwarfs. In: Paths to Exploding Stars: Accretion and Eruption, KITP conference, Kavli Institute, Santa Barbara; March 2007.
19. Rappaport S, Pfahl E, Rasio FA, Podsiadlowski P. Formation of compact binaries in globular clusters. In: Podsiadlowski P, Rappaport S, King AR, D'Antona F, Burderi L, editors. Evolution of binary and multiple star systems. ASP Conf Ser. 2001;229:409.
20. Popov SB, Pons JA, Miralles JA, Boldin PA, Posselt B. Population synthesis studies of isolated neutron stars with magnetic field decay. MNRAS. 2010;401:2675.
21. Faucher-Giguère C-A, Kaspi VM. Birth and evolution of isolated radio pulsars. ApJ. 2006;643:332.
22. Burgay M, et al. An increased estimate of the merger rate of double neutron stars from observations of a highly relativistic system. Nature. 2003;426:531–33.
23. Lyne AG, et al. A double-pulsar system: a rare laboratory for relativistic gravity and plasma physics. Science. 2004;303:1153–7.
24. Nice DJ, Sayer RW, Taylor JH. PSR J1518+4904: a mildly relativistic binary pulsar system. ApJ. 1996;466:L87–90.
25. Wolszczan A. PSR 1257+12 and PSR 1534+12. IAU Circ., 5073, 1, 1990; (Green DWE, editor).
26. Stairs IH, et al. Measurement of relativistic orbital decay in the PSR B1534+12 binary system. ApJ. 1998;505:352–7.
27. Faulkner AJ, et al. PSR J1756−2251: a new relativistic double neutron star system. ApJ. 2005;618:L119–22.
28. Lyne AG, et al. The Parkes multibeam pulsar survey: PSR J1811−1736—a pulsar in a highly eccentric binary system. MNRAS. 2000;312:698–702.
29. Corongiu A, et al. The binary pulsar PSR J1811−1736: evidence of a low amplitude supernova kick. A&A. 2007;462:703–9.
30. Champion DJ, et al. PSR J1829+2456: a relativistic binary pulsar. MNRAS. 2004;350: L61–5.
31. Lorimer DR, et al. Arecibo pulsar survey using ALFA. II. The young, highly relativistic binary pulsar J1906+0746. ApJ. 2006;640:428–34.
32. Hulse RA, Taylor JH. Discovery of a pulsar in a binary system. ApJ. 1975;195:L51–3.
33. Weisberg JM, Taylor JH. The relativistic binary pulsar B1913+16. In: Bailes M, Nice DJ, Thorsett S, editors. Radio pulsars. San Francisco: Astronomical Society of the Pacific; 2003. p. 93–8.
34. Anderson SB, Gorham PW, Kulkarni SR, Prince TA. Discovery of two radio pulsars in the globular cluster M15. Nature. 1990;346:42–4.

35. Jacoby BA, et al. Measurement of orbital decay in the double neutron star binary PSR B2127+11C. ApJ. 2006;644:L113–6.
36. Antoniadis J, van Kerkwijk MH, Koester D, et al. The relativistic pulsar-white dwarf binary PSR J1738+0333—I. Mass determination and evolutionary history, MNRAS. 2012;423:3316.
37. Freire PPC, Wex N, Esposito-Farese G, et al. The relativistic pulsar-white dwarf binary PSR J1738+0333—II. The most stringent test of scalar-tensor gravity. MNRAS. 2012;423:3328.
38. Ransom SM, Stairs IH, Archibald AM, et al. A millisecond pulsar in a stellar triple system. Nature. 2014;505:520.
39. Tauris T, van den Heuvel E. Formation of the galactic millisecond pulsar triple system PSR J0337+1715A neutron star with two orbiting white dwarfs. ApJ. 2014;781:L13.
40. Lorimer DR, Kramer M. Handbook of pulsar astronomy. Cambridge: Cambridge University Press; 2005.
41. Lyne AG, Smith G. Pulsar astronomy. 3rd ed. Cambridge: Cambridge University Press; 2005.
42. Manchester RN, Taylor JH. Pulsars. San Francisco: Freeman & Co.; 1977.
43. E.g. http://ssd.jpl.nasa.gov/?planet_eph_export. Accessed 30 Sep 2015.
44. E.g. http://www.iers.org/IERS/EN/Publications/Bulletins/bulletins.html. Accessed 30 Sep 2015.
45. Shapiro II. Fourth test of general relativity. Phys Rev Lett. 1964;13:789.
46. van Straten W, Bailes M, Britton M, Kulkarni SR, Anderson SB, Manchester RN, Sarkissian J. A test of general relativity from the three-dimensional orbital geometry of a binary pulsar. Nature. 2001;412:158.
47. Anderson PW, Itoh N. Pulsar glitches and restlessness as a hard superfluidity phenomenon. Nature. 1975;256:25.
48. Alpar MA, Nandkumar R, Pines D. ApJ. 1986;311:197.
49. Shabanova TV. Slow glitches in the pulsar B1822-09. Astrophys Space Sci. 2007;308:591.
50. Lyne AG, Hobbs G, Kramer M, Stairs I, Stappers B. Switched magnetospheric regulation of pulsar spin-down. Science. 2010;329:408.
51. Hobbs G, Lyne AG, Kramer M. An analysis of the timing irregularities for 366 pulsars. MNRAS. 2010;402:1027.
52. Cognard I, Backer DC. A microglitch in the millisecond pulsar PSR B1821-24 in M28. ApJ. 2004;612:125.
53. Shannon RM, Cordes JM. Assessing the role of spin noise in the precision timing of millisecond pulsars. ApJ. 2010;725:1607.
54. Will C. Theory and experiments in gravitational physics. Cambridge: Cambridge University Press; 1993.
55. Will C. The confrontation between general relativity and experiment. In: Ciufolini I, Matzner, editors. General relativity and John Archibald Wheeler. Springer Netherlands: Astrophysics and Space Science Library. 2010;367;73.
56. Damour T, Esposito-Farèse G. Testing gravity to second post-Newtonian order: a field-theory approach. Phys Rev D. 1996;53:5541.
57. Will C. The confrontation between general relativity and experiment. LLR. 2006;9:3 (http://relativity.livingreviews.org/Articles).
58. Will C, Nordvedt KJ. Conservation laws and preferred frames in relativistic gravity. I. Preferred-frame theories and an extended PPN formalism. ApJ. 1972;177:757.
59. Damour T, Esposito-Farèse G. Tensor-multi-scalar theories of gravitation. Class Quantum Gravity. 1992;9:2093.
60. Stairs I. Testing general relativity with pulsar timing. LLR. 2003;6:5. http://relativity.livingreviews.org/Articles. Accessed 30 Sep 2015.
61. Nordtvedt K. Testing relativity with laser ranging of the moon. Phys Rev. 1968;170:1186.
62. Stairs IH, et al. Discovery of three wide-orbit binary pulsars: implications for binary evolution and equivalence principles. ApJ. 2005;632:1060.
63. Gonzales ME, et al. High-precision timing of five millisecond pulsars: space velocities, binary evolution and equivalence principles. Astrophys J. 2011;743:102.

64. Damour T, Schäfer G. New tests of the strong equivalence principle using binary pulsar data. Phys Rev Lett. 1991;66:2549.
65. Freire PCC, Kramer M, Wex N. Tests of the universality of free fall for strongly self-gravitating bodies with radio pulsars. Class Quantum Gravity. 2012;29:184007.
66. Shao L, Wex N. New tests of local Lorentz invariance of gravity with small-eccentricity binary pulsars. Class Quantum Gravity. 2012;29:215018.
67. Shao L, Caballero RN, Kramer M, et al. A new limit on local Lorentz invariance violation of gravity from solitary pulsars. Class Quantum Gravity. 2013;30:165019.
68. Müller J, Williams JG, Turyshev SG. Lunar laser ranging contributions to relativity and geodesy. In: Dittus H, Lammerzahl C, Turyshev SG, editors. Lasers, clocks and drag-free control: exploration of relativistic gravity in space. Berlin Springer: Astrophysics and Space Science Library. 2008;349:457.
69. Nordtvedt K. Probing gravity to the second post-Newtonian order and to one part in 107 using the spin axis of the sun. ApJ. 1987;320:871.
70. Shao L, Wex N. New limits on the violation of local position invariance of gravity. Class Quantum Gravity. 2013;30:165020.
71. Warburton RJ, Goodkind JM. Search for evidence of a preferred reference frame. ApJ. 1976;208:881.
72. Will CM. Is momentum conserved? A test in the binary system PSR B1913+16. ApJ. 1992;393:L59.
73. Damour T, Deruelle N. General relativistic celestial mechanics of binary systems. I. The post-Newtonian motion. Ann Inst H Poincaré. (Phys Théor.) 1985;43:107.
74. Damour T, Deruelle N. General relativistic celestial mechanics of binary systems. II. The post-Newtonian timing formula. Ann Inst H Poincaré. (Phys Théor.) 1986;44:263.
75. Taylor JH, Weisberg JM. Further experimental tests of relativistic gravity using the binary pulsar PSR 1913+16. ApJ. 1989;345:434.
76. Damour T, Taylor JH. Strong-field tests of relativistic gravity and binary pulsars. Phys Rev D. 1992;45:1840.
77. Hulse RA, Taylor JH. Discovery of a pulsar in a binary system. ApJ. 195:L51.
78. Weisberg JM, Nice DJ, Taylor JH. Timing measurements of the relativistic binary pulsar PSR B1913+16. ApJ. 2010;722:1030.
79. Damour T. The problem of motion in Newtonian and Einsteinian gravity. In: Hawking SW, Israel W, editors. Three hundred years of gravitation. Cambridge: Cambridge University Press; 1987. p. 128.
80. Damour T, Taylor JH. On the orbital period change of the Binary Pulsar PSR 1913+13. ApJ. 1991;366:501.
81. Wolszczan A. A nearby 37.9 ms radio pulsar in a relativistic binary system. Nature. 1991;350:688.
82. Stairs IH, Thorsett SE, Taylor JH, Wolszczan A. Studies of the relativistic binary pulsar PSR B1534+12: I. Timing analysis. ApJ. 2002;581:501.
83. Burgay M, et al. The Parkes high latitude pulsar survey. MNRAS. 2006;368:283.
84. Kalogera V, et al. The cosmic coalescence rates for double neutron star binaries. ApJ. 2004;601:L179.
85. Stella L, Possenti A. Lense–Thirring precession in the astrophysical context. Space Sci Rev. 2009;148:105.
86. Breton RP, et al. Relativistic spin precession in the double pulsar. Science. 2008;321:104.
87. Barker BM, O'Connell RF. Gravitational two-body problem with arbitrary masses, spins, and quadrupole moments. Phys Rev D. 1975;12:329.
88. Stairs IH, Thorsett SE, Arzoumanian Z. Measurement of gravitational spin-orbit coupling in a binary-pulsar system. Phys Rev Lett. 2004;93:141101.
89. Kramer M. Determination of the geometry of the PSR B1913+16 system by geodetic precession. ApJ. 1998;509:856.
90. Weisberg JM, Taylor JH. General relativistic geodetic spin precession in binary pulsar B1913+16: mapping the emission beam in two dimensions. ApJ. 2002;576:942.

91. Clifton T, Weisberg JM. A simple model for pulse profiles from precessing pulsars, with special application to relativistic binary PSR B1913+16. ApJ. 2008;679:687.
92. Manchester RN, et al. Observations and modeling of relativistic spin precession in PSR J1141-6545. ApJ. 2010;710:1694.
93. Manchester RN, et al. The mean pulse profile of PSR J0737-3039A. ApJ. 2005;621:49.
94. Ferdman RD, et al. The double pulsar: evolutionary constraints from the system geometry. In 40 years of pulsars. AIP Conf Proc. 2008;983:474.
95. Kramer M, Stairs I. The double pulsar. Ann Rev A&A. 2008;46:541.
96. Kramer M, et al. Tests of general relativity from timing the double pulsar. Science. 2006;314:97.
97. Perera BBP, et al. The evolution of PSR J0737−3039B and a model for relativistic spin precession. ApJ. 2010;721:1193.
98. Deller AT, Bailes M, Tingay SJ. Implications of a VLBI distance to the double pulsar J0737-3039A/B. Science. 2009;323:1327.
99. Wex N, Kramer M. A characteristic observable signature of preferred-frame effects in relativistic binary pulsars. MNRAS. 2007;380:455.
100. Damour T, Schaefer G. Higher-order relativistic periastron advances and binary pulsars. Nuovo Cim. 1988;101:127.
101. Lattimer JM, Schutz BF. Constraining the equation of state with moment of inertia measurements. ApJ. 2005;629:979.
102. Johnston S, et al. Discovery of a very bright, nearby binary millisecond pulsar. Nature. 1993;361:613.
103. Deller A, Tingay SJ, Bailes M, Reynolds JE. Precision southern hemisphere VLBI pulsar astrometry. II. Measurement of seven parallaxes. ApJ. 2009;701:1243.
104. Kopeikin SM. On possible implications of orbital parallaxes of wide orbit binary pulsars and their measurability. ApJ. 1995;439:L5.
105. Kopeikin SM. Proper motion of binary pulsars as a source of secular variation of orbital parameters. ApJ. 1996;467:L93.
106. van Stratten W, et al. A test of general relativity from the three-dimensional orbital geometry of a binary pulsar. Nature. 2001;412:158.
107. Verbiest J, et al. Precision timing of PSR J0437-4715: an accurate pulsar distance, a high pulsar mass, and a limit on the variation of Newton's gravitational constant. ApJ. 2008;679:675.
108. Kaspi VM, et al. Discovery of a young radio pulsar in a relativistic binary orbit. ApJ. 2000;543:321–7.
109. Antoniadis J, Bassa CG, Wex N, Kramer M, Napiwotzki R. A white dwarf companion to the relativistic pulsar PSR J1141–6545. MNRAS. 2011;412:580.
110. Tauris TM, Sennels T. Formation of the binary pulsars PSR B2303+46 and PSR J1141-6545. Young neutron stars with old white dwarf companions. A&A. 2000;355:236.
111. Bhat NDR, Bailes M, Verbiest JPW. Gravitational-radiation losses from the pulsar white-dwarf binary PSR J1141 6545. Phys Rev D. 2008;77.12:124017.
112. Damour T, Esposito-Farèse G. Tensor-scalar gravity and binary pulsar experiments. Phys Rev D. 1996;54:1474.
113. Brans C, Dicke RH. Mach's principle and a relativistic theory of gravitation. Phys Rev. 1961;124:925.
114. Will CM, Zaglauer HW. Gravitational radiation, close binary systems, and the Brans–Dicke theory of gravity. ApJ. 1989;346:366.
115. Esposito-Farèse G. Binary-pulsars tests of strong-field gravity and gravitational radiation damping. In: Proceedings of X Marcel Grossman meeting, Rio de Janeiro, 2003 July; 2005. p. 647.
116. Jacoby BA. Thesis at California Institute of Technology: California—USA, Publication Number: AAT 3161140; 2005.
117. Esposito-Farèse G. Motion in alternative theories of gravity. Lecture given at School on Mass, Orléans (France), June 2008, Arxiv:0905.2575; 2009.

118. Boyles J, et al. The Green Bank telescope 350 MHz drift-scan survey. I. Survey observations and the discovery of 13 pulsars. Astrophys J. 2013;763:80.
119. Lynch RS, et al. The Green Bank telescope 350 MHz drift-scan survey. II: Data analysis and the timing of 10 new pulsars, including a relativistic binary. Astrophys J. 2013;763:81.
120. Ferdman RD, et al. The European pulsar timing array: current efforts and a LEAP toward the future. Class Quant Gravity. 2010;27:084014.
121. Keith MJ, et al. The high time resolution universe pulsar survey I: system configuration and initial discoveries. MNRAS. 2010;409:619.
122. Boyles J, et al. Timing derived properties of the first discoveries from the GBT 350-MHz pulsar survey. Bull AAS. 2010;42:464.
123. Beroiz M, et al. Current results at PALFA Pulsar Survey at Arecibo Observatory. Bull AAS. 2010;42:460.
124. Abdo AA, et al. Detection of 16 gamma-ray pulsars through blind frequency searches using the Fermi LAT. Science. 2009;325:840.
125. Smits R. Pulsar searches and timing with the square kilometre array. A&A. 2009;493:1161.
126. Smits R. Pulsar science with the five hundred metre aperture spherical telescope. A&A. 2009;506:919.
127. Deneva IS. Elusive neutron star populations: galactic center and intermittent pulsars. Ph.D. dissertation, Cornell University, USA; 2010.
128. Pfahl E, Podsiadlowski P, Rappaport S. Relativistic binary pulsars with black hole companions. ApJ. 2005;628:343.
129. Devecchi B, Colpi M, Mapelli M, Possenti A. Millisecond pulsars around intermediate-mass black holes in globular clusters. MNRAS. 2007;380:691.
130. Kramer M, et al. Strong-field tests of gravity using pulsars and black holes. New Astr Rev. 2004;48:993.
131. Liu K, Wex N, Kramer M, Cordes JM, Lazio TJW. Prospects for probing the spacetime of Sgr A* with pulsars. ApJ. 2012;747:L1.
132. Hawking SW, Penrose R. The singularities of gravitational collapse and cosmology. Royal Soc Proc Ser A. 1970;314:529.
133. Foster R, Backer D. Constructing a pulsar timing array. ApJ. 1990;361:300.
134. Hellings R, Downs G. Upper limits on the isotropic gravitational radiation background from pulsar timing analysis. ApJ. 1983;265:L39.
135. Hobbs G. The Parkes pulsar timing array. Class Quantum Gravity. 2013;30:224007.
136. Kramer M, Champion D. The European pulsar timing array and the large European array for pulsars. Class Quantum Gravity. 2013;30:224009.
137. McLaughlin MA. The North American nanohertz observatory for gravitational waves. Class Quantum Gravity. 2013;30:224008.
138. Hobbs G, et al. The international pulsar timing array project: using pulsars as a gravitational wave detector. Class Quantum Gravity. 2010;27:084013.
139. van Haasteren R, et al. Placing limits on the stochastic gravitational-wave background using European timing array data. MNRAS. 2011;414:3117.
140. Demorest PB, et al. Limits on the stochastic gravitational wave background from the North American nanohertz observatory for gravitational waves. ApJ. 2013;762:94.
141. Shannon R, et al. Gravitational-wave limits from pulsar timing constrain supermassive black hole evolution. Science. 2013;342:334.
142. Verbiest J, et al. Timing stability of millisecond pulsars and prospects for gravitational-wave detection. MNRAS. 2009;400:951.
143. Yardley DRB, et al. The sensitivity of the Parkes pulsar timing array to individual sources of gravitational waves. MNRAS. 2010;407:669.
144. Thorne KS. Multipole expansions of gravitational radiation. Rev Mod Phys. 1980;52:299.
145. Lee K, et al. Detecting massive gravitons using pulsar timing arrays. ApJ. 2010;722:1589.
146. Sesana A, Vecchio A. Measuring the parameters of massive black hole binary systems with pulsar timing array observations of gravitational waves. Phys Rev D. 2010;81:104008.
147. Deng X, Finn LS. Pulsar timing array observations of gravitational wave source timing parallax. MNRAS. 2011;414:50.

Part III
Frontiers

Non-inertial Frames in Special and General Relativity

Luca Lusanna

Abstract A theory of global non-inertial frames in special relativity is developed by taking into account relativistic metrology and by avoiding the coordinate singularities of the rotating disk and of Fermi coordinates. Then this theory is used in general relativity for the treatment of globally hyperbolic, asymptotically Minkowskian space-times solution of Einstein's equations. Also some comments on the problem of dark matter, connecting it to the relativistic metrology used in astrophysics, are made.

1 Introduction

The aim of this contribution is to clarify what is known about non-inertial frames in special relativity (SR) and general relativity (GR). This topic is rarely discussed and till recently there was no attempt to develop a consistent general theory. All the results of the standard model of elementary particles are defined in the inertial frames of Minkowski space-time. Only at the level of neutron, atomic, and molecular physics one needs a local study of non-inertial frames in SR, for instance the rotating ones for the Sagnac effect.

Moreover, relativistic metrology [1] and space physics around the Earth and in the solar system [2] must take into account the gravitational field and the post-Newtonian (PN) limit of GR, a theory in which global inertial frames are forbidden by the equivalence principle. In Einstein's GR, the gauge group of its Lagrangian formulation, the diffeomorphism group, implies that the 4-coordinates of the space-time (and therefore the local non-inertial frames) are *gauge variables*. As a consequence, one would like to describe the effects of the physical degrees of freedom of the gravitational field by means of 4-scalars. This is an open theoretical problem. The praxis of experimentalists, who do not know which is the correct formulation of GR among the existing ones, is completely different.

Inside the solar system, the experimental localization of macroscopic classical objects is unavoidably done by choosing some *convention* for the local 4-coordinates of space-time. Atomic physicists, NASA engineers, and astronomers have chosen a

L. Lusanna (✉)
Polo Scientifico, Sezione INFN di Firenze, Via Sansone 1, 50019 Sesto Fiorentino, Florence, Italy
e-mail: lusanna@fi.infn.it

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_9

series of reference frames and standards of time and length suitable for the existing technology [1,2]. These conventions determine certain post-Minkowskian (PM) 4-coordinate systems of an asymptotically Minkowskian space-time, in which the instantaneous 3-spaces are not strictly Euclidean. Then these reference frames are seen as a local approximation of an International Celestial Reference Frame (ICRS), where however the space-time has become a cosmological Friedman–Robertson–Walker (FRW) one, which is only conformally asymptotically Minkowskian at spatial infinity. A search of a consistent patching of the 4-coordinates from inside the solar system to the rest of the universe will start when the data from the future GAIA mission [3] for the cartography of the Milky Way will be available. This will allow a PM definition of a Galactic Reference System containing at least our galaxy. Let us remark that notwithstanding the FRW instantaneous 3-spaces are not strictly Euclidean, all the books on galaxy dynamics describe the galaxies by means of Kepler theory in Galilei space-time.

A well-posed formulation of a PM ICRS (a global non-inertial frame for the 3-universe) would also be needed to face the main open problem of astrophysics, namely the dominance of *dark* entities, the dark matter, and the dark energy, in the existing description of the universe given by the standard ΛCDM cosmological model [4] based on the cosmological principle (homogeneity and isotropy of the space-time), which selects the class of FRW space-times. After the transition from quantum cosmology to classical astrophysics, with the Heisenberg cut (the interface between the quantum and classical worlds) roughly located at a suitable cosmic time ($\approx 10^5$ years after the big bang) and at the recombination surface identified by the cosmic microwave background (CMB), one has a description of the universe in which the known forms of baryonic matter and radiation contribute only with a few percents of the global budget. One has a great variety of models trying to explain the composition of the universe in accelerated expansion (based on data on high red-shift supernovae, galaxy clusters, and CMB): WIMPS (mainly supersymmetric particles), $f(R)$ modifications of Einstein's gravity (with a modified Newton's potential), MOND (with a modification of Newton's law), ...for dark matter; cosmological constant, string theory, back-reaction (spatial averages, nonlinearity of Einstein's equations), inhomogeneous space-times (Lemaitre–Tolman–Bondi, Szekeres), scalar fields (quintessence, k-essence, phantom), fluids (Chaplygin fluid), ...for dark energy.

A PM ICRS would allow to interpret the astronomical data (luminosity, light spectrum, angles) on the two-dimensional sky vault in a more realistic way (taking into account the inhomogeneities in the 3-universe) than in the nearly flat 3-spaces (as required by CMB data) of FRW space-times. In particular, one needs new standards of time and length like the cosmic time and the luminosity distance extending the standard relativistic metrology inside the solar system.

All these open problems justify the following description of what is known about non-inertial frames in SR and GR.

2 Non-inertial Frames in Special Relativity

In nonrelativistic (NR) Newtonian physics isolated systems are described in Galilei space-time, where both time and the instantaneous Euclidean 3-spaces are absolute quantities. As a consequence, the transition from the description of the system in NR inertial frames to its description in rigid non-inertial frames can be done by defining the following 3-coordinate transformation

$$x^i = y^i(t) + \sigma^r \, R_{ri}(t). \tag{1}$$

Here x^i's are inertial Cartesian 3-coordinates centered on an inertial observer, while σ^r's are rigid non-inertial 3-coordinates centered on an arbitrary observer whose trajectory is described by the Cartesian 3-coordinates $y^i(t)$ in the inertial frame. This accelerated observer has a 3-velocity, which can be conveniently written in the form $v^i(t) = R_{ij}(t) \frac{dy^j(t)}{dt}$. $R(t)$ is a time-dependent rotation matrix ($R^{-1} = R^T$), which can be parametrized with three time-dependent Euler angles. The angular velocity of the rotating frame is $\omega^i(t) = \frac{1}{2} \varepsilon^{ijk} \, \Omega_{jk}(t)$ with $\Omega_{jk}(t) = -\Omega_{kj}(t) = \left(\frac{dR(t)}{dt} \, R^T(t)\right)_{jk}$.

A particle of mass m_o with inertial Cartesian 3-coordinates $x^i_o(t)$ is described in the non-inertial frame by 3-coordinates $\eta^r(t)$ such that

$$x^i_o(t) = y^i(t) + \eta^r(t) \, R_{ri}(t). \tag{2}$$

As shown in every book on Newtonian mechanics, a particle satisfying the equation of motion $m_o \frac{d^2 \, \mathbf{x}_o(t)}{dt^2} = -\frac{\partial \, V(t, x^j_o(t))}{\partial \, \mathbf{x}_o}$, if an external potential $V(t, x^k_o(t)) = \tilde{V}(t, \eta^r(t))$ is present, will satisfy the following equation of motion in the rigid non-inertial frame

$$\begin{aligned} m_o \, \frac{d^2 \, \boldsymbol{\eta}(t)}{dt^2} = & - \frac{\partial \, \tilde{V}(t, \eta^r(t))}{\partial \, \boldsymbol{\eta}_o} \\ & - m_o \Big[\frac{d\mathbf{v}(t)}{dt} + \boldsymbol{\omega}(t) \times \mathbf{v}(t) + \frac{d\boldsymbol{\omega}(t)}{dt} \times \boldsymbol{\eta}(t) \\ & + 2 \, \boldsymbol{\omega}(t) \times \frac{d\boldsymbol{\eta}(t)}{dt} + \boldsymbol{\omega}(t) \times [\boldsymbol{\omega}(t) \times \boldsymbol{\eta}(t)] \Big], \end{aligned} \tag{3}$$

where the standard Euler, Jacobi, Coriolis, and centrifugal *inertial forces* associated with the linear acceleration of the non-inertial observer and with the angular velocity of the rotating frame are present.

In Ref. [5] there is the extension to nonrigid non-inertial frames in which Eq. (1) is replaced by $x^i = \mathcal{A}^i(t, \sigma^r)$ with $\mathcal{A}^i$ arbitrary functions well behaved at spatial infinity. For instance, a differentially rotating non-inertial frame is described by Eq. (1) with a point-dependent rotation matrix $R(t, \sigma^r)$.

To go from NR inertial frames to the nonrigid non-inertial ones one has to replace the group of Galilei transformations, connecting the NR inertial frames, with some subgroup of the group of 3-diffeomorphisms of the Euclidean 3-space.

The transition to SR is highly nontrivial because, due to the Lorentz signature of Minkowski space-time, time is no more absolute and there is no notion of instantaneous 3-space: the only intrinsic structure is the conformal one, that is, the light cone as the locus of incoming and outgoing radiation. *A convention on the synchronization of clocks is needed to define an instantaneous 3-space*. For instance, the *1-way velocity of light* from one observer A to an observer B has a meaning only after a choice of a convention for synchronizing the clock in A with the one in B. Therefore, the crucial quantity in SR is the *2-way (or round trip) velocity of light* c involving only one clock. It is this velocity which is isotropic and constant in SR and replaces the standard of length in relativistic metrology[1].

Einstein convention for the synchronization of clocks in Minkowski space-time uses the 2-way velocity of light to identify the Euclidean 3-spaces of the inertial frames centered on an inertial observer A by means of only its clock. The inertial observer A sends a ray of light at x^o_i towards the (in general accelerated) observer B; the ray is reflected towards A at a point P of B world-line and then reabsorbed by A at x^o_f; by convention, P is synchronous with the mid-point between emission and absorption on A's world-line, that is, $x^o_P = x^o_i + \frac{1}{2}(x^o_f - x^o_i) = \frac{1}{2}(x^o_i + x^o_f)$. This convention selects the Euclidean instantaneous 3-spaces $x^o = ct = const.$ of the inertial frames centered on A. Only in this case the 1-way velocity of light between A and B coincides with the 2-way one, c. However if the observer A is accelerated, the convention can break down due to the possible appearance of coordinate singularities.

The existing coordinatizations, like either Fermi or Riemann normal coordinates, hold only locally. They are based on the *1 + 3 point of view*, in which only the world-line of a time-like observer is given. In each point of the world-line, the observer 4-velocity determines an orthogonal three-dimensional space-like tangent hyperplane, which is identified with an instantaneous 3-space. However, these tangent planes intersect at a certain distance from the world-line (the so-called acceleration length depending upon the 4-acceleration of the observer [6]), where 4-coordinates of the Fermi type develop a coordinate singularity. Another type of coordinate singularity is developed in rigidly rotating coordinate systems at a distance r from the rotation axis where $\omega r = c$ (ω is the angular velocity and c the 2-way velocity of light). This is the so-called horizon problem of the rotating disk: a time-like 4-velocity becomes a null vector at $\omega r = c$, like it happens on the horizon of a black hole. See Ref. [7] for a classification of the possible pathologies of non-inertial frames and on how to avoid them.

As a consequence, a theory of global non-inertial frames in Minkowski space-time has to be developed in a metrology-oriented way to overcome the pathologies of the $1 + 3$ point of view. This has been done in the papers of Ref. [7] by using the $3 + 1$ *point of view* in which, besides the world-line of a time-like observer, one gives a global nice foliation of the space-time with instantaneous 3-spaces.

[1] See Ref. [1] for an updated review on relativistic metrology on Earth and in the solar system.

Assume that the world-line $x^\mu(\tau)$ of an arbitrary time-like observer carrying a standard atomic clock is given: τ is an arbitrary monotonically increasing function of the proper time of this clock. Then one gives an admissible $3+1$ splitting of Minkowski space-time, namely a nice foliation with space-like instantaneous 3-spaces Σ_τ. It is the mathematical idealization of a protocol for clock synchronization: all the clocks in the points of Σ_τ sign the same time of the atomic clock of the observer[2]. On each 3-space Σ_τ one chooses curvilinear 3-coordinates σ^r having the observer as origin. These are the Lorentz-scalar and observer-dependent *radar 4-coordinates* $\sigma^A = (\tau;\sigma^r)$, first introduced by Bondi [8].

If $x^\mu \mapsto \sigma^A(x)$ is the coordinate transformation from the Cartesian 4-coordinates x^μ of a reference inertial observer to radar coordinates, its inverse $\sigma^A \mapsto x^\mu = z^\mu(\tau, \sigma^r)$ defines the *embedding* functions $z^\mu(\tau, \sigma^r)$ describing the 3-spaces Σ_τ as embedded 3-manifolds into Minkowski space-time. The induced 4-metric on Σ_τ is the following functional of the embedding ${}^4g_{AB}(\tau, \sigma^r) = [z^\mu_A\, \eta_{\mu\nu}\, z^\nu_B](\tau, \sigma^r)$, where $z^\mu_A = \partial\, z^\mu / \partial\, \sigma^A$ and ${}^4\eta_{\mu\nu} = \varepsilon\,(+---)$ is the flat metric[3]. While the 4-vectors $z^\mu_r(\tau, \sigma^u)$ are tangent to Σ_τ, so that the unit normal $l^\mu(\tau, \sigma^u)$ is proportional to $\varepsilon^\mu{}_{\alpha\beta\gamma}\, [z^\alpha_1\, z^\beta_2\, z^\gamma_3](\tau, \sigma^u)$, one has $z^\mu_\tau(\tau, \sigma^r) = [N\, l^\mu + N^r\, z^\mu_r](\tau, \sigma^r)$ with $N(\tau, \sigma^r) = \varepsilon\, [z^\mu_\tau\, l_\mu](\tau, \sigma^r) = 1 + n(\tau, \sigma^r)$ and $N_r(\tau, \sigma^r) = -\varepsilon\, g_{\tau r}(\tau, \sigma^r)$ being the lapse and shift functions, respectively.

As a consequence, the components of the 4-metric ${}^4g_{AB}(\tau, \sigma^r)$ have the following expression:

$$\begin{aligned}
&\varepsilon\, {}^4g_{\tau\tau} = N^2 - N_r\, N^r, \qquad -\varepsilon\, {}^4g_{\tau r} = N_r = {}^3g_{rs}\, N^s,\\
&{}^3g_{rs} = -\varepsilon\, {}^4g_{rs} = \sum_{a=1}^{3} {}^3e_{(a)r}\, {}^3e_{(a)s}\\
&\qquad = \tilde{\phi}^{2/3} \sum_{a=1}^{3} e^{2\, \sum_{\bar{b}=1}^{2}\, \gamma_{\bar{b}a}\, R_{\bar{b}}}\, V_{ra}(\theta^i)\, V_{sa}(\theta^i),
\end{aligned} \tag{4}$$

where ${}^3e_{(a)r}(\tau, \sigma^u)$ are cotriads on Σ_τ, $\tilde{\phi}^2(\tau, \sigma^r) = det\, {}^3g_{rs}(\tau, \sigma^r)$ is the 3-volume element on Σ_τ, $\lambda_a(\tau, \sigma^r) = [\tilde{\phi}^{1/3}\, e^{\sum_{\bar{b}=1}^{2}\, \gamma_{\bar{b}a}\, R_{\bar{b}}}](\tau, \sigma^r)$ are the positive eigenvalues of the 3-metric ($\gamma_{\bar{a}a}$ are suitable numerical constants) and $V(\theta^i(\tau, \sigma^r))$ are diagonalizing rotation matrices depending on three Euler angles.

Therefore, starting from the *four* independent embedding functions $z^\mu(\tau, \sigma^r)$ one obtains the *ten* components ${}^4g_{AB}$ of the 4-metric (or the quantities N, N_r, $\tilde{\phi}$, $R_{\bar{a}}$, θ^i), which play the role of the *inertial potentials* generating the relativistic apparent forces in the non-inertial frame. It can be shown [7] that the usual NR Newtonian inertial potentials are hidden in the functions N, N_r, and θ^i. The extrinsic curvature tensor ${}^3K_{rs}(\tau, \sigma^u) = \left[\frac{1}{2N}\, (N_{r|s} + N_{s|r} - \partial_\tau\, {}^3g_{rs})\right](\tau, \sigma^u)$ ("|"

[2] It is the nonfactual idealization required by the Cauchy problem generalizing the existing protocols for building a coordinate system inside the future light cone of a time-like observer.

[3] $\varepsilon = \pm 1$ according to either the particle physics $\varepsilon = 1$ or the GR $\varepsilon = -1$ convention.

denotes the covariant derivative inside the 3-spaces), describing the *shape* of the instantaneous 3-spaces of the non-inertial frame as embedded 3-sub-manifolds of Minkowski space-time, is a secondary inertial potential, functional of the ten inertial potentials ${}^4g_{AB}$. Now a relativistic positive-energy scalar particle with world-line $x^\mu_o(\tau)$ is described by 3-coordinates $\eta^r(\tau)$ defined by $x^\mu_o(\tau) = z^\mu(\tau, \eta^r(\tau))$, satisfying equations of motion containing relativistic inertial forces whose NR limit reproduces Eq. (3) as shown in Refs. [5, 7].

The foliation is nice and admissible if it satisfies the following conditions:

1. $N(\tau, \sigma^r) > 0$ in every point of Σ_τ so that the 3-spaces never intersect, avoiding the coordinate singularity of Fermi coordinates;
2. $\varepsilon\, {}^4g_{\tau\tau}(\tau, \sigma^r) > 0$, so to avoid the coordinate singularity of the rotating disk, and with the positive-definite 3-metric ${}^3g_{rs}(\tau, \sigma^u) = -\varepsilon\, {}^4g_{rs}(\tau, \sigma^u)$ having three positive eigenvalues (these are the Møller conditions [9]);
3. all the 3-spaces Σ_τ must tend to the same space-like hyperplane at spatial infinity with a unit normal ε^μ_τ, which is the time-like 4-vector of a set of asymptotic orthonormal tetrads ε^μ_A. These tetrads are carried by asymptotic inertial observers and the spatial axes ε^μ_r are identified by the fixed stars of star catalogues. At spatial infinity the lapse function tends to 1 and the shift functions vanish.

By using the asymptotic tetrads ε^μ_A one can give the following parametrization of the embedding functions

$$z^\mu(\tau, \sigma^r) = x^\mu(\tau) + \varepsilon^\mu_A\, F^A(\tau, \sigma^r), \qquad F^A(\tau, 0) = 0,$$

$$x^\mu(\tau) = x^\mu_o + \varepsilon^\mu_A\, f^A(\tau), \tag{5}$$

where $x^\mu(\tau)$ is the world-line of the observer. The functions $f^A(\tau)$ determine the 4-velocity $u^\mu(\tau) = \dot{x}^\mu(\tau)/\sqrt{\varepsilon\, \dot{x}^2(\tau)}$ $\left(\dot{x}^\mu(\tau) = \frac{dx^\mu(\tau)}{d\tau}\right)$ and the 4-acceleration $a^\mu(\tau) = \frac{du^\mu(\tau)}{d\tau}$ of the observer.

The Møller conditions are nonlinear differential conditions on the functions $f^A(\tau)$ and $F^A(\tau, \sigma^r)$, so that it is very difficult to construct explicit examples of admissible $3+1$ splittings. When these conditions are satisfied, Eq. (5) describes a global non-inertial frame in Minkowski space-time.

Till now the solution of Møller conditions is known in the following two cases in which the instantaneous 3-spaces are parallel Euclidean space-like hyperplanes not equally spaced due to a linear acceleration.

A) *Rigid non-inertial reference frames with translational acceleration*. For example, the following embeddings:

$$z^\mu(\tau, \sigma^u) = x^\mu_o + \varepsilon^\mu_\tau\, f(\tau) + \varepsilon^\mu_r\, \sigma^r,$$

$$g_{\tau\tau}(\tau, \sigma^u) = \varepsilon \left(\frac{df(\tau)}{d\tau}\right)^2, \quad g_{\tau r}(\tau, \sigma^u) = 0, \quad g_{rs}(\tau, \sigma^u) = -\varepsilon\, \delta_{rs}. \tag{6}$$

This is a foliation with parallel hyperplanes with normal $l^\mu = \varepsilon^\mu_\tau = const.$ and with the time-like observer $x^\mu(\tau) = x^\mu_o + \varepsilon^\mu_\tau\, f(\tau)$ as origin of the 3-coordinates. The hyperplanes have translational acceleration $\ddot{x}^\mu(\tau) = \varepsilon^\mu_\tau\, \ddot{f}(\tau)$, so that they are not uniformly distributed like in the inertial case $f(\tau) = \tau$.

B) *Differentially rotating non-inertial frames* without the coordinate singularity of the rotating disk. The embedding defining these frames is

$$z^\mu(\tau, \sigma^u) = x^\mu(\tau) + \varepsilon^\mu_r\, R^r{}_s(\tau, \sigma)\, \sigma^s \;\to_{\sigma\to\infty}\; x^\mu(\tau) + \varepsilon^\mu_r\, \sigma^r,$$

$$R^r{}_s(\tau, \sigma) = R^r{}_s(\alpha_i(\tau, \sigma)) = R^r{}_s(F(\sigma)\, \tilde{\alpha}_i(\tau)),$$

$$0 < F(\sigma) < \frac{1}{A\, \sigma}, \qquad \frac{d\, F(\sigma)}{d\sigma} \neq 0 \text{ (Møller conditions)},$$

$$z^\mu_\tau(\tau, \sigma^u) = \dot{x}^\mu(\tau) - \varepsilon^\mu_r\, R^r{}_s(\tau, \sigma)\, \delta^{sw}\, \varepsilon_{wuv}\, \sigma^u\, \frac{\Omega^v(\tau, \sigma)}{c},$$
$$z^\mu_r(\tau, \sigma^u) = \varepsilon^\mu_k\, R^k{}_v(\tau, \sigma)\left(\delta^v_r + \Omega^v_{(r)u}(\tau, \sigma)\, \sigma^u\right), \tag{7}$$

where $\sigma = |\boldsymbol{\sigma}|$ and $R^r{}_s(\alpha_i(\tau, \sigma))$ is a rotation matrix satisfying the asymptotic conditions $R^r{}_s(\tau, \sigma) \to_{\sigma\to\infty} \delta^r_s$, $\partial_A\, R^r{}_s(\tau, \sigma) \to_{\sigma\to\infty} 0$, whose Euler angles have the expression $\alpha_i(\tau, \boldsymbol{\sigma}) = F(\sigma)\, \tilde{\alpha}_i(\tau)$, $i = 1, 2, 3$. The unit normal is $l^\mu = \varepsilon^\mu_\tau = const.$ and the lapse function is $1 + n(\tau, \sigma^u) = \varepsilon \left(z^\mu_\tau\, l_\mu\right)(\tau, \sigma^u) = \varepsilon\, \varepsilon^\mu_\tau\, \dot{x}_\mu(\tau) > 0$. In Eq. (7) one uses the notations $\Omega_{(r)}(\tau, \sigma) = R^{-1}(\tau, \boldsymbol{\sigma})\, \partial_r\, R(\tau, \sigma)$ and $\left(R^{-1}(\tau, \sigma)\, \partial_\tau\, R(\tau, \sigma)\right)^u{}_v = \delta^{um}\, \varepsilon_{mvr}\, \frac{\Omega^r(\tau, \sigma)}{c}$, with $\Omega^r(\tau, \sigma) = F(\sigma)\, \tilde{\Omega}(\tau, \sigma)\, \hat{n}^r(\tau, \sigma)$[4] being the angular velocity. The angular velocity vanishes at spatial infinity and has an upper bound proportional to the minimum of the linear velocity $v_l(\tau) = \dot{x}_\mu\, l^\mu$ orthogonal to the space-like hyperplanes. When the rotation axis is fixed and $\tilde{\Omega}(\tau, \sigma) = \omega = const.$, a simple choice for the function $F(\sigma)$ is $F(\sigma) = \frac{1}{1+\frac{\omega^2\, \sigma^2}{c^2}}$[5].

To evaluate the NR limit for $c \to \infty$, where $\tau = c\, t$ with t the absolute Newtonian time, one chooses the gauge function $F(\sigma) = \frac{1}{1+\frac{\omega^2\, \sigma^2}{c^2}} \to_{c\to\infty} 1 - \frac{\omega^2\, \sigma^2}{c^2} + O(c^{-4})$. This implies that the corrections to rigidly rotating non-inertial frames coming from Møller conditions are of order $O(c^{-2})$ and become important at the distance from the rotation axis where the horizon problem for rigid rotations appears.

[4] $\hat{n}^r(\tau, \sigma)$ defines the instantaneous rotation axis and $0 < \tilde{\Omega}(\tau, \sigma) < 2\, max\left(\dot{\tilde{\alpha}}(\tau), \dot{\tilde{\beta}}(\tau), \dot{\tilde{\gamma}}(\tau)\right)$.

[5] Nearly rigid rotating systems, like a rotating disk of radius σ_o, can be described by using a function $F(\sigma)$ approximating the step function $\theta(\sigma - \sigma_o)$.

As shown in the first paper in Ref. [7], *global rigid rotations are forbidden in relativistic theories*, because if one uses the embedding $z^\mu(\tau,\sigma^u) = x^\mu(\tau) + \varepsilon^\mu_r\, R^r{}_s(\tau)\,\sigma^s$ describing a global rigid rotation with angular velocity $\Omega^r = \Omega^r(\tau)$, then the resulting $g_{\tau\tau}(\tau,\sigma^u)$ violates Møller conditions, because it vanishes at $\sigma = \sigma_R = \frac{1}{\Omega(\tau)}\left[\sqrt{\dot{x}^2(\tau) + [\dot{x}_\mu(\tau)\,\varepsilon^\mu_r\, R^r{}_s(\tau)\,(\hat{\sigma}\times\hat{\Omega}(\tau))^r]^2} - \dot{x}_\mu(\tau)\,\varepsilon^\mu_r\, R^r{}_s(\tau)\,(\hat{\sigma}\times\hat{\Omega}(\tau))^r\right]$ ($\sigma^u = \sigma\,\hat{\sigma}^u$, $\Omega^r = \Omega\,\hat{\Omega}^r$, $\hat{\sigma}^2 = \hat{\Omega}^2 = 1$). At this distance from the rotation axis the tangential rotational velocity becomes equal to the velocity of light. This is the *horizon problem* of the rotating disk (the horizon is often named the *light cylinder*). Let us remark that even if in the existing theory of rotating relativistic stars [10] one uses differential rotations, notwithstanding that in the study of the magnetosphere of pulsars often the notion of light cylinder is still used.

The search of admissible $3+1$ splittings with non-Euclidean 3-spaces is much more difficult. The simplest case is the following parametrization of the embeddings (4) in terms of Lorentz matrices $\Lambda^A{}_B(\tau,\sigma) \to_{\sigma\to\infty} \delta^{A\,6}_B$ with $\Lambda^A{}_B(\tau,0)$ finite. The Lorentz matrix is written in the form $\Lambda = \mathcal{B}\,\mathcal{R}$ as the product of a boost $\mathcal{B}(\tau,\sigma)$ and a rotation $\mathcal{R}(\tau,\sigma)$ like the one in Eq. (7) ($\mathcal{R}^\tau{}_\tau = 1$, $\mathcal{R}^\tau{}_r = 0$, $\mathcal{R}^r{}_s = R^r{}_s$). The components of the boost are $\mathcal{B}^\tau{}_\tau(\tau,\sigma) = \gamma(\tau,\sigma) = 1/\sqrt{1-\boldsymbol{\beta}^2(\tau,\sigma)}$, $\mathcal{B}^\tau{}_r(\tau,\sigma) = \gamma(\tau,\sigma)\,\beta_r(\tau,\sigma)$, $\mathcal{R}^r{}_s(\tau,\sigma) = \delta^r_s + \frac{\gamma\,\beta^r\,\beta_s}{1+\gamma}(\tau,\sigma)$, with $\beta^r(\tau,\sigma) = G(\sigma)\,\beta^r(\tau)$, where $\beta^r(\tau)$ is defined by the 4-velocity of the observer $u^\mu(\tau) = \varepsilon^\mu_A\,\beta^A(\tau)/\sqrt{1-\boldsymbol{\beta}^2(\tau)}$, $\beta^A(\tau) = (1;\beta^r(\tau))$. The Møller conditions are restrictions on $G(\sigma) \to_{\sigma\to\infty} 0$ with $G(0)$ finite, whose explicit form is still under investigation.

The embedding (7) has been studied in details in Ref. [11] for the development of quantum mechanics in non-inertial frames.

See the second paper of Ref. [7] for the description of the electromagnetic field and of phenomena like the Sagnac effect and the Faraday rotation in this framework for non-inertial frames.

The previous approach based on the $3+1$ point of view has allowed a complete reformulation of relativistic particle mechanics in SR [7, 12, 13]. By means of *parametrized Minkowski theories* [7, 12], one can get the description of arbitrary isolated systems (particles, strings, fluids, fields) admitting a Lagrangian formulation in arbitrary non-inertial frames[7]. To get it, the Lagrangian is coupled to an external gravitational field and then the gravitational 4-metric is replaced with the 4-metric ${}^3g_{AB}(\tau,\sigma^r)$, a functional of the embedding $z^\mu(\tau,\sigma^r)$, induced by an admissible $3+1$ splitting of Minkowski space-time. The new Lagrangian, a function of the

[6] It corresponds to the *locality hypothesis* of Ref. [6], according to which at each instant of time the detectors of an accelerated observer give the same indications as the detectors of the instantaneously comoving inertial observer.

[7] See Ref. [5] for the definition of parametrized Galilei theories in NR mechanics.

matter and of the embedding, is invariant under the frame-preserving diffeomorphisms of Ref. [14][8]. This kind of general covariance implies that the embeddings are *gauge variables*, so that the transition among non-inertial frames is described as a *gauge transformation*: only the appearances change, not the physics.

This framework allows us to define the *inertial and non-inertial rest frames* of the isolated systems, where to develop the rest-frame instant form of the dynamics and to build the explicit form of the Lorentz boosts for interacting systems. While the inertial rest frames have their Euclidean 3-spaces defined as space-like 3-manifolds of Minkowski space-time orthogonal to the conserved 4-momentum of the isolated system, the non-inertial rest frames are admissible non-inertial frames whose 3-spaces tend to those of some inertial rest frames at spatial infinity, where the 3-space becomes orthogonal to the conserved 4-momentum.

This makes possible to study the problem of the relativistic center of mass with the associated external and internal (i.e., inside the 3-space) realizations of the Poincaré algebra [15], relativistic bound states [16–18], relativistic kinetic theory, and relativistic micro-canonical ensemble [19] and various other systems [20, 21]. Moreover, a Wigner-covariant relativistic quantum mechanics [22], with a solution of all the known problems introduced by SR, has been developed after some preliminary work done in Ref. [11]. This allows the beginning of the study of relativistic entanglement taking into account all the consequences of the Lorentz signature of Minkowski space-time. As shown in Ref. [22], in SR the relativistic center of mass is a nonlocal nonmeasurable quantity: only relative variables have an operational meaning and this implies a spatial non-separability, that is, some form of weak relationism in which all the objects know each other, differently from the NR case where the center of mass is a measurable quantity.

See Ref. [23] for an extended review of this approach both in SR and in GR. In the next section, there will be a sketch of the known results in GR.

3 Non-inertial Frames in General Relativity

In GR, global inertial frames are forbidden by the equivalence principle. Therefore, gravitational physics has to be described in non-inertial frames.

While in SR Minkowski space-time is an absolute notion, unifying the absolute notions of time and 3-space of the NR Galilei space-time, in Einstein's GR also the space-time is a dynamical object [24] and the gravitational field is described by the metric structure of the space-time, namely by the ten dynamical fields ${}^4g_{\mu\nu}(x)$ (x^μ are world 4-coordinates) satisfying Einstein's equations.

The ten dynamical fields ${}^4g_{\mu\nu}(x)$ are not only a (pre)potential for the gravitational field (like the electromagnetic and Yang–Mills fields are the potentials for

[8] This is the only paper known to us where there is an attempt to formulate a theory of non-inertial frames in SR.

electromagnetic and non-Abelian forces) but also determine the *chrono-geometrical structure of space-time* through the line element $ds^2 = {}^4g_{\mu\nu}\, dx^\mu\, dx^\nu$. Therefore, the 4-metric teaches relativistic causality to the other fields: it says to massless particles like photons and gluons which are the allowed world-lines in each point of space-time. The Atoitic Clock Ensemble in Space (ACES) mission of European Space Agency (ESA) [25] will give the first precision measurement of the gravitational redshift of the geoid, namely of the $1/c^2$ deformation of Minkowski light cone caused by the geopotential.

The metrology-oriented solution of the problem of clock synchronization used in SR can be extended to GR if Einstein's space-times are restricted to the class of globally hyperbolic, topologically trivial, asymptotically Minkowskian space-times without super-translations[9].

As shown in the first paper of Ref. [26], in the chosen class of space-times the 4-metric ${}^4g_{\mu\nu}(x)$ tends in a suitable way to the flat Minkowski 4-metric ${}^4\eta_{\mu\nu}$ at spatial infinity and the ten *strong* asymptotic Arnowitt–Deser–Misner (ADM) Poincaré generators P^A_{ADM}, J^{AB}_{ADM} (they are fluxes through a 2-surface at spatial infinity) are well-defined functionals of the 4-metric fixed by the boundary conditions at spatial infinity.

These properties do not hold in generic asymptotically flat space-times, because they have the SPI group of asymptotic symmetries (direction-dependent asymptotic Killing symmetries at SPatial Infinity) [27] and this is an obstruction to the existence of asymptotic Lorentz generators for the gravitational field [28]. However, if one restricts the class of space-times to those *not containing super-translations* [29], then the SPI group reduces to the asymptotic ADM Poincaré group [30]: these space-times are *asymptotically Minkowskian*, they contain an asymptotic Minkowski 4-metric (to be used as an *asymptotic background* at spatial infinity in the linearization of the theory), and they have asymptotic inertial observers at spatial infinity whose spatial axes may be identified by means of the fixed stars of star catalogues[10]. Moreover, in the limit of vanishing Newton's constant ($G = 0$), the asymptotic ADM Poincaré generators become the generators of the special relativistic Poincaré group describing the matter present in the space-time. This is an important condition for the inclusion into GR of the classical version of the standard model of particle physics, whose properties are all connected with the representations of this group in the inertial frames of Minkowski space-time. In absence of matter, a subclass of these space-times is the (singularity-free) family of Christodoulou–Klainermann solutions of Einstein's equations [31] (they are near to Minkowski space-time in a norm sense and contain gravitational waves).

In the first paper of Ref. [26], it is also shown that the boundary conditions on the 4-metric required by the absence of super-translations imply that the only admissible

[9] At this preliminary level these space-times must also be without Killing symmetries, because, otherwise, at the Hamiltonian level one should introduce complicated sets of extra Dirac constraints for each existing Killing vector.

[10] The fixed stars can be considered as an empirical definition of spatial infinity of the observable universe.

$3+1$ splittings of space-time (i.e., the allowed global non-inertial frames) are the *non-inertial rest frames*: their 3-spaces are asymptotically orthogonal to the weak ADM 4-momentum. Therefore, one gets $\hat{P}^r_{ADM} \approx 0$ as the rest-frame condition of the 3-universe with a mass and a rest spin fixed by the boundary conditions. Like in SR the 3-universe can be visualized as a decoupled non-covariant (nonmeasurable) external relativistic center of mass plus an internal non-inertial rest-frame 3-space containing only relative variables (see the first paper in Ref. [32]).

In these space-times one can define global non-inertial frames by using the same admissible $3+1$ splittings, centered on a time-like observer, and the observer-dependent radar 4-coordinates $\sigma^A = (\tau;\sigma^r)$ employed in SR. This will allow to separate the *inertial* (gauge) degrees of freedom of the gravitational field (playing the role of inertial potentials) from the dynamical *tidal* ones at the Hamiltonian level.

In GR the dynamical fields are the components ${}^4g_{\mu\nu}(x)$ of the 4-metric and not the embeddings $x^\mu = z^\mu(\tau,\sigma^r)$ defining the admissible $3+1$ splittings of space-time like in the parametrized Minkowski theories of SR. Now the gradients $z^\mu_A(\tau,\sigma^r)$ of the embeddings give the transition coefficients from radar to world 4-coordinates, so that the components ${}^4g_{AB}(\tau,\sigma^r) = z^\mu_A(\tau,\sigma^r)\, z^\nu_B(\tau,\sigma^r)\, {}^4g_{\mu\nu}(z(\tau,\sigma^r))$ of the 4-metric will be the dynamical fields in the ADM action. Like in SR the 4-vectors $z^\mu_A(\tau,\sigma^r)$, tangent to the 3-spaces Σ_τ, are used to define the unit normal $l^\mu(\tau,\sigma^r) = z^\mu_A(\tau,\sigma^r)\, l^A(\tau,\sigma^r)$ to Σ_τ, while the 4-vector $z^\mu_\tau(\tau,\sigma^r)$ has the lapse function as component along the unit normal and the shift functions as components along the tangent vectors.

Since the world-line of the time-like observer can be chosen as the origin of a set of the spatial world coordinates, that is, $x^\mu(\tau) = (x^o(\tau);0)$, it turns out that with this choice the space-like surfaces of constant coordinate time $x^o(\tau) = const.$ coincide with the dynamical instantaneous 3-spaces Σ_τ with $\tau = const.$ By using asymptotic flat tetrads $\varepsilon^\mu_A = \delta^\mu_o\, \delta^\tau_A + \delta^\mu_i\, \delta^i_A$ (with ε^A_μ denoting the inverse flat cotetrads) and by choosing a coordinate world time $x^o(\tau) = x^o_o + \varepsilon^o_\tau\, \tau = x^o_o + \tau$, one gets the following preferred embedding corresponding to these given world 4-coordinates

$$x^\mu = z^\mu(\tau,\sigma^r) = x^\mu(\tau) + \varepsilon^\mu_r\, \sigma^r = \delta^\mu_o\, x^o_o + \varepsilon^\mu_A\, \sigma^A. \tag{8}$$

This choice implies $z^\mu_A(\tau,\sigma^r) = \varepsilon^\mu_A$ and ${}^4g_{\mu\nu}(x = z(\tau,\sigma^r)) = \varepsilon^A_\mu\, \varepsilon^B_\nu\, {}^4g_{AB}(\tau,\sigma^r)$.

As shown in Ref. [24], the dynamical nature of space-time implies that each solution (i.e., an Einstein 4-geometry) of Einstein's equations (or of the associated ADM Hamilton's equations) dynamically selects a preferred $3+1$ splitting of the space-time, namely in GR the instantaneous 3-spaces are dynamically determined in the chosen world coordinate system, modulo the choice of the 3-coordinates in the 3-space and modulo the trace of the extrinsic curvature of the 3-space as a space-like sub-manifold of the space-time. Equation (8) can be used to describe this $3+1$ splitting and then by means of 4-diffeomorphisms the solution can be written in an arbitrary world 4-coordinate system in general not adapted to the dynamical $3+1$ splitting. This gives rise to the 4-geometry corresponding to the given solution.

To define the canonical formalism the Einstein–Hilbert action for metric gravity (depending on the second derivative of the metric) must be replaced with the ADM

action (the two actions differ for a surface tern at spatial infinity). As shown in the first paper of Ref. [26], the Legendre transform and the definition of a consistent canonical Hamiltonian require the introduction of the DeWitt surface term at spatial infinity: the final canonical Hamiltonian turns out to be the *strong* ADM energy (a flux through a 2-surface at spatial infinity), which is equal to the *weak* ADM energy (expressed as a volume integral over the 3-space) plus constraints.

Therefore, there is not a frozen picture like in the "spatially compact space-times without boundaries" used in loop quantum gravity[11], but an evolution generated by a Dirac Hamiltonian equal to the weak ADM energy plus a linear combination of the first-class constraints. Also, the other strong ADM Poincaré generators are replaced by their weakly equivalent weak form $\hat{P}^A_{ADM}$, $\hat{J}^{AB}_{ADM}$.

To take into account the fermion fields present in the standard particle model one must extend ADM gravity to ADM tetrad gravity. Since our class of space-times admits orthonormal tetrads and a spinor structure [34], the extension can be done by simply replacing the 4-metric in the ADM action with its expression in terms of tetrad fields, considered as the basic 16 configurational variables substituting the 10 metric fields. This can be achieved by decomposing the 4-metric on cotetrad fields (by convention, a sum on repeated indices is assumed)

$$ {}^4g_{AB}(\tau, \sigma^r) = E^{(\alpha)}_A(\tau, \sigma^r)\, {}^4\eta_{(\alpha)(\beta)}\, E^{(\beta)}_B(\tau, \sigma^r) \tag{9} $$

by putting this expression into the ADM action and by considering the resulting action, a functional of the 16 fields $E^{(\alpha)}_A(\tau, \sigma^r)$, as the action for ADM tetrad gravity. In Eq. (9), (α) are flat indices and the cotetrad fields $E^{(\alpha)}_A$ are the inverse of the tetrad fields $E^A_{(\alpha)}$, which are connected to the world tetrad fields by $E^\mu_{(\alpha)}(x) = z^\mu_A(\tau, \sigma^r)\, E^A_{(\alpha)}(z(\tau, \sigma^r))$ by the embedding of Eq. (8).

This leads to an interpretation of gravity based on a congruence of time-like observers endowed with orthonormal tetrads: In each point of space-time, the time-like axis is the unit 4-velocity of the observer, while the spatial axes are a (gauge) convention for observer's gyroscopes. This framework was developed in the second and third paper of Ref. [26].

Even if the action of ADM tetrad gravity depends upon 16 fields, the counting of the physical degrees of freedom of the gravitational field does not change, because this action is invariant not only under the group of 4-diffeomorphisms but also under the O(3,1) gauge group of the Newman–Penrose approach [35] (the extra gauge freedom acting on the tetrads in the tangent space of each point of space-time).

[11] In these space-times the canonical Hamiltonian vanishes and the Dirac Hamiltonian is a combination of first-class constraints, so that it only generates Hamiltonian gauge transformations. In the reduced phase space, quotient with respect to the Hamiltonian gauge group, the reduced Hamiltonian is zero and one has a *frozen picture* of dynamics. This class of space-times fits well with Machian ideas (no boundary conditions) and with interpretations in which there is no physical time like the one in Ref. [33]. However, it is not clear how to include in this framework the standard model of particle physics.

The cotetrads $E^{(\alpha)}_A(\tau,\sigma^r)$ are the new configuration variables. They are connected to cotetrads ${}^4\overset{\circ}{E}{}^{(\alpha)}_A(\tau,\sigma^r)$ adapted to the $3+1$ splitting of space-time, namely such that the inverse adapted time-like tetrad ${}^4\overset{\circ}{E}{}^A_{(o)}(\tau,\sigma^r)$ is the unit normal to the 3-space Σ_τ, by a standard Wigner boost for time-like Poincaré orbits with parameters $\varphi_{(a)}(\tau,\sigma^r)$, $a=1,2,3$

$$E^{(\alpha)}_A = L^{(\alpha)}{}_{(\beta)}(\varphi_{(a)})\, \overset{o}{E}{}^{(\beta)}_A, \qquad {}^4g_{AB} = {}^4\overset{\circ}{E}{}^{(\alpha)}_A\, {}^4\eta_{(\alpha)(\beta)}\, {}^4\overset{\circ}{E}{}^{(\beta)}_B,$$

$$L^{(\alpha)}{}_{(\beta)}(\varphi_{(a)}) \overset{def}{=} L^{(\alpha)}{}_{(\beta)}(V(z(\sigma));\, \overset{\circ}{V}) = \delta^{(\alpha)}_{(\beta)} + 2\varepsilon\, V^{(\alpha)}(z(\sigma))\, \overset{\circ}{V}_{(\beta)} - \\ -\varepsilon\, \frac{(V^{(\alpha)}(z(\sigma)) + \overset{\circ}{V}{}^{(\alpha)})\,(V_{(\beta)}(z(\sigma)) + \overset{\circ}{V}_{(\beta)})}{1+V^{(o)}(z(\sigma))}. \tag{10}$$

In each tangent plane to a point of Σ_τ this point-dependent standard Wigner boost sends the unit future-pointing time-like vector $\overset{o}{V}{}^{(\alpha)} = (1;0)$ into the unit time-like vector $V^{(\alpha)} = {}^4E^{(\alpha)}_A\, l^A = \left(\sqrt{1+\sum_a \varphi^2_{(a)}};\varphi^{(a)} = -\varepsilon\,\varphi_{(a)}\right)$. As a consequence, the flat indices (a) of the adapted tetrads and cotetrads and of the triads and cotriads on Σ_τ transform as Wigner spin-1 indices under point-dependent SO(3) Wigner rotations $R_{(a)(b)}(V(z(\sigma));\, \Lambda(z(\sigma)))$ associated with Lorentz transformations $\Lambda^{(\alpha)}{}_{(\beta)}(z)$ in the tangent plane to the space-time in the given point of Σ_τ. Instead, the index (o) of the adapted tetrads and cotetrads is a local Lorentz scalar index.

The adapted tetrads and cotetrads have the expression

$$\begin{aligned} &{}^4\overset{\circ}{E}{}^A_{(o)} = \frac{1}{1+n}\left(1; -\sum_a n_{(a)}\, {}^3e^r_{(a)}\right) = l^A, \qquad {}^4\overset{\circ}{E}{}^A_{(a)} = (0;{}^3e^r_{(a)}), \\ &{}^4\overset{\circ}{E}{}^{(o)}_A = (1+n)\,(1;\mathbf{0}) = \varepsilon\, l_A, \qquad {}^4\overset{\circ}{E}{}^{(a)}_A = (n_{(a)};{}^3e_{(a)r}), \end{aligned} \tag{11}$$

where ${}^3e^r_{(a)}$ and ${}^3e_{(a)r}$ are triads and cotriads on Σ_τ and $n_{(a)} = n_r\, {}^3e^r_{(a)} = n^r\, {}^3e_{(a)r}$[12] are adapted shift functions. In Eq. (11), $N(\tau,\boldsymbol{\sigma}) = 1+n(\tau,\boldsymbol{\sigma}) > 0$, with $n(\tau,\boldsymbol{\sigma})$ vanishing at spatial infinity (absence of super-translations), so that $N(\tau,\boldsymbol{\sigma})\,d\tau$ is positive from Σ_τ to $\Sigma_{\tau+d\tau}$, is the lapse function; $N^r(\tau,\boldsymbol{\sigma}) = n^r(\tau,\boldsymbol{\sigma})$, vanishing at spatial infinity (absence of super-translations), are the shift functions.

The adapted tetrads ${}^4\overset{\circ}{E}{}^A_{(a)}$ are defined modulo SO(3) rotations ${}^4\overset{\circ}{E}{}^A_{(a)} = \sum_b R_{(a)(b)}(\alpha_{(e)})\, {}^4\overset{\circ}{\bar{E}}{}^A_{(b)}$, ${}^3e^r_{(a)} = \sum_b R_{(a)(b)}(\alpha_{(e)})\, {}^3\bar{e}^r_{(b)}$, where $\alpha_{(a)}(\tau,\boldsymbol{\sigma})$ are three point-dependent Euler angles. After having chosen an arbitrary point-dependent

[12] Since one uses the positive-definite 3-metric $\delta_{(a)(b)}$, one will use only lower flat spatial indices. Therefore, for the cotriads, one uses the notation ${}^3e^{(a)}_r \overset{def}{=} {}^3e_{(a)r}$ with $\delta_{(a)(b)} = {}^3e^r_{(a)}\, {}^3e_{(b)r}$.

origin $\alpha_{(a)}(\tau, \sigma) = 0$, one arrives at the following adapted tetrads and cotetrads $[\bar{n}_{(a)} = \sum_b n_{(b)} R_{(b)(a)}(\alpha_{(e)}), \sum_a n_{(a)} {}^3e^r_{(a)} = \sum_a \bar{n}_{(a)} {}^3\bar{e}^r_{(a)}]$

$$
\begin{aligned}
&{}^4\overset{\circ}{\bar{E}}{}^A_{(o)} = {}^4\overset{\circ}{E}{}^A_{(o)} = \frac{1}{1+n}\left(1; -\sum_a \bar{n}_{(a)} {}^3\bar{e}^r_{(a)}\right) = l^A, \qquad {}^4\overset{\circ}{\bar{E}}{}^A_{(a)} = (0; {}^3\bar{e}^r_{(a)}),\\
&{}^4\overset{\circ}{\bar{E}}{}^{(o)}_A = {}^4\overset{\circ}{E}{}^{(o)}_A = (1+n)\,(1;\mathbf{0}) = \varepsilon\, l_A, \qquad {}^4\overset{\circ}{\bar{E}}{}^{(a)}_A = (\bar{n}_{(a)}; {}^3\bar{e}_{(a)r}),
\end{aligned}
\tag{12}
$$

which one will use as a reference standard.

The expression for the general tetrad

$$
\begin{aligned}
{}^4E^A_{(\alpha)} &= {}^4\overset{\circ}{E}{}^A_{(\beta)}\, L^{(\beta)}{}_{(\alpha)}(\varphi_{(a)}) = {}^4\overset{\circ}{\bar{E}}{}^A_{(o)}\, L^{(o)}{}_{(\alpha)}(\varphi_{(c)})\\
&\quad + \sum_{ab} {}^4\overset{\circ}{\bar{E}}{}^A_{(b)}\, R^T_{(b)(a)}(\alpha_{(c)})\, L^{(a)}{}_{(\alpha)}(\varphi_{(c)})
\end{aligned}
\tag{13}
$$

shows that every point-dependent Lorentz transformation Λ in the tangent planes may be parametrized with the (Wigner) boost parameters $\varphi_{(a)}$ and the Euler angles $\alpha_{(a)}$ being the product $\Lambda = R\,L$ of a rotation and a boost.

The future-oriented unit normal to Σ_τ and the projector on Σ_τ are $l_A = \varepsilon\,(1+n)\left(1;0\right)$, ${}^4g^{AB}\, l_A\, l_B = \varepsilon$, $l^A = \varepsilon\,(1+n)\,{}^4g^{A\tau} = \frac{1}{1+n}\left(1; -n^r\right) = \frac{1}{1+n}\left(1; -\sum_a \bar{n}_{(a)} {}^3\bar{e}^r_{(a)}\right)$, ${}^3h^B_A = \delta^B_A - \varepsilon\, l_A\, l^B$.

The 4-metric has the following expression:

$$
\begin{aligned}
&{}^4g_{\tau\tau} = \varepsilon\,[(1+n)^2 - {}^3g^{rs}\, n_r\, n_s] = \varepsilon\left[(1+n)^2 - \sum_a \bar{n}^2_{(a)}\right],\\
&{}^4g_{\tau r} = -\varepsilon\, n_r = -\varepsilon \sum_a \bar{n}_{(a)} {}^3\bar{e}_{(a)r},\\
&{}^4g_{rs} = -\varepsilon\, {}^3g_{rs} = -\varepsilon \sum_a {}^3e_{(a)r}\, {}^3e_{(a)s} = -\varepsilon \sum_a {}^3\bar{e}_{(a)r}\, {}^3\bar{e}_{(a)s},
\end{aligned}
$$

$$
\begin{aligned}
&{}^4g^{\tau\tau} = \frac{\varepsilon}{(1+n)^2}, \qquad {}^4g^{\tau r} = -\varepsilon\,\frac{n^r}{(1+n)^2} = -\varepsilon\,\frac{\sum_a {}^3\bar{e}^r_{(a)}\,\bar{n}_{(a)}}{(1+n)^2},\\
&{}^4g^{rs} = -\varepsilon\,({}^3g^{rs} - \frac{n^r\, n^s}{(1+n)^2}) = -\varepsilon \sum_{ab} {}^3\bar{e}^r_{(a)}\, {}^3\bar{e}^s_{(b)}\,(\delta_{(a)(b)} - \frac{\bar{n}_{(a)}\,\bar{n}_{(b)}}{(1+n)^2}),
\end{aligned}
$$

$$
\begin{aligned}
&\sqrt{-g} = \sqrt{|{}^4g|} = \frac{\sqrt{{}^3g}}{\sqrt{\varepsilon\, {}^4g^{\tau\tau}}} = \sqrt{\gamma}\,(1+n) = {}^3e\,(1+n),\\
&{}^3g = \gamma = ({}^3e)^2, \quad {}^3e = det\, {}^3e_{(a)r}.
\end{aligned}
\tag{14}
$$

The 3-metric ${}^3g_{rs}$ has signature $(+++)$, so that one may put all the flat 3-indices *down*. One has ${}^3g^{ru}\, {}^3g_{us} = \delta^r_s$.

After having introduced the kinematical framework for the description of non-inertial frames in GR, we must study the dynamical aspects of the gravitational field to understand which variables are dynamically determined and which are the inertial effects hidden in the general covariance of the theory. Since at the Lagrangian level it is not possible to identify which components of the 4-metric tensor are connected with the gauge freedom in the choice of the 4-coordinates and which ones describe the dynamical degrees of freedom of the gravitational field, one must restrict himself to the class of globally hyperbolic, asymptotically flat space-times allowing a Hamiltonian description starting from the description of Einstein's GR in terms of the ADM action [36] instead in terms of the Einstein–Hilbert one. In canonical ADM gravity one can use Dirac theory of constraints [37] to describe the Hamiltonian gauge group, whose generators are the first-class constraints[13] of the model. The basic tool of this approach is the possibility to find so-called Shanmugadhasan canonical transformations [38], which identify special canonical bases adapted to the first-class constraints (and also to the second-class ones when present). In these special canonical bases the vanishing of certain momenta (or of certain configurational coordinates) corresponds to the vanishing of well-defined Abelianized combinations of the first-class constraints (Abelianized because the new constraints have exactly zero Poisson brackets even if the original constraints were not in strong involution). As a consequence, the variables conjugate to these Abelianized constraints are inertial Hamiltonian gauge variables describing the Hamiltonian gauge freedom.

Starting from the ADM action for tetrad gravity, one defines the Hamiltonian formalism in a phase space containing 16 configurational field variables and 16 conjugate moments. One identifies the 14 first-class constraints of the system and finds that the canonical Hamiltonian is the weak ADM energy (it is given as a volume integral over 3-space). The existence of these 14 first-class constraints implies that 14 components of the tetrads (or of the conjugate momenta) are Hamiltonian gauge variables describing the *inertial* aspects of the gravitational field (6 of these inertial variables describe the extra gauge freedom in the choice of the tetrads and in their transport along world-lines). Therefore, there are only $2+2$ degrees of freedom for the description of the *tidal* dynamical aspects of the gravitational field. The asymptotic ADM Poincaré generators can be evaluated explicitly. Till now, the type of matter studied in this framework [32] consists of the electromagnetic field and of N charged scalar particles, whose signs of the energy and electric charges are Grassmann-valued to regularize both the gravitational and electromagnetic self-energies (it is both an ultraviolet and an infrared regularization).

[13] The gauge invariance of actions like the ADM one imply the existence of equations, named constraints, restricting the coordinates and momenta of the model to a sub-manifold (named the constraint manifold) of phase space. One of these equations is named a first-class constraint, if it determines one canonical variable with the conjugate variable completely undetermined (arbitrary gauge variable). Two of these equations form a pair of second-class constraints, if they determine a pair of conjugate canonical variables (redundant variables, for instance present to implement manifest covariance of the model).

The remaining $2+2$ conjugate variables describe the dynamical tidal degrees of freedom of the gravitational field (the two polarizations of gravitational waves in the linearized theory). If one would be able to include all the constraints in the Shanmugadhasan canonical basis, these $2+2$ variables would be the *Dirac observables* of the gravitational field, invariant under the Hamiltonian gauge transformations. However, such Dirac observables are not known: one only has statements about their existence [39]. Moreover, in general, they are not 4-scalar observables. The problem of the connection between the 4-diffeomorphism group and the Hamiltonian gauge group was studied in Ref. [40] by means of the inverse Legendre transformation and of the notion of dynamical symmetry. The conclusion is that on the space of solutions of Einstein's equations there is an overlap of the two types of observables: There should exist special Shanmugadhasan canonical bases in which the $2+2$ Dirac observables become 4-scalars when restricted to the space of solutions of the Einstein's equations. In any case the identification of the inertial gauge components of the 4-metric is what is needed to make a fixation of 4-coordinates as required by relativistic metrology.

It can be shown that there is a Shanmugadhasan canonical transformation [41] (implementing the so-called York map [42] and diagonalizing the York–Lichnerowicz approach [43]) to a so-called York canonical basis adapted to 10 of the 14 first-class constraints. Only the super-Hamiltonian and super-momentum constraints, whose general solution is not known, are not included in the basis, but it is clarified which variables are to be determined by their solution, namely the 3-volume element (the determinant of the 3-metric) of the 3-space Σ_τ and the three momenta conjugated to the 3-coordinates on Σ_τ. The 14 inertial gauge variables turn out to be: (a) the six configurational variables $\varphi_{(a)}$ and $\alpha_{(a)}$ of the tetrads describing their O(3,1) gauge freedom; (b) the lapse and shift functions; (c) the 3-coordinates on the 3-space (their fixation implies the determination of the shift functions); (d) the York time [44] ${}^3K$, that is, the trace of the extrinsic curvature of the 3-spaces as 3-manifolds embedded into the space-time (its fixation implies the determination of the lapse function). It is the only gauge variable which is a momentum in the York canonical basis[14]: This is due to the Lorentz signature of space-time, because the York time and three other inertial gauge variables can be used as 4-coordinates of the space-time (see Ref. [24] for this topic and for its relevance in the solution of the hole argument[15]). In this way, an identification of the inertial gauge variables to be fixed to get a 4-coordinate system in relativistic metrology was found. While in SR all the components of the tetrads and their conjugate momenta are inertial gauge variables, in GR the two eigenvalues of the 3-metric with determinant one and their

[14] Instead in Yang–Mills theory all the gauge variables are configurational.

[15] The hole argument concerns the problem of how to make a consistent individuation of the points of a space-time, solution of Einstein's equations, taking into account the general covariance under both active and passive diffeomorphisms. If all the solutions (i.e., the mathematical space-times) connected by such diffeomorphisms correspond to a unique Einstein's space-time (equivalence class of gauge-transformed space-times), then the points must be identified with special coordinates depending upon the 4-metric.

conjugate momenta describe the physical tidal degrees of freedom of the gravitational field. In the first paper of Ref. [32] is shown the expression of the Hamilton's equations for all the variables of the York canonical basis.

An important remark is that in the framework of the York canonical basis the natural family of gauges is not the harmonic one, but the family of 3-orthogonal Schwinger time gauges in which the 3-metric in the 3-spaces is diagonal.

Both in SR and GR an admissible $3+1$ splitting of space-time has two associated congruences of time-like observers [7], geometrically defined and not to be confused with the congruence of the world-lines of fluid elements, when relativistic fluids are added as matter in GR [45–47]. One of the two congruences, with zero vorticity, is the congruence of the Eulerian observers, whose 4-velocity field is the field of unit normals to the 3-spaces. This congruence allows us to reexpress the nonvanishing momenta of the York canonical basis in terms of the expansion ($\theta = -{}^3K$) and of the shear of the Eulerian observers. This allows us to compare the Hamilton's equations of ADM canonical gravity with the usual first-order non-Hamiltonian ADM equations deducible from Einstein's equations given a $3+1$ splitting of space-time but without using the Hamiltonian formalism. As a consequence, one can extend our Hamiltonian identification of the inertial and tidal variables of the gravitational field to the Lagrangian framework and use it in the cosmological (conformally asymptotically flat) space-times: in them, it is not possible to formulate the Hamiltonian formalism, but the standard ADM equations are well defined. The time inertial gauge variable needed for relativistic metrology is now the expansion of the Eulerian observers of the given $3+1$ splitting of the globally hyperbolic cosmological space-time.

4 Conclusion

In conclusion we now have a framework for non-inertial frames in GR and an identification of the inertial gauge variables in asymptotically Minkowskian and also cosmological space-times.

See Refs. [23, 32] for the possibility that dark matter is only a relativistic inertial effect induced by the inertial gauge variable ${}^3K$ (the York time): A suitable choice of the 3-space in PM International Celestial Reference Frame could simulate the effects explained with dark matter. As shown in the third paper of Ref. [32], in the PM equations of motion for point particles (with a suitable regularization of the gravitational self-energies), there are *inertial forces* depending upon the York time. In the PN limit there are residual inertial forces of this type at the order $1/c$ and they may explain the three existing signatures of dark matter (rotation curves of galaxies, mass of galaxies from weak gravitational lensing, and mass of clusters of galaxies from the virial theorem). Since ${}^3K$ determines how non-Euclidean is the 3-space at each instant, dark matter can be reinterpreted as a metrological standard for the deviation of 3-space from the Euclidean one in astrophysics.

Moreover in Ref. [23], at a preliminary level, it is also shown that the York time is connected also with dark energy in cosmological space-times [4]. In the standard FRW space-times, the Killing symmetries connected with homogeneity and isotropy imply (τ is the cosmic time, $a(\tau)$ the scale factor) that the York time is no more a gauge variable but coincides with the Hubble constant: ${}^3K(\tau) = -\frac{\dot{a}(\tau)}{a(\tau)} = -H(\tau)$. However, at the first order in cosmological perturbations (see Ref. [48] for a review) one has ${}^3K = -H + {}^3K_{(1)}$ with ${}^3K_{(1)}$ being again an inertial gauge variable to be fixed with a metrological convention. Therefore, the York time has a central role also in cosmology and one needs to know the dependence on it of the main quantities, like the redshift and the luminosity distance from supernovae [49], which require the introduction of the notion of dark energy to explain the 3-universe and its accelerated expansion in the framework of the standard ΛCDM cosmological model.

In particular, it will be important to study inhomogeneous space-times without Killing symmetries like the Szekeres ones [50], where the York time remains an arbitrary inertial gauge variable, to see whether it is possible to find a 3-orthogonal gauge in them (at least in a PM approximation) in which the convention on the inertial gauge variable York time allows one to eliminate both dark matter and dark energy through the choice of a 4-coordinate system in a consistent PM reformulation of ICRS and simultaneously to save the main good properties of the standard ΛCDM cosmological model due to the inertial and dynamical properties of the space-time.

References

1. Lusanna L. Relativistic metrology: from earth to astrophysics, In-Tech E-Book Modern Metrology Concerns, 2012 (ISBN 978-953-51-0584-8).
2. Soffel M, Klioner SA, Petit G, Wolf P, Kopeikin SM, Bretagnon P, Brumberg VA, Capitaine N, Damour T, Fukushima T, Guinot B, Huang T, Lindegren L, Ma C, Nordtvedt K, Ries J, Seidelmann PK, Vokroulicky' D, Will C, Xu Ch. The IAU 2000 resolutions for astrometry, celestial mechanics and metrology in the relativistic framework: explanatory supplement. Astron J. 2003;126:2687–2706 (arXiv astro-ph/0303376).
   IERS Conventions (2003). In: McCarthy DD, Petit G, editors. IERS TN 32. Verlag des BKG; 2004.
   Kaplan GH. The IAU resolutions on astronomical reference systems, time scales and earth rotation models. U.S. Naval Observatory circular No. 179 (2005) (arXiv astro-ph/0602086).
   Moyer TD. Formulation for observed and computed values of deep space network data types for navigation. New York: Wiley; 2003.
3. Jordan S. The GAIA project: technique, performance and status. Astron Nachr. 2008;329:875. doi:10.1002/asna.200811065.
4. Bartelmann M. The dark universe. Rev Mod Phys. 2010;82:331 (arXiv 0906.5036).
   Bean R, TASI 2009. Lectures on Cosmic Acceleration (arXiv 1003.4468).
5. Alba D. Quantum mechanics in noninertial frames with a multitemporal quantization scheme: II. Nonrelativistic particles. Int J Mod Phys. 2006;A21;3917 (arXiv hep-th/0504060).
6. Mashhoon B. The hypothesis of locality and its limitatins. In: Rizzi G, Ruggiero ML, editors. Relativity in rotating frames. Dordrecht: Kluwer; 2003 (arXiv gr-qc/0303029).
   Mashhoon B, Muench U. Length Measurements in Accelerated Systems. Ann Phys (Leipzig). 2002;11:532.

7. Alba D, Lusanna L. Charged particles and the electro-magnetic field in non-inertial frames: I. Admissible $3+1$ splittings of Minkowski spacetime and the non-inertial rest frames. Int J Geom Methods Mod Phys. 2010;7:33 (arXiv 0908.0213) and II. Applications: rotating frames, sagnac effect, faraday rotation, wrap-up effect. Int J Geom Methods Mod Phys. 2010;7:185 (arXiv 0908.0215); Generalized radar 4-coordinates and equal-time cauchy surfaces for arbitrary accelerated observers (2005). Int J Mod Phys. 2007;D16:1149 (arXiv gr-qc/0501090).
   Bini D, Lusanna L, Mashhoon B. Limitations of radar coordinates. Int J Mod Phys. 2005;D14:1 (arXiv gr-qc/0409052).
8. Bondi H. Assumption and myth in physical theory. Cambridge: Cambridge University Press; 1967.
   D'Inverno R. Introducing Einstein relativity. Oxford: Oxford University Press; 1992.
9. Møller CM. The theory of relativity. Oxford: Oxford University Press; 1957.
   Møller CM. Sur la dinamique des systèmes ayant un moment angulaire interne. Ann Inst H Poincaré. 1949;11:251.
10. Stergioulas N. Rotating stars in relativity. Living Rev Relativ. 2003;6:3.
11. Alba D, Lusanna L. Quantum mechanics in noninertial frames with a multitemporal quantization scheme: I. Relativistic particles. Int J Mod Phys. 2006;A21:2781 (arXiv hep-th/0502194).
12. Lusanna L. The N- and 1-time classical descriptions of N-body relativistic kinematics and the electromagnetic interaction. Int J Mod Phys. 1997;A12:645.
13. Lusanna L. The chrono-geometrical structure of special and general relativity: a re-visitation of canonical geometrodynamics, lectures at 42nd Karpacz Winter School of Theoretical Physics: Current Mathematical Topics in Gravitation and Cosmology, Ladek, Poland, 6-11 Feb 2006. Int J Geom Methods Mod Phys. 2007;4:79 (arXiv gr-qc/0604120).
   Lusanna L. The chronogeometrical structure of special and general relativity: towards a background-independent description of the gravitational field and elementary particles (2004). In: Reiner A, editor. General relativity research trends (Horizon in World Physics vol. 249); New York: Nova Science; 2005 (arXiv gr-qc/0404122).
14. Schmutzer E, Plebanski J. Quantum mechanics in noninertial frames of reference. Fortschr Phys. 1978;25:37.
15. Alba D, Lusanna L, Pauri M. New directions in non-relativistic and relativistic rotational and multipole kinematics for N-Body and continuous systems (2005). In: Ping YL, editor. Atomic and molecular clusters: new research. New York: Nova Science; 2006 (arXiv hep-th/0505005).
   Alba D, Lusanna L, Pauri M. Centers of mass and rotational kinematics for the relativistic N-Body problem in the rest-frame instant form. J Math Phys. 2002;43:1677–1727 (arXiv hep-th/0102087).
   Alba D, Lusanna L, Pauri M. Multipolar expansions for closed and open systems of relativistic particles. J Math Phys. 2004;46:062505, 1–36 (arXiv hep-th/0402181).
16. Crater HW, Lusanna L. The rest-frame Darwin potential from the Lienard–Wiechert solution in the radiation gauge. Ann Phys. (N.Y.) 2001;289:87.
   Alba D, Crater HW, Lusanna L. The semiclassical relativistic Darwin potential for spinning particles in the rest frame instant form: two-body bound states with spin 1/2 constituents. Int J Mod Phys. 2001;A16:3365–3478 (arXiv hep-th/0103109).
   Alba D, Lusanna L. The Lienard–Wiechert potential of charged scalar particles and their relation to scalar electrodynamics in the rest-frame instant form. Int J Mod Phys. 1998;A13:2791 (arXiv hep-th/0708156).
17. Alba D, Crater HW, Lusanna L. Towards relativistic atom physics. I. The rest-frame instant form of dynamics and a canonical transformation for a system of charged particles plus the electro-magnetic field. Can J Phys. 2010;88:379 (arXiv 0806.2383) and II. Collective and relative relativistic variables for a system of charged particles plus the electro-magnetic field. Can J Phys. 2010;88:425 (arXiv 0811.0715).
18. Alba D, Crater HW, Lusanna L. Hamiltonian relativistic two-body problem: center of mass and orbit reconstruction. J Phys. 2007;A40:9585 (arXiv gr-qc/0610200).

19. Alba D, Crater HW, Lusanna L. On the relativistic micro-canonical ensemble and relativistic kinetic theory for N relativistic particles in inertial and non-inertial rest frames. Int J Geom Methods Mod Phys. 2015;12:1550049 (arXiv 1202.4667).
20. Bini D, Lusanna L. Spin-rotation couplings: spinning test particles and Dirac fields. Gen Rel Grav. 2008;40:1145 (arXiv 0710.0791).
Bigazzi F, Lusanna L. Spinning particles on spacelike hypersurfaces and their rest-frame description. Int J Mod Phys. 1999;A14:1429 (arXiv hep-th/9807052); Dirac fields on spacelike hypersurfaces, their rest-frame description and Dirac observables. Int J Mod Phys. 1999;A14:1877 (arXiv hep-th/9807054).
21. Alba D, Lusanna L. The classical relativistic quark model in the rest-frame Wigner-covariant coulomb gauge. Int J Mod Phys. 1998;A13:3275 (arXiv hep-th/9705156).
Alba D, Crater HW, Lusanna L. Massless particles plus matter in the rest-frame instant form of dynamics. J Phys. 2010;A43:405203 (arXiv 1005.5521).
Alba D, Crater HW, Lusanna L. The rest-frame instant form and Dirac observables for the open Nambu String. Eur Phys J Plus. 2011;126:26 (arXiv 1005.3659).
Alba D, Crater HW, Lusanna L. A relativistic version of the two-level atom in the rest-frame instant form of dynamics. J Phys. 2013;A46:195303 (arXiv 1107.1669).
22. Alba D, Crater HW, Lusanna L. Relativistic quantum mechanics and relativistic entanglement in the rest-frame instant form of dynamics. J Math Phys. 2011;52:062301 (arXiv 0907.1816).
23. Lusanna L. From clock synchronization to dark matter as a relativistic inertial effect, Lecture at the Black Objects in Supergravity School BOSS2011, Frascati, 9–13 May 2011 (arXiv 1205.2481), Springer Proc. in Phys. 144, S.Bellucci editor, Springer 2013.
24. Lusanna L, Pauri M. Explaining Leibniz equivalence as difference of non-inertial appearances: dis-solution of the Hole Argument and physical individuation of point-events. Hist Philos Mod Phys. 2006;37:692 (arXiv gr-qc/0604087); The physical role of gravitational and gauge degrees of freedom in general relativity. I: Dynamical synchronization and generalized inertial effects; II: Dirac versus Bergmann observables and the objectivity of space-time. Gen Rel Grav. 2006;38, 187 and 229 (arXiv gr-qc/0403081 and 0407007); Dynamical emergence of instantaneous 3-spaces in a class of models of general relativity, to appear in the book Relativity and the Dimensionality of the World, Springer, Berlin. In: van der Merwe A, editor. (Springer Series Fundamental Theories of Physics 153) (arXiv gr-qc/0611045).
25. Cacciapuoti L, Salomon C. ACES: Mission Concept and Scientific Objective, 28/03/2007, ESA document, Estec (ACES_Science_v1_printout.doc).
Blanchet L, Salomon C, Teyssandier P, Wolf P. Relativistic theory for time and frequency transfer to order $1/c^3$. Astron Astrophys. 2000;370:320.
Lusanna L. Dynamical emergence of 3-space in general relativity: implications for the ACES mission, in Proceedings of the 42th Rencontres de Moriond Gravitational Waves and Experimental Gravity, La Thuile (Italy), 11–18 March 2007.
26. Lusanna L. The rest-frame instant form of metric gravity. Gen Rel Grav. 2001;33:1579 (arXiv gr-qc/0101048).
Lusanna L, Russo S. A new parametrization for tetrad gravity. Gen Rel Grav. 2002;34:189 (arXiv gr-qc/0102074).
DePietri R, Lusanna L, Martucci L, Russo S. Dirac's observables for the rest-frame instant form of tetrad gravity in a completely fixed 3-orthogonal gauge. Gen Rel Grav. 2002;34:877 (arXiv gr-qc/0105084).
Agresti J, De Pietri R, Lusanna L, Martucci L. Hamiltonian linearization of the rest-frame instant form of tetrad gravity in a completely fixed 3-orthogonal gauge: a radiation gauge for background-independent gravitational waves in a post-Minkowskian Einstein spacetime. Gen Rel Grav. 2004;36:1055 (arXiv gr-qc/0302084).
27. Ashtekar A. Asymptotic structure of the gravitational field at spatial infinity. In: Held A, editor. General relativity and gravitation. Vol. 2. New York: Plenum; 1980.
28. McCarthy PJ. Asymptotically flat space-times and elementary particles. Phys Rev Lett. 1972;29:817; Structure of the Bondi–Metzner–Sachs group. J Math Phys. 1972;13:1837.

29. Beig R, Ó Murchadha N. The Poincaré group as the symmetry group of canonical general relativity. Ann Phys (NY). 1987;174:463.
30. Fatibene L, Ferraris M, Francaviglia M, Lusanna L. ADM pseudotensors, conserved quantities and covariant conservation laws in general relativity. Ann Phys (NY). 2012;327:1593 (arXiv 1007.4071).
Jaramillo JL, Gourgoulhon E. Mass and angular momentum in general relativity. In: Blanchet L, Spallicci A, Whiting B, editors. Mass and motion in general relativity. (Fundamental Theories in Physics Vol. 162, p. 87) (Springer, Berlin, 2011) (arXiv 1001.5429).
Szabados LB. On the roots of Poincaré structure of asymptotically flat spacetimes. Class Quant Grav. 2003;20:2627 (arXiv gr-qc/0302033).
31. Christodoulou D, Klainerman S. The global nonlinear stability of the Minkowski space. Princeton: Princeton University Press; 1993.
32. Alba D, Lusanna L. The Einstein–Maxwell-particle system in the York canonical basis of ADM tetrad gravity: I) The equations of motion in arbitrary Schwinger time gauges., 2009 (arXiv 0907.4087); II) The weak field approximation in the 3-orthogonal gauges and Hamiltonian post-Minkowskian gravity: the N-body problem and gravitational waves with asymptotic background., (arXiv 1003.5143); III) The post-Minkowskian N-body problem, its post-Newtonian limit in non-harmonic 3-orthogonal gauges and dark matter as an inertial effect. (arXiv 1009.1794); Can J Phys. 2012;90, 1017 and 1077 and 1131.
33. Rovelli C. Quantum gravity. Cambridge: Cambridge University Press; 2004.
34. Geroch R. Spinor structure of space-times in general relativity I. J Math Phys. 1968;9:1739.
35. Stewart J. Advanced general relativity. Cambridge: Cambridge University Press; 1993.
36. Arnowitt R, Deser S, Misner CW. The dynamics of general relativity. In: Witten L, editor. Gravitation: an introduction to current research. New York: Wiley; 1962 (arXiv gr-qc/0405109).
37. Dirac PAM. Lectures on Quantum Mechanics, Belfer Graduate School of Science, Monographs Series. New York: Yeshiva University; 1964.
Henneaux M, Teitelboim C. Quantization of Gauge Systems. Princeton: Princeton University Press; 1992.
38. Shanmugadhasan S. Canonical formalism for degenerate Lagrangians. J Math Phys. 1973;14:677.
Lusanna L. The Shanmugadhasan canonical transformation, function groups and the second Noether theorem. Int J Mod Phys. 1993;A8:4193; The relevance of canonical transformations in gauge theories and general relativity, Lecture Notes of "Seminario Interdisciplinare di Matematica" (Basilicata University) 5, 125 (2006).
Chaichian M, Louis Martinez D, Lusanna L. Dirac's constrained systems: the classification of second-class constraints. Ann Phys (NY) 1994;232:40.
39. Dittrich B. Partial and complete observables for Hamiltonian constrained systems. Gen Rel Grav. 2007;39:1891 (arXiv gr-qc/0411013); Partial and complete observables for canonical general relativity. Class Quantum Grav. 2006;23:6155 (arXiv gr-qc/0507106).
Thiemann T. Reduced phase space quantization and Dirac observables. Class Quantum Grav. 2006;23:1163 (arXiv gr-qc/0411031).
Pons JM, Salisbury DC, Sundermeyer KA. Revisiting observables in generally covariant theories in the light of gauge fixing methods. Phys Rev. 2009;D80:084015 (0905.4564); Observables in classical canonical gravity: folklore demystified. J Phys Conf Series. 2010;222:012018 (arXiv 1001.2726).
40. Moncrief V. Space-time symmetries and linearization stability of the Einstein equations. I J Math Phys. 1975;16:493; Decompositions of gravitational perturbations. J Math Phys. 1975;16:1556; Space-time symmetries and linearization stability of the Einstein equations. II. J Math Phys. 1976;17:1893; Invariant states and quantized gravitational perturbations. Phys Rev. 1978;D18:983.
Choquet-Bruhat Y, Fischer AE, Marsden JE. Maximal hypersurfaces and positivity of mass, LXVII E. Fermi Summer School of Physics Isolated Gravitating Systems in General Relativity. In: Ehlers J, editor. North-Holland Amsterdam; 1979.

Fisher AE, Marsden JE. The initial value problem and the dynamical formulation of general relativity. In: Hawking SW, Israel W, editors. General relativity. An Einstein centenary survey. Cambridge: Cambridge University Press; 1979.

41. Alba D, Lusanna L. The York map as a Shanmugadhasan canonical transformation in tetrad gravity and the role of non-inertial frames in the geometrical view of the gravitational field. Gen Rel Grav. 2007;39:2149 (arXiv gr-qc/0604086, v2).
42. Isenberg J, Marsden JE. The York map is a canonical transformation. J Geom Phys. 1984;1:85.
43. Ciufolini I, Wheeler JA. Gravitation and inertia. Princeton: Princeton University Press; 1995.
44. York JW Jr. Kinematics and dynamics of general relativity. In: Smarr LL, editor. Sources of gravitational radiation. Battelle-Seattle workshop 1978; Cambridge: Cambridge University Press; 1979.
    Quadir A, Wheeler JA. In: York's cosmic time versus proper time. In: Gotsma E, Tauber G, editors. From SU(3) to gravity, Y.Ne'eman's festschrift. Cambridge: Cambridge University Press; 1985.
    Beig R. The classical theory of canonical general relativity. In: Ehlers J, Friedrich H, editors. Canonical gravity: from classical to quantum. Berlin: Springer; 1994 (Bad Honnef 1993, Lecture Note in Phys. 434).
45. Alba D, Lusanna L. Dust in the York canonical basis of ADM tetrad gravity: the problem of vorticity. Int J Geom Methods Mod Phys. 2015;12:1550076 (arXiv 1106.0403).
46. Lusanna L, Nowak-Szczepaniak D. The rest-frame instant form of relativistic perfect fluids with equation of state $\rho = \rho(\eta, s)$ and of nondissipative elastic materials. Int J Mod Phys. 2000;A15:4943.
    Alba D, Lusanna L. Generalized Eulerian coordinates for relativistic fluids: Hamiltonian rest-frame instant form, relative variables, rotational kinematics. Int J Mod Phys. 2004;A19:3025 (arXiv hep-th/020903).
47. Brown JD. Action functionals for relativistic perfect fluids. Class Quantum Grav. 1993;10:1579.
48. Nakamura K. Second-order gauge-invariant cosmological perturbation theory: current status. Adv Astron. 2010;2010:576273 (arXiv 1001.2621).
49. Durrer R. What do we really know about dark energy? Phil Trans Roy Soc London. 2011;A369:5102 (arXiv 1103.5331).
    Bonvin C, Durrer R. What galaxy surveys really measure. Phys Rev. 2011;D84:063505 (arXiv 1105.5280).
50. Bonnor WB, Sulaiman AH, Tanimura N. Szekeres's space-times have no Killing vectors. Gen Rel Grav. 1977;8:549.
    Berger BK, Eardley DM, Olson DW. Notes on the spacetimes of Szekeres. Phys Rev. 1977;D16:3086.
    Nwankwo A, Ishak M, Thomson J. Luminosity distance and redshift in the Szekeres inhomogeneous cosmological models. J Cosm Astro Part. 2011;2011:28 (arXiv 1005.2989).
    Pleban'ski J, Krasin'ski A. An introduction to general relativity and cosmology. Cambridge: Cambridge University Press; 2006.
    Celerier MN. Effects of inhomogeneities on the expansion of the Universe: a challenge to dark energy? J Phys Conf Series. 2014;484:012005 (arXiv 1203.2814).

The Acceleration Scale, Modified Newtonian Dynamics and Sterile Neutrinos

Antonaldo Diaferio and Garry W. Angus

Abstract General relativity is able to describe the dynamics of galaxies and larger cosmic structures only if most of the matter in the universe is dark, namely, it does not emit any electromagnetic radiation. Intriguingly, on the scale of galaxies, there is strong observational evidence that the presence of dark matter appears to be necessary only when the gravitational field inferred from the distribution of the luminous matter falls below an acceleration of the order of 10^{-10} m s^{-2}. In the standard model, which combines Newtonian gravity with dark matter, the origin of this acceleration scale is challenging and remains unsolved. On the contrary, the full set of observations can be neatly described, and were partly predicted, by a modification of Newtonian dynamics, dubbed MOND, that does not resort to the existence of dark matter. On the scale of galaxy clusters and beyond, however, MOND is not as successful as on the scale of galaxies, and the existence of some dark matter appears unavoidable. A model combining MOND with hot dark matter made of sterile neutrinos seems to be able to describe most of the astrophysical phenomenology, from the power spectrum of the cosmic microwave background anisotropies to the dynamics of dwarf galaxies. Whether there exists a yet unknown covariant theory that contains general relativity and Newtonian gravity in the weak field limit and MOND as the ultra-weak field limit is still an open question.

1 Introduction

The dynamical properties of galaxies and larger cosmic structures should be described by their mass content if gravity is the dominant force field. In 1932, Oort noticed a shortage of mass required to describe the velocity of stars in the solar neighbourhood [1]. In 1933, Zwicky found the very same problem by applying the

A. Diaferio (✉)
Dipartimento di Fisica, Università degli Studi di Torino, and Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
e-mail: diaferio@ph.unito.it

G. W. Angus
Astrophysics, Cosmology and Gravity Centre, University of Cape Town, Private Bag X3, Rondebosch 7700, South Africa
e-mail: angus.gz@gmail.com

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_10

virial theorem to the Coma cluster, based on the radial velocities of a few galaxies derived from their optical spectra [2]. These early discoveries were revived in the 70s, when the existence of flat rotation curves in disk galaxies [3] and the requirement of the dynamical stability of galactic disks [4, 5] showed that a large fraction of mass, in addition to the observed luminous mass, was necessary to describe the dynamics of galaxies. Ever since, the evidence of missing mass in cosmic structures has become overwhelming [6].

General relativity (GR) has encountered such a plethora of successes in the solar system that the scientific community finds it hard to suppose that GR, and its Newtonian weak field limit, might fail on cosmic scales. It is more natural to suppose that this missing mass is actually some form of dark matter (DM) that does not emit any electromagnetic radiation but only acts gravitationally.

This idea meets the demand of particle physics: going beyond the standard model of particle physics requires the existence of still unknown elementary particles like supersymmetric particles, axions, Kaluza–Klein excitations or sterile neutrinos. These particles may naturally play the role of the astrophysical DM [7]: they formed in the early universe and their relic abundance can fill in the fraction of mass that is required to describe both the internal dynamics of structures and their formation by gravitational instability.

In addition, most of these particles have the advantage of being non-baryonic, namely, in astrophysical jargon, they are neither electrons nor particles made of quarks. Non-baryonic matter includes neutrinos and the hypothetical weakly interacting massive particles (WIMPs). So far, only neutrinos have been detected; all other non-baryonic matter is still hypothetical.

Being non-baryonic is advantageous if gravitational instability drives the formation of the cosmic structure. In fact, in the inflationary scenario of the standard hot Big Bang cosmology, small perturbations due to quantum fluctuations are inflated to cosmic scales by the ~ 100 e-folding expansion of the universe and provide the initial conditions of the matter density field. However, this scenario is contradicted by the cosmic microwave background (CMB) anisotropies if the matter density field is mostly baryonic. In fact, the observation of temperature anisotropies $\delta T/T = \delta/3$, with δ the baryonic matter density fluctuations, in the CMB, which formed by redshift $z \sim 10^3$, yields $\delta T/T \sim 10^{-5}$ on $\theta \sim 7^\circ$ angular scales, corresponding to superclusters and larger structures [8]. Gravitational instability yields a growth rate $\propto (1+z)^{-1}$ or slower. Thus, superclusters would have matter overdensities δ of the order $\sim 10^{-2}$ today, rather than the observed $\delta \sim 1-10$. Non-baryonic DM decouples from the radiation field much earlier than baryons and its density perturbations can start growing at the time of equivalence, when the radiation and matter energy densities are equal. At the time of the baryon-radiation decoupling, DM perturbations have already grown to $\delta \sim 10^{-2} - 10^{-3}$ on supercluster scales and they will keep growing to $\delta \sim 1-10$ by today.

Further evidence of the non-baryonic nature of DM is the abundance of light elements which are synthesized in the early universe. Measures of the primordial

abundance of deuterium, for example, which is particularly sensitive to the photon-to-baryon ratio, implies a baryon density $\Omega_b h^2 = 0.0214 \pm 0.0020$[1] [9], which agrees with the baryon density $\Omega_b h^2 = 0.02229 \pm 0.00073$ implied by the CMB anisotropies [10], yet it is smaller than the total matter density $\Omega_m h^2 = 0.114$ [11].

By combining all these pieces of evidence, the standard interpretation of the observations of the cosmic structure, from galactic scales to the CMB, suggests that only 17 % of the matter in the universe is baryonic, whereas the remaining 83 % is required to be non-baryonic cold DM [11]. The entire amount of matter, however, yields only 27 % of the matter-energy density required to make the universe geometrically flat, as suggested by the CMB power spectrum [12]. The missing 73 % can be easily described by a cosmological constant [13], but a large fraction of the cosmological community is searching for more sophisticated models, some of them including an additional dark energy (DE) fluid [6].

The current standard ΛCDM model thus contains non-baryonic cold DM and a cosmological constant Λ. However, we are in the uneasy situation where only 4.6 % of the matter-energy density of the universe is made of matter we can detect in laboratories on Earth. The rest is hypothetical (DM) and has unusual properties, like negative pressure (Λ or DE). To avoid this situation, the alternative solution is to assume that GR breaks down at some scale. A number of extended or modified theories of gravity have been suggested. Among them, Milgrom [14] proposed modified Newtonian dynamics (MOND), an empirical law that modifies Newtonian dynamics and actually goes beyond a simple alternative theory of gravity. MOND can elegantly deal with the galaxy scale phenomenology without DM but is less successful at describing the dynamical properties of galaxy clusters. Famaey and McGaugh [15] have recently extensively reviewed this model. Here, we summarize the successes of MOND and mention how the combination of MOND with the existence of sterile neutrinos of mass in the range 11 eV$-$1 keV can provide a reasonable description of the observed phenomenology, from galaxy scales to the CMB.

2 The Acceleration Scale

In 1983, MOND was originally introduced to explain the observed rotation curves of spiral galaxies [14]: The observed velocities of the stars are larger than the velocities that Newtonian dynamics would predict based on the distribution of the luminous matter [3]. As we mentioned above, this observational piece of evidence suggested the existence of some hidden mass, or DM, responsible for a stronger gravitational field and hence larger velocities. An alternative solution is that, on these cosmic scales, Newtonian gravity does not hold. Milgrom [14] went beyond the suggestion of a modified theory of gravity that becomes relevant beyond some large length

[1] We use the standard parametrization for the Hubble constant today $H_0 = 100\,h$ km s^{-1} Mpc^{-1}.

scale. Rather, he suggested that Newtonian dynamics breaks down below an acceleration scale a_0. In the framework of classical physics, MOND assumes the following relation between the acceleration $\mathbf{a}$ actually felt by a point mass and the acceleration $\mathbf{a}_\mathrm{N}$ expected in Newtonian dynamics:

$$\mu\left(\frac{|\mathbf{a}|}{a_0}\right)\mathbf{a} = \mathbf{a}_\mathrm{N}, \tag{1}$$

where $\mu(x)$ is an unknown interpolating function that satisfies the conditions $\mu(x \to +\infty) = 1$ and $\mu(x \to 0) = x$.[2] With the acceleration scale $a_0 \approx 10^{-10}$ m s^{-2}, this relation enables the description of the observed rotation curves of spiral galaxies without requiring the existence of any DM [16]. Moreover, Milgrom showed that this modification of Newtonian dynamics also drastically reduces the amount of DM required in groups and clusters of galaxies [17].

More importantly, Milgrom's suggestion provided clean predictions [16] that were confirmed in later years. Most notably (1) the zero point of the Tully–Fisher relation, (2) the one-to-one correspondence between features in the baryonic distribution and the rotation curve, (3) a larger mass discrepancy, when interpreted with Newtonian gravity, in low-surface brightness dwarf spheroidal galaxies and (4) the validity of the Tully–Fisher relation for low-surface brightness disk galaxies.

It also became clear very soon that there are different numerical coincidences that might suggest that a_0 is indeed a fundamental quantity (see [18] for a recent review). For example, it is rather intriguing that a_0 is related to the Hubble constant H_0 and the cosmological constant Λ with the relations $a_0 = cH_0/2\pi$ and $a_0^2 = c^2\Lambda/2\pi$, where c is the speed of light. A compilation of kinematic and photometric data of disk galaxies very clearly shows the validity of this *ansatz* on the acceleration scale [19]: Disk galaxies do not require DM beyond a given length scale, but rather below a given acceleration scale. Figure 1 shows the ratio between the observed velocity V and the velocity V_b expected in Newtonian gravity from the distribution of the visible matter. The square of this ratio is proportional to the ratio between the total mass and the visible mass in a spherical system. Evidently, $(V/V_b)^2 > 1$ reflects a mass discrepancy and the requirement of DM; the larger the ratio the larger is the contribution of DM. Figure 1 clearly shows that the need for DM increases with decreasing Newtonian acceleration $g_N = V_b^2/r$ which is proportional to the

[2] A year later, Bekenstein and Milgrom [20] showed that, in spherical symmetry, Eq. (1) is equivalent to a modified theory of gravity where the standard Poisson equation is replaced by

$$\nabla\cdot\left[\mu\left(\frac{|\nabla\Phi|}{a_0}\right)\nabla\Phi\right] = 4\pi G\rho, \tag{2}$$

where Φ is the gravitational potential and ρ the mass distribution. A more recent modified theory of gravity that reproduces Eq. (1) is the quasi-linear formulation of MOND (QUMOND) [21]. This theory has the advantage of involving only linear differential equations and one nonlinear algebraic step. See [15] for a comprehensive review of the various formulations of MOND as modified dynamics or modified gravity.

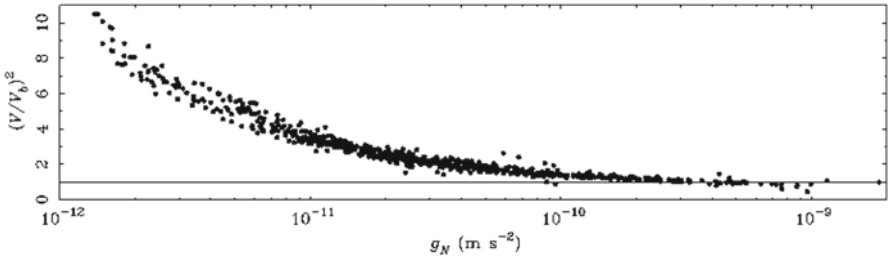


Fig. 1 $(V/V_b)^2$ versus the Newtonian gravity $g_N = V_b^2/r$, derived from the baryonic surface density in almost 100 spiral galaxies. The need for DM appears when $(V/V_b)^2 > 1$. Observations clearly show that this only happens below a specific acceleration of the order of 10^{-10} m s^{-2}. (Courtesy of B. Famaey and S. McGaugh. Reproduced from [15], with permission)

baryonic surface density. The MOND prediction is thus that low-surface brightness (LSB) galaxies are more DM-dominated than high-surface brightness (HSB) galaxies in Newtonian gravity. However, the Tully–Fisher relation, that we describe below, remains valid for both classes of galaxies. In the DM paradigm, it is unclear how the tight correlation shown in Fig. 1 can emerge from the combination of intrinsically chaotic processes, including the history of the DM halo formation, the star formation history and feedback processes.

If this acceleration scale does indeed exist, we have two relevant consequences: (1) MOND has to break the strong equivalence principle, which states that all laws of physics are independent of velocity and location in spacetime and (2) Birkhoff's theorem is not valid.

MOND is not required to break the weak equivalence principle which asserts the equality between inertial and gravitational masses. In GR, the weak quivalence principle translates into the assumption that all matter fields are coupled with the geodesic metric $g_{\mu\nu}$, whereas the strong equivalence principle is guaranteed by assuming that the same metric $g_{\mu\nu}$ obeys the Einstein–Hilbert action $S_{\rm EH} = c^4/(16\pi G)\int g^{\mu\nu}R_{\mu\nu}(-g)^{1/2}{\rm d}^4x$, where $R_{\mu\nu}$ is the Ricci tensor and g the determinant of the metric. Therefore, theories where the Einstein and geodesic metrics are distinct do not necessarily satisfy the strong equivalence principle.

The nonvalidity of Birkhoff's theorem clearly complicates the interpretation of the dynamical properties of self-gravitating systems. In Newtonian gravity, any object is subject to the gravitational attraction of all the other objects in the universe, but in Newtonian linear dynamics a star in a galaxy or a galaxy in a galaxy cluster is virtually isolated from the gravitational field felt by the parent system, unless this latter external field varies across the parent system and originates the well-known effect of tides. In MOND, the external field affects the internal motions of the objects even when this field is constant, because it is the total acceleration felt by the object that determines the dynamical, MONDian or Newtonian, regime. We will mention below that this issue can be relevant in the investigation of the internal dynamics of elliptical galaxies, which are mostly located in regions of high galaxy density, and dwarf spheroidal galaxies and globular clusters, which are satellites of larger galaxies.

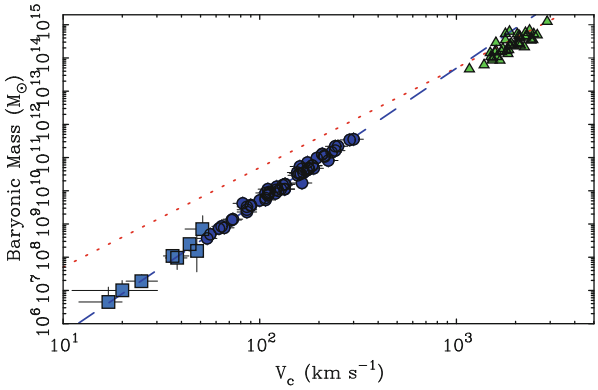


Fig. 2 Baryonic mass M_b versus circular velocity V_c in dwarf *(squares)* and spiral *(circles)* galaxies [23] and in clusters of galaxies (*triangles*; [24]). The *blue dashed line* is the fit to the spiral galaxies alone $M_b = 50(V_c/\mathrm{km\,s^{-1}})^4 M_\odot$. The *red dotted line* is the simplest standard model expectation if all the baryons in each DM halo are identified. (Courtesy of S. McGaugh. Reproduced from [25], with permission)

In Sect. 4, we describe an example of how these issues enter the construction of a covariant theory of MOND.

3 Self-gravitating Systems

3.1 Disk Galaxies

On the scales of galaxies there are a number of observations that were either predicted or easily explained by MOND. In the DM paradigm, these very same observations require extreme fine-tuning between baryonic and non-baryonic physics or even yet undiscovered mechanisms.

Tully and Fisher [22] showed that, in disk galaxies, the maximum rotation velocity v_{rot} is proportional to the galaxy luminosity. MOND predicts that at large radii $a^2/a_0 \sim a_N = GM_b/r^2$ (Eq. 1) and from $a = v_{\mathrm{rot}}^2/r$ we obtain

$$v_{\mathrm{rot}}^4 = GM_b a_0, \tag{3}$$

where M_b is the baryonic mass. Figure 2 shows the measured baryonic mass M_b as a function of the measured circular velocity V_c for astrophysical systems on different scales, from dwarf galaxies (left bottom corner) to clusters of galaxies (upper right corner). V_c is a characteristic circular velocity measured in the outer region of the system [25]. The blue dashed line shows the relation $M_b = 50(V_c/\mathrm{km\,s^{-1}})^4 M_\odot$. The MOND prediction clearly agrees with both the observed slope and normalization. In addition, the observed spread is consistent with the uncertainties. This result implies that the relation $v_{\mathrm{rot}}^4 = GM_b a_0$ holds exactly.

In the standard model, disk galaxies are embedded in DM halos whose average density within their virial radius is basically independent of the halo mass. It is thus usual to define r_{200} as the radius within which the average density is 200 times the critical density of the universe. It follows that for M_{200}, the mass within r_{200}, we have $M_{200} \propto r_{200}^3$ and the circular velocity $V_c = (GM_{200}/r_{200})^{1/2} \propto M_{200}^{1/3}$. This relation is shown as the red dotted line in Fig. 2. The circular velocity is not necessarily identical to the disk rotation velocity $v_{\rm rot}$, because of the complex interplay between the merger history of the DM halo and the star formation history and energy feedback of the galaxy. For example, we could easily recover the correct slope $v_{\rm rot}^4 \propto M_b$, by assuming that luminosity traces the total mass ($L \propto M_b \propto M_{\rm tot}$) and that the density and scale height of the galaxy disk is roughly constant in disk galaxies; this latter assumption implies $M_{\rm tot} \propto R^2$ where R is the disk size. From $v_{\rm rot}^2 = GM_{\rm tot}/R$, we correctly derive $v_{\rm rot}^4 \propto M_b$. A rigorous comparison between observations and the standard model is not trivial [26], but recent analyses, where properly balanced contributions of the various physical and observational effects are carefully blended, seem to bring the ΛCDM Tully–Fisher relation in better agreement with observations [27, 28].

However, in the standard model, the predicted scatter remains larger than observed, because, unlike MOND, we do expect that the galaxy merger and star formation history mentioned above introduce an intrinsic scatter. In addition, a priori, we do not have any reason to expect that (1) the observed relation extends over five orders of magnitude in baryonic mass, from dwarf galaxies to massive disk galaxies, again with basically no scatter, as shown in Fig. 2; and that (2) the LSB galaxies, which should presumably have a star formation efficiency lower than normal galaxies [29, 30], also perfectly fit into this relation. This latter result was originally predicted by Milgrom [16], 15 years before the measurements of the rotation curves of LSB galaxies [31]. MOND goes beyond the description of global properties of disk galaxies. The open symbols with error bars in Fig. 3 show the observed rotation curves of an HSB galaxy and an LSB galaxy. The rotation curves expected in Newtonian gravity from the distribution of baryonic matter (black solid lines) severely underestimate the observations. The underestimate increases with distance R from the galaxy centre and is larger for the LSB galaxy. In the standard model, this observation is explained by increasing the DM contribution with increasing R and decreasing galaxy luminosity. This solution is usually motivated by a lower star formation efficiency at larger radii, as suggested by extensive surveys of neutral hydrogen in nearby galaxies [32, 33].

What is more remarkable is that the small-scale features of the rotation curves mirror the distribution of the baryonic matter in the disk. This characteristic is shared by the rotation curves expected in Newtonian gravity from the distribution of baryonic matter (black solid lines in Fig. 3). In the standard model, this property of the total rotation curve is unexpected, because this rotation curve should be mostly determined by the dynamically dominant DM distribution and the baryonic distribution should play a very minor role. On the contrary, MOND describes the observations with impressive accuracy (blue solid lines), including the small-scale features of the curves (see also [34]).

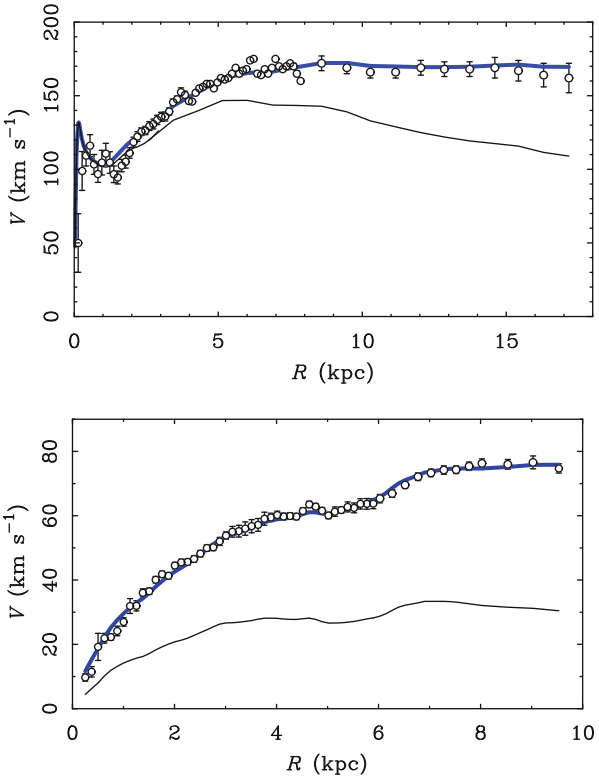


Fig. 3 Examples of modification of Newtonian dynamics (MOND) rotation curve fits of a high-surface brightness (HSB) galaxy (NGC 6946, *top panel*) and a low-surface brightness (LSB) galaxy (NGC 1560, *bottom panel*). The *black lines* show the Newtonian rotation curve expected from the observed distribution of stars and gas. The *blue lines* are the MOND fits with best fit stellar mass-to-light ratios in the K-band 0.37 $M_\odot/L_\odot$ (NGC 6946) and 0.18 $M_\odot/L_\odot$ (NGC 1560). (Courtesy of B. Famaey and S. McGaugh. Reproduced from [15], with permission)

The MOND curves are derived with a single free parameter, the stellar mass-to-light ratio of the disk (namely the ratio between the disk mass in stars and the disk luminosity), in addition to the distance to the galaxy required to determine its length and hence, acceleration scales; the best fit values of the mass-to-light ratios are in perfect agreement with values derived from stellar population synthesis models. Moreover, redder galaxies require larger mass-to-light ratios, as expected on completely independent astrophysical grounds [35]. This limited freedom of MOND should be compared with the standard model that requires two parameters for the DM halo of each galaxy and a global mass-to-light ratio of the galaxy that depends on R and is unrelated to the mass-to-light ratios of the stellar populations.

The agreement shown in Fig. 3 is common to 75 nearby galaxies to date [15]. Among these, an interesting case is NGC 7814: This galaxy is almost perfectly edge-on and the uncertainties on the rotation curve deriving from the disk inclination are

negligible; in addition, the distance to the galaxy is accurate to 5 % and is basically not a free parameter any longer. This galaxy provides a stringent test for MOND: high-quality infrared photometric observations [36] enable, for the first time, the construction, from an accurate bulge-disk decomposition, of a three-dimensional model of the galaxy whose gravitational potential is inferred by numerically solving the MONDian Poisson equation. The comparison of the model rotation curve with the observed one allows the derivation of the mass-to-light ratios for both the disk and the bulge components. Both ratios are found to be in excellent agreement with the expected values [37].

The data expected in the upcoming GAIA mission and other future surveys will provide unprecedented possibilities to test MOND with the Milky Way dynamics [15]. Pioneering work with current data has already shown that the rotation curve and the surface density of the inner disk of the Milky Way are fully consistent with each other within the MOND framework [38, 39]. Similarly, the escape velocity from the solar neighbourhood agrees with current estimates if the external gravitational field in which the Milky Way is embedded is $a \sim 10^{-2} a_0$: This value is indeed compatible with the actual field [40]. An additional test involves the velocity ellipsoid tilt angle within the meridional galactic plane. The angles expected in MOND and in the standard Newtonian gravity with DM agree with each other and with observations at galactic heights $z = 1$ kpc; however, the discrepancy between the predicted angles in the two models increases with z and the measure of the velocity ellipsoid tilt angle will thus be a relevant test to discriminate between the models [41].

3.2 Elliptical Galaxies

The role played by the acceleration scale is also apparent in elliptical galaxies, dwarf spheroidal galaxies (dSphs) and globular clusters.

In elliptical galaxies, which are pressure-supported systems, Faber and Jackson [42] observed a relation similar to the Tully–Fisher one that is valid for disk galaxies: The galaxy luminosity L correlates with the stellar velocity dispersion σ in the galaxy's central region, according to the power law $L \propto \sigma^4$. If we assume that ellipticals are isothermal spheres that, in MOND, have finite mass with asymptotically decreasing density $\rho \propto r^\alpha$, with $\alpha = d \ln \rho / d \ln r = -4$, we find $\sigma^4 = G M_b a_0 / \alpha^2$ [43]. Unlike the Tully–Fisher relation, the observed Faber–Jackson relation is not expected to be exact in MOND, because ellipticals are not strictly isothermal and their velocity field is not isotropic; the velocity anisotropy parameter must actually vary with radius to match the observed fundamental plane of ellipticals [44]. The MOND fundamental plane is actually slightly tilted compared to observations, but this problem might be removed by including the external field effect [45].

Elliptical galaxies pose some challenges to the standard model, because they should be embedded, like disk galaxies, in DM halos. X-ray-emitting hot gas coronae are expected signatures of a DM halo and are indeed observed in many early type galaxies [46], including isolated ellipticals, like NGC 1521 [47]. However, there are cases that are unexpected in the DM paradigm: Accurate observations, based on planetary nebulae, of the kinematics of the outer parts of three ordinary ellipticals show very little evidence, if any, of the presence of DM [48], but are in good agreement with MOND, because the large masses implied by the high-surface brightnesses indicate that the gravitational field is in the Newtonian regime $a > a_0$ [49,50].

The kinematics of the outskirts of ellipticals, where $a < a_0$, can be probed with spectroscopic observations of their globular cluster systems: For example, the galaxy NGC 4636 in the Virgo cluster is surrounded by 460 globular clusters with measured velocity, and this sample represents one of the largest currently available. The MONDian predictions agree with the kinematic data and NGC 4636 also appears to fall onto the baryonic Tully–Fisher relation shown in Fig. 2 [51].

The X-ray data available for NGC 1521 and NGC 720 offer a unique test of MOND in elliptical galaxies: These galaxies are not embedded in groups or clusters and the X-ray data extend to large radii. Similar to disk galaxies, one can thus test MOND on the outskirts of galaxies where the gravitational acceleration due to the luminous matter is smaller than a_0 and the external field effect is negligible. MOND describes the distribution of the baryonic mass in these galaxies with mass-to-light ratios fully consistent with stellar population synthesis models [52].

3.3 Dwarf Spheroidals

At the low-mass end of the galaxy mass function, dSphs also pose a challenge to the standard model (see [53] for a recent review of dSphs in MOND). dSphs have low-surface brightnesses and, according to MOND, based on what is shown in Fig. 1 for disk galaxies, Milgrom predicted that they should be DM-dominated and have mass discrepancies larger than ten when analysed with Newtonian gravity [16]. This prediction was impressively confirmed when the first measures of the stellar velocity dispersion in the central regions of these galaxies were available a decade later [54].

More recently, intense observational programs provided velocity dispersion profiles of the dSphs orbiting the Milky Way [55]. This piece of information enables the estimate of the combination of the mass profiles of the galaxies with the profiles of their velocity anisotropy parameter. The most recent detailed dynamical analysis [56] confirms that the mass-to-light ratios in the V band are in the range $1 \div 3\ M_\odot/L_\odot$, and are therefore consistent with stellar population synthesis models. This analysis solved an open issue raised earlier [57]: Unbound stars can contaminate the velocity dispersion and artificially inflate the estimate of the mass-to-light ratio. When the unbound stars are properly removed with the caustic technique [58, 59], Sculptor and Sextans do indeed show the sensible mass-to-light ratios

1.8 $M_\odot/L_\odot$ and 2.7 $M_\odot/L_\odot$ respectively, whereas Carina still shows a too large $\sim 6\ M_\odot/L_\odot$.

This discrepancy might originate from (1) the uncertainties of the luminosity distribution that is challenging to estimate accurately enough in LSB galaxies and (2) the ellipticity of Carina that significantly departs from the spherical symmetry assumed to derive the mass-to-light ratio. However, this discrepant mass-to-light ratio might have a deeper origin due to the specific feature of MOND that we mentioned in Sect. 2, the external gravitational field effect: A star in a dSph, that is a satellite of the Milky Way, moves according to both the dSph mass and the gravitational field exerted by the Milky Way; only if this latter external acceleration is negligible compared to the acceleration internal to the dSph is the dSph mass derived from the stellar velocity dispersion accurate. Carina is one of the least luminous, and presumably least massive, dSphs and one of the closest to the Milky Way. It is therefore reasonable to suspect that the Milky Way's gravitational field can play a role in inflating the velocity dispersion of this dSph. Although this suggestion still awaits a quantitative confirmation, this same effect appears to be responsible for the deviation of some dSphs from the expected Tully–Fisher relation shown in Fig. 2 [60].

An additional piece of evidence, which is problematic for the standard model, is the phase-space distribution of the dSphs that are satellites of the Milky Way. These dSphs are distributed over an extended thin disk whose thickness is between 10 and 30 kpc and radius $\sim$ 200 kpc. An invoked solution is that these dSphs fell into the Milky Way halo as a small group of galaxies who kept their orbits correlated [61]. However, recent measures of the dSph proper motions indicate that this scenario is untenable because, according to these measures, four dSphs must have fallen in at least 5 Gyr ago [62] and N-body simulations of the standard model show that the orbit correlation cannot be preserved for such a long time [63].

In MOND, dSphs may form as tidal debris during close encounters of large galaxies [64], and the orbit correlation would thus be a natural consequence of this formation process [65,66]. The formation of tidal dwarf galaxies has been observed in interacting galaxies for the last 20 years [67], since the first detection in the Antennae [68]. The observed stellar velocities of these systems agree with MOND in the majority of the systems using sensible values of the two available free but constrained parameters, the mass-to-light ratio and the velocity anisotropy parameter β [69,70]. In the standard model, these systems are somewhat challenging because, in this case, the observed velocities require a factor of a few more mass than the observed luminous mass, despite the fact that there is no physical reason for these LSB tidal dwarfs to drag large amounts of DM [71].

The formation of dwarf galaxies as tidal debris is also likely to solve the missing satellite problem of the standard model that predicts a factor of ten more satellites in the halos of large galaxies than is actually observed [72]. In MOND, the rate of galaxy encounters is small enough that it might provide the right number of satellites with the correct dynamics. Contrarily, in the standard model, a conspiracy of processes regulating the star formation efficiency is required so that most DM satellites form no stars [73]. In addition, the possible supersonic relative velocity

between baryons and DM before reionization might be responsible for the inhibition of the formation of half the expected luminous satellites [74]. If we take into account the ultra-faint galaxies surrounding the Milky Way, the problem can also be partly alleviated, but there would still be a factor of four too few dwarfs [75]. In addition, these ultra-faint galaxies show indications of tidal disruption, although one would expect that, with mass-to-light ratios of the order of 1000 $M_\odot/L_\odot$, the large DM halo within which they are embedded should be sufficient to screen the stellar components from the external tidal field of the Milky Way. The properties of dwarf galaxies thus remain challenging for the DM paradigm [76–78].

3.4 Globular Clusters

The good agreement between the expected dSph dynamics in MOND and observations is more intriguing when we consider globular clusters. These stellar systems and dSphs roughly have the same baryonic mass but different surface brightnesses, stellar populations and ages. In the standard model, globular clusters are devoid of DM, whereas dSphs are the cosmic structure with the largest fraction of DM, with mass-to-light ratios in the range $\sim 10 \div 1000\ M_\odot/L_\odot$ [55, 79]. These discrepancies between globular clusters and dSphs are solved by invoking low formation efficiencies in low-mass DM halos [80, 81] and two completely different formation processes for the two kinds of systems. Contrarily, in MOND, the different observed internal velocities are exactly what we expect from the different surface brightnesses (Fig. 1) [82].

Nevertheless, there is a system that requires careful consideration. NGC 2419 is a globular cluster in the outer halo of the Milky Way with low enough surface brightness to be in the MOND regime. It is far enough from the gravitational field of the galaxy that the MOND external field effect might not play a relevant role. Similarly to the analyses of the dSphs, it is possible to estimate the mass-to-light ratio of the cluster from measurements of the projected velocity dispersion of the stars [83]. In MOND, the mass-to-light ratio required to describe the stellar kinematics is a factor of a few lower than expected (the opposite case as for the Carina dSph). Invoking a variable velocity anisotropy does not help improve the comparison between the MOND fit and the data, but non-isothermal polytropic models seem to provide a MONDian description of the kinematic and photometric observations of this cluster [84, 85]. An additional solution is that the lifetime of the globular cluster is long enough that mass segregation has already taken place. In this case, the distribution of Red-Giant-Branch (RGB) and luminous upper main sequence stars used in the Jeans analysis is expected to be more centrally concentrated than the other less massive stars of the clusters; therefore, the true gravitational potential of the cluster might be different from the gravitational potential actually derived with the Jeans equation. Another possibility is that globulars have nonstandard initial stellar mass functions. More data and more testable globular clusters are clearly called for.

Another system that has seemed to be challenging for MOND is the five globular clusters orbiting the Fornax dSph. Fornax is the most luminous dwarf spheroidal (by about a factor of ten) and is the only classical dwarf of the Milky Way with a system of globulars. It was suggested that the surprisingly stronger dynamical friction due to the dwarf galaxy's low-density component of stars in MOND, relative to the dynamical friction from the more dense DM halo in Newtonian gravity [86], would cause the globular clusters to lose their orbital angular momentum in a period of time much shorter than the Hubble time and create a stellar nucleus in Fornax which is not observed [87, 88]. However, Angus and Diaferio [89] used an orbit integrator with accurate mass models of Fornax both in MOND and Newtonian dynamics to convincingly demonstrate that the situation is relatively easy to explain: The globular clusters can orbit for a Hubble time as long as their orbits start near the tidal radius. This solution does not apply to the most massive globular cluster; however, this globular cluster is a statistics of one and could have a sizeable line of sight distance where the dynamical friction is negligible.

3.5 Groups and Clusters of Galaxies

The impressive agreement between observations and MOND predictions on the scale of galaxies, based on the introduction of the acceleration scale a_0, is not shared by the galaxy cluster data [90,91]. Clusters do not perfectly obey the relation between circular velocity and baryonic mass of Eq. (3) (green triangles in Fig. 2) but seem to require more mass than actually observed. In fact, in the core of clusters $a > a_0$ and the luminous matter should be enough to describe the observed dynamics. The evidence suggests the opposite is true, because the amount of observed mass is a factor of two too small. The shortage of mass, which is confined to the central regions of clusters and progressively disappears in the cluster outskirts [92], is not as large as the factor of five or larger we have in the standard model, but it clearly poses a challenge to MOND. An equivalent, more conventional way to look at Fig. 2 for clusters is to consider the relation between the cluster mass and the X-ray temperature of the hot intracluster gas. According to the relation between mass and circular velocity, MOND makes a clear prediction for the mass–temperature relation. By assuming the same argument used for the Faber–Jackson relation in elliptical galaxies, the relation between the velocity dispersion σ of the galaxies in a cluster and the baryonic mass M_b of the cluster is $\sigma^4 = GM_b a_0/\alpha^2$, where α is the logarithmic slope of the mass density profile. The temperature T of the intracluster medium is a measure of the kinetic energy of the galaxies and $T \propto \sigma^2$. Therefore, in MOND, we have $M_b \propto T^2$. Figure 4 shows that the data agree with this scaling relation but not with the normalization, because of the mass discrepancy in the cluster cores mentioned above.

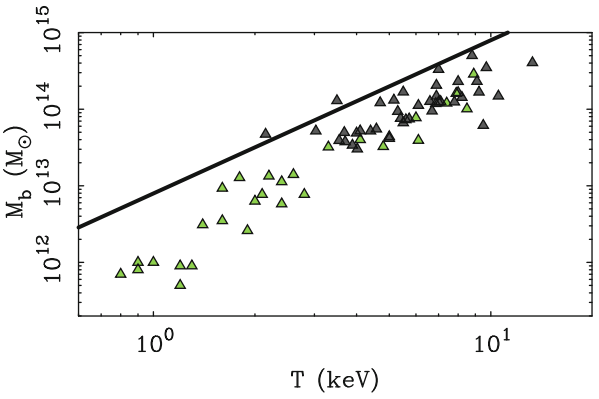


Fig. 4 The baryonic mass–X-ray temperature relation for rich clusters *(gray triangles [93])* and groups of galaxies *(green triangles [92])*. The *solid line* is the MOND prediction $M_b \propto T^2$. (Courtesy of B. Famaey and S. McGaugh. Reproduced from [15], with permission)

This relation is not at odds with the standard model relation $M \propto T^{3/2}$,[3] which is known to agree with observations [94], for two reasons. First, the $M \propto T^{3/2}$ relation is between the X-ray temperature and the total cluster mass, which includes DM, rather than the baryonic mass alone that we have in the MOND $M_b \propto T^2$ relation. Second, the mass M is inferred by assuming valid Newtonian gravity whereas M_b shown in Fig. 4 is derived with MOND [92]. Nevertheless, we can neglect the latter reason and still see that the two scaling relations are roughly consistent with each other. In fact, in the standard model, we can write $M = M_b/f_b \propto T^{3/2}$, where f_b is the baryon fraction of the cluster total mass. Therefore, the observed MONDian $M_b \propto T^2$ should imply $f_b \propto T^{1/2}$. This expectation is broadly consistent with observations: By combining X-ray Chandra groups and HIFLUGCS [95] over the temperature range [0.6, 15] keV, Eckmiller et al. [96] find $f_b \sim T^{0.79\pm0.09}$ at r_{500}; by limiting the sample to clusters with T in the range [1, 15] keV, the slope is 0.83 ± 0.42, but it appears to be shallower at smaller radii.

The apparent mass discrepancy in X-ray bright groups and clusters in MOND [92, 98], clearly shown by the difference between the observed and predicted normalizations of the mass–temperature relation shown in Fig. 4, does not seem to appear in groups of galaxies where the X-ray emission is negligible or absent

[3] In the standard model, the cluster virial mass scales as $M_{\rm vir} \propto \rho_c(z)\Delta_c(z)R^3$, where R is the cluster size in proper units (not comoving), $\rho_c(z) = 3H^2(z)/(8\pi G)$ is the critical density of the universe and $\Delta_c(z)$ is the cluster density in units of $\rho_c(z)$. A widely used approximation is

$$\Delta_c(z) = 18\pi^2 + \begin{cases} 60w - 32w^2, & \Omega_m \le 1, \quad \Omega_\Lambda = 0 \\ 82w - 39w^2, & \Omega_m + \Omega_\Lambda = 1, \end{cases} \tag{4}$$

where $w = \Omega_m(z) - 1$ [97]. Now, $\rho_c(z)$ scales with redshift z as $\rho_c(z) \propto E^2(z) = \Omega_m(1+z)^3 + (1 - \Omega_m - \Omega_\Lambda)(1+z)^2 + \Omega_\Lambda$. The cluster size thus scales as $R \propto M_{\rm vir}^{1/3}\Delta_c^{-1/3}(z)E^{-2/3}(z)$, and the temperature as $T \propto M_{\rm vir}/R \propto M_{\rm vir}^{2/3}\Delta_c^{1/3}(z)E^{2/3}(z)$.

[99,100]. This piece of evidence might suggest that the mass discrepancy in MOND can be solved by assuming the presence of undetected baryonic mass in the form of cold, dense gas clouds that can form and be stable because of the presence of the ionised hot gas [101].

A different solution is that we might well have some DM that clusters on cluster scales but not on galaxy scale. The piece of evidence that is usually claimed to be difficult to interpret without assuming the existence of some form of DM comes from colliding galaxy clusters. During the collision of two clusters, the galaxies and the two halos of DM, if DM exists and is collisionless, keep their momentum, whereas the intracluster medium is shock-heated and slows down. The DM and galaxy components thus separate from the gas component. This separation is clearly visible by combining X-ray and optical images of individual colliding clusters: for example, the so-called Bullet cluster 1E 0657-558 [102] and the cluster MACS J0024.4-1222 [103].

In the absence of DM, most of the matter resides in the gas rather than in the galaxies. The only way to measure how the total mass is distributed is to derive a map of the gravitational potential with weak gravitational lensing. In GR, this procedure is now standard and in the two systems mentioned above, the mass appears to be concentrated where the galaxies are and not where the gas is: This is in striking agreement with the presence of some form of collisionless DM dominating the mass content of the cluster.

Unfortunately, a similar gravitational lensing analysis is not trivial in alternative theories of gravity where the effect of gravity on the light path is not treated properly. MOND belongs to this group of theories because it is a classical theory and gravitational lensing can only be described by resorting to one of the different covariant extensions of MOND. We will describe the results of this approach in Sect. 4. Here, we wish to emphasize that the location of the gravitational peaks is not an observational fact, but derives from the assumption of GR's validity. A priori, it is possible that the nonlinearity of a gravity theory alternative to GR is sufficient to mimic the observed gravitational lensing distortion with a mass distribution different from that required by GR. For example, in the covariant MOND theory named TeVeS, particular matter distributions can yield nonzero convergence on sky positions where there is no projected mass [104]. Weyl gravity provides another clear non-MONDian example of this phenomenology [105, 106].

There is a final cautionary comment that cannot be omitted: The interpretation of the observed properties of colliding clusters is far from being clear, because these systems can also be challenging for the standard model. The Abell cluster A 520 has the same morphology of the Bullet cluster and MACS J0024.4-1222: Two clouds of galaxies on the opposite sides of a cloud of hot gas [107, 108]. However, the GR weak lensing analysis indicates that, in addition to the two gravitational potential peaks at the location of the galaxies, there is a significant peak where the gas is. This peak is difficult to explain in the standard model, because it requires the existence of a massive DM halo devoid of galaxies, but is associated with the cluster hot gas. A number of possible solutions have been suggested, but none of them appears to be fully convincing [108].

In addition, colliding clusters appear to have relative velocities that are unlikely in the standard model: A system like the Bullet cluster, with a relative velocity derived from the shockwave of $\sim$ 4700 km s^{-1} [102], requires an initial infall velocity of the two clusters of $\sim$ 3000 km s^{-1} [109] that has a probability smaller than 3.6×10^{-9} to occur in ΛCDM [110]. On the contrary, the enhanced intensity of MOND gravity may naturally produce these large relative velocities [111, 112]. Similarly, the coherent motion of galaxy clusters on large scales, measured with a technique based on the kinematic Sunyaev–Zeldovich effect produced by Compton scattering of the CMB photons, appears to be challenging for the standard gravitational instability paradigm: These bulk flows are a factor of five larger than predicted by the standard model and might require a modification of the theory of gravity, among other possible solutions [113].

4 TeVeS and Gravitational Lensing

MOND, as described in Sect. 2, is a classical empirical law that we would like to recover as the weak-field limit of a covariant theory. This theory should contain GR, which excellently describes the gravitational phenomenology of the solar system. A covariant theory for MOND is required if we wish to build a cosmological model and describe the gravitational lensing phenomenology within the MONDian framework. A covariant theory is thus essential for the validation of MOND.

The attempts to build a covariant theory containing MOND are numerous. A recent overview of these attempts is provided by Famaey and McGaugh [15]. Here, we briefly outline one of them, namely TeVeS (see [114] for a recent review), whose Lagrangian contains a time-like vector field and a scalar field in addition to the tensor field representing the metric, hence the name Te(nsor)Ve(ctor)S(calar). Halle et al. [115] have shown that a very general Lagrangian for a decaying vector field unifies most of the popular gravity theories, including quintessence, $f(R)$, Einstein–Aether theories and TeVeS itself, among others. We stress that TeVeS by no means is the only possible covariant theory that yields MOND in the weak-field limit. Moreover, it is clear that any problem or failure belonging to one of these covariant theories are not necessarily problems or failures of MOND.

TeVeS was first proposed by Bekenstein [116]. At the classical level, MOND introduces an acceleration scale. This issue poses an immediate problem for building a generally covariant theory, because the acceleration scale is played by the affine connection $\Gamma^{\kappa}_{\mu\nu}$, that involves the first derivatives of the metric tensor $g_{\mu\nu}$, and $\Gamma^{\kappa}_{\mu\nu}$ is not a tensor. As we mentioned in Sect. 2, one way to bypass this problem is to distinguish between the Einstein metric $g_{\mu\nu}$, which enters the Einstein–Hilbert action, and the geodesic metric $\tilde{g}_{\mu\nu}$, which enters the matter action. In GR, the two metrics coincide. In TeVeS, the two metrics are related by the following equation[4] [117]

$$\tilde{g}_{\mu\nu} = e^{-2\varphi} g_{\mu\nu} - 2 \sinh(2\varphi) U_\mu U_\nu, \tag{5}$$

[4] In this section, we use units where the speed of light $c = 1$.

where φ is a scalar field and U_μ is a normalized vector field with $g^{\mu\nu}U_\mu U_\nu = -1$. Both fields are dynamical. Therefore, the TeVeS action is the sum of three terms: the standard Einstein–Hilbert action, the action term for the scalar field

$$S_\varphi = -\frac{1}{2k^2\ell^2 G}\int \mathcal{F}(k\ell^2 h^{\alpha\beta}\varphi_{,\alpha}\varphi_{,\beta})(-g)^{1/2}\mathrm{d}^4x, \tag{6}$$

where $\mathcal{F}$ is an arbitrary positive function, k is a dimensionless coupling constant, ℓ a constant scale length, and $h^{\alpha\beta} = g^{\alpha\beta} - g^{\alpha\mu}g^{\beta\nu}U_\mu U_\nu$, and the action term for the vector field

$$\begin{aligned} S_U = -\frac{1}{32\pi G}\int [K g^{\alpha\beta}g^{\mu\nu}U_{[\alpha,\mu]}U_{[\beta,\nu]} + \bar{K}(g^{\alpha\beta}U_{\alpha;\beta})^2 \\ - 2\lambda(g^{\mu\nu}U_\mu U_\nu + 1)](-g)^{1/2}\mathrm{d}^4x, \end{aligned} \tag{7}$$

where the square brackets indicate antisymmetrisation, K and $\bar{K}$ are dimensionless coupling costants and λ is a Lagrange multiplier to guarantee the normalization of U_μ.

TeVeS violates the local Lorentz invariance, because at each point in spacetime there is a preferred frame in which the time coordinate aligns with U_μ. The violation of Lorentz invariance derives from the invalidity of the strong equivalence principle anticipated in Sect. 2. Clearly, this violation has to be smaller than current experimental bounds.

In the limit $K, \bar{K}, 1/\ell \to 0$, with $k \sim \ell^{-2/3}$, for quasi-static systems and homogeneous cosmology, TeVeS corresponds to GR. To recover MOND in the nonrelativistic ultra-weak-field limit (Eq. 1), we need to choose the function $\mathcal{F}$. The arbitrariness of the function $\mathcal{F}$ makes TeVeS a family of models, rather than a single model. The function $\mathcal{F}$ proposed by Bekenstein [116] yields the MOND acceleration scale

$$a_0 = \frac{\sqrt{3k}}{4\pi\ell}. \tag{8}$$

As we have just mentioned, we recover GR in the limit $K, \bar{K} \to 0$, whereas to recover Newtonian gravity in the weak-field limit $\ell \to \infty$ suffices. Therefore, in principle, TeVeS and GR might differ in the strong field regime [118].

In the weak-field limit, and quasi-static system, the geodesic metric becomes

$$\tilde{g}_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -(1+2\Phi)\mathrm{d}t^2 + (1+2\Psi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j. \tag{9}$$

This metric is formally identical to GR, where $\Phi = -\Psi = \phi_N$, and ϕ_N is the Newtonian gravitational potential. In this limit, in TeVeS we also have $\Phi = -\Psi$, but $\Phi = \Xi\phi_N + \varphi$, where $\Xi = (1 - K/2)^{-1}e^{-2\varphi_c} \sim 1$, with φ_c the asymptotic boundary value of φ. It appears clear that the scalar field φ plays the role that DM plays in the standard model.

Equation (9) enables us to derive the gravitational lensing equations in TeVeS. In Sect. 3, we have seen that MOND excellently describes the dynamics of galaxies but

requires some additional DM on the scales of clusters. One can thus anticipate that the description of gravitational lensing with TeVeS fits the observations on galaxy scales without the need of any matter in addition to the observed baryonic matter, but it does not do as well on cluster scales.

In fact, the TeVeS gravitational lensing equations applied to the strong lensing regime on galaxy scales reproduce the lensing morphology with the observed baryonic matter alone, both in simple spherically symmetric models of the lens [119, 120] and in models departing from spherical symmetry [121]. These analyses self-consistently include TeVeS cosmology, because of the cosmological distances of the lenses: It is easy to get erroneous results if one uses hybrid theories like MOND combined with GR rather than TeVeS itself [122]. In addition, claimed inadequacies of TeVeS [123] may be easily healed by assuming different, but still perfectly reasonable, mass models for the lens [124]. When analysed properly, to date no lensing system on galaxy scale appears to be problematic, including lensed quasars [124–126]. TeVeS even returns a measure of the Hubble constant consistent with other independent estimates when its time-delay formula is applied to lensed variable quasars [127]. Nevertheless, the debate on the adequacy of TeVeS to describe lensing data on galaxy scales is still open and lively [128].

Lensing of clusters returns the mass discrepancy that one encounters with the analysis of galaxy kinematics or X-ray emission. Because of the complexity of the covariant theories of MOND, the lensing analysis usually assumes spherical symmetry and quasi-stationarity. Relatively relaxed clusters, where these assumptions are substantially valid, show the expected mass discrepancies [129, 130]. DM halos are required, and it was suggested that DM in MOND could be made of neutrinos [93]. However, the neutrino phase-space density must be smaller than the Tremaine–Gunn limit we describe below in Sect. 5.1. The combination of strong and weak gravitational lensing data shows that the DM central density of MONDian clusters is larger than the Tremaine–Gunn limit for neutrino masses smaller than 7 eV [130]; this mass exceeds by more than a factor of three the experimental upper limits on the neutrino mass, and this result suggests that neutrinos are inadequate MONDian DM particles.

The Bullet cluster is usually held as the definitive proof of the existence of collisionless DM [102]. Unfortunately, it is a difficult lens to model in MOND covariant theories because of its large departure from the assumptions mentioned above of spherical symmetry and quasi-stationarity. Figure 5 shows the results of one of the first TeVeS analyses of the Bullet cluster [131]: The convergence map is computed by assuming four spherically symmetric mass distributions at the location of the two galaxy distributions and the two X-ray-emitting gas clouds. The model agrees with observations if the galaxies are embedded in additional halos of collisionless matter. Angus et al. [131] concluded that massive neutrinos with 2 eV mass accounting for two thirds of the total mass of the system is sufficient to reproduce the lensing signal. This amount of additional mass agrees with the results of the dynamical analysis of other clusters in MOND reviewed in Sect. 3.5, but, it became clear later [92, 130], as we mention earlier, that the conclusion about the 2 eV mass neutrinos is erroneous. The nonlinearity of TeVeS does not seem to be conducive to removing the demand

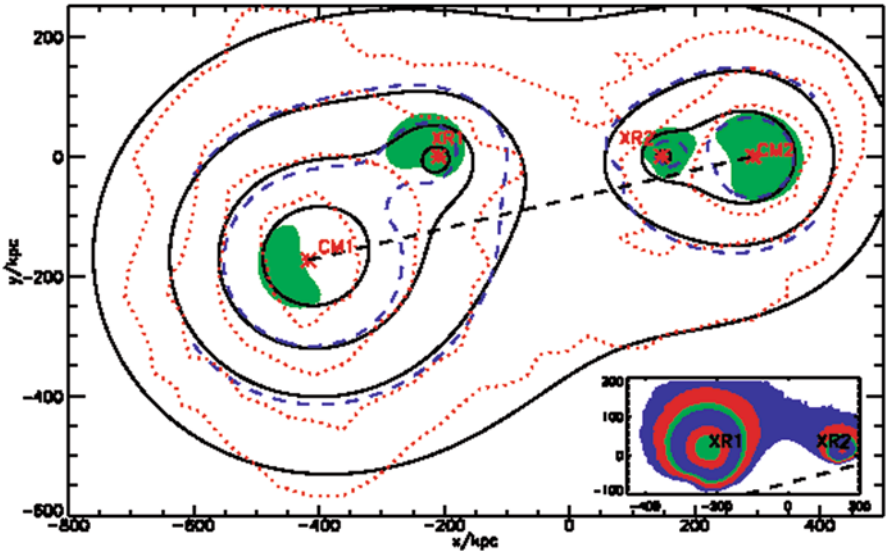


Fig. 5 The *solid black contours* show the MONDian convergence map of the Bullet cluster [131]. The *dotted red contours* show the GR convergence map [102]. The contour levels are [0.37, 0.30, 0.23, 0.16]. The *red stars* indicate the centres of the four potentials used. The *blue dashed lines* show the contours of surface density $[4.8, 7.2]\times 10^2$ $M_\odot$ pc^{-2} for the MOND standard μ function. In the *green shaded* region, the matter density is larger than 1.8×10^{-3} $M_\odot$ pc^{-3} and indicates the clustering of 2 eV neutrinos. *Inset:* The surface density of the gas in the Bullet cluster predicted by a collisionless matter subtraction method for the standard μ-function described in [131]. The contour levels are [30, 50, 80, 100, 200, 300] $M_\odot pc^{-2}$. The origin is $[06^h58^m24^s.38, -55^o56'.32]$. (Reproduced from [131])

for DM in the Bullet cluster [132]. On the other hand, the effect of the external gravitational field, that plays a role in the internal dynamics of dSphs, or the non-trivial features that the additional TeVeS fields can induce in the lensing phenomena remain to be investigated [133].

5 MOND and Sterile Neutrinos

MOND assumes the existence of an acceleration scale below which an ultra-weak-field limit of the theory of gravity sets in. From this very simple *ansatz*, MOND describes with impressive success the dynamics of galaxies and smaller systems. On the scale of galaxy clusters, MOND partly fails (Fig. 2) and the observed kinematics require either a new gravity law or the existence of some baryonic or non-baryonic DM. This failure is bound to be shared by any covariant theory, like TeVeS, that is conceived to yield MOND in the ultra-weak-field limit.

For most cosmologists, this failure implies the end of MOND as a successful description of the universe. However, the DM paradigm currently is far from being

satisfactory on galactic scales [78]. Therefore, if we are interested in finding a theory that describes the universe with the minimum number of assumptions, we have to consider the possibility that MOND can indeed be a valid description of the observed phenomenology and look for possible solutions for its shortcomings on larger scales.

Similarly to the standard model, a natural solution is the introduction of some form of DM. However, we might have an advantage over the standard model. In a MONDian model with DM, DM has to be hot enough to freely stream out of galactic systems, to preserve the excellent description of the galactic dynamics without DM, but cold enough to cluster on the scale of galaxy clusters. Unlike the hypothetical cold DM particles, we know that an elementary particle that can play the role of hot DM does exist: the neutrino.

The properties of neutrinos are currently constrained by various experimental results. In 1979, Tremaine and Gunn [134], by considering the maximum mass density that DM halos made of light leptons can reach, set a lower limit of ~ 1 MeV to the mass of the light leptons that can make the DM halos of galaxies and clusters of galaxies. Below, we briefly review this argument for its relevance to the subsequent discussion.

5.1 The Tremaine–Gunn Limit

Neutrinos are collisionless particles, and, according to Liouville's theorem, the phase-space fluid they form is incompressible. In practice, if neutrinos make the DM in self-gravitating systems, like galaxy clusters, this theorem sets an upper limit to the observable coarse-grained phase-space density.

If self-gravitating systems form by violent relaxation, the neutrino coarse-grained phase-space density is $f(\mathbf{x}, \mathbf{v}) = f(\varepsilon) = f_0\{1 + \exp[\beta(\varepsilon - \chi)]\}^{-1}$, where $\varepsilon = \mathbf{v}^2/2 + \phi(\mathbf{x})$, $\phi(\mathbf{x})$ is the gravitational potential, $\sigma^2(\mathbf{x}) = 1/\beta$ is the one-dimensional velocity dispersion, χ is the neutrino chemical potential, and $f_0 = g_\nu m_\nu^4 h^{-3}$ is the mass phase-space density of an occupied microcell; $g_\nu = 2$ is the number of degrees of freedom, which includes the anti-particles, m_ν is the neutrino mass and h the Planck constant.

For a nondegenerate neutrino fluid, we have $f(\varepsilon) \ll f_0$ which implies $\beta(\phi - \chi) \gg 0$ and the phase-space density must be smaller than the Maxwell–Boltzmann distribution $f_{\rm MB}(\varepsilon) = f_0 \exp[-\mathbf{v}^2/2\sigma^2(\mathbf{x})]$. Therefore, for the neutrino mass density $\rho_\nu(\mathbf{x})$, we must have

$$\rho_\nu(\mathbf{x}) \le 4\pi \int_0^{+\infty} v^2 f_{\rm MB}(\varepsilon) dv = f_0 [2\pi\sigma^2(\mathbf{x})]^{3/2} . \tag{10}$$

This relation implies that

$$\max\{f\} = f_0 \ge \frac{\rho_\nu(\mathbf{x})}{[2\pi\sigma^2(\mathbf{x})]^{3/2}} . \tag{11}$$

However, clusters form from the relic neutrino background that has the Fermi distribution $f(p) = f_0[1+\exp(pc/k_BT)]^{-1}$, with p the momentum and $T = 1.95$ K the neutrino temperature today. The initial maximum coarse-grained phase-space density is therefore $\max\{f\} = f(p = 0) = f_0/2$. According to Liouville's theorem, this upper limit cannot increase, and we must thus have

$$\rho_\nu(\mathbf{x}) \leq \frac{f_0}{2}[2\pi\sigma^2(\mathbf{x})]^{3/2}. \tag{12}$$

For a fully degenerate gas, all the microcells with $v < v_{\rm lim}$ are occupied and $f(\varepsilon) = f_0$, whereas all the microcells with $v > v_{\rm lim}$ are empty and $f(\varepsilon) = 0$. Therefore, in this case, the neutrino mass density is

$$\rho_\nu(\mathbf{x}) \leq f_0 4\pi \int_0^{v_{\rm lim}} v^2 dv = f_0 \frac{4\pi v_{\rm lim}^3}{3}. \tag{13}$$

By considering only bound states with $\varepsilon = \mathbf{v}^2/2 + \phi(\mathbf{x}) \leq 0$, we have $\mathbf{v}^2 \leq 2|\phi| \equiv v_{\rm lim}^2$, and thus $\rho_\nu(\mathbf{x}) \leq f_0 4\pi |2\phi|^{3/2}/3$. By assuming $3\sigma^2 = -\phi$, we obtain $\rho_\nu(\mathbf{x}) \leq f_0[2\pi\sigma^2(\mathbf{x})]^{3/2}4(3/\pi)^{1/2}$, which implies that the maximum phase-space density is

$$\max\{f\} = f_0 \geq \frac{\rho_\nu(\mathbf{x})}{4[2\pi\sigma^2(\mathbf{x})]^{3/2}} \left(\frac{\pi}{3}\right)^{1/2}. \tag{14}$$

However, applying Liouville's theorem again with the initial maximum phase-space density $f_0/2$, we obtain the more stringent upper limit to the neutrino density

$$\rho_\nu(\mathbf{x}) \leq \frac{f_0}{2}[2\pi\sigma^2(\mathbf{x})]^{3/2}4\left(\frac{3}{\pi}\right)^{1/2}, \tag{15}$$

which is $4(3/\pi)^{1/2} \approx 3.91$ times larger than the density upper limit obtained in the nondegenerate case [135, 136]. Therefore, the nondegenerate case yields the most restrictive upper limit and is usually called the Tremaine–Gunn limit.

We can write this limit in astrophysical interesting units as

$$\rho_\nu \leq 2.16 \times 10^2 \left(\frac{m_\nu}{\rm eV}\right)^4 \left(\frac{\sigma}{c}\right)^3 \frac{M_\odot}{\rm pc^3} \tag{16}$$

or, by considering the relation $k_BT/\mu m_p = \sigma^2$ between temperature and velocity dispersion, with m_p the proton mass and $\mu = 0.6$ the mean atomic weight for a fully ionized gas of solar abundance,

$$\rho_\nu \leq 4.64 \times 10^{-7} \left(\frac{m_\nu}{\rm eV}\right)^4 \left(\frac{k_BT}{\rm keV}\right)^{3/2} \frac{M_\odot}{\rm pc^3}. \tag{17}$$

5.2 The Role of an 11-eV Sterile Neutrino

The mass contribution of the three ordinary (active) neutrinos of the standard model, ν_e, ν_μ and ν_τ, could, in principle, solve the mass discrepancy in MONDian galaxy clusters. At the end of the 90s, laboratory experiments yielded an upper bound limit to the mass of ordinary neutrinos of 2.2 eV [137]. Therefore, neutrinos with mass 2 eV were proposed as DM in MONDian models [93]. However, when their contributions to the properties of astrophysical sources, namely, galaxy clusters, CMB and massive galaxies, are analysed in detail, the ordinary neutrinos with mass smaller than this limit are shown to be inadequate.

Sanders [93] and Pointecouteau and Silk [138] pointed out the relevance of the Tremaine–Gunn limit in galaxy clusters in MOND. An extensive analysis of 26 X-ray emitting groups and clusters of galaxies [92] considered neutrinos with the maximum mass allowed by the current upper limits. These neutrinos have a Tremaine–Gunn limit that is at least a factor of two smaller than the DM density within 100 kpc that is required to describe the thermal properties of the intra-cluster medium in MOND. Therefore, either neutrinos are more massive, that is excluded by laboratory measurements, or they are not the major contribution to the MOND DM in clusters.

The three species of ordinary neutrinos with mass in the range $1 \div 2$ eV also are problematic when we attempt to reproduce the CMB power spectrum. These neutrinos suppress the third peak by $\sim$ 25% when compared to observations, because of their free-streaming imposed by the Tremaine–Gunn limit [139]. We can obtain a CMB power spectrum consistent with observations with these neutrinos in TeVeS if we substantially increase the MOND acceleration scale a_0 at the time of recombination [140]. However, this redshift dependence of a_0 remains unproved observationally.

Finally, the amplitude of the weak lensing signal around luminous galaxies ($L > 10^{11} L_\odot$), extracted from the Red-Sequence Cluster Survey and the Sloan Digital Sky Survey, suggests mass-to-light ratios larger than $\sim 10\ M_\odot/L_\odot$ in the MOND framework [141]. It thus follows that DM is also required on the scale of massive galaxies. This DM cannot be made of ordinary neutrinos that are not massive enough to cluster on these small scales.

The inadequacies of ordinary neutrinos with any mass smaller than the current upper limits to properly describe this astrophysical phenomenology forces us to conclude that ordinary neutrinos as the MONDian DM are ruled out.

MOND can still be viable if we resort to hot DM made of sterile neutrinos. Sterile neutrinos do not have standard weak interactions and are right-handed, unlike the three ordinary neutrinos, and are motivated by a number of anomalies observed in neutrino experiments (see [142] for an extensive review). For example, the existence of one sterile neutrino in addition to the three ordinary neutrinos can elegantly explain the disappearance of electron neutrinos from the low-energy beam measured in short-baseline neutrino oscillation experiments [143–146].

If sterile neutrinos are more massive than ordinary neutrinos, they can have larger Tremaine–Gunn limits, and thus eventually solve the problems we mentioned above. The analysis of the MiniBooNE experiment data does indeed favour a sterile neutrino mass in the $4 \div 18$ eV range [147]. This mass range is inconsistent with cosmological constraints in the ΛCDM framework that would require the mass to be smaller than 1 eV (see, e.g., [144] and references therein). However, this mass limit derives from the enhanced gravity on galactic scales that is required by the standard model but not by MOND. Therefore, in MOND, the $4 \div 18$ eV mass range might be perfectly reasonable.

It is actually quite impressive that a universe with baryonic matter and one massive sterile neutrino alone reproduces very well the observed CMB power spectrum if the sterile neutrino mass is $m_{\nu_s} = 11$ eV [148], a value that is fully consistent with the mass range inferred from the MiniBooNE data. In addition, the sterile neutrino mass must be within $\sim 10\%$ of the 11 eV value if we wish to keep the good match with the observed CMB power spectrum. This strong constraint is mostly due to the fact that, in this case, the contribution of the sterile neutrinos to the density of the universe is $\Omega_{\nu_s} h^2 = 0.117$. This value is comparable to the contribution of CDM in the standard model that we know to describe the CMB power spectrum very well. It is important to remind that this analysis assumes that the sterile neutrinos are fully thermalised at the time of decoupling; more massive sterile neutrinos are possible but they must not be fully thermalised [149].

Figure 6 shows that the CMB power spectra of the standard ΛCDM model and of a model with baryonic matter and an 11-eV sterile neutrino are indistinguishable. Note that at the time of recombination, the average gravitational field is strong enough that the universe is not in the MOND regime, with a_0 kept constant to the present day value; therefore, the CMB power spectrum can be estimated with the standard theory of gravity. It is only later that the existence of such a massive neutrino would be problematic for the formation of structure in the standard model.

On the contrary, in MOND this massive sterile neutrino would be adequate to solve the mass discrepancy on the cluster scale. In fact, the study of the hydrostatic equilibrium configuration of 30 groups and clusters of galaxies, including the two clumps forming the Bullet cluster system, by the analysis of the profiles of the galaxy velocity dispersion and hot gas temperature, shows that the mass density profile of DM made of 11-eV neutrinos always reaches the Tremaine–Gunn limit in the cluster centre [150]. Some of the dynamical mass must be provided, as expected, by the central galaxies, but the required amount implies a mass-to-light ratio of $1.2\, M_\odot/L_\odot$ in the K band, in agreement with stellar population synthesis models. An 11-eV sterile neutrino is also consistent with the straight arc, originated by a strong lensing effect, observed in the cluster A 2390 [151].

It is remarkable that the 11-eV sterile neutrino required to match the CMB power spectrum also solves the completely independent problem of the mass discrepancy of MONDian clusters without requiring any additional free parameter or adjustment of the model. It is also intriguing that the Tremaine–Gunn limit is always reached in the cluster centres, implying that the dynamical properties of clusters are uniquely set by the mass of the sterile neutrino. This relation is completely unknown to the

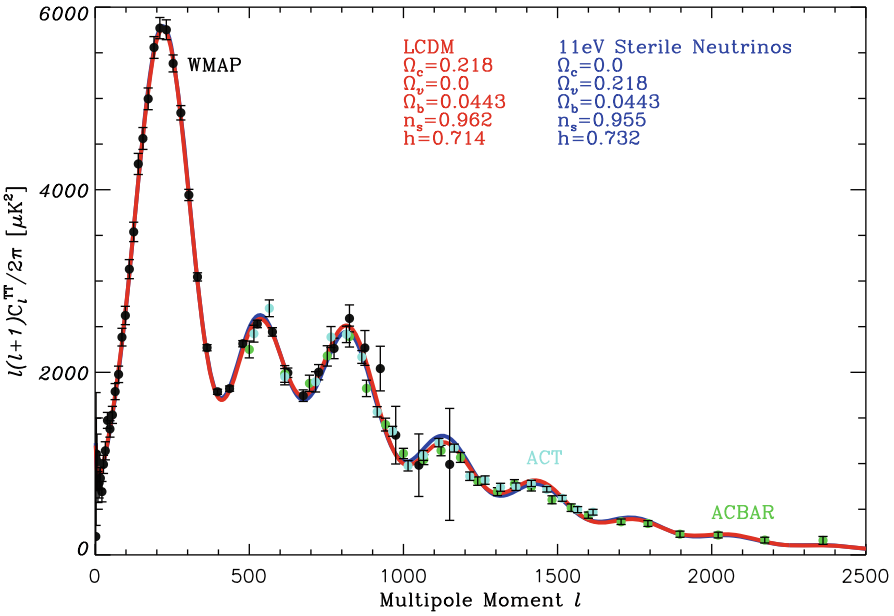


Fig. 6 Cosmic microwave background (CMB) angular power spectra for a cosmological model with baryons and an 11-eV sterile neutrinos *(blue line)*, and for the ΛCDM model *(red line)*. The points show data from the Wilkinson Microwave Anisotropy Probe (WMAP) year 7 *(black)*, Atacama Cosmology Telescope (ACT) *(turquoise)* and Arcminute Cosmology Bolometer Array Receiver (ACBAR) *(green)*. (Reproduced from [112])

standard DM paradigm, where the mass of the cold DM particle does not have any role in the dynamical properties of clusters.

It remains to be seen whether gravitational instability in a universe filled with baryonic matter and one species of sterile neutrino with 11 eV mass can form the observed cosmic structure at the correct pace, although we remind that the ability to explain the cluster mass discrepancy does not directly imply that MOND combined with 11-eV sterile neutrinos can form clusters in a cosmological context.

The investigation of structure formation in this model requires an efficient N-body code and proper cosmological initial conditions. Previous N-body simulations of MOND do not consider the presence of any non-baryonic DM and the adopted initial conditions might not be consistent for a universe filled with baryons alone. These previous N-body simulations show an overproduction of cosmic structure in a baryon-only MONDian universe, in obvious disagreement with observations [152, 153].

A first attempt of N-body simulation of structure formation in a MONDian universe with baryons and an 11-eV sterile neutrino has been performed with a particle-mesh cosmological N-body code that solves the modified Poisson equation of the quasi-linear formulation of MOND (QUMOND) and with initial conditions

appropriate to DM made of 11-eV neutrinos [112]. The simulation evolved a box of $512\,h^{-1}$ Mpc on a side with 256^3 particles from redshift $z = 250$ to the present time.

This MONDian hot DM model does indeed produce galaxy clusters with the correct order of magnitude of the abundance of observed X-ray clusters, unlike hot DM models with standard gravity that have structure formation suppressed on small scales [154]. However, the model overproduces the X-ray luminous clusters by a factor of three and underproduces the low luminosity clusters with $T < 4.5$ keV by at least a factor of ten. Nevertheless, the density profiles of the simulated clusters are compatible with the observed profiles of MONDian clusters. In addition, the frequency of relative velocities larger than 3000 km s^{-1} of cluster pairs is large enough to make likely, unlike the standard model [110], the occurrence of systems like the Bullet cluster.

Overall, the results of this simulation are somewhat unsatisfactory. However, the simulation has very poor mass and length resolutions $\sim 10^{12} M_\odot$ and ~ 2 Mpc, respectively; therefore, clustering on small scales is artificially suppressed and this can partly explain the severe underestimate of the abundance of low-luminous X-ray clusters. The underproduction of low massive clusters could also be quelled by swapping the 11-eV sterile neutrino for a more massive, up to 1-keV, sterile neutrino, because the free-streaming scale decreases with increasing mass of the DM particle. As we mentioned above, however, these higher mass sterile neutrinos would not be fully thermalised prior to decoupling. The overproduction of massive clusters and supercluster size objects actually is a problem that MOND has with any mass of sterile neutrino or any other DM particle.

In addition, the expansion of the universe is assumed to coincide with the standard Friedmann–Robertson–Walker model with a cosmological constant. It is doubtful that a self-consistent relativistic version of MOND will be close enough to the standard model to be consistent with current observational limits on the expansion history of the universe, but discrepant enough to yield the correct structure formation rate. Structure formation is affected by the acceleration scale a_0 which sets the enhancement of the intensity of the gravitational field compared to standard gravity. Currently, we do not have any observational constraints on the redshift dependence of a_0, but if a_0 is lower at higher redshift, the overproduction of cosmic structure could be suppressed and the abundance of massive clusters could become consistent with observations. However, this feature would suppress galaxy formation and is therefore not ideal.

Clearly, the phenomenological formulation of MOND is not conceived to yield a self-consistent cosmological model, and a covariant model including MOND needs to be implemented to test its predictions of structure formation robustly. For example, the linear perturbation theory of the density field in TeVeS has already been outlined [155], but the nonlinear evolution of cosmic structure, with or without a non-baryonic DM component, still needs to be investigated.

6 Conclusion

MOND is based on the *ansatz* that Newtonian dynamics is modified when the gravitational field determined by the distribution of baryonic matter drops below the acceleration $a_0 \sim 10^{-10}$ m s^{-2} [14]. The observational evidence for the existence of this acceleration scale has been rapidly accumulating over the last decade. The success of MOND at describing the dynamics of self-gravitating systems up to the scale of galaxies with fewer parameters than the standard DM paradigm is undeniable: In the standard model, the additional gravitational force that is required by the observed kinematics is supplied by the presence of proper amounts of DM, but this solution requires, unlike MOND, a number of fine-tunings and coincidences to explain the existence of the acceleration scale that emerges at different length scales. MOND provides a far simpler explanation of these observations.

In addition, on the scale of galaxies, MOND has an impressive predictive power that is alien to the DM paradigm. Numerous observations predicted by MOND were confirmed years later. One of the most striking was that, if interpreted in Newtonian dynamics, the kinematics of LSB galaxies imply that DM dominates the dynamics of these systems more than in any other system in the universe [16].

However, on the scale of galaxy clusters, MOND still requires some form of DM, although not as much as in the DM paradigm [92]. Reproducing the power spectrum of the CMB beyond the second peak also requires some non-baryonic DM [139].

In principle, neutrinos might represent a possible candidate for this non-baryonic DM. However, current upper limits on the ordinary neutrino mass show that they would not be massive enough to condense in the core of clusters and preserve the height of the third peak of the CMB power spectrum, as required by observations.

A viable alternative candidate is a sterile neutrino. This particle seems to be required to interpret a number of anomalous results from neutrino experiments [142]. A sterile neutrino with mass in the range 11 eV$\div$1 keV would ensure DM hot enough to stream out of galaxies but cold enough to cluster on the scale of massive galaxies and beyond. It is thus, in principle, capable of explaining the phenomenology of astrophysical systems preserving the success of MOND on small scales and, in principle, the success of the standard model on large scales [148]. Such a model, where we have a modification of gravity and an exotic DM particle, might be more attractive than the current standard ΛCDM model, because of its elegance on the scale of galaxies, the physical motivation of the existence of its DM particle from experiments on Earth and, eventually, the fact that it requires fewer free parameters than ΛCDM. However, it remains to be investigated whether this model can reproduce the full phenomenology of the large-scale structure formation and evolution.

Moreover, this model rests on MOND, which is still a classical theory, not a covariant theory. The parent covariant theory that gives MOND in the proper limit is still unknown. Many possibilities have been proposed, but none of them appears yet to be fully convincing.

The exciting conclusion is that, whatever model is going to be correct, the existence of the acceleration scale, that the data on galaxy scales indicates with astonishing regularity, has to be explained naturally by the correct model, unless a number of surprising coincidences happen to fool us.

Acknowledgements We thank Benoit Famaey and Stacy McGaugh for useful suggestions and for providing us with Figs. 1, 2, 3 and 4. We thank Ana Laura Serra for a careful reading of the manuscript and Luisa Ostorero for enlightening and encouraging discussions. AD gratefully acknowledges partial support form INFN grant PD51 and PRIN-MIUR-2008 grant 2008NR3EBK_003 "Matter-antimatter asymmetry, dark matter and dark energy in the LHC era". GWA is supported by the Claude Leon Foundation and a University Research Committee Fellowship from the University of Cape Town. This research has made use of NASA's Astrophysics Data System.

References

1. Oort JH. Bull Astron Inst Neth. 1932;6:249.
2. Zwicky F. Helv Phys Acta. 1933;6:110.
3. Roberts MS, Rots AH. A&A. 1973;26:483.
4. Ostriker JP, Peebles PJE. ApJ. 1973;186:467.
5. Einasto J, Kaasik A, Saar E. Nature. 1974;250:309.
6. Diaferio A. The evidence for unusual gravity from the large-scale structure of the Universe, Proceedings of the 1st AFI symposium in from the vacuum to the universe (Innsbruck, Austria, October 2007), In: Bass SD, Schallhart F, Tasser B, editors. Innsbruck University Press; 2008. p. 71-85 (arXiv:0802.2532).
7. Bertone G, Hooper D, Silk J. Phys Rep. 2005;405:279.
8. Smoot GF, Bennett CL, Kogut A, et al. ApJL. 1992;396:L1.
9. Kirkman D, Tytler D, Suzuki N, O'Meara JM, Lubin D. ApJS. 2003;149:1.
10. Spergel DN, Bean R, Doré O, et al. ApJS. 2007;170:377.
11. Komatsu E, Smith KM, Dunkley J, et al. ApJS. 2011;192:18.
12. de Bernardis P, Ade PAR, Bock JJ, et al. ApJ. 2002;564:559.
13. Bianchi E, Rovelli C. 2010. arXiv:1002.3966.
14. Milgrom M. ApJ. 1983;270:365.
15. Famaey B, McGaugh S. Living reviews in relativity 15. 2012. http://www.livingreviews.org/lrr-2012-6. Accessed Feb 2015.
16. Milgrom M. ApJ. 1983;270:371.
17. Milgrom M. ApJ. 1983;270:384.
18. Bernal T, Capozziello S, Cristofano G, de Laurentis M. Mod Phys Lett A. 2011;26:2677.
19. McGaugh SS. ApJ. 2004;609:652.
20. Bekenstein J, Milgrom M. ApJ. 1984;286:7.
21. Milgrom M. MNRAS. 2010;403:886.
22. Tully RB, Fisher JR. A&A. 1977;54:661.
23. McGaugh SS. ApJ. 2005;632:859.
24. Reiprich TH. Ph.D. Thesis, Ludwig-Maximilians-Universität. 2001.
25. McGaugh SS. IAU symposium. 2008;244:136.
26. Steinmetz M, Navarro JF. ApJ. 1999;513:555.
27. Desmond H. 2012. arXiv:1204.1497.
28. Dutton AA. 2012. arXiv:1206.1855.
29. Wyder TK, Martin DC, Barlow TA, et al. ApJ. 2009;696:1834.
30. van der Kruit PC, Freeman KC. ARA&A. 2011;49:301.
31. McGaugh SS, de Blok WJG. ApJ. 1998;499:66.

32. Leroy AK, Walter F, Brinks E, et al. AJ. 2008;136:2782.
33. Bigiel F, Leroy A, Walter F, et al. AJ. 2010;140:1194.
34. Gentile G, Famaey B, Zhao H, Salucci P. Nature. 2009;461:627.
35. Bell EF, McIntosh DH, Katz N, Weinberg MD. ApJS. 2003;149:289.
36. Fraternali F, Sancisi R, Kamphuis P. A&A. 2011;531:A64.
37. Angus GW, van der Heyden K, Diaferio A. A&A. 2012. in press.
38. McGaugh SS. ApJ. 2008;683:137.
39. Famaey B, Binney J. MNRAS. 2005;363:603.
40. Famaey B, Bruneton J-P, Zhao H. MNRAS. 2007;377:L79.
41. Bienaymé O, Famaey B, Wu X, Zhao HS, Aubert D. A&A. 2009;500:801.
42. Faber SM, Jackson RE. ApJ. 1976;204:668.
43. Milgrom M. ApJ. 1984;287:571.
44. Sanders RH. MNRAS. 2000;313:767.
45. Cardone VF, Angus G, Diaferio A, Tortora C, Molinaro R. MNRAS. 2011;412:2617.
46. O'Sullivan E, Forbes DA, Ponman TJ. MNRAS. 2001;328:461.
47. Humphrey PJ, Buote DA, O'Sullivan E, Ponman TJ. 2012. arXiv:1204.3095.
48. Romanowsky AJ, et al. Science. 2003;301:1696.
49. Milgrom M, Sanders RH. ApJL. 2003;599:L25.
50. Tiret O, Combes F, Angus GW, Famaey B, Zhao HS. A&A. 2007;476:L1.
51. Schuberth Y, Richtler T, Hilker M, Salinas R, Dirsch B, Larsen SS. arXiv:1205.2093.
52. Milgrom M. 2012. arXiv:1205.1308.
53. Kosowsky A. Adv Astron. 2010;2010:4.
54. Milgrom M. ApJ. 1995;455:439.
55. Walker MG, Mateo M, Olszewski EW, et al. ApJL. 2007;667:L53.
56. Serra AL, Angus GW, Diaferio A. A&A. 2010;524:A16.
57. Angus GW. MNRAS. 2008;387:1481.
58. Diaferio A. MNRAS. 1999;309:610.
59. Serra AL, Diaferio A, Murante G, Borgani S. MNRAS. 2011;412:800.
60. McGaugh SS, Wolf J. ApJ. 2010;722:248.
61. Li Y-S, Helmi A. MNRAS. 2008;385:1365.
62. Angus GW, Diaferio A, Kroupa P. MNRAS. 2011;416:1401.
63. Klimentowski J, Łokas EL, Knebe A, et al. MNRAS. 2010;402:1899.
64. Kroupa P. Famaey B, de Boer KS, et al. A&A. 2010;523:A32.
65. Pawlowski MS, Kroupa P, de Boer KS. A&A. 2011;532:A118.
66. Pawlowski MS, Kroupa P, Angus G, et al. 2012. arXiv:1204.6039.
67. Duc P-A. 2012. arXiv:1205.2297.
68. Mirabel IF, Dottori H, Lutz D. A&A. 1992;256:L19.
69. Milgrom M. ApJL. 2007;667:L45.
70. Gentile G, Famaey B, Combes F, et al. A&A. 2007;472:L25.
71. Bournaud F. Adv Astron. 2010;1. arXiv:0907.3831.
72. Moore B, Ghigna S, Governato F, et al. ApJL. 1999;524:L19.
73. Guo Q, White S, Boylan-Kolchin M, et al. MNRAS. 2011;413:101.
74. Bovy J, Dvorkin C. 2012. arXiv:1205.2083.
75. Simon JD, Geha M. ApJ. 2007;670:313.
76. Boylan-Kolchin M, Bullock JS, Kaplinghat M. MNRAS. 2012;422:1203.
77. Hensler G. 2012. arXiv:1205.1243.
78. Kroupa P. 2012. arXiv:1204.2546.
79. Walker MG. 2012. arXiv:1205.0311.
80. Kravtsov AV, Gnedin OY. ApJ. 2005;623:650.
81. Kravtsov A. Adv Astron. 2010;2010:8.
82. Gentile G, Famaey B, Angus G, Kroupa P. A&A. 2010;509:A97.
83. Ibata R, Sollima A, Nipoti C, et al. ApJ. 2011;738:186.
84. Sanders RH. MNRAS. 2012;419:L6.
85. Sanders RH. MNRAS. 2012;422:L21.

86. Ciotti L, Binney J. MNRAS. 2004;351:285.
87. Sánchez-Salcedo FJ, Reyes-Iturbide J, Hernandez X. MNRAS. 2006;370:1829.
88. Cole DR, Dehnen W, Read JI, Wilkinson MI. 2012. arXiv:1205.6327.
89. Angus GW, Diaferio A. MNRAS. 2009;396:887.
90. The LS, White SDM. AJ. 1988;95:1642.
91. Gerbal D, Durret F, Lachieze-Rey M, Lima-Neto G. A&A. 1992;262:395.
92. Angus GW, Famaey B, Buote DA. MNRAS. 2008;387:1470.
93. Sanders RH. MNRAS. 2003;342:901.
94. Vikhlinin A, Kravtsov A, Forman W, et al. ApJ. 2006;640:691.
95. Zhang Y-Y, Andernach H, Caretta CA, et al. A&A. 2011;526:A105.
96. Eckmiller HJ, Hudson DS, Reiprich TH. A&A. 2011;535:A105.
97. Bryan GL, Norman ML. ApJ. 1998;495:80.
98. Aguirre A, Schaye J, Quataert E. ApJ. 2001;561:550.
99. Milgrom M. ApJL. 1998;496:L89.
100. Milgrom M. ApJL. 2002;577:L75.
101. Milgrom M. New Astron Rev. 2008;51:906.
102. Clowe D, Bradač M, Gonzalez AH, et al. ApJL. 2006;648:L109.
103. Bradač M, Allen SW, Treu T, et al. ApJ. 2008;687:959.
104. Angus GW, Famaey B, Zhao HS. MNRAS. 2006;371:138.
105. Pireaux S. Class Quantum Gravity. 2004;21:1897.
106. Pireaux S. Class quantum gravity. 2004;21:4317.
107. Mahdavi A, Hoekstra H, Babul A, Balam DD, Capak PL. ApJ. 2007;668:806.
108. Jee MJ, Mahdavi A, Hoekstra H, et al. ApJ. 2012;747:96.
109. Mastropietro C, Burkert A. MNRAS. 2008;389:967.
110. Lee J, Komatsu E. ApJ. 2010;718:60.
111. Angus GW, McGaugh SS. MNRAS. 2008;383:417.
112. Angus GW, Diaferio A. MNRAS. 2011;417:941.
113. Kashlinsky A, Atrio-Barandela F, Ebeling H. 2012. arXiv:1202.0717.
114. Bekenstein JD. 2012. arXiv:1201.2759.
115. Halle A, Zhao H, Li B. ApJS. 2008;177:1.
116. Bekenstein JD. Phys Rev D. 2004;70:083509.
117. Sanders RH. ApJ. 1997;480:492.
118. Lasky PD, Sotani H, Giannios D. Phys Rev D. 2008;78:104019.
119. Chiu M-C, Ko C-M, Tian Y. ApJ. 2006;636:565.
120. Zhao H, Bacon DJ, Taylor AN, Horne K. MNRAS. 2006;368:171.
121. Shan HY, Feix M, Famaey B, Zhao H. MNRAS. 2008;387:1303.
122. Ferreras I, Sakellariadou M, Yusaf MF. Phys Rev Lett. 2008;100:031302.
123. Mavromatos NE, Sakellariadou M, Yusaf MF. Phys Rev D. 2009;79:081301.
124. Chiu M-C, Ko C-M, Tian Y, Zhao H. Phys Rev D. 2011;83:063523.
125. Chen D-M, Zhao H. ApJL. 2006;650:L9.
126. Chen D-M. JCAP. 2008;1:6.
127. Tian Y, Ko C-M, Chiu M-C. 2012. arXiv:1204.6359.
128. Ferreras I, Mavromatos N, Sakellariadou M, Furqaan Yusaf M. 2012. arXiv:1205.4880.
129. Takahashi R, Chiba T. ApJ. 2007;671:45.
130. Natarajan P, Zhao H. MNRAS. 2008;389:250.
131. Angus GW, Shan HY, Zhao HS, Famaey B. ApJL. 2007;654:L13.
132. Feix M, Fedeli C, Bartelmann M. A&A. 2008;480:313.
133. Ferreira PG, Starkman GD. Science. 2009;326:812.
134. Tremaine S, Gunn JE. Phys Rev Lett. 1979;42:407.
135. Kull A, Treumann RA, Böhringer H. ApJL. 1996;466:L1.
136. Treumann RA, Kull A, Böhringer H. New J Phys. 2000;2:11.
137. Groom DE. Eur Phys J C. 2000;15:358.
138. Pointecouteau E, Silk J. MNRAS. 2005;364:654.
139. McGaugh SS. ApJ. 2004;611:26.

140. Skordis C, Mota DF, Ferreira PG, Bœhm C. Phys Rev Lett. 2006;96:011301.
141. Tian L, Hoekstra H, Zhao H. MNRAS. 2009;393:885.
142. Abazajian KN, Acero MA, Agarwalla SK, et al. 2012. arXiv:1204.5379.
143. Giunti C, Laveder M. Phys Rev D. 2011;83:053006.
144. Giunti C, Laveder M. Phys Rev D. 2011;84:093006.
145. Kuflik E, McDermott SD, Zurek KM. 2012. arXiv:1205.1791.
146. Nelson AE. Phys Rev D. 2011;84:053001.
147. Giunti C, Laveder M. Phys Rev D. 2008;77:093002.
148. Angus GW. MNRAS. 2009;394:527.
149. Boyanovsky D. Phys Rev D. 2008;78:103505.
150. Angus GW, Famaey B, Diaferio A. MNRAS. 2010;402:39.
151. Feix M, Zhao H, Fedeli C, Pestaña JLG, Hoekstra H. Phys Rev D. 2010;82:124003.
152. Nusser A. MNRAS. 2002;331:909.
153. Llinares C, Knebe A, Zhao H. MNRAS. 2008;391:1778.
154. Cen R, Ostriker JP. ApJ. 1992;399:331.
155. Skordis C. Phys Rev D. 2006;74:103513.

Lorentz Breaking Effective Field Theory Models for Matter and Gravity: Theory and Observational Constraints

Stefano Liberati and David Mattingly

Abstract A number of different approaches to quantum gravity are at least partly phenomenologically characterized by their treatment of Lorentz symmetry, in particular whether the symmetry is exact or modified/broken at the smallest scales. For example, string theory generally preserves Lorentz symmetry while analog gravity and Lifshitz models break it at microscopic scales. In models with broken Lorentz symmetry, there are a vast number of constraints on departures from Lorentz invariance that can be established with low-energy experiments by employing the techniques of effective field theory in both the matter and gravitational sectors. We shall review here the low-energy effective field theory approach to Lorentz breaking in these sectors, and present various constraints provided by available observations.

1 Introduction

Our understanding of the observed laws of nature is currently based on two different theories: the Standard Model (SM) of particle physics, and general relativity (GR). However, in spite of their phenomenological successes, SM and GR leave many fundamental theoretical questions unanswered. First of all, part of the success of the SM has been the recognition that symmetry breaking is an important part of modern physics and that what appear to be multiple forces at low energies can often be described in a unified manner. The prime example of this is the Glashow–Salam–Weinberg theory of electroweak interactions. Since such unification is possible, and since many physicists feel that our understanding of the fundamental laws of nature would be deeper and more accomplished if we are able to reduce the number of degrees of freedom in a theory, much effort has been spent trying to construct unified theories in which not only all the subnuclear forces are seen as different aspects of a unique interaction but also gravity is included in a consistent manner as merely part of the overall structure.

S. Liberati (✉)
SISSA and INFN, Via Bonomea 265, Trieste, Italy
e-mail: liberati@sissa.it

D. Mattingly
University of New Hampshire, Durham, NH 03824, USA
e-mail: dyo7@unh.edu

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_11

Another important reason why we seek for a new theory of gravity comes directly from the gravity side. We know that GR fails to be a predictive theory in some regimes. Indeed, many solutions of Einstein's equations are singular in some region, and GR is not able to make any prediction in those regions of space-time. Moreover, there are honest classical solutions of the Einstein's equations that contain closed time-like curves, which would allow traveling back and forth in time with the associated causal paradoxes. Finally, the problem of black hole evaporation considered just within the framework of semiclassical gravity clashes with quantum mechanical unitary evolution.

This long list of puzzles spurred intense research toward a quantum theory of gravity that started almost immediately after Einstein's proposal of GR and which is still one of the most active areas of theoretical physics. The quantum gravity (QG) problem is not only conceptually and technically challenging but has also been an almost metaphysical pursuit for several decades, in that most progress has been on the theoretical side and the experimental aspect has been (for good reasons) neglected. Indeed, we expect QG effects at experimentally/observationally accessible energies to be extremely small, due to suppression by the Planck scale $M_{\rm pl} \equiv \sqrt{\hbar c/G_{\rm N}} \simeq 1.22 \times 10^{19}\ {\rm GeV}/c^2$. In this sense it has been considered (and it is still considered by many) that only ultrahigh-precision (or Planck scale energy) experiments would be able to test QG models.

It was however realized (mainly over the course of the past decade) that the situation is not quite as bleak as it appears. In fact, quantum gravitational models beyond GR have shown that there can be several of what we term low-energy "relic signatures" of quantum gravitational effects which would lead to deviations from the standard theory predictions (SM plus GR) in specific regimes. Some of these new phenomena, which comprise what is often termed "QG phenomenology," include:

- Quantum decoherence and state collapse [1]
- QG imprint on initial cosmological perturbations [2]
- Cosmological variation of couplings [3, 4]
- TeV black holes, related to extra dimensions [5]
- Violation of discrete symmetries [6]
- Violation of space-time symmetries [7]

In this chapter we will focus upon the phenomenology of violations of space-time symmetries, and in particular of local Lorentz invariance (LLI), a pillar both of quantum field theory as well as GR (LLI is a crucial part of the Einstein's equivalence principle on which metric theories of gravity are based).

2 A Brief History of Lorentz Breaking

Contrary to the common perception, explorations of the possible breakdown of LLI have a long standing history. It is however undeniable that the past 20 years have witnessed a striking acceleration in the development both of theoretical ideas as

well as of phenomenological tests previously unimagined. We shall here present an admittedly incomplete review of these developments.

2.1 Early Works

The possibility that Lorentz invariance violation (LV) could play a role in physics dates back at least 60 years [8–13] and in the 1970s and 1980s there was already a well-established literature investigating the possible phenomenological consequences of LV (see, e.g., [14–19]).

The relative scarcity of these studies in the field was due to the general expectation that new effects would only appear in particle interactions where the particle energies were of the order of the Planck scale, as only at those energies would the natural QG suppression by powers of the Planck scale be overcome. However, it was only in the 1990s that it was clearly realized that there are special situations in which new effects, even if highly suppressed, can have observational consequences. These situations were termed "windows on QG."

2.2 The Dawn of QG Phenomenology

At first glance, it appears hopeless to search for effects suppressed by the Planck scale. Even the most energetic particles ever detected (ultrahigh-energy cosmic rays (UHECR), see, e.g., [20, 21]) have $E \lesssim 10^{11}$ GeV $\sim 10^{-8} M_{\rm pl}$. However, tiny corrections can be magnified into an observational effect if the physics in question involves not just the Planck scale and the energy scale of the particle, but also another scale such as the mass of a light particle or a cosmological travel time. In these situations observables can be constructed that leverage the scales against one another to create an actually measurable effect at achievable energies (i.e., anything from the energies in tabletop optical experiments to cosmic rays. See, e.g., [7, 22] for an extensive review).

A partial list of these *windows on QG* includes:

- Sidereal signal variations as a laboratory apparatus such as an optical cavity moves with respect to a preferred frame or direction
- Cumulative effects: long baseline dispersion and vacuum birefringence (e.g., of signals from gamma-ray bursts (GRBs), active galactic nuclei (AGN), and pulsars)
- Anomalous (normally forbidden) threshold reactions allowed by LV terms (e.g., photon decay, vacuum Čerenkov (VC) effect)
- Shifting of existing threshold reactions (e.g., photon annihilation from blazars, ultrahigh-energy (UHE) protons pion production)
- Lorentz violation induced decays not characterized by a threshold (e.g., decay of a particle from one helicity to the other or photon splitting)

- Maximum attainable particle velocities different from c (e.g., synchrotron peak from supernova remnants)
- Dynamical effects of LV background fields (e.g., gravitational coupling and additional wave modes)

It is rare one can assign a definitive "paternity" to a field, and our so-called QG phenomenology is no exception. However, among the papers commonly accepted as seminal we can cite the one by Kostelecký and Samuel [23] that already in 1989 envisaged, within a string field theory framework, the possibility of nonzero vacuum expectation values (VEV) for some Lorentz breaking operators. This work led later on to the development of a systematic extension of the SM (what was later on called "minimal Standard Model extension," mSME) incorporating all possible Lorentz breaking and power counting renormalizable operators (i.e., of mass dimension ≤ 4) by Colladay and Kostelecký [24]. This provided a framework for computing, in effective field theory (EFT), the observable consequences for many experiments and led to much experimental work setting limits on the LV parameters in the Lagrangian (see, e.g., [25] for a review).

Another seminal paper was that of Amelino-Camelia and collaborators [26] which highlighted the possibility to cast observational constraints on Planck-suppressed violations of Lorentz invariance in the photon dispersion relation by examining the propagation of light from remote astrophysical sources like GRBs and AGN. Finally, we also mention the influential papers by Coleman and Glashow [27–29] which brought the subject of systematic tests of Lorentz violation to the attention of the broader community of particle physicists.

Let us stress that this is necessarily an incomplete account of the literature that investigated departures from special relativity. Several papers appeared in the same period and some of them anticipated many important results, see, for example, [30, 31]; unfortunately at the time of their appearance they were hardly noticed (and seen by many as too "exotic").

In the first decade after 2000 the field reached a concrete maturity and many papers pursued both a systematization of the various frameworks and the available constraints (see ,e.g., [32–34]). In this sense another crucial contribution was the development of an EFT approach also for higher order (mass dimension greater than four), naively non-power counting renormalizable, operators [1]. This was first done for rotationally invariant dimension-five operators in quantum electrodynamics (QED) [38] by Myers and Pospelov which was later on extended to larger sections of the SM by Bolokhov and Pospelov [39] and dimension-six operators by Mattingly [40]. The general set of higher-dimension operators for free photons and fermions has been recently cataloged by Kostelecky and Mewes [41,42].

[1] Anisotropic scaling [35–37] techniques were recently recognized to be the most appropriate way of handling higher-order operators in Lorentz breaking theories and in this case the highest-order operators are indeed crucial in making the theory power counting renormalizable. This is why we shall adopt sometimes the expression "naively non-renormalizable."

Why did all this attention to Lorentz breaking frameworks and observations develop in the late 1990s and in the first decade of the new century? The answer is twofold as it is related to important developments coming from experiments and observation as well as from theoretical investigations. Observationally, there were a number of puzzling observations related to gravity that spawned a corresponding growth in the zoo of QG models/scenarios with a low-energy phenomenology. For example, in cosmology these are the years of the striking realization that our universe is undergoing an accelerated expansion phase [43, 44] which apparently requires a new exotic cosmological fluid, called dark energy, which violates the strong energy condition (to be added to the already well known, and still mysterious, dark matter component).

Also in the same period high-energy astrophysics provided some new observational puzzles directly related to Lorentz symmetry, first with the apparent absence of the Greisen–Zatsepin–Kuzmin (GZK) cutoff [45, 46], a suppression of the high-energy tail of the UHECR spectrum due to UHECR interaction with cosmic microwave background (CMB) photons, as claimed by the Japanese experiment AGASA [47], and later on with the so-called TeV-gamma rays crisis, that is, the apparent detection of a reduced absorption of TeV gamma rays emitted by AGN [48]. Both these "crises" later on subsided or at least alternative, more orthodox explanations for them were advanced. However, their existence undoubtedly boosted the research in the field at that time.

It is perhaps this past "training" that made several collaborations within the QG phenomenology community strongly emphasize the apparent incompatibility of the recent CERN–LNGS based experiment OPERA [49] measurement of superluminal propagation of muonic neutrinos with Lorentz violating EFT (see, e.g., [50–53]). There is now evidence that the OPERA measurement might be flawed due to unaccounted experimental errors and furthermore has been refuted by a similar measurement of the ICARUS collaboration [54]. Nonetheless, this claim propelled a new burst of activity in Lorentz breaking phenomenology which might still provide useful insights for future searches.

Parallel to these exciting developments on the experimental/observational side, theoretical investigations provided new motivations for Lorentz breaking searches and constraints. Indeed, specific hints of LV arose from various approaches to QG. Among the many examples are the abovementioned string theory tensor VEVs [23] and space-time foam models [26, 55–58], then semiclassical spin-network calculations in Loop QG [59], noncommutative geometry [60–62], some brane-world backgrounds [63].

More recently, it cannot be omitted the role associated with the development of Lorentz breaking theories of gravity from early studies [64–68] to more systematic approaches such as Einstein-aether [69–71] and Hořava–Lifshitz [36] gravity. Finally, there was the vigorous development over the same time of the so-called condensed matter analogues of "emergent gravity" [72], which showed how approximate LLI can arise from a fundamental Galilean theory.

Many of these approaches yield a low energy description of Lorentz violation in terms of EFT, and so before we delve into the specific operators that people have

considered we remark on some generic aspects of embedding phenomenologically acceptable Lorentz violation in a low-energy EFT.

3 Modified Dispersion Relations and Their Naturalness

We will concentrate here on the free field modifications to EFT from LV and hence talk primarily about modifications to dispersion relations. This restriction is not as limiting as it might seem. For example, the necessary LV interaction terms generated when one wants to maintain gauge covariance in an LV theory [24] are controlled by the same LV coefficients that control the free theory, so no new LV coefficients are introduced. Since most experimental work is sensitive to free field behavior, constraints on those coefficients are generated from the phenomenology of the free part of the Lagrangian rather than the interaction part. One can of course add other LV interaction terms by hand—such terms have not been as extensively studied in the literature and so we will not focus on them here. There is one caveat: When one sees constraints in the literature from modified particle decay mechanisms there is usually an assumption that the interactions of the SM hold with only small modifications if any. This is reasonable, as if there is only a tiny modification to the free field equations then by the argument above about how gauge generated LV interaction terms are controlled by the same coefficients as the free field part, the corresponding modifications to the interaction terms would also be small. This would, in general, modify the rate of various reactions but the rate difference, given the constraints detailed below, would be unobservable.

Turning back to the issues of modified dispersion relations in LV EFT, one has to first recognize that calculations of a specific dispersion modification from a specific QG theory are in general problematic and that cases where one can do so [55] are the exception rather than the norm. A more reasonable approach is to, therefore, simply consider a generic momentum expansion of a dispersion relation in a specified observer's frame of

$$E^2 = \mathbf{p}^2 + m^2 + \sum_{N=1}^{\infty} \eta^{(N)}_{\alpha_1 \alpha 2 \ldots \alpha_N} p^{\alpha_1} p^{\alpha_2} \ldots p^{\alpha_N}, \tag{1}$$

where the low energy speed of light $c = 1$ and each $\eta^{(N)}$ is an arbitrary rank N tensor with mass dimension $2 - N$ and the α_i indices run over space-time coordinates.

Let us note that this ansatz assumes that the propagating mass eigenstates are also eigenstates of the Lorentz violating physics. This does not need to be the case, and having the eigenstates of Lorentz violation not match the mass eigenstates can be useful when trying to analyze the effects of Lorentz violation on neutrinos. However, since such a mismatch in other sectors would introduce oscillations and other unseen effects for particles other than neutrinos, that is, those particles we are primarily concerned about, we shall not consider such possibilities outside of the neutrino sector.

The $\eta^{(N)}$'s can be mapped on to coefficients in a corresponding Lagrangian, although we note for the reader that in a generic Lagrangian there are many other coefficients that do not influence free field dispersions (c.f. [41]). Needless to say, as N increases, there are a multitude of possible dispersions for even low N as the number of components of $\eta^{(N)}$ scales as 4^N. Testing all such possible combinations can be done, of course, and would be the most systematic way to evaluate the possibility of Lorentz violation.

One can, however, simplify the possible set of dispersions by making various assumptions such as Charge/parity/time-reversal (CPT) invariance, rotational invariance, etc. Rotational invariance is one of the most common assumptions made, for three reasons [2]. The first is that it dramatically reduces the number of possible dispersion terms while still preserving interesting phenomenology. Second, in many instances rotational invariance is more stringently tested than boost invariance so it provides a good model for testing boost invariance alone. Finally, many QG models out there that do not predict exact Lorentz invariance still preserve rotational invariance. (See [34] for further discussion about this assumption.)

With rotational invariance in mind in some frame, a common assumption for the dispersion relation is then

$$E^2 = p^2 + m^2 + \sum_{N=1}^{\infty} \tilde{\eta}_N p^N \ , \tag{2}$$

where p is the magnitude of the three-momentum. This type of expansion assumes high-energy particles and that the corrections are small, so anywhere E would have appeared on the right-hand side of (2) it is replaced by p. In general one can allow the LV parameters $\tilde{\eta}_N$ to depend on the particle type, and indeed it turns out that they *must* sometimes be different but related in certain ways for photon polarization states, and for particle and antiparticle states, if the framework of EFT is adopted. The lowest-order LV terms (p, p^2, p^3, p^4) have been the terms that have generated the most attention (c.f. [7,73] and references therein)[3].

3.1 The Naturalness Problem

From an EFT point of view the only relevant operators should be the lowest order ones, that is, those of mass @@dimension 3, 4 corresponding to terms of order

[2] A notable exception to this assumption is the SME and associated tests. Rotational invariance is not assumed in the program of the SME as it considers all terms at each mass dimension.

[3] We disregard here the possible appearance of dissipative terms [74] in the dispersion relation, as this would correspond to a theory with unitarity loss and to a more radical departure from standard physics than that envisaged in the framework discussed herein (albeit a priori such dissipative scenarios are logically consistent and even plausible within some quantum/emergent gravity frameworks).

p and p^2 in the dispersion relation. Situations in which higher-order operators contribute to the dispersion as much as the lowest order ones at some energy are only possible at the cost of a severe, indeed arbitrary, fine-tuning of the coefficients $\tilde{\eta}_N$ (which we discuss below). However, as we shall see current observational constraints are incredibly tight on dimension-three operators and very severe on dimension four ones. This is kind of obvious, given that these operators would end up modifying the dispersion relation of elementary particles at easily achievable energies. Dimension 3 operators would dominate as $p \to 0$ while the dimension four ones would generically induce a species dependent, constant shift in the limit speed for elementary particles. Hence, one is left with two less than perfect approaches.

First, one can assume the standard EFT hierarchy, stop testing at operators of mass dimension three and four, and, due to the tightness of limits, argue that Lorentz invariance is likely an exact symmetry of nature and that QG/emergent models that do not respect the symmetry should be discounted. Or, one can avoid assumptions about whether any additional new physics comes into play between everyday energies and the QG scale and so not assume a particular hierarchy between operators of various mass dimension. This is the approach many phenomenologists take: Simply start at the lowest-dimension operators, derive constraints, and work upward in N as far as one can observationally go without imposing any necessary hierarchy.

Not assuming a necessary hierarchy (3) and simply constraining the coefficients $\tilde{\eta}_N$ at each order is perfectly good phenomenologically, and we will take that approach going forward, but it is important that the reader understand why it is theoretically unnatural. The reason such a hierarchy is unnatural is simple: In EFT, radiative corrections will generically allow the percolation of higher-dimension Lorentz violating terms into the lower-dimension terms due to the interactions of particles [75, 76].

In EFT, loop integrals will be naturally cutoff at the EFT breaking scale, if such scale is as well the Lorentz breaking scale the two will effectively cancel leading to unsuppressed, coupling dependent, contributions to the base dimension four kinetic terms that generate the usual propagators. Hence, radiative corrections will not allow a dispersion relation with only p^3 or p^4 Lorentz breaking terms but will automatically induce extra unsuppressed LV terms in p and p^2 which will be naturally dominant. One could argue that renormalization group (RG) effects might naturally suppress the sizes of these coefficients at low energies. As we shall see, in specific models where the RG flow has been calculated, the running of LV coefficients is only logarithmic and so there is no indication that RG flow will actually drive coefficients to zero quickly in the infrared.

Several ideas have been advanced in order to resolve such a "naturalness problem" (see, e.g., [34]). While it would be cumbersome to review all the proposals here, we point out two of the more prominent ideas, both of which involve introducing a new scale into the problem in addition to the Lorentz violating scale. If there are two scales M and μ involved, there can be a hierarchy of LV coefficients different than the naive one, for example,

$$\tilde{\eta}_1 = \eta_1 \frac{\mu^2}{M}, \qquad \tilde{\eta}_2 = \eta_2 \frac{\mu}{M}, \qquad \tilde{\eta}_3 = \eta_3 \frac{1}{M} \tag{3}$$

or

$$\tilde{\eta}_2 = \eta_2 \frac{\mu^2}{M^2}, \qquad \tilde{\eta}_4 = \eta_4 \frac{1}{M^2}, \tag{4}$$

where M is the QG scale, usually taken to be the Planck scale, μ is some other far lower-energy scale, and η_N is a de-dimensionalized coefficient usually assumed to be O(1). The exact hierarchy, whether it involves terms of every mass dimension as in (3), or only even dimension as in (4) (which can be accomplished up to $N = 4$ by imposing CPT) is model dependent. We now briefly describe two ideas that have been put forward that would generate such scales.

3.1.1 A New Symmetry

Most of the aforementioned proposals implicitly assume that the Lorentz breaking scale is the Planck scale. One then needs the EFT scale (which can be naively identified with what we called previously μ) to be different from the Planck scale and actually sufficiently small so that the lowest order "induced" coefficients can be suppressed by suitable small ratios of the kind μ^p/M^q where p, q are some positive powers.

A possible solution in this direction can be provided by introducing what is commonly called a "custodial symmetry" — a symmetry other than Lorentz that forbids lower-dimension operators and is broken at the low-scale μ. The most plausible candidate for this role is supersymmetry (SUSY) [77, 78]. SUSY is by definition a symmetry relating fermions to bosons, that is, matter with interaction carriers. As a matter of fact, SUSY is intimately related to Lorentz invariance. Indeed, it can be shown that the composition of at least two SUSY transformations induces space-time translations. However, SUSY can still be an exact symmetry even in presence of LV and can actually serve as a custodial symmetry preventing certain operators to appear in LV field theories.

The effect of SUSY on LV is to prevent dimension ≤ 4, renormalizable LV operators to be present in the Lagrangian. Moreover, it has been demonstrated [77, 78] that the RG equations for supersymmetric QED plus the addition of rotationally invariant dimension-five operators [38] do not generate lower-dimensional operators if SUSY is unbroken. However, this is not the case for our low-energy world, of which SUSY is definitely not a symmetry.

The effect of soft SUSY breaking was also investigated in [77, 78]. It was found there that, as expected, when SUSY is broken the renormalizable operators are generated. In particular, dimension κ ones arise from the percolation of dimension $\kappa + 2$

LV operators[4]. The effect of SUSY soft breaking is, however, to introduce a suppression of order $m_s^2/M_{\rm Pl}$ ($\kappa = 3$) or $(m_s/M_{\rm Pl})^2$ ($\kappa = 4$), where $m_s \simeq 1$ TeV is the scale of SUSY soft breaking. Although, given present constraints, the theory with $\kappa = 3$ needs a lot of fine-tuning to be viable, since the SUSY-breaking-induced suppression is not enough powerful to kill linear modifications in the dispersion relation of electrons, if $\kappa = 4$ then the induced dimension-four terms are suppressed enough, provided $m_s < 100$ TeV. Current lower bounds from the Large Hadron Collider are at most around 950 GeV for the most simple models of SUSY [79] (the so called constrained minimal supersymmetric standard model, CMSSM).

Finally, it is also interesting to note that the analogue model of gravity can be used as a particular implementation of the abovementioned mechanism for avoiding the so called naturalness problem via a custodial symmetry. This was indeed the case of multi-Bose–Einstein condensate (or multi-BEC) [80, 81].

3.1.2 Gravitational Confinement of Lorentz Violation

The alternative to an extra symmetry is to turn the problem upside down and posit that the Lorentz breaking scale (the M appearing in the above dispersion relations) is not set by the Planck scale, but is instead the lower-scale μ. If one does this and begins with a theory which has higher-order Lorentz violating operators only in the gravitational sector, then one can hope that the gravitational coupling $G_N \sim M_{\rm Pl}^{-2}$ will suppress the "percolation" to the matter sector where the constraints are strongest. Matter Lorentz violating terms will all possess factors of the order $(\mu/M_{\rm Pl})^2$ which can become strong suppression factors if $\mu \ll M_{\rm Pl}$. This is the idea underlying the work presented in [82] which applies it to the special case of Hořava–Lifshitz gravity. There it was shown that indeed a workable low-energy limit of the theory can be derived through this mechanism which apparently is fully compatible with existing constraints on Lorentz breaking operators in the matter sector. In our opinion, this new route deserves further attention and should be more deeply explored in the future.

4 Dynamical Frameworks I: Rotationally Invariant CPT Even Standard Model Extension with Mass Dimension 5 and 6 Operators

We now turn from theoretical considerations about naturalness and retreat to the phenomenological approach of "constrain everything as best as one can do observationally." There are various systematic frameworks and approaches for this. The

[4] We consider here only $\kappa = 3, 4$, for which these relationships have been demonstrated.

SME contains all possible Lorentz violating tensors that can be coupled to SM fields without changing the field content or violating gauge symmetry. The SME can be split into the 'mSME" [23] of Kostelecky et. al., which contains only renormalizable operators, and the full SME, which contains the infinite tower of non-renormalizable higher-mass dimension operators. As one can imagine, there are dozens of possible operators even for the renormalizable case, and the number of operators up to mass dimension six is in the hundreds. Many of the operators can be constrained by similar methods, so for the purposes of this introduction we will concentrate on the most well-studied sector of the SME, that of rotationally invariant QED. In particular, we will concentrate on the interactions of photons, electrons, and protons.

Usually when defining a field theory, one starts with the renormalizable operators and proceeds in increasing mass dimension. Here we start with higher-mass dimension and work backwards for purely pedagogical reasons—many of the reactions and constraints we describe in detail for the higher-dimension operators will also be useful for setting constraints on various lower-dimension operators. The first mass dimension we explore is mass dimension six, as this is the highest-mass dimension where all the operators have been classified and significantly studied. We break the operators into CPT even and odd classes as a different set of observations can be used to constraint CPT odd operators.

4.1 The Model

A list of CPT even, rotationally invariant mass dimension five- and six-LV terms was computed in [40] through the same procedure used by Myers and Pospelov for dimension-five LV (see below), and this has been extended to nonrotationally invariant operators in [41, 42]. With rotation invariance all LV tensors must reduce to suitable products of a time-like vector field, usually denoted by u^α. This is usually taken to be unit, so that in the frame of the observer whose world line is tangent to u^α, u^α has components $(1, 0, 0, 0)$. This allows us to express constraints solely in terms of the numerical coefficient involved in any u^α-matter interaction term. Of course, the actual direction of u^α is technically arbitrary. However, the common choice, which we make here, is to define u^α to be aligned with the rest frame of the cosmic microwave background. In terms of u^α the known mass dimension-six fermion operators are

$$-\frac{i}{M_{\rm Pl}^2}\overline{\psi}(u\cdot D)^3(u\cdot\gamma)(\alpha_L^{(6)}P_L+\alpha_R^{(6)}P_R)\psi \qquad (5)$$
$$-\frac{i}{M_{\rm Pl}^2}\overline{\psi}(u\cdot D)\Box(u\cdot\gamma)(\tilde{\alpha}_L^{(6)}P_L+\tilde{\alpha}_R^{(6)}P_R)\psi\ ,$$

where $P_{R,L}$ are the usual left- and right-spin projectors $P_{R,L} = (1 \pm \gamma^5)/2$ and D is the gauge covariant derivative. All the coefficients α are dimensionless because

we factorize suitable powers of the Planck mass out explicitly. In addition there is a CPT even dimension-five term [40]

$$-\frac{1}{M_{\rm Pl}}\overline{\psi}(u\cdot D)^2(\alpha_L^{(5)}P_L+\alpha_R^{(5)}P_R)\psi. \tag{6}$$

The known photon operator is

$$-\frac{1}{2M_{\rm Pl}^2}\beta_\gamma^{(6)}F^{\mu\nu}u_\mu u^\sigma(u\cdot\partial)F_{\sigma\nu}\ . \tag{7}$$

From these operators, the dispersion relations of fermions and photons can be computed, yielding

$$\begin{aligned} E^2-p^2-m^2 &= \frac{\alpha_R^{(6)}E^3}{M_{\rm Pl}^2}(E+sp)+\frac{\alpha_L^{(6)}E^3}{M_{\rm Pl}^2}(E-sp)+\frac{m}{M_{\rm Pl}}(\alpha_R^{(5)}+\alpha_L^{(5)})p^2 \\ &\quad +\alpha_R^{(5)}\alpha_L^{(5)}\frac{p^4}{M_{\rm Pl}^2} \\ \omega^2-k^2 &= \beta^{(6)}\frac{k^4}{M_{\rm Pl}^2}\ , \end{aligned} \tag{8}$$

where m is the electron mass and $s=\sigma\cdot\mathbf{p}/|\mathbf{p}|$ is the helicity of the fermions. The $\tilde{\alpha}$ terms contribute as $m^2/M_{\rm Pl}^2$, that is, are highly suppressed, and so will be neglected.

In general, a LV dispersion for a particle with a certain set of quantum numbers (mass, spin, etc.) will be of the form $E^2=p^2+f^{(n)}p^n/M_{\rm Pl}^{n-2}$, and so we will often refer to type "n" LV. For example, because the high-energy fermion states are almost exactly chiral, we can further simplify the fermion dispersion relation Eq. (9) (with $R=+, L=-$)

$$E^2=p^2+m^2+\frac{m}{M_{\rm Pl}}\eta^{(2)}p^2+\eta_\pm^{(4)}\frac{p^4}{M_{\rm Pl}^2}, \tag{9}$$

where $\eta^{(n)}$ is the dispersion coefficient of the LV p^n term in the dispersion relation for the fermion. We choose η as the coefficient as this nomenclature is common in the literature. Similarly, $\xi^{(n)}$ will refer to the generic dispersion coefficient for a photon (so in the case above $\xi^{(4)}=\beta^{(6)}$). As it is suppressed by a factor of order $m/M_{\rm Pl}$, we will drop the quadratic modification generated by the dimension-five operator. Indeed this can be safely neglected, provided that $E>\sqrt{mM_{\rm Pl}}$. Let us stress however, that this is exactly an example of a dimension-four LV term with a natural suppression, which for electron is of order $m_e/M_{\rm Pl}\sim 10^{-22}$. Therefore, any limit larger than 10^{-22} placed on this term would not have to be considered as an effective constraint (to date, the best constraint for a rotationally invariant electron LV term of dimension four is $O(10^{-16})$ [83]). Note that modulo this the CPT even dimension-five operator for fermions has the same effect on the dispersion as the CPT even dimension six in that it generates a p^4 term, so we will generally just write

constraints directly on $\eta_{\pm}^{(6)}$. It may seem puzzling that in a CPT invariant theory we distinguish between different fermion helicities in (9). However, although they are CPT invariant, some of the LV terms displayed in Eq. (7) are odd under P and T.

CPT invariance allows us to determine a relationship between the LV coefficients of fermions and antifermions. Indeed, to obtain these we simply realize that, by CPT, the dispersion relation of the antifermion is given by (9), with the replacements $s \to -s$ and $p \to -p$. If $q, \overline{q}$ denote a charge fermion and antifermion, then the relevant antifermion coefficient $\eta_{\overline{q}}^{(6)}$ is such that $\eta_{\overline{q}_{\pm}}^{(6)} = \eta_{q_{\mp}}^{(6)}$, where $\overline{q}_{\pm}$ indicates an antifermion of positive/negative helicity (and similarly for the $q_{\pm}$). Let us anticipate that the same argument used above leads to the conclusions that for dispersion relations with odd powers of "n" (e.g., p^3-type dispersion relations) one obtains $\eta_{\overline{q}_{\pm}}^{(n_{odd})} = -\eta_{q_{\mp}}^{(n_{odd})}$. Hence, for arbitrary "n" one would expect $\eta_{\overline{q}_{\pm}}^{(n)} = (-1)^n \eta_{q_{\mp}}^{(n)}$. This different behavior between even and odd powers "n"-type dispersion relations leads to quite distinct phenomenologies as we shall see later.

4.2 Constraints

4.2.1 Threshold Reactions

Threshold reactions with UHECR provide the only significant constraints on the above operators. A threshold reaction is a reaction that does not occur above a certain energy scale, which we call the "threshold energy." An interesting and useful phenomenology of threshold reactions is introduced by LV in EFT; also, threshold theorems can be rederived [33]. Sticking to the present case of rotational invariance and monotonic dispersion relations (see [84] for a generalization to more complex situations), the kinematics of threshold reactions yield a number of useful phenomenological facts about these reactions [32]:

- Threshold configurations still correspond to head-on incoming particles and parallel outgoing ones.
- The threshold energy of existing threshold reactions can shift, and upper thresholds (i.e., maximal incoming momenta at which the reaction can happen in any configuration) can appear.
- Pair production can occur with unequal outgoing momenta.
- New, normally forbidden reactions can be viable.

LV corrections are surprisingly important in threshold reactions because the LV term (which as a first approximation can be considered as an additional energy-dependent "mass") should be compared not to the momentum of the involved particles, but rather to the (invariant) mass of the heaviest particle in the interaction. Thus, an estimate for the threshold energy is

$$p_{\text{th}} \simeq \left(\frac{m^2 M_{\text{Pl}}^{n-2}}{\eta^{(n)}} \right)^{1/n} , \tag{10}$$

Table 1 Values of $p_{\rm th}$, according to Eq. (10), for different particles involved in the reaction: neutrinos, electrons, and protons. Here we assume $\eta^{(n)} \simeq 1$

	$m_\nu \simeq 0.1$ eV	$m_e \simeq 0.5$ MeV	$m_p \simeq 1$ GeV
$n = 2$	0.1 eV	0.5 MeV	1 GeV
$n = 3$	500 MeV	14 TeV	2 PeV
$n = 4$	33 TeV	74 PeV	3 EeV

where m is the mass of the heaviest particle involved in the reaction. Interesting values for $p_{\rm th}$ are discussed, for example, in [32] and given in Table 1. Reactions involving neutrinos are the best candidate for observation of LV effects, whereas electrons and positrons can provide results for $n = 3$ theories but cannot readily be accelerated by astrophysical objects up to the required energy for $n = 4$. In this case reactions of protons can be very effective, because cosmic rays can have measured energies well above 3 EeV. We now discuss three threshold reactions are of particular use when constraining $n = 4$ LV.

LV-Allowed Threshold Reactions: γ-decay

The decay of a photon into an electron/positron pair is made possible by LV because energy-momentum conservation may now allow otherwise forbidden reactions to occur. Since the decay is a reaction described by the fundamental QED vertex, the rate once above threshold will be quite fast. The threshold for this process is set by the condition [34]

$$k_{th} \approx \left(\frac{m^2 M_{\rm Pl}^{n-2}}{(F(\eta^{(n)}, \xi^{(n)}))^{n-2}} \right)^{1/n}, \tag{11}$$

where $F(\eta^{(n)}, \xi^{(n)})$ is a linear combination of $\eta^{(n)}, \xi^{(n)}$. Notably, the electron–positron pair can be created with slightly different outgoing momenta (asymmetric pair production). Furthermore, the decay rate is extremely fast above threshold [34] and is of the order of $(10\ {\rm ns})^{-1}$ ($n = 3$) or $(10^{-6}\ {\rm ns})^{-1}$ ($n = 4$) if the LV coefficients are of O(1).

LV-Allowed Threshold Reactions: Vacuum Čerenkov

In the presence of LV, the process of VC radiation $q^{\pm} \to q^{\pm}\gamma$, where q is a charged fermion, can occur as this is just a rotated diagram of γ-decay. The threshold energy of the reaction is roughly the same and so is also given by

$$E_{th} \approx \left(\frac{m^2 M_{\rm Pl}^{n-2}}{(F(\eta^{(n)}, \xi^{(n)})^{n-2}} \right)^{1/n}. \tag{12}$$

Just above threshold this process is also an extremely efficient method of energy loss. Note that while γ-decay destroys the incoming photon, the VC effect merely is an energy loss process.

LV-Modified Threshold Reactions: Photon Absorption
A process related to photon decay is photon absorption, $\gamma\gamma \to e^+e^-$. Unlike photon decay, this is allowed in Lorentz invariant QED and it plays a crucial role in making our universe opaque to gamma rays above tens of TeVs.

If one of the photons has energy ω_0, the threshold for the reaction occurs in a head-on collision with the second photon having the momentum (equivalently energy) $k_{\rm LI} = m^2/\omega_0$. For example, if $k_{\rm LI} = 10$ TeV (the typical energy of inverse Compton-generated photons in some AGN) the soft photon threshold ω_0 is approximately 25 meV, corresponding to a wavelength of 50 μ.

In the presence of Lorentz violating dispersion relations the threshold for this process is in general altered, and the process can even be forbidden. Moreover, as firstly noticed by Kluźniak [85], in some cases there is an upper threshold beyond which the process does not occur. Physically, this means that at sufficiently high momentum the photon does not carry enough energy to create a pair and simultaneously conserve energy and momentum. Note also, that an upper threshold can only be found in regions of the parameter space in which the γ-decay is forbidden, because if a single photon is able to create a pair, then a fortiori two interacting photons will do [32].

Let us exploit the abovementioned relation $\eta_{\pm}^{e^-} = (-1)^n \eta_{\mp}^{e^+}$ between the electron–positron coefficients, and assume that on average the initial state is unpolarized. In this case, using the energy-momentum conservation, the kinematics equation governing pair production is the following [34]:

$$\frac{m^2}{k^n y\,(1-y)} = \frac{4\omega_b}{k^{n-1}} + \tilde{\xi} - \tilde{\eta}\left(y^{n-1} + (-1)^n\,(1-y)^{n-1}\right), \tag{13}$$

where $\tilde{\xi} \equiv \xi^{(n)}/M^{n-2}$ and $\tilde{\eta} \equiv \eta^{(n)}/M^{n-2}$ are respectively the photon's and electron's LV coefficients divided by powers of M, $0 < y < 1$ is the fraction of momentum carried by either the electron or the positron with respect to the momentum k of the incoming high-energy photon and ω_b is the energy of the target photon. The analysis is more complicated than simple one-particle initial-state decay or radiative processes. In particular it becomes necessary to sort out whether the thresholds are lower or upper ones and whether they occur with the equal or different pair momenta.

4.2.2 Constraints from GZK Secondaries

One of the most interesting features related to the physics of UHECR is the GZK cutoff [45,46], a suppression of the high-energy tail of the UHECR spectrum arising from interactions with CMB photons, according to $p\gamma \to \Delta^+ \to p\pi^0(n\pi^+)$. This process has a (LI) threshold energy $E_{\rm th} \simeq 5 \times 10^{19}\ (\omega_b/1.3\ {\rm meV})^{-1}$ eV (ω_b is the target photon energy). Experimentally, the presence of a suppression of the UHECR flux was claimed only recently [20, 21]. Although the cutoff could be also due to the finite acceleration power of the UHECR sources, the fact that it occurs at the

expected energy favors the GZK explanation. The results presented in [86] seemed to further strengthen this hypothesis (but see further discussion below).

Rather surprisingly, significant limits on $\xi = \xi^{(6)}$ and $\eta = \eta^{(6)}$ for the proton can be derived by considering UHE photons generated as secondary products of the GZK reaction [87, 88]. This can be used to further improve the constraints on dimension-five LV operators and provide a first robust constraint of QED with dimension-six CPT even LV operators.

These UHE photons originate because the GZK process leads to the production of neutral pions that subsequently decay into photon pairs. These photons are mainly absorbed by pair production onto the CMB and radio background. Thus, the fraction of UHE photons in UHECRs is theoretically predicted to be less than 1 % at 10^{19} eV [89]. Several experiments imposed limits on the presence of photons in the UHECR spectrum. In particular, the photon fraction is less than 2.0, 5.1, 31, and 36 % (95 % C.L) at $E = 10, 20, 40, 100$ EeV, respectively [90, 91].

However, we have just seen that pair production can be strongly affected by LV. In particular, the (lower) threshold energy can be slightly shifted and in general an upper threshold can be introduced [32]. If the upper threshold energy is lower than 10^{20} eV, then UHE photons are no longer attenuated by the CMB and can reach the Earth, constituting a significant fraction of the total UHECR flux and thereby violating experimental limits [87, 88, 92].

Moreover, it has been shown [88] that the γ-decay process can also imply a significant constraint. Indeed, if some UHE photon ($E_\gamma \simeq 10^{19}$ eV) is detected by experiments (and the Pierre Auger Observatory (PAO) will be able to do so in few years [90]), then γ-decay must be forbidden above 10^{19} eV.

In conclusion we show in Fig. 1 the overall picture of the constraints of QED dimension-six LV operators, where the green dotted lines do not correspond to real constraints, but to the ones that will be achieved when PAO will observe, as expected, some UHE photon.

5 Dynamical Frameworks II: Rotationally Invariant CPT Odd Standard Model Extension with Mass Dimension 5 Operators

5.1 The Model

Myers and Pospelov [38] found that there are essentially only three operators of dimension five, quadratic in the fields, that can be added to the QED Lagrangian that give rise to dispersion modifications of type $n = 3$.[5] These extra-terms are the

[5] Actually these criteria allow the addition of other (CPT even) terms, but these would not lead to modified dispersion relations (they can be thought of as extra, Planck suppressed, interaction terms) [39].

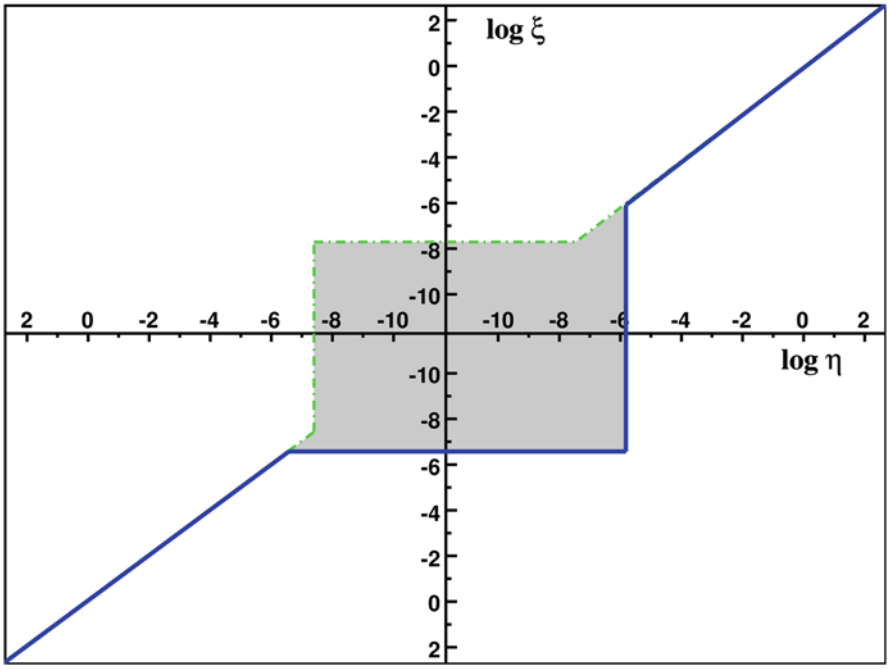


Fig. 1 LV induced by dimension-six operators. The LV parameter space is shown. The allowed regions are shaded *grey*. *Green dotted lines* represent values of (η, ξ) for which the γ-decay threshold $k_{\gamma-dec} \simeq 10^{19}$ eV. *Solid, blue lines* indicate pairs (η, ξ) for which the pair production upper threshold $k_{\rm up} \simeq 10^{20}$ eV. Picture taken from [93]

following:

$$-\frac{\xi}{2M_{\rm Pl}} u^m F_{ma}(u\cdot D)(u_n \tilde{F}^{na}) + \frac{1}{2M_{\rm Pl}} u^m \overline{\psi}\gamma_m(\zeta_1 + \zeta_2\gamma_5)(u\cdot D)^2\psi \ , \qquad (14)$$

where $\tilde{F}$ is the dual of F and ξ, $\zeta_{1,2}$ are dimensionless parameters. All these terms also violate CPT symmetry. More recently, this construction has been extended to the whole SM [39,41,42].

From (14) the dispersion relations of the fields are modified as follows. For the photon one has

$$\omega_\pm^2 = k^2 \pm \frac{\xi^{(3)}}{M_{\rm Pl}} k^3 \ , \qquad (15)$$

where $\xi^{(3)} = \xi$ and the $+$ and $-$ signs denote right and left circular polarization, while for the fermion (with the $+$ and $-$ signs now denoting positive and negative helicity states)

$$E_\pm^2 = p^2 + m^2 + \eta_\pm^{(3)} \frac{p^3}{M_{\rm Pl}} \ , \qquad (16)$$

with $\eta_{\pm}^{(3)} = 2(\zeta_1 \pm \zeta_2)$. For the antifermion, it can be shown by simple "hole interpretation" arguments that the same dispersion relation holds, with $\eta_{\pm}^{(3),af} = -\eta_{\mp}^{(3),f}$ where af and f superscripts denote respectively antifermion and fermion coefficients [34, 94]. Note that if CPT ends up being a fundamental symmetry of nature without Lorentz symmetry it would forbid all of the above mentioned CPT odd operators.

5.2 An Aside on Naturalness

With the specific dimension-five operators in hand, we can now return to a previously mentioned point, that RG flow does not significantly suppress the sizes of operators at low energies. Let us consider the evolution of the dimension-five LV parameters. Bolokhov and Pospelov [39] addressed the problem of calculating the RG equations for QED and the SM extended with dimension 5 operators that violate Lorentz symmetry. In the framework defined above, assuming that no extra physics enters between the low energies at which we have modified dispersion relations and the Planck scale at which the full theory is defined, the evolution equations for the LV terms in Eq. (14) that produce modifications in the dispersion relations, can be inferred as

$$\frac{d\zeta_1}{dt} = \frac{25}{12}\frac{\alpha}{\pi}\zeta_1 \,, \quad \frac{d\zeta_2}{dt} = \frac{25}{12}\frac{\alpha}{\pi}\zeta_2 - \frac{5}{12}\frac{\alpha}{\pi}\xi \,, \quad \frac{d\xi}{dt} = \frac{1}{12}\frac{\alpha}{\pi}\zeta_2 - \frac{2}{3}\frac{\alpha}{\pi}\xi \,, \tag{17}$$

where $\alpha = e^2/4\pi \simeq 1/137$ ($\hbar = 1$) is the fine structure constant and $t = \ln(\mu^2/\mu_0^2)$ with μ and μ_0 two given energy scales. (Note that the above formulae are given to lowest order in powers of the electric charge, which allows one to neglect the running of the fine structure constant.)

These equations show that the running is only logarithmic and therefore low-energy constraints are robust: $O(1)$ parameters at the Planck scale are still $O(1)$ at lower energy. Moreover, they also show that $\eta_{+}^{(3)}$ and $\eta_{-}^{(3)}$ cannot, in general, be equal at all scales. Similar calculations in the context of the renormalizable SME give equivalent results.

5.3 Constraints

We now detail some of the constraints that can be put on the CPT odd dimension-five operators. For a thorough review of these constraints, see also [93].

5.3.1 Photon Time of Flight

Although photon time-of-flight constraints from high-energy photons propagating from cosmologically distant objects currently provide limits several orders of

magnitude weaker than other constraints, they have been widely adopted in the astrophysical community. They were one of the first to be proposed in the seminal paper [26]. More importantly, given their purely kinematical nature, they may be applied to a broad class of frameworks, even beyond EFT with LV.

In general, a photon dispersion relation in the form of (15) implies that photons of different colors (wave vectors k_1 and k_2) travel at slightly different speeds. Let us first ignore any birefringence effects, and just consider some coefficient $\xi^{(n)}$ that is universal for all photons. Then, upon propagation on a cosmological distance d, the effect of energy dependence of the photon group velocity produces a time delay

$$\Delta t^{(n)} = \frac{n-1}{2} \frac{k_2^{n-2} - k_1^{n-2}}{M_{\rm Pl}^{n-2}} \xi^{(n)} d \ , \tag{18}$$

which clearly increases with d and with the energy difference as long as $n > 2$. The largest systematic error affecting this method is the uncertainty about whether photons of different energy are produced simultaneously in the source.

One way to alleviate systematic uncertainties is to apply EFT which gives more information than just a modified dispersion for photons. In particular, one knows that the term generated by $\xi^{(3)}$, in an EFT context implies birefringence. Furthermore, photon beams generally are not circularly polarized; thus, they are a superposition of super and subluminal circularly polarized modes. Hence, one can remove any systematic uncertainty relating to source dynamics or measured energy by measuring the velocity difference between the two polarization states at a single energy, corresponding to

$$\Delta t = 2|\xi^{(3)}| k \, d / M_{\rm Pl} \ . \tag{19}$$

This bound would require that both polarizations be observed and that no spurious helicity-dependent mechanism (such as, for example, propagation through a birefringent medium) affects the relative propagation of the two polarization states.

Note that one does not have to actually measure the actual polarization — a double peak in the high-energy spectra of GRB's with a separation that scaled linearly with distance would be a smoking gun for birefringent theories. However, if the polarization states do not fully separate then extracting a signal becomes more complicated. Since current time-of-flight constraints compare low- to high-energy photons, birefringence destroys this. A GRB would not have low-energy photons arriving first with high energy following (or vice versa). Instead the structure of the burst for a birefringent theory would be high-energy photons first, low-energy photons following, and then more high-energy photons at the end. Therefore, the net effect of this superposition may be to partially or completely erase the time-delay effect as it is usually calculated.

In order to compute this modulating effect on a generic photon beam in a birefringent theory, let us describe a beam of light by means of the associated electric field, and let us assume that this beam has been generated with a Gaussian width

$$\mathbf{E} = A \left(e^{i(\Omega_0 t - k^+(\Omega_0) z)} \, e^{-(z - v_g^+ t)^2 \delta \Omega_0^2} \hat{e}_+ + e^{i(\Omega_0 t - k^-(\Omega_0) z)} \, e^{-(z - v_g^- t)^2 \delta \Omega_0^2} \hat{e}_- \right) \ , \tag{20}$$

where Ω_0 is the wave frequency, $\delta\Omega_0$ is the gaussian width of the wave, $k^{\pm}(\Omega_0)$ is the "momentum" corresponding to the given frequency according to (15) and $\hat{e}_{\pm} \equiv (\hat{e}_1 \pm i\hat{e}_2)/\sqrt{2}$ are the helicity eigenstates. Note that by complex conjugation $\hat{e}_+^* = \hat{e}_-$. Also, note that $k^{\pm}(\omega) = \omega \mp \xi\omega^2/M_{\rm Pl}$. Thus,

$$\mathbf{E} = A\, e^{i\Omega_0(t-z)} \left(e^{i\xi\Omega_0^2/M_{\rm Pl}z}\, e^{-(z-v_g^+t)^2\delta\Omega_0^2}\hat{e}_+ + e^{-i\xi\Omega_0^2/M_{\rm Pl}z}\, e^{-(z-v_g^-t)^2\delta\Omega_0^2}\hat{e}_- \right) . \tag{21}$$

The intensity of the wave beam can be computed as

$$\begin{aligned} \mathbf{E}\cdot\mathbf{E}^* &= |A|^2 \left(e^{2i\xi\Omega_0^2/M_{\rm Pl}z} + e^{-2i\xi\Omega_0^2/M_{\rm Pl}z} \right) e^{-\delta\Omega_0^2\left((z-v_g^+t)^2+(z-v_g^-t)^2\right)} \\ &= 2|A|^2 e^{-2\delta\Omega_0^2(z-t)^2} \cos\left(2\xi\,\frac{\Omega_0}{M_{\rm Pl}}\Omega_0 z \right) e^{-2\xi^2\frac{\Omega_0^2}{M^2}(\delta\Omega_0 t)^2} . \end{aligned} \tag{22}$$

This shows that there is an effect even on a linearly-polarized beam. The effect is a modulation of the wave intensity that depends quadratically on the energy and linearly on the distance of propagation. In addition, for a Gaussian wave packet, there is a shift of the packet center, that is controlled by the square of $\xi^{(3)}/M_{\rm Pl}$; and hence, is strongly suppressed with respect to the cosinusoidal modulation. Hence, by looking for modulation of the signal with energy and distance one can in principle determine if the LV is birefringent from time-of-flight information even if the photons arrive as part of one "burst."

So far, the most robust constraints on $\xi^{(3)}$, derived from time-of-flight differences, have been obtained within the D-brane model (discussed in Sect. 9.1) from a statistical analysis applied to the arrival times of sharp features in the intensity at different energies from a large sample of GRBs with known redshifts [95], leading to limits $\xi^{(3)} \leq O(10^3)$. The importance of systematic uncertainties can be found in [96], where the strongest limit $f^{(3)} < 47$ is found by looking at a very strong flare in the TeV band of the AGN Markarian 501. Finally, an extremely strong limit (for this method at least) of $\xi^{(3)} < 0.8$ has been obtained from the short, high-energy GRB 090510 [97].

5.3.2 Photon Polarization

Since electromagnetic waves with opposite circular polarizations have slightly different group velocities in rotationally invariant EFT LV when CPT is violated, the polarization vector of a linearly polarized plane wave with energy k rotates. During the wave propagation over a distance d, the rotation angle for $n = 3$ dispersion

modifications is [34] [6]

$$\theta(d) = \frac{\omega_+(k) - \omega_-(k)}{2} d \simeq \xi^{(3)} \frac{k^2 d}{2\, M_{\rm Pl}} \,. \tag{24}$$

Observations of polarized light from a distant source can then lead to a constraint on $|\xi^{(3)}|$ that, depending on the amount of available information—both on the observational and on the theoretical (i.e., astrophysical source modeling) side—can be cast in two different ways [98]:

1. Because detectors have a finite-energy bandwidth, Eq. (24) is never probed in real situations. Rather, if some net amount of polarization is measured in the band $k_1 < E < k_2$, an order of magnitude constraint arises from the fact that if the angle of polarization rotation (24) differed by more than $\pi/2$ over this band, the detected polarization would fluctuate sufficiently for the net signal polarization to be suppressed [94, 99]. From (24), this constraint is

$$\xi^{(3)} \lesssim \frac{\pi\, M_{\rm Pl}}{(k_2^2 - k_1^2) d(z)}. \tag{25}$$

   This constraint requires that any intrinsic polarization (at source) not be completely washed out during signal propagation. It thus relies on the mere detection of a polarized signal; there is no need to consider the observed polarization degree. A more refined limit can be obtained by calculating the maximum observable polarization degree, given the maximum intrinsic value [100]:

$$\Pi(\xi) = \Pi(0) \sqrt{\langle \cos(2\theta) \rangle_{\mathcal{P}}^2 + \langle \sin(2\theta) \rangle_{\mathcal{P}}^2}, \tag{26}$$

   where $\Pi(0)$ is the maximum intrinsic degree of polarization, θ is defined in Eq. (24) and the average is weighted over the source spectrum and instrumental efficiency, represented by the normalized weight function $\mathcal{P}(k)$ [99]. Conservatively, one can set $\Pi(0) = 100\,\%$, but a lower value may be justified on the basis of source modeling. Using (26), one can then cast a constraint by requiring $\Pi(\xi)$ to exceed the observed value.
2. Suppose that polarized light measured in a certain energy band has a position angle $\theta_{\rm obs}$ with respect to a fixed direction. At fixed energy, the polarization vector rotates by the angle (24) [7]; if the position angle is measured by averaging over

[6] Note that for an object located at cosmological distance (let z be its redshift), the distance d becomes

$$d(z) = \frac{1}{H_0} \int_0^z \frac{1 + z'}{\sqrt{\Omega_\Lambda + \Omega_m (1 + z')^3}}\, dz' \,, \tag{23}$$

where $d(z)$ is not exactly the distance of the object as it includes a $(1 + z)^2$ factor in the integrand to take into account the redshift acting on the photon energies.

[7] Faraday rotation is negligible at these energies.

a certain energy range, the final net rotation $\langle \Delta\theta \rangle$ is given by the superposition of the polarization vectors of all the photons in that range:

$$\tan\left(2\,\langle \Delta\theta \rangle\right) = \frac{\langle \sin(2\theta) \rangle_{\mathcal{P}}}{\langle \cos(2\theta) \rangle_{\mathcal{P}}}\,, \tag{27}$$

where θ is given by (24). If the position angle at emission $\theta_{\rm i}$ in the same energy band is known from a model of the emitting source, a constraint can be set by imposing

$$\tan\left(2\,\langle \Delta\theta \rangle\right) < \tan\left(2\theta_{\rm obs} - 2\theta_{\rm i}\right)\,. \tag{28}$$

Although this limit is tighter than those based on Eqs. (25) and (26), it clearly hinges on assumptions about the nature of the source, which may introduce significant uncertainties.

In the case of the Crab Nebula (CN), a $(46 \pm 10)\,\%$ degree of linear polarization in the 100 keV − 1 MeV band has recently been measured by the INTEGRAL mission [101, 102]. This measurement uses all photons within the Spectrometer on INTEGRAL (SPI) instrument energy band. However, the convolution of the instrumental sensitivity to polarization with the detected number counts as a function of energy, $\mathcal{P}(k)$, is maximized and approximately constant within a narrower energy band (150–300 keV) and falls steeply outside this range [103]. For this reason we shall, conservatively, assume that most polarized photons are concentrated in this band. Given $d_{\rm Crab} = 1.9$ kpc, $k_2 = 300$ keV and $k_1 = 150$ keV, Eq. (25) leads to the order of magnitude estimate $|\xi| \lesssim 2 \times 10^{-9}$. A more accurate limit follows from (26). In the case of the CN there is a robust understanding that photons in the range of interest are produced via the synchrotron process, for which the maximum degree of intrinsic linear polarization is about 70 % (see, e.g., [104]). Figure 2 illustrates the dependence of Π on ξ (see Eq. (26)) for the distance of the CN and for $\Pi(0) = 70\,\%$. The requirement $\Pi(\xi) > 16\%$ (taking account of a 3σ offset from the best fit value 46 %) leads to the constraint (at 99 % confidence level)

$$|\xi| \lesssim 6 \times 10^{-9}\,. \tag{29}$$

It is interesting to notice that X-ray polarization measurements of the CN already available in 1978 [105] set a constraint $|\xi| \lesssim 5.4 \times 10^{-6}$, only one order of magnitude less stringent than that reported in [106].

Constraint (29) can be tightened by exploiting the current astrophysical understanding of the source. The CN is a cloud of relativistic particles and fields powered by a rapidly rotating, strongly magnetized neutron star. Both the Hubble Space Telescope and the Chandra X-ray satellite have imaged the system, revealing a jet and torus that clearly identify the neutron star rotation axis [107]. The projection of this axis on the sky lies at a position angle of $124.0^\circ \pm 0.1^\circ$ (measured from north in anticlockwise direction). The neutron star itself emits pulsed radiation at its rotation frequency of 30 Hz. In the optical band these pulses are superimposed on a fainter steady component with a linear polarization degree of 30 % and direction precisely aligned with that of the rotation axis [108]. The direction of polarization measured

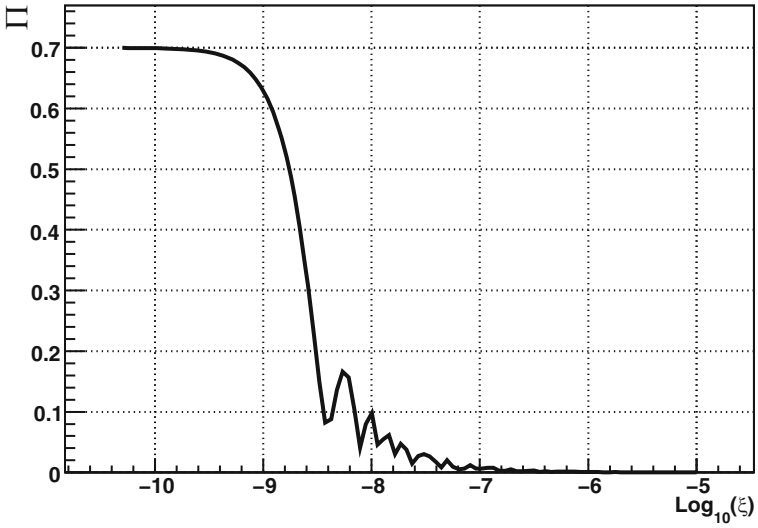


Fig. 2 Constraint for the polarization degree. Dependence of Π on ξ for the distance of the CN and photons in the 150–300 keV range, for a constant instrumental sensitivity $\mathcal{P}(k)$. Picture taken from [98]

by INTEGRAL-SPI in the γ-rays is $\theta_{\rm obs} = 123^\circ \pm 11^\circ$ (1σ error) from the north, thus also closely aligned with the jet direction and remarkably consistent with the optical observations.

This compelling (theoretical and observational) evidence allows us to use Eq. (28). Conservatively assuming $\theta_{\rm i} - \theta_{\rm obs} = 33^\circ$ (i.e., 3σ from $\theta_{\rm i}$, 99 % confidence level), this translates into the limit

$$|\xi^{(3)}| \lesssim 9 \times 10^{-10} \,, \tag{30}$$

and $|\xi^{(3)}| \lesssim 6 \times 10^{-10}$ for a 2σ deviation (95 % confidence level).

Polarized light from GRBs has also been detected and given their cosmological distribution they could be ideal sources for improving the abovementioned constraints from birefringence. Attempts in this sense were done in the past [94, 109] (but later on the relevant observation [110] appeared controversial) but so far we do not have sources for which the polarization is detected and the spectral redshift is precisely determined. In [111] this problem was circumvented by using indirect methods (the same used to use GRBs as standard candles) for the estimate of the redshift. This leads to a possibly less robust but striking constraints $|\xi^{(3)}| \lesssim 2.4 \times 10^{-14}$.

Remarkably this constraint was recently further improved by using the INTEGRAL-IBIS observation of the GRB 041219A, for which a luminosity distance of 85 Mpc ($z \approx 0.02$) was derived thanks to the determination of the GRB's host galaxy. In this case a constraint $|\xi^{(3)}| \lesssim 1.1 \times 10^{-14}$ was derived [112].[8]

5.3.3 Synchrotron Radiation

Synchrotron emission is strongly affected by LV, however for Planck scale LV and observed energies, it is a relevant "window" only for CPT odd dimension-five LV QED (and dimension-four LV QED, which we describe below). We shall work out here the details of CPT-odd dimension-five QED ($n = 3$) for illustrative reasons, for the lower-dimension case see [113].

In both LI and LV cases [34], most of the radiation from an electron of energy E is emitted at a critical frequency

$$\omega_c = \frac{3}{2} e B \frac{\gamma^3(E)}{E}, \tag{31}$$

where $\gamma(E) = (1 - v^2(E))^{-1/2}$, and $v(E)$ is the electron group velocity.

However, in the LV case, and assuming specifically $n = 3$, the electron group velocity is given by

$$v(E) = \frac{\partial E}{\partial p} = \left(1 - \frac{m_e^2}{2p^2} + \eta^{(3)} \frac{p}{M}\right) . \tag{32}$$

Therefore, $v(E)$ can exceed 1 if $\eta > 0$ or it can be strictly less than 1 if $\eta < 0$. This introduces a fundamental difference between particles with positive or negative LV coefficient η.

If η is negative, the group velocity of the electrons is strictly less than the (low energy) speed of light. This implies that, at sufficiently high energy, $\gamma(E)_- < E/m_e$, for all E. As a consequence, the critical frequency $\omega_c^-(\gamma, E)$ is always less than a maximal frequency $\omega_c^{\max}$. Then, if synchrotron emission up to some maximal frequency $\omega_{\rm obs}$ is observed, one can deduce that the LV coefficient for the corresponding leptons cannot be more negative than the value for which $\omega_c^{\max} = \omega_{\rm obs}$, leading to the bound [34]

$$\eta^{(3)} > -\frac{M}{m_e} \left(\frac{0.34\, e B}{m_e\, \omega_{\rm obs}}\right)^{3/2} . \tag{33}$$

[8] The same paper also claims a strong constraint on the parameter $\xi^{(4)}$. Unfortunately, such a claim is based on the erroneous assumption that the EFT order six operators responsible for this term imply opposite signs for opposite helicities of the photon. We have instead seen that the CPT evenness of the relevant dimension-six operators imply a helicity independent dispersion relation for the photon (see Eq. (9)).

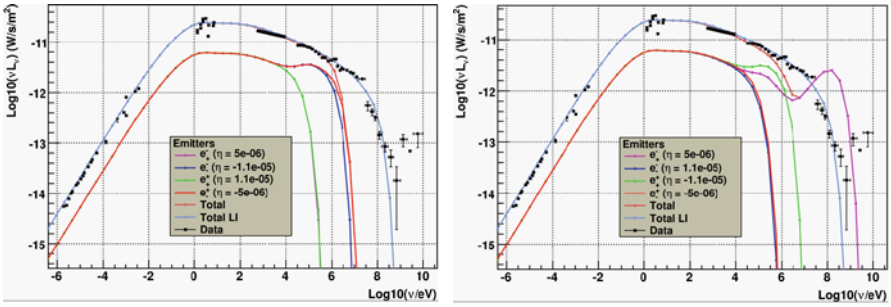


Fig. 3 Comparison between observational data, the LI model and a LV one with $\eta_+ \cdot \eta_- < 0$ *(left)* and $\eta_+ \cdot \eta_- > 0$ *(right)*. The values of the LV coefficients, reported in the insets, show the salient features of the LV modified spectra. The leptons are injected according to the best fit values $p = 2.4$, $E_c = 2.5$ PeV. The individual contribution of each lepton population is shown. Picture taken from [115]

If η is instead positive the leptons can be superluminal. One can show that at energies $E_c \gtrsim 8\ \text{TeV}/\eta^{1/3}$, $\gamma(E)$ begins to increase faster than E/m_e and reaches infinity at a finite energy which corresponds to the threshold for soft VC emission. The critical frequency is thus larger than the LI one and the spectrum shows a characteristic bump due to the enhanced ω_c.

How the synchrotron emission processes at work in the CN would appear in a "LV world" has been studied in [114, 115]. There the role of LV in modifying the characteristics of the Fermi mechanism (which is thought to be responsible for the formation of the spectrum of energetic electrons in the CN [116]) and the contributions of VC and helicity decay (HD) were investigated for $n = 3$ LV. This procedure requires fixing most of the model parameters using radio to soft X-rays observations, which are basically unaffected by LV.

Given the dispersion relations (15) and (16), clearly only two configurations in the LV parameter space are truly different: $\eta_+ \cdot \eta_- > 0$ and $\eta_+ \cdot \eta_- < 0$, where η_+ is assumed to be positive for definiteness. The configuration wherein both $\eta_\pm$ are negative is the same as the ($\eta_+ \cdot \eta_- > 0$, $\eta_+ > 0$) case, whereas that whose signs are scrambled is equivalent to the case ($\eta_+ \cdot \eta_- < 0$, $\eta_+ > 0$). This is because positron coefficients are related to electron coefficients through $\eta_\pm^{af} = -\eta_\mp^f$ [34]. Examples of spectra obtained for the two different cases are shown in Fig. 3.

A χ^2 analysis has been performed to quantify the agreement between models and data [115]. From this analysis, one can conclude that the LV parameters for the leptons are both constrained, at 95 % confidence level, to be $|\eta_\pm| < 10^{-5}$, as shown by the red vertical lines in Fig. 4. Although the best fit model is not the LI one, a careful statistical analysis (performed with present-day data) shows that it is statistically indistinguishable from the LI model at 95 % confidence level [115].

5.3.4 Constraints from GZK Secondaries

The same reasoning that established constraints on the CPT even higher-dimension sector from GZK secondaries can also be applied to further strengthen the available constraints on CPT odd dimension-five LV QED. In this case the absence of relevant UHE photon flux strengthens by (at most) two orders of magnitude the constraint on the photon coefficient $\xi^{(3)}$ while an eventual detection of the expected flux of UHE photons would constrain $\eta^{(3)}$ for the electron/positron at the level of $|\eta^{(3)}| \lesssim 10^{-16}$ (see [88,93] for further details) by limiting the γ-decay process. Note however that in this case, unlike the CPT even case, one cannot exclude that only one photon helicity survives and hence a detailed flux reconstruction would be needed.

5.3.5 Summary of Constraints

Constraints on LV QED at $n = 3$ are summarized in Fig. 4 where also the constraints — coming from the observations of up to 80 TeV γ-rays from the CN [117] (which imply no γ-decay for these photons neither VC at least up to 80 TeV for the electrons producing them via inverse Compton scattering)—are plotted for completeness.

5.4 Other Threshold Processes

Once one realizes the power of the VC effect it is natural to explore other threshold processes that might yield useful constraints. The constraints in Fig. 4 are the best available constraints. We detail below some other processes that have been used to set constraints in the past.

5.4.1 Helicity Decay

A slightly different version of the VC process is that of HD, $e_{\mp} \rightarrow e_{\pm}\gamma$. If $\eta_+ \neq \eta_-$, an electron can flip its helicity by emitting a suitably polarized photon. This reaction does not have a real threshold, but rather an effective one [34]—$p_{\rm HD} = (m_e^2 M/\Delta\eta)^{1/3}$, where $\Delta\eta = |\eta_+^{(3)} - \eta_-^{(3)}|$—at which the decay lifetime τ_{HD} changes how it scales with $\Delta\eta$. For $\Delta\eta \approx O(1)$ this effective threshold is around 10 TeV. Below threshold the lifetime $\tau_{\rm HD}$ is given by $\tau_{\rm HD} > \Delta\eta^{-3}(p/10\ {\rm TeV})^{-8}\, 10^{-9}$s, while above threshold $\tau_{\rm HD}$ becomes independent of $\Delta\eta$ [34] and is given by $\tau_{\rm HD} \approx 10^{-10} \times (10\ {\rm Tev})/E$ s. It is more difficult, however, to use HD to set constraints as we do not have polarization measurements of high-energy cosmic rays and the flux above a certain energy scale is only halved in the case of HD, rather than almost completely removed as in the VC effect. One therefore requires more detailed knowledge of the source spectrum to properly apply HD constraints.

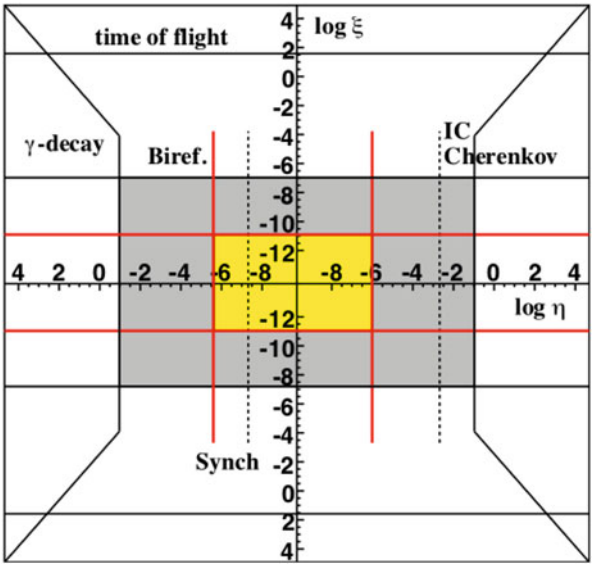


Fig. 4 Summary of the constraints on LV QED at order $n = 3$. The *red lines* are related to the constraints derived from the detection of polarized synchrotron radiation from the CN as discussed in the text. For further reference the constraints that can be derived from the detection of 80 TeV photons from the CN: the *solid black lines* symmetric w.r.t. the ξ axis are derived from the absence of gamma decay, the *dashed vertical line* cutting the η axis at about 10^{-3} refers to the limit on the VC effect coming from the inferred 80 TeV inverse Compton electrons are also shown. The *dashed vertical line* on the negative side of the η axis is showing the first synchrotron-based constraint derived in [114]

Nonetheless, HD can play a crucial role in lepton/antilepton propagation by basically leaving only a survival helicity state for each particle type. This mechanism, for example, played a pivotal role in the reconstruction of the CN synchrotron spectrum [115].

5.4.2 Photon Splitting and Lepton Pair Production

VC radiation is an effect that would ordinarily be forbidden, however one can also look for modifications of normally allowed threshold reactions that are especially relevant for high-energy astrophysics. It is rather obvious that once photon decay and VC are allowed also the related reactions in which an outgoing lepton/antilepton pair is replaced by two or more photons, $\gamma \to 2\gamma$ and $\gamma \to 3\gamma$, etc., or the outgoing photons are replaced by an electron-positron pair, $e^- \to e^- e^- e^+$, are also allowed.

LV-allowed Reactions: Photon Splitting
This is forbidden for $\xi^{(n)} < 0$ while it is always allowed if $\xi^{(n)} > 0$ [32]. When allowed, the relevance of this process is simply related to its rate. The most relevant cases are $\gamma \to \gamma\gamma$ and $\gamma \to 3\gamma$, because processes with more photons in the final state are suppressed by higher powers of the fine structure constant.

The $\gamma \to \gamma\gamma$ process is forbidden in QED because of kinematics and CP conservation. In LV EFT both the kinematics changes and CP is not necessarily a good symmetry. However, we can argue that $\gamma \to \gamma\gamma$ is suppressed by an additional power of the Planck mass with respect to $\gamma \to 3\gamma$. In fact, in LI QED the matrix element is zero due to the exact cancellation of fermionic and antifermionic loops. In LV EFT this cancellation is not exact and the matrix element is expected to be proportional to at least $(\xi E/M_{\rm Pl})^p$, $p > 0$, as it is induced by LV and must vanish in the limit $M_{\rm Pl} \to \infty$.

Therefore we have to deal only with $\gamma \to 3\gamma$. This process has been studied in [32, 118]. In particular, in [118] it was found that, if the "effective photon mass" $m_\gamma^2 \equiv \xi E_\gamma^n / M_{\rm Pl}^{n-2} \ll m_e^2$, then the splitting lifetime of a photon is approximately $\tau^{n=3} \simeq 0.025\, \xi^{-5} f^{-1} \left(50\ {\rm TeV}/E_\gamma\right)^{14}$ s, where f is a phase space factor of order 1. This rate was rather higher than the one obtained via dimensional analysis in [32] because, due to integration of loop factors, additional dimensionless contributions proportional to m_e^8 enhance the splitting rate at low energy.

This analysis, however, does not apply for the most interesting case of UHE photons around 10^{19} eV given that at these energies $m_\gamma^2 \gg m_e^2$ if $\xi^{(3)} > 10^{-17}$ and $\xi^{(4)} > 10^{-8}$. Hence the abovementioned loop contributions are at most logarithmic, as the momentum circulating in the fermionic loop is much larger than m_e. Moreover, in this regime the splitting rate depends only on m_γ, the only energy scale present in the problem. One then expects the analysis proposed in [32] to be correct and the splitting time scale to be negligible at $E_\gamma \simeq 10^{19}$ eV, which therefore makes it not particularly competitive with other constraints.

Lepton Pair Production
The process $e^- \to e^- e^- e^+$ is similar to VC radiation or HD with the final photon replaced by an electron-positron pair. Various combinations of helicities for the different fermions can be considered individually. If we choose the particularly simple case (and the only one we shall consider here) where all electrons have the same helicity and the positron has the opposite helicity, then the threshold energy will depend on only one LV parameter. In [32] the threshold for this reaction was derived for electron pair production, and it was found that the rate is a factor of $\sim$ 2.5 times higher than that for soft VC radiation. Therefore, since the rate for

the reaction is high as well, constraints may be imposed using just the value of the threshold energy. [9]

6 Dynamical Frameworks III: Rotationally Invariant Minimal Standard Model Extension in QED

6.1 The Model

The mSME, is the set of renormalizable operators that generate LV but maintain the existing particle content of the SM and do not violate gauge invariance. A subset of the mSME, the rotationally invariant LV operators are

$$-au_\mu\overline{\psi}\gamma^\mu\psi - bu_\mu\overline{\psi}\gamma_5\gamma^\mu\psi + \frac{1}{2}icu_\mu u_\nu\overline{\psi}\gamma^\mu \overleftrightarrow{D^\nu}\psi + \frac{1}{2}idu_\mu u_\nu\overline{\psi}\gamma_5\gamma^\mu \overleftrightarrow{D^\nu}\psi \tag{34}$$

for fermions and

$$-\frac{1}{4}(k_F)u_\kappa\eta_{\lambda\mu}u_\nu F^{\kappa\lambda}F^{\mu\nu} + \frac{1}{4}k_{AF}u^\kappa\varepsilon_{\kappa\alpha\beta\gamma}A^\alpha F^{\beta\gamma} \tag{35}$$

for photons. The dimension three, CPT odd k_{AF} term generates an instability in the theory and so we will set it to zero from here on out. The a term can be absorbed by shifting the phase of the fermion field and so we will ignore it temporarily (we shall return to this term when we deal with gravity). Note however that while the a term can be absorbed in QED, it can be measured in matter sectors where the phase of the fermion is important. For example, neutrinos can constrain differences in a between species [42].

The corresponding high-energy ($M_{\rm Pl} \gg E \gg m$) dispersion relations for QED can be expressed as (see [7] and references therein for more details)

$$E_{\rm el}^2 = m_e^2 + p^2 + f_e^{(1)}p + f_e^{(2)}p^2 \quad \text{electrons} \tag{36}$$

$$E_\gamma^2 = (1 + f_\gamma^{(2)})p^2 \quad \text{photons,} \tag{37}$$

where $f_e^{(1)} = -2bs$, $f_e^{(2)} = -(c - ds)$, and $f_\gamma^{(2)} = k_F/2$ with $s = \pm 1$ the helicity state of the fermion [7]. The antifermion dispersion relation is the same as (36) with the replacement $p \to -p$, which will change only the $f_e^{(1)}$ term.

[9] One could of course consider any lepton/antilepton pair as produced, for example, the reaction $e^- \to e^-\nu\overline{\nu}$. While standard particle physics arguments imply that the rate will be roughly equivalent to the e^-e^+ pair production case [119] given the same order of coefficients, and the threshold will be slightly lower, the constraints are on a higher-dimensional parameter space and so less useful.

Note that the typical energy at which new phenomenology should start to appear is quite low. In fact, taking for example $f_e^{(2)} \sim O(1)$, one finds that the corresponding extra term is comparable to the particle mass m precisely at $p \simeq m$. Even worse, for the linear modification to the dispersion relation, we would have, in the case in which $f_e^{(1)} \simeq O(1)$, that $p_{\rm th} \sim m^2/M_{\rm Pl} \sim 10^{-17}$ eV for electrons. (Notice that this energy corresponds by chance to the present upper limit on the photon mass, $m_\gamma \lesssim 10^{-18}$ eV [120].)

6.2 Constraints

In contrast to the higher-dimension operators, LV due to the mSME does not grow with energy. Therefore, astrophysics with its higher energies does not necessarily provide a tremendous advantage over laboratory and terrestrial experiments when testing the mSME. We list below the best constraints currently available on the QED sector of the mSME. For a recent listing of all the constraints as well as the non-rotationally invariant case, see [73].

6.2.1 Spin Polarized Torsion Constraints on *b* for Electrons

Both spin polarized and unpolarized torsion balances can place limits on the mSME. Spin polarized torsion balances place limits on the electron sector of the mSME [121], while unpolarized torsion balances constrain the gravitational sector [122]. Spin polarized torsion balances constrain the electron sector, as the torsion balances are constructed to have a large number of aligned electron spins (e.g., a simple magnet attached on a torsion fiber is a very crude spin polarized torsion balance). Usually of course, the magnet design is optimized to search for LV. For example, a vertical stack of an octagonally symmetric pattern of magnets constructed to have an overall spin polarization in the octagon's plane has been used as a torsion balance with a net electron spin polarization of 10^{23} electron spins [121]. The mSME coefficients give rise to an interaction potential for nonrelativistic electrons which produces an orientation-dependent torque on the torsion balance which is measured using the twist of the torsion fiber. The torsion balance in its sealed vacuum chamber is mounted on a rotating turntable, which allows for very sensitive detection of any anomalous torque as a function of the rotation frequency (and the earth's rotation and motion in the solar system). Current limits on b from torsion balances are of the order of 10^{-27} GeV [121].

6.2.2 Accelerator Bounds on *c* for Electrons

Accelerator beams of various subatomic particles are produced with (roughly) time-independent energies. Energy loss mechanisms such as the VC effect $q \to q + \gamma$

(see Sect. 4.2.1) for charged fermions are incompatible with a constant beam energy as long as the energy loss rate is high enough. Limits can therefore be set on the c coefficient for accelerated charged particles, as if c is too much larger than zero then above the $n = 2$ threshold energy (12) radiative energy losses become significant over the beam time. The large electron-positron (LEP) experiment accelerates electrons and positrons to energies of roughly 100 GeV in the lab frame. The beam energy is carefully measured and it is known that the beam does not lose any significant energy to VC radiation. Additionally, the synchrotron emission from the beam has been measured. Since synchrotron emission is sensitive to the Lorentz factor of the accelerated particles, the existence of synchrotron radiation of a certain spectrum given a source particle energy and path provides constraints on any LV present (see Sect. 5.3.3). Using the characteristics of the LEP beam and the associated synchrotron radiation spectrum limits c for electrons to be of order $|c| < 10^{-15}$.

6.2.3 Astrophysical Inverse Compton Bounds on d for Electrons

The parameter d causes a spin-dependent change to the dispersion relation for a fermion. As such, one of the helicity states for a fermion is always subluminal and one cannot trivially apply simple astrophysical arguments based on the VC effect, etc. without knowing the helicity of the measured fermion. However, there is a way around this by considering the dynamics of sources of high-energy photons produced by inverse Compton scattering. We receive energies from the radio up to 80 TeV from astrophysical sources such as the CN [117]. The overall spectrum of these sources is well understood and the high-energy emission is dominated by inverse Compton scattering of accelerated electrons off of lower-energy photons. In order to constrain d one cannot simply require the absence of VC radiation for the electrons, as half of any population will still be present under the influence of a d term and available to inverse Compton scatter. However, the magnetic field present in the same sources causes the electrons to precess, thereby destroying any initial polarization of the electrons. Therefore, as argued in [123] both helicities of electrons must be present and stable if there is to be inverse Compton radiation. The VC effect is therefore forbidden for both helicities, which allows one to put double-sided constraints on d of the order of $|d| < O(10^{-12})$ [123].[10]

6.2.4 Cosmic Ray and HESS Bounds on k_F

Bounds on the fermion sector that use processes involving the fundamental QED vertex implicitly assume that the photon sector is unmodified. The sector which is

[10] Note that the bounds presented here are weaker by a factor of 10^{-3}, as we have used the CMB frame as the rest frame rather than the Sun-centered frame and therefore the strengths of the bounds are weakened by the v/c of the Sun with respect to the CMB.

the "unmodified" sector is arbitrary, in that the limiting speed of one of the sectors can be defined to be the "speed of light." When constraining photon coefficients, we simply make the opposite assumption—we assume the fermion sector is unmodified and constrain k_F. Constraints on k_F can be generated for protons and photons using the same effects as above, that is, the necessary absence of VC and photon decay if we see high-energy cosmic ray protons and TeV gamma rays. In [124] the authors used the necessary absence of $p \to p + \gamma$ and a Pierre Auger event in conjunction with the excess of high-energy TeV gamma rays observed by the HESS telescope (which forbids $\gamma \to p + \overline{p}$) to produce a two-sided bound on k_F of $-9 \times 10^{-16} < k_f < 6 \times 10^{-20}$.

7 Dynamical Frameworks IV: Gravity

7.1 The Model

The SME is constructed by coupling matter terms to non-zero LV tensors in vacuum. If we left the tensors as constants, without any sort of dynamics, then one would break general covariance. It may be that for other reasons one may want to change the underlying symmetry structure for gravity in just this way; such is the case with Hořava–Lifshitz gravity [36]. Alternatively, if one wants to preserve general covariance one can do so by promoting any LV tensors to dynamical fields. Dynamical Lorentz breaking is also sometimes called "spontaneous Lorentz violation," although this is a bit of a misnomer as there are models where there really is no Lorentz invariant phase. There are many ways that a dynamical field can generate a LV term in vacuum. If we restrict ourselves to rotational invariance, then it is natural to generate LV couplings by including in the action either a scalar or a timelike vector field that takes a VEV. In the case of a scalar, one can use a shift symmetry ($\phi(x) \to \phi(x) + \phi_0$) to construct actions for which the derivative of the scalar takes a nonzero value (c.f. [125]). In the vector case, one simply puts a potential for the vector field such that the vector acquires a vev. We concentrate on the vector case here as it is the simplest model that allows for rotationally invariant Lorentz violation [126]. Just as the SME is a derivative expansion in derivatives of matter fields, one can treat the vector field in the same way. If we still denote the vector field by u^α then we can write the low-energy action for gravity and the vector as

$$\mathcal{S} = \mathcal{S}_{\text{EH}} + \mathcal{S}_{\sqcap} = \frac{\infty}{\infty \not{\mathcal{G}}_{\text{æ}}} \int \mathrm{d}^{\triangle}\S \sqrt{-\mathsf{g}} \ (\mathcal{R} + \mathscr{L}_{\sqcap}) \,, \tag{38}$$

where $\mathscr{L}_u$ is given by

$$\mathscr{L}_u = -Z^{\alpha\beta}_{\gamma\delta}(\nabla_\alpha u^\gamma)(\nabla_\beta u^\delta) + V(u). \tag{39}$$

The tensor $Z^{\alpha\beta}_{\gamma\delta}$ is defined as [71][11]

$$Z^{\alpha\beta}_{\gamma\delta} = c_1 \mathsf{g}^{\alpha\beta}\mathsf{g}_{\gamma\delta} + c_2 \delta^{\alpha}_{\gamma}\delta^{\beta}_{\delta} + c_3 \delta^{\alpha}_{\delta}\delta^{\beta}_{\gamma} - c_4 u^{\alpha}u^{\beta}\mathsf{g}_{\gamma\delta}\,, \tag{40}$$

where c_i, $i = 1, \ldots, 4$ are simple coefficients of the various kinetic terms and $V(u)$ is a potential term that generates a nonzero VEV for u^{α}. An additional term, $R_{ab}u^a u^b$, is a combination of the above terms when integrated by parts, and hence is not explicitly included here.

In general, such a theory possesses four additional degrees of freedom. One of the vector components, however, will necessarily have a wrong sign for its kinetic term, thereby generating a ghost excitation. This can be remedied, at least at low energies, by choosing the potential $V(u) = \lambda(u^2 + 1)$, where λ is a Lagrange multiplier. This fixes the norm of u^{α} and removes the ghost excitation [12] so the excitations of the vector all have positive norm (c.f. the discussions in [70] and [127]). With this form of the potential, the vector theory is known as "Einstein-aether theory" [126]. We will refer to the vector field as an "aether" field, vector field models that couple to the SM Lagrangian are also sometimes nicknamed "bumblebee" models. In reality of course the Lagrange multiplier is likely simply an approximation to a potential that generates dynamics such that the ghost only has ultraviolet (UV) effects. For the purposes of constraints, we will treat the Lagrange multiplier as "the" potential term and so neglect any possible ghost excitation.

The matter couplings between aether and the SM field content are the same as they were before, only now the field has dynamics so there can be position-dependent violations of Lorentz symmetry. Interestingly, some of the couplings to matter that are unobservable for a single fermion field can have relevant effects when the aether varies. For example, the $-au_{\mu}\overline{\psi}\gamma^{\mu}\psi$ term in the mSME could be removed by making a phase change for the fermion. However, once u^{α} is dynamical and varies with position, only a single component of the term can actually be removed by a phase change [128]. This leads to a new type of term which requires gravitational/position-dependent tests in the matter sector [128].

Additionally, just as constraints can be put on the coupling of matter fields to the LV vector field, constraints can also be put on the c_i coefficients themselves by examining the relevant parameterized post-Newtonian (PPN) parameters and comparing to known PPN limits from solar system and other observations.

[11] Note the index symmetry $Z^{\beta\alpha}_{\delta\gamma} = Z^{\alpha\beta}_{\gamma\delta}$.

[12] Note that one could have fixed the norm to be different than -1, however, one can simply scale u^{α} to have norm -1 and absorb the scaling into the coefficients.

7.2 Constraints

7.2.1 Constraints on the Aether Kinetic Terms

Constraints from PPN Analysis

There are two primary methods to constrain the aether kinetic terms. First, one can compute the PPN coefficients for the theory and compare to observational tests. The only two nonvanishing PPN parameters are α_1, α_2 which describe preferred-frame effects. α_1 and α_2 have been calculated [129] as

$$\alpha_1 = \frac{-8(c_3^2 + c_1 c_4)}{2c_1 - c_1^2 + c_3^2}, \tag{41}$$

$$\alpha_2 = \frac{\alpha_1}{2} - \frac{(c_1 + 2c_3 - c_4)(2c_1 + 3c_2 + c_3 + c_4)}{(c_1 + c_2 + c_3)(2 - c_1 - c_4)}. \tag{42}$$

Current constraints are $\alpha_1 < 10^{-4}$ and $\alpha_2 < 4 \times 10^{-7}$ [130] and so from a PPN analysis alone there is still a large two-dimensional region of parameter space that remains consistent with available tests of GR.

Constraints from Gravity-Aether Wave Modes

The aether introduces three new excitations into the gravity sector (there are naively four, but the unit constraint removes one). These excitations strongly couple to the metric via the constraint. The combined aether-metric modes consist of the two usual transverse traceless graviton modes, a vector mode, and a scalar mode [131]. Gravitational wave detectors can in principle test for polarizations [132], however, there is an additional possibility to constrain the kinetic terms. The speeds of each of the modes can differ from the speed of light. Hence if the speeds are less than unity, high-energy cosmic rays will emit vacuum gravitational Čerenkov radiation [133]. If we denote the speeds of the spin-2, spin-1 and spin-0 modes by s_2, s_1, s_0 then we have [131]

$$s_2^2 = (1 - c_1 - c_3)^{-1}, \tag{43}$$

$$s_1^2 = \frac{2c_1 - c_1^2 + c_3^2}{2(c_1 + c_4)(1 - c_1 - c_3)}, \tag{44}$$

$$s_0^2 = \frac{(c_1 + c_2 + c_3)(2 - c_1 - c_4)}{(c_1 + c_4)(1 - c_1 - c_3)(2 + 3c_2 + c_1 + c_3)}. \tag{45}$$

Requiring all these speeds to be greater than unity therefore places constraints on a combination of the c_i coefficients.

Combined Constraints

Even after imposing the above constraints there is still a large region of parameter space allowed. To get an estimate of the size of the space, one can set α_1 and α_2

equal to zero and solve the resulting equations for c_2 and c_4 in terms of c_1 and c_3. If we define $c_+ = c_1 + c_3$ and $c_- = c_1 - c_3$ then the PPN and gravitational Čerenkov constraints are all satisfied provided [71]

$$0 \leq c_+ \leq 1 \tag{46}$$

$$0 \leq c_- \leq \frac{c_+}{3(1-c_+)}, \tag{47}$$

which shows that the gravitational sector is only minimally constrained compared to aether-matter couplings.

7.2.2 Gravitational Constraints on Aether-Matter Couplings

There are couplings between aether and matter that can only be strongly tested when the aether is dynamical. The prime example is the coupling $-au_\mu \overline{\psi}\gamma^\mu\psi$ in the mSME. As said, when gravity is neglected and we are dealing with an observation that involves only one fermion field, then the au_μ term can be absorbed by making a phase transformation of the fermion $\psi \rightarrow e^{iau\cdot x}\psi$. This cannot be done when u is a function of position, which would necessarily be the case if u is a dynamical field strongly coupled to gravity (as we have done with the constraint above). One can therefore use gravitational experiments to measure the position dependence of u^μ, similar to how redshift experiments measure the gravitational field. For a larger overview of constraints, see [73,134]. We will concentrate here on a recently applied method, that of atom interferometry [134].

Atom interferometry involves splitting of a beam of cold atoms such as Cs along two beam arms and comparing the phase difference at a point of the Cs atoms that travel along each beam. Since gravity and the matter sector of the mSME affect the phase of the atomic wavefunctions, one can look for an anomalous sidereal pattern. By orienting the interferometer vertically and horizontally, one can adjust the sensitivity to the gravitational and matter sectors of the mSME. Since one can also use other limits to control for any possible matter mSME effect on the signal, one can thereby isolate solely gravitational effects. Hence, atom interferometers have been used with great success on constraining the gravitational sector of the mSME.

Consider an interferometer that is constructed to measure phase shifts of atoms moving vertically versus horizontally. There will be a contribution to the total phase shift $\Delta\phi$ that comes from the gravitational redshift, that is, some $\Delta\phi_{red}$. In the Newtonian limit where u_μ is proportional to the timelike Killing vector, u^μ is given by $u^\mu = (1 - U, 0, 0, 0)$ [135], where U is the Newtonian gravitational potential. If we couple u^μ to matter, one can easily see that the effect of the coupling via an au^μ or $cu^\mu u^\nu$ term is to change the redshift by an additional amount proportional to the gravitational potential, that is, the change in frequency of a wave is given by $\Delta\omega = (1+\beta)\Delta U$ where β is an experimental parameter related to the fundamental Lagrangian parameters (c.f. [136] for a discussion). The phase shift generated by this extra gravitational redshift is then just $\Delta\phi_{red} = (1+\beta)\Delta\phi_0$, where $\Delta\phi_0$ is

the expected gravitational redshift. Constraints on β therefore limit the fundamental Lagrangian parameters. β is limited to be zero within a few parts per billion, which then generates fairly strong limits on position dependence of the a and c couplings [134].

8 Neutrinos

While we have concentrated primarily on QED, we would be remiss to not mention at least in passing the important role of neutrinos in tests of Lorentz violation. More so than other particles, neutrinos are uniquely suited to test various aspects of Lorentz symmetry. They are copiously produced in both terrestrial experiments and astrophysical objects, so there are solar neutrinos at < 1 MeV, controlled beams of neutrinos at roughly 10 GeV, atmospheric neutrinos up to 10 TeV, and (theoretically) UHE neutrinos up to 10^{18} eV and above. Neutrinos are weakly interacting, so they propagate over long distances which allows for detailed time of flight and threshold reaction analyses. Neutrinos have a very small mass, which make threshold analyses even more sensitive. Finally, neutrinos oscillate between flavor eigenstates, which constrains interspecies Lorentz violation.

The problem, of course, is that neutrinos are rather difficult to detect. However, thanks to the numerous ongoing neutrino experiments, many significant experimental results have been published that can be adapted to constraining Lorentz violation in the neutrino sector. We now turn to the theoretical framework and current and future constraints on the neutrino sector.

8.1 Neutrinos I: Species Dependent Lorentz Violation

8.1.1 The Model

Neutrinos come in three distinct flavors/masses, and there is *a priori* no reason to believe that any Lorentz violating coefficients are the same for each species. It may even be that any Lorentz violation is not diagonal in either basis. We make a couple simplifying assumptions on this point — that the neutrinos are Dirac neutrinos and that any Lorentz violation is diagonal in the mass basis. This is a reasonable starting point, as one natural idea is that since any theory of QG must reduce to GR in the infrared, any Lorentz violation induced by QG would be primarily controlled by the charges that couple to gravity. However, this does not mean that the coefficients for each mass eigenstate are the same. Indeed, one would expect that, due to RG effects, even if the coefficients were the same for all eigenstates at one energy they would not be the same for the large range of energies we can test in neutrino experiments.

With these assumptions, the Lorentz violating terms (written in the mass basis) are exactly those for the QED fermions,

$$-a_i u_\mu \overline{\psi}_i \gamma^\mu \psi_i - b_i u_\mu \overline{\psi}_i \gamma_5 \gamma^\mu \psi_i + \frac{1}{2} i c_i u_\mu u_\nu \overline{\psi}_i \gamma^\mu \overleftrightarrow{\partial^\nu} \psi_i \qquad (48)$$
$$+\frac{1}{2} i d_i u_\mu u_\nu \overline{\psi}_i \gamma_5 \gamma^\mu \overleftrightarrow{\partial^\nu} \psi_i + \frac{1}{2M_{\rm Pl}} u^m \overline{\psi}_i \gamma_m (\zeta_{i,1} + \zeta_{i,2}\gamma_5)(u\cdot\partial)^2 \psi_i$$
$$-\frac{i}{M_{\rm Pl}^2} \overline{\psi}_i (u\cdot\partial)^3 (u\cdot\gamma)(\alpha_{i,L}^{(6)} P_L + \alpha_{i,R}^{(6)} P_R)\psi_i ,$$

where i is the mass index. We have dropped the gauge covariant derivative above, as it is irrelevant and couples in the flavor basis so merely would add needless complication. Also note that we have included both right and left projection operators. However, since SM interactions only produce left-handed neutrinos, the constraints will primarily be on the corresponding left-handed operators and so we drop all P_R terms in our discussion of oscillations (although we will need to return to them when we discuss neutrino splitting). Finally, in contrast to the QED case, we cannot drop the a_i term as this gives a contribution to the oscillation pattern.

The above terms and the usual Dirac Lagrangian for the neutrino yield a high-energy neutrino dispersion relation of

$$E_i^2 = p^2 + N_i^2 \qquad (49)$$
$$N_i^2 = m_i^2 + 2(a_i + b_i)p - (c_i + d_i)p^2 + \frac{m_i}{M_{\rm Pl}} \alpha_{i,L}^{(5)} p^2$$
$$+2(\zeta_{i,1} - \zeta_{i,2})\frac{p^3}{M_{\rm Pl}} + 2\frac{\alpha_{i,L}^{(6)} p^4}{M_{\rm Pl}^2} .$$

The tightest constraint on species dependence comes from neutrino flavor oscillation measurements. Neutrino oscillations depend on the differences in $E - p$ between different neutrino eigenstates. In standard neutrino oscillations, this difference is governed by the squared mass differences between the energy eigenstates. With Lorentz violation and our assumption that the Lorentz violating eigenstates are the mass eigenstates, oscillations are governed by the differences in the effective mass squared. Therefore, neutrino oscillations do not probe any absolute Lorentz violation in the neutrino sector, but rather the differences in the dispersion relations between different neutrino states.

Now let us consider a neutrino produced via a particle reaction in a definite flavor eigenstate I with momentum p. The amplitude for this neutrino to be in a particular mass eigenstate i is represented by the matrix U_{Ii}, where $\sum U_{Ji}^\dagger U_{Ii} = \delta_{IJ}$. The amplitude for the neutrino to be observed in another flavor eigenstate J at some distance L from the source, after some time T is then

$$A_{IJ} = \sum_i U_{Ji}^\dagger e^{-i(ET - pL)} U_{Ii} \approx \sum_I U_{Ji}^\dagger e^{-iLN_i^2/(2E)} U_{Ii} . \qquad (50)$$

The transition probability can then be written as

$$P_{IJ} = \delta_{IJ} - \sum_{i,j>i} 4F_{IJij} \sin^2\left(\frac{\delta N_{ij}^2 L}{4E}\right) + 2G_{IJij} \sin^2\left(\frac{\delta N_{ij}^2 L}{2E}\right), \tag{51}$$

with $\delta N_{ij}^2 = N_i^2 - N_j^2$ and F_{IJij} and G_{IJij} are functions of the mixing matrices. With this formalism one can compare with existing neutrino oscillation experiments that measure P_{IJ}. Many of these experiments quote results on a deviation of the neutrino speed from that of light, that is,

$$\left(\frac{\Delta c}{c}\right)_{ij}^{LIV} = E^{-2}(\delta N_{ij}^2 - \delta m_{ij}^2) \tag{52}$$

which can be easily translated into a constraint on the coefficients in the Lagrangian, as we do below.

8.1.2 Constraints

We now turn to the constraints that can be put on the neutrino sector. For the renormalizable terms governed by a_i, b_i, c_i, d_i and the non-renormalizable $\alpha_{i,L}^{(5)}$ term that contributes a similar term in the dispersion relation the best known limits come from the miniBoone experiment [137]. They limit various combinations of coefficients at the 10^{-20} GeV level, which implies that the order of magnitude constraints (assuming no cancellations) on these coefficients are

$$\begin{aligned} |a_i|, |b_i| &\lesssim 10^{-20}\text{GeV}, \\ |c_i|, |d_i| &\lesssim 10^{-20}, \\ \alpha_{i,L}^{(5)} &\lesssim 10^{9}, \end{aligned} \tag{53}$$

where we have used the mass of the neutrino as approximately 0.1 eV and the energy of the neutrinos in the miniBoone experiment as approximately 1 GeV (it is actually slightly less). As one can see, both the renormalizable operators are very tightly constrained, while the non-renormalizable operator is essentially free.

For the higher-dimension operators that contribute higher-order corrections to the dispersion relation the best constraints to date come from the survival of atmospheric muon neutrinos observed by the former IceCube detector AMANDA-II in the energy range 100 GeV to 10 TeV [138]. AMANDA-II searched for a generic LIV in the neutrino sector [139] and achieved $(\Delta c/c)_{ij} \leq 2.8 \times 10^{-27}$ at 90 % confidence level assuming maximal mixing for some of the combinations i, j. Using the low end of the energy band (100 GeV) to be conservative, this yields order of magnitude constraints on the Lorentz violating coefficients of

$$\begin{aligned} |\zeta_{i,1}|, |\zeta_{i,2}| &\lesssim 10^{-10} \\ |\alpha_{i,L}^{(6)}| &\lesssim 10^{7}. \end{aligned} \tag{54}$$

Given that IceCube does not distinguish neutrinos from antineutrinos, the same constraints apply to the corresponding antiparticles. Of interest is that due to the strong energy dependence of the dimension-six term, as more data is taken at the 10-TeV range the constraint will drop below $O(1)$ (at 10 TeV for the energy the constraint is already $\alpha_{i,L}^{(6)} < 0.1$). The IceCube detector is expected to improve this constraint to $(\Delta c/c)_{ij} \leq 9 \times 10^{-28}$ in the next few years [140]. We also note that the lack of sidereal variations in the atmospheric neutrino flux also yields comparable constraints on some combinations of SME parameters [141], which can be translated into the framework above. Finally, a nice summary of neutrino oscillation observations, with particular attention to LIV, can be found in [142]. For a comprehensive listing of constraints on terms in the neutrino sector of the SME see [42, 73].

8.2 Neutrinos II: Species Independent Lorentz Violation

Neutrino oscillation depends on δN_{ij} and so if all the Lorentz violating terms are species independent there is no contribution to the oscillation pattern from Lorentz violation. Hence other methods must be used to constrain these terms. The model is almost the same as above. The only differences are that all neutrinos have the same Lorentz violating coefficients and therefore the a_i term can be dropped as in QED. Hence we proceed directly to the relevant constraints.

8.2.1 Constraints

Time-of-Flight

For pure time-of-flight constraints we have to date only two observations to rely on, the supernova SN1987a neutrino burst and the ICARUS experiment. We deal first with SN1987a, which was a unique event that generated the almost simultaneous (within a few hours) arrival of electronic antineutrinos and photons. Although only few electronic antineutrinos at MeV energies were detected by the experiments KamiokaII, IMB, and Baksan, it was enough to establish a constraint $(\Delta c/c)^{TOF} \lesssim 10^{-8}$ [143] or $(\Delta c/c)^{TOF} \lesssim 2 \times 10^{-9}$ [144] by looking at the difference in arrival time between antineutrinos and optical photons over a baseline distance of 1.5×10^5 light years. Further analyses of the time structure of the neutrino signal strengthened this constraint down to $\sim 10^{-10}$ [145, 146].

The scarcity of the detected neutrino did not allow the reconstruction of the full energy spectrum and of its time evolution; in this sense one should probably consider constraints purely based on the difference in the arrival time with respect to photons more conservative and robust. Adopting $\Delta c/c \lesssim 10^{-8}$, the supernova constraint implies the following order of magnitude constraints:

$$
\begin{aligned}
|b| &\lesssim 10^{-11}\text{GeV} \\
|c|, |d_i| &\lesssim 10^{-8} \\
\alpha_L^{(5)} &\lesssim 10^{21} \\
|\zeta_1|, |\zeta_2| &\lesssim 10^{13} \\
|\alpha_L^{(6)}| &\lesssim 10^{34}.
\end{aligned}
\tag{55}
$$

Hence, time-of-flight constraints are quite tight for renormalizable operators but leave the non-renormalizable operators effectively unconstrained. The Imaging Cosmic And Rare Underground Signals (ICARUS) experiment measured the time of flight of neutrinos traveling from CERN to Gran Sasso in a repeat of the OPERA experiment. ICARUS found that the arrival time was consistent with zero and within approximately 10 ns of the expected light travel time [54]. The light travel time between CERN and Gran Sasso is roughly 2.4 ms, so the $\Delta c/c$ from ICARUS is of the order of 10^{-5}, consistent with previous measurements made by the MINOS detector [147]. Therefore the SN1987A neutrinos remain the tightest constraint on Lorentz violation from time-of-flight experiments.

Threshold Reactions

Threshold reactions also can be used to cast constraints on the neutrino sector. Several processes are of interest: neutrino Čerenkov emission $\nu \to \gamma\,\nu$, neutrino splitting $\nu \to \nu\,\nu\overline{\nu}$, and neutrino electron/positron pair production $\nu \to \nu\,e^-e^+$. Let us consider for illustration the latter process as the others work similarly. Neglecting possible LV modification in the electron/positron sector (on which we have seen we have already strong constraints) the threshold energy for a dispersion modification that scales as p^n is

$$
E_{th,(n)}^2 = \frac{4m_e^2}{\delta_{(n)}}\,, \tag{56}
$$

with $\delta_{(n)} = \xi_\nu (E_{th}/M)^{n-2}$.

The rate of this reaction was first computed in [51] for $n = 2$ but can be easily generated to arbitrary n [52]. The generic energy loss time-scale then reads (dropping purely numerical factors)

$$
\tau_{\nu-\text{pair}} \simeq \frac{m_Z^4 \cos^4\theta_w}{g^4 E^5} \left(\frac{M}{E}\right)^{3(n-2)}\,, \tag{57}
$$

where g is the weak coupling and θ_w is Weinberg's angle. ICARUS found no electron-positron pair creation from the CERN neutrino beam as it passed through their detector [148], but the best constraint comes from the observation of upward-going atmospheric neutrinos up to 400 TeV by the IceCube experiment. Since the neutrinos propagated through the entire Earth to reach the IceCube detector the free path of these particles is at least longer than the Earth radius. This measurement has

been used to establish constraints on $|\zeta_1|, |\zeta_2| < 30$ while c, d are constrained at the 10^{-12} level [52]. No effective constraint can be optioned for the dimension-six operators, however in this case neutrino splitting (which has the further advantage to be purely dependent on LV on the neutrinos sector) could be used on the "cosmogenic" neutrino flux. This is supposedly created via the decay of charged pions produced by the aforementioned GZK effect. The neutrino splitting should modify the spectrum of the UHE neutrinos by suppressing the flux at the highest energies and enhancing it at the lowest ones. In [149] it was shown that future experiments like ARIANNA [150] will achieve the required sensitivity to cast a constraint of order $|\alpha_L^{(6)}| \lesssim 10^{-4}$. Note however, that the rate for neutrino splitting computed in [149] was recently recognized to be underestimated by a factor $O(E/M)^2$ [119]. Hence, the future constraints here mentioned should be recomputed and one should be able to strengthen the constraint by a few orders of magnitude.

9 Other Frameworks

Specifying which dynamical framework is employed is crucial when discussing the phenomenology of Lorentz violations. The most robust and well-motivated framework is that which we have been discussing, EFT. However, it is not the only one and, in fact, there are reasonable arguments from holography that a QG theory should not necessarily be a local field theory in the UV (cf. the discussion in [151]). Hence Lorentz violation may enter into low-energy physics in novel ways. In addition, if one believes that the fundamental QG theory should not be Lorentz invariant, then one might look for ways in which Lorentz violation might appear outside the realm of EFT and so avoid many of the constraints that exist in the EFT framework. For completeness, and because the EFT approach is nothing more than a highly reasonable, but rather arbitrary "assumption," it is worth studying and constraining additional models, given that they may evade the majority of the constraints discussed in this review.

9.1 D-brane Models

Ellis et al. developed a model [56, 58] in which modified dispersion relations derive from the Liouville string approach to quantum space-time [152]. Liouville string models [152] motivate corrections to the usual relativistic dispersion relations that are first order in the particle energies and that correspond to a vacuum refractive index $\eta = 1 - (E/M_{\rm Pl})^{\alpha}$, where $\alpha = 1$. Models with quadratic dependencies of the vacuum refractive index on energy, $\alpha = 2$, have also been considered [63].

In particular, the D-particle realization of the Liouville string approach predicts that only gauge bosons such as photons, and not charged matter particles such as electrons, might have QG-modified dispersion relations. This difference occurs

since excitations which are charged under the gauge group are represented by open strings with their ends attached to the D-brane [153], and that only neutral excitations are allowed to propagate in the bulk space transverse to the brane [154]. Thus, if we consider photons and electrons, in this model the parameter η is forced to be null, whereas ξ is free to vary. Even more importantly, the theory is CPT even, implying that vacuum is not birefringent for photons ($\xi_+ = \xi_-$).

9.2 New Relativity Theories

Lorentz invariance of physical laws relies on only a few assumptions: the principle of relativity, stating the equivalence of physical laws for non-accelerated observers, isotropy (no preferred direction) and homogeneity (no preferred location) of space-time, and a notion of precausality, requiring that the time ordering of co-local events in one reference frame be preserved [155–162].

All the realizations of LV we have discussed so far explicitly violate the principle of relativity by introducing a preferred reference frame. This may seem a high price to pay to include QG effects in low-energy physics. For this reason, it is worth exploring an alternative possibility that keeps the relativity principle but that relaxes one or more of the above postulates.

For example, relaxing the space isotropy postulate leads to the so-called very special relativity framework [163], which was later on understood to be described by a Finslerian-type geometry [164–166]. In this example, however, the generators of the new relativity group number fewer than the usual ten associated with Poincaré invariance. Specifically, there is an explicit breaking of the $O(3)$ group associated with rotational invariance.

One may wonder whether there exist alternative relativity groups with the same number of generators as special relativity. Currently, we know of no such generalization in coordinate space. However, it has been suggested that, at least in momentum space, such a generalization is possible, and it was termed "doubly" or "deformed" (to stress the fact that it still has ten generators) special relativity (DSR). Even though DSR aims at consistently including dynamics, a complete formulation capable of doing so is still missing, and present attempts face major problems. Thus, at present DSR is only a kinematic idea.

Finally, we cannot omit the recent development of what one could perhaps consider a spin-off of DSR that is relative locality, which is based on the idea that the invariant arena for classical physics is a curved momentum space rather than space-time (the latter being a derived concept) [167].

DSR and relative locality are still a subject of active research and debate (see, e.g., [168–171]); nonetheless, they have not yet attained the level of maturity needed to cast robust constraints[13].

[13] Note however, that some knowledge of DSR phenomenology can be obtained by considering that, as in special relativity, any phenomenon that implies the existence of a preferred reference

10 Discussion and Perspectives

As we have seen, for rotationally invariant QED, Lorentz symmetry is *extremely* well tested with strong constraints all the way up to dimension-six operators. Lest the reader get a false impression, we also should mention that many other sectors of the SM, from neutrons to mesons both with and without rotation invariance also have tight limits set on any possible deviation from Lorentz invariance.

There are two areas where there remains immediate work to be done. First, there is one current caveat in regards to the dimension-six operator constraints that needs to be resolved. As we have seen, the dimension-six constraints mostly rely on the physics of the GZK feature of the UHECR spectrum. More specifically, UHECR constraints rest upon the hypothesis, not in contrast with any previous experimental evidence, that protons constitute the majority of UHECRs above 10^{19} eV. Recent PAO [172] and Yakutsk [173] observations, however, showed hints of an increase of the average mass composition with rising energies up to $E \approx 10^{19.6}$ eV, although still with large uncertainties mainly due to the proton-air cross section at ultrahigh energies. Hence, *experimental data suggest that heavy nuclei can possibly account for a substantial fraction of UHECR arriving on Earth. Furthermore, the evidence for correlations between UEHCR events and their potential extragalactic sources [86]—such as AGN (mainly blazars)—has not improved with increasing statistics. This might be interpreted as a further hint that a relevant part of the flux at very high energies should be accounted for by heavy ions (mainly iron) which are much more deviated by the extra- and intergalactic magnetic fields due to their larger charge with respect to protons (an effect partially compensated by their shorter mean free path at very high energies). If consequently one conservatively decides to momentarily suspend his/her judgment about the evidence for a GZK feature, then he/she would lose the constraints at $n = 4$ on the QED sector[14] as well as very much weaken the constraints on the hadronic one.

Assuming that current hints for a heavy composition at energies $E \sim 10^{19.6}$ eV [172] may be confirmed in the future, that some UHECR is observed up to $E \sim 10^{20}$ eV [175], and that the energy and momentum of the nucleus are the sum of energies and momenta of its constituents (so that the parameter in the modified dispersion relation of the nucleus is the same of the elementary nucleons) one could place a first constraint on the absence of spontaneous decay for nuclei which could

frame is forbidden. Thus, the detection of such a phenomenon would imply the falsification of both special and DSR. An example of such a process is the decay of a massless particle.

[14] This is a somewhat harsh statement given that it was shown in [174] that a substantial (albeit reduced) high-energy gamma-ray flux is still expected also in the case of mixed composition, so that in principle the previously discussed line of reasoning based on the absence of upper threshold for UHE gamma rays might still work.

not spontaneously decay without LV.[15] Such a constraint would place bounds on subluminal protons, because in this case the energy of the emitted nucleon is lowered with respect to the LI case until it "compensates" the binding energy of the nucleons in the initial nucleus in the energy-momentum conservation. A complementary constraint, an upper limit on superluminal protons, can be obtained from the absence of VC emission. If UHECR are mainly iron at the highest energies the constraint is given by $\zeta_p \lesssim 2 \times 10^2$ for nuclei observed at $10^{19.6}$ eV (and $\zeta_p \lesssim 4$ for 10^{20} eV), while for helium it is $\zeta_p \lesssim 4 \times 10^{-3}$ [176].

Second, the gravitational sector of Lorentz symmetry violation is currently less constrained. A large region of the parameter space of Einstein-aether theory remains unconstrained, while atom interferometry is starting to probe matter-tensor couplings that are sensitive to the dynamics of the Lorentz violating tensors. Useful modified gravity theories that also evade solar system tests, such as galileons [177], or allow for better UV behavior, such as the aforementioned Hořava–Lifshitz gravity, yield interesting Lorentz symmetry violating gravitational phenomenology. Hence, fully exploring Lorentz violation in the gravitational sector is currently an important area that requires further progress.

Acknowledgements We wish to thank Luca Maccione for useful insights, discussions, and feedback on the manuscript preparation.

References

1. Mavromatos NE. CPT violation and decoherence in quantum gravity. Lect Notes Phys. 2005;669:245–320.
2. Weinberg S. Quantum contributions to cosmological correlations. Phys Rev. 2005;D72:043514. doi:10.1103/PhysRevD.72.043514.
3. Damour T, Polyakov AM. The string dilaton and a least coupling principle. Nucl Phys. 1994;B423:532–58. doi:10.1016/0550-3213(94)90143-0.
4. Barrow JD. Zichichi QA, editor. Current topics in astrofundamental physics: primordial cosmology. 1998. pp. 269.
5. Bleicher M, Hofmann S, Hossenfelder S, Stoecker H. Black hole production in large extra dimensions at the Tevatron: a chance to observe a first glimpse of TeV scale gravity. Phys Lett. 2002;B548:73–6. doi:10.1016/S0370-2693(02)02732-6.
6. Kostelecky VA. Gravity, Lorentz violation, and the standard model. Phys Rev. 2004;D69:105009. doi:10.1103/PhysRevD.69.105009.
7. Mattingly D. Modern tests of lorentz invariance. Living Rev Rel. 2005;8:5.
8. Dirac PAM. Is there an aether? Nature. 1951;168:906–7. doi:10.1038/168906a0.

[15] UHE nuclei suffer mainly from photodisintegration losses as they propagate in the intergalactic medium. Because photodisintegration is indeed a threshold process, it can be strongly affected by LV. According to [176], and in the same way as for the proton case, the mean free paths of UHE nuclei are modified by LV in such a way that the final UHECR spectra after propagation can show distinctive LV features. However, a quantitative evaluation of the propagated spectra has not been performed yet.

9. Bjorken JD. A dynamical origin for the electromagnetic field. Ann Phys. 1963;24:174–87. doi:10.1016/0003-4916(63)90069-1.
10. Phillips PR. Is the Graviton a Goldstone boson? Phys Rev. 1966;146:966. doi:10.1103/PhysRev.146.966–73.
11. Blokhintsev DI. Basis for special relativity theory provided by experiments in high energy physics. (T). Soviet Physics Uspekhi. 1966;9:405. (1966). doi:10.1070/PU1966v009n03ABEH002890.
12. Pavlopoulos TG. Breakdown of Lorentz invariance. Phys Rev. 1967;159:1106–10. doi:10.1103/PhysRev.159.1106.
13. Rédei LB. Validity of special relativity at small distances and the velocity dependence of the muon lifetime. Phys Rev. 1967;162:1299–1300. doi:10.1103/PhysRev.162.1299.
14. Nielsen HB, Ninomiya M. β-Function in a non-covariant Yang-Mills theory. Nucl Phys B. 1978;141:153–77. doi:10.1016/0550-3213(78)90341-3.
15. Ellis J, Gaillard MK, Nanopoulos DV, Rudaz S. Uncertainties in the proton lifetime. Nucl Phys B. 1980;176:61–99. doi:10.1016/0550-3213(80)90064-4.
16. Zee A. Perhaps proton decay violates Lorentz invariance. Phys Rev. 1982;D25:1864. doi:10.1103/PhysRevD.25.1864.
17. Nielsen HB, Picek I. Redei like model and testing Lorentz invariance. Phys Lett. 1982;B114:141. doi:10.1016/0370-2693(82)90133-2.
18. Chadha S, Nielsen HB. Lorentz invariance as a low energy phenomenon. Nucl Phys B. 1983;217:125–44. doi:10.1016/0550-3213(83)90081-0.
19. Nielsen HB, Picek I. Lorentz non-invariance. Nucl Phys B. 1983;211:269–96. doi:10.1016/0550-3213(83)90409-1.
20. Roth M. Measurement of the UHECR energy spectrum using data from the Surface Detector of the Pierre Auger Observatory. International Cosmic Ray Conference. 2008;4:327–30.
21. Abbasi R, et al. First observation of the Greisen-Zatsepin-Kuzmin suppression. Phys Rev Lett. 2008;100(10):101101. doi:10.1103/PhysRevLett.100.101101.
22. Bluhm R. Overview of the SME: implications and phenomenology of Lorentz violation. Lect Notes Phys. 2006;702:191–226. doi:10.1007/3-540-34523-X-8.
23. Kostelecky VA, Samuel S. Spontaneous breaking of lorentz symmetry in string theory. Phys Rev. 1989;D39:683.
24. Colladay D, Kostelecky V. Lorentz violating extension of the standard model. Phys Rev. 1998;D58:116002. doi:10.1103/PhysRevD.58.116002.
25. Kostelecky E, Alan V. Data tables for Lorentz and CPT violation. Rev Mod Phys. 2011;83:11. (ARXIV:0801.0287).
26. Amelino-Camelia G, Ellis JR, Mavromatos NE, Nanopoulos DV, Sarkar S. Potential sensitivity of gamma-ray burster observations to wave dispersion in vacuo. Nature. 1998;393:763–5.
27. Coleman SR, Glashow SL. Cosmic ray and neutrino tests of special relativity. Phys Lett. 1997;B405:249–52. doi:10.1016/S0370-2693(97)00638-2.
28. Coleman S, Glashow SL. Evading the GZK cosmic-ray cutoff. ArXiv High Energy Physics-Phenomenology, 1998. e-prints: hep-ph/9808446.
29. Coleman SR, Glashow SL. High-energy tests of Lorentz invariance. Phys Rev. 1999;D59:116008. doi:10.1103/PhysRevD.59.116008.
30. Gonzalez-Mestres L. Physical and cosmological implications of a possible class of particles able to travel faster than light. ArXiv High Energy Physics-Phenomenology e-prints: hep-ph/9610474 1996.
31. Gonzalez-Mestres L. In 19th Texas Symposium on Relativistic Astrophysics and Cosmology, ed. by J. Paul, T. Montmerle, E. Aubourg, 1998.
32. Jacobson T, Liberati S, Mattingly D. Threshold effects and Planck scale Lorentz violation: combined constraints from high energy astrophysics. Phys Rev. 2003;D67:124011. doi:10.1103/PhysRevD.67.124011.
33. Mattingly D, Jacobson T, Liberati S. Threshold configurations in the presence of Lorentz violating dispersion relations. Phys Rev. 2003;D67:124012. doi:10.1103/PhysRevD.67.124012.

34. Jacobson T, Liberati S, Mattingly D. Lorentz violation at high energy: concepts, phenomena and astrophysical constraints. Annals Phys. 2006;321:150–96.
35. Anselmi D. Weighted scale invariant quantum field theories. JHEP. 2008;0802:051. doi:10.1088/1126-6708/2008/02/051.
36. Horava P. Quantum gravity at a Lifshitz point. Phys Rev. 2009;D79:084008. doi:10.1103/PhysRevD.79.084008.
37. Visser M. Power-counting renormalizability of generalized Horava gravity. ArXiv e-prints: hep-th/0912.4757 2009.
38. Myers RC, Pospelov M. Experimental challenges for quantum gravity. Phys Rev Lett. 2003;90:211601.
39. Bolokhov PA, Pospelov M. Classification of dimension 5 Lorentz violating interactions in the standard model. Phys Rev. 2008;D77:025022. doi:10.1103/PhysRevD.77.025022.
40. Mattingly D. Have we tested Lorentz invariance enough? PoS QG-PH. 2007;026.
41. Kostelecky VA, Mewes M. Electrodynamics with Lorentz-violating operators of arbitrary dimension. Phys Rev. 2009;D80:015020. doi:10.1103/PhysRevD.80.015020.
42. Kostelecky A, Mewes M. Neutrinos with Lorentz-violating operators of arbitrary dimension. Phys Rev. 2012;D85:096005.
43. Riess AG, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. Astron J. 1998;116:1009–38. (1998). doi:10.1086/300499.
44. Perlmutter S, et al. Measurements of Omega and Lambda from 42 high redshift supernovae. Astrophys J. 1999;517:565–86. doi:10.1086/307221. (The Supernova Cosmology Project).
45. Greisen K. End to the cosmic ray spectrum? Phys Rev Lett. 1966;16:748–50. doi:10.1103/PhysRevLett.16.748.
46. Zatsepin GT, Kuz'min VA. On the interaction of cosmic rays with photons. Cosmic rays, Moscow, No. 11, pp. 45–7, vol. 11, 1969.
47. Takeda M, Hayashida N, Honda K, Inoue N, Kadota K, et al. Extension of the cosmic ray energy spectrum beyond the predicted Greisen-Zatsepin-Kuz'min cutoff. Phys Rev Lett. 1998;81:1163–6. doi:10.1103/PhysRevLett.81.1163.
48. Protheroe R, Meyer H. An infrared background TeV gamma-ray crisis? Phys Lett. 2000;B493:1–6. doi:10.1016/S0370-2693(00)01113-8.
49. Adam T, et al. Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. 2011. http://press.web.cern.ch/press/pressreleases/Releases2011/PR19.11E.html Press Release.
50. Amelino-Camelia G, Gubitosi G, Loret N, Mercati F, Rosati G, et al. OPERA-reassessing data on the energy dependence of the speed of neutrinos. Int J Mod Phys. 2011;D20:2623–40. doi:10.1142/S0218271811020780. (Some references added/ Figs.1 and 2 redrawn for better visibility of the effects of bias1979/ some sentences which had been written a bit too 'densely' in V1 are now more readable in this V2).
51. Cohen AG, Glashow SL. Pair creation constrains superluminal neutrino propagation. Phys Rev Lett. 2011;107:181803. doi:10.1103/PhysRevLett.107.181803.
52. Maccione L, Liberati S, Mattingly DM. Violations of Lorentz invariance in the neutrino sector after OPERA. JCAP. 2013;1303:039. doi:10.1088/1475-7516/2013/03/039.
53. Carmona JM, Cortes JF. Constraints from Neutrino Decay on superluminal velocities. ArXiv e-prints: hep-th/1110.0430 2011.
54. Antonello M, et al. A search for the analogue to Cherenkov radiation by high energy neutrinos at superluminal speeds in ICARUS. Phys Lett. 2012;B713:17. doi:10.1016/j.physletb.2012.05.033.
55. Amelino-Camelia G, Ellis JR, Mavromatos NE, Nanopoulos DV. Distance measurement and wave dispersion in a liouville- string approach to quantum gravity. Int J Mod Phys. 1997;A12:607–24.
56. Ellis JR, Mavromatos NE, Nanopoulos DV. A microscopic recoil model for light-cone fluctuations in quantum gravity. Phys Rev. 2000;D61:027503.
57. Ellis JR, Mavromatos NE, Nanopoulos DV. Dynamical formation of horizons in recoiling D-branes. Phys Rev. 2000;D62:084019.

58. Ellis JR, Mavromatos NE, Sakharov AS. Synchrotron radiation from the Crab Nebula discriminates between models of space-time foam. Astropart Phys. 2004;20:669–82. doi:10.1016/j.astropartphys.2003.12.001.
59. Gambini R, Pullin J. Nonstandard optics from quantum spacetime. Phys Rev. 1999;D59:124021.
60. Carroll SM, Harvey JA, Kostelecky VA, Lane CD, Okamoto T. Noncommutative field theory and Lorentz violation. Phys Rev Lett. 2001;87:141601.
61. Lukierski J, Ruegg H, Zakrzewski WJ. Classical quantum mechanics of free kappa relativistic systems. Ann Phys. 1995;243:90–116.
62. Amelino-Camelia G, Majid S. Waves on noncommutative spacetime and gamma-ray bursts. Int J Mod Phys. 2000;A15:4301–24
63. Burgess CP, Cline J, Filotas E, Matias J, Moore GD. Loop-generated bounds on changes to the graviton dispersion relation. JHEP. 2002;03:043.
64. Gasperini M. Inflation and broken Lorentz symmetry in the very early universe. Phys Lett. 1985;B163:84. doi:10.1016/0370-2693(85)90197-2.
65. Gasperini M. Broken Lorentz symmetry and the dimension of space-time. Phys Lett. 1986;B180;221–4. doi:10.1016/0370-2693(86)90299-6.
66. Gasperini M. Singularity prevention and broken Lorentz symmetry. Class Quant Gra. 1987;4:485–94.
67. Gasperini M. Lorentz noninvariance and the universality of free fall in Quasiriemannian gravity. Erice School Cosmology 1987:0181.
68. Gasperini M. Repulsive gravity in the very early universe. Gen Rel Grav. 1998;30:1703–9. doi:10.1023/A:1026606925857.
69. Mattingly D, Relativistic gravity with a dynamical preferred frame. In: Kostelecký VA, editor. CPT and Lorentz symmetry., pp. 331–5. doi:10.1142/9789812778123-0042.
70. Eling C, Jacobson T, Mattingly D. Einstein-aether theory. ArXiv General Relativity and Quantum Cosmology e-prints: gr-qc/0410001 2004.
71. Jacobson T. Einstein-aether gravity: a status report. PoS. 2007;QG-PH:020.
72. Barcelo C, Liberati S, Visser M. Analogue gravity. Living Rev Rel. 2005;8:12.
73. Kostelecky V, Russell N. Data tables for Lorentz and CPT violation. Rev Mod Phys. 2011;83:11. doi:10.1103/RevModPhys.83.11.
74. Parentani R. Constructing QFT's wherein Lorentz invariance is broken by dissipative effects in the UV. PoS. 2007;QG-PH:031.
75. Collins J, Perez A, Sudarsky D, Urrutia L, Vucetich H. Lorentz invariance: an additional fine-tuning problem. Phys Rev Lett. 2004;93:191301. doi:10.1103/PhysRevLett.93.191301.
76. Polchinski J. Comment on [arXiv:1106.1417] 'Small Lorentz violations in quantum gravity: do they lead to unacceptably large effects?'. Class Quant Grav. 2012;29:088001. doi:10.1088/0264-9381/29/8/088001.
77. Groot Nibbelink S, Pospelov M. Lorentz violation in supersymmetric field theories. Phys Rev Lett. 2005;94:081601. doi:10.1103/PhysRevLett.94.081601.
78. Bolokhov PA, Nibbelink SG, Pospelov M. Lorentz violating supersymmetric quantum electrodynamics. Phys Rev. 2005;D72:015013. doi:10.1103/PhysRevD.72.015013.
79. ATLAS-Collaboration. Public data release. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/CombinedSummary-Plots.
80. Liberati S, Visser M, Weinfurtner S. Naturalness in emergent spacetime. Phys Rev Lett. 2006;96:151301. doi:10.1103/PhysRevLett.96.151301.
81. Liberati S, Visser M, Weinfurtner S. Analogue quantum gravity phenomenology from a two-component Bose-Einstein condensate. Class Quant Grav. 2006;23:3129–54. doi:10.1088/0264-9381/23/9/023.
82. Pospelov M, Shang Y. On Lorentz violation in Horava-Lifshitz type theories. Phys Rev. 2012;D85:105001. doi:10.1103/PhysRevD.85.105001.
83. Stecker F, Glashow SL. New tests of Lorentz invariance following from observations of the highest energy cosmic gamma-rays. Astropart Phys. 2001;16:97–9. doi:10.1016/S0927-6505(01)00137-2.

84. Baccetti V, Tate K, Visser M. Lorentz violating kinematics: threshold theorems. JHEP. 2012;1203:087. doi:10.1007/JHEP 03(2012)087.
85. Kluzniak W. Transparency of the universe to TeV photons in some models of quantum gravity. Astropart Phys. 1999;11:117–8. doi:10.1016/S0927-6505(99)00070-5.
86. Abraham J, et al. Correlation of the highest energy cosmic rays with nearby extragalactic objects. Science. 2007;318:938–43. (2007). doi:10.1126/science.1151124.
87. Galaverni M, Sigl G. Lorentz violation in the photon sector and ultra-high energy cosmic rays. Phys Rev Lett. 2008;100:021102. doi:10.1103/PhysRevLett.100.021102.
88. Maccione L, Liberati S. GZK photon constraints on Planck scale Lorentz violation in QED. JCAP. 2008;0808:027. doi:10.1088/1475-7516/2008/08/027.
89. Gelmini G, Kalashev OE, Semikoz DV. GZK photons as ultra high energy cosmic rays. J Exp Theor Phys. 2008;106:1061–82. doi:10.1134/S106377610806006X.
90. Abraham J, et al. Upper limit on the cosmic-ray photon flux above 10**19 eV using the surface detector of the Pierre Auger observatory. Astropart Phys. 2008;29:243–56. doi:10.1016/j.astropartphys.2008.01.003.
91. Rubtsov GI, et al. Upper limit on the ultra-high-energy photon flux from AGASA and Yakutsk data. Phys Rev. 2006;D73:063009. doi:10.1103/PhysRevD.73.063009.
92. Galaverni M, Sigl G. Lorentz violation and ultrahigh-energy photons. Phys Rev. 2008;D78:063003. doi:10.1103/PhysRevD.78.063003.
93. Liberati S, Maccione L. Lorentz Violation: Motivation and new constraints. Ann Rev Nucl Part Sci. 2009;59:245–67. doi:10.1146/annurev.nucl.010909.083640.
94. Jacobson TA, Liberati S, Mattingly D, Stecker FW. New limits on Planck scale Lorentz violation in QED. Phys Rev Lett. 2004;93:021101.
95. Ellis JR, Mavromatos NE, Nanopoulos DV, Sakharov AS, Sarkisyan EKG. Robust limits on Lorentz violation from gamma-ray Bursts. Astropart Phys. 2006;25:402–11.
96. Albert J, et al. Probing quantum gravity using photons from a flare of the active galactic nucleus Markarian 501 Observed by the MAGIC telescope. Phys Lett. 2008;B668:253–7. doi:10.1016/j.physletb.2008.08.053.
97. Ackermann M, et al. A limit on the variation of the speed of light arising from quantum gravity effects. Nature. 2009;462:331–4. doi:10.1038/nature08574.
98. Maccione L, Liberati S, Celotti A, Kirk JG, Ubertini P. Gamma-ray polarization constraints on Planck scale violations of special relativity. 2008.
99. Gleiser RJ, Kozameh CN. Astrophysical limits on quantum gravity motivated birefringence. Phys Rev. 2001;D64:083007.
100. McMaster WH. Matrix representation of polarization. Rev Mod Phys. 1961;33(1):8. doi 10.1103/RevModPhys.33.8.
101. Parmar AN, Winkler C, Barr P, Hansson L, Kuulkers E, Much R, Orr A. INTEGRAL mission. In: X-ray and gamma-ray telescopes and instruments for astronomy. Edited by Joachim E. Truemper, Harvey D. Tananbaum. Proceedings of the SPIE, Volume 4851, pp. 1104–12; 2003.
102. Dean AJ, Clark DJ, Stephen JB, McBride VA, Bassani L, Bazzano A, Bird AJ, Hill AB, Shaw SE, Ubertini P. Science. Polarized gamma-ray emission from the crab. 2008;321:1183. doi:10.1126/science.1149056.
103. McGlynn S, et al. Polarisation studies of the prompt gamma-ray emission from GRB 041219a using the spectrometer aboard INTEGRAL. Astron Astrophys 2007;466:895–904. doi:10.1051/0004-6361:20066179.
104. Petri J, Kirk JG. Polarization of high-energy pulsar radiation in the striped wind model. Astrophys J. 2005;627:L37–40.
105. Weisskopf MC, Silver EH, Kestenbaum HL, Long KS, Novick R. A precision measurement of the X-ray polarization of the Crab Nebula without pulsar contamination. ApJL. 1978;220:L117–21. doi:10.1086/182648.
106. Fan YZ, Wei DM, Xu D. Gamma-ray burst UV/optical afterglow polarimetry as a probe of quantum gravity. Mon Not Roy Astron Soc. 2006;376:1857–60.
107. Ng CY, Romani RW. Fitting pulsar wind tori. Astrophys J. 2004;601:479–84.

108. Kanbach G, Slowikowska A, Kellner S, Steinle H. New optical polarization measurements of the Crab pulsar. AIP Conf Proc. 2005;801:306–11.
109. Mitrofanov IG. Astrophysics (communication arising): a constraint on canonical quantum gravity? Nature. 2003;426:139.
110. Coburn W, Boggs SE. Polarization of the prompt γ-ray emission from the γ-ray burst of 6 December 2002. Nature. 2003;423:415–7.
111. Stecker FW. A new Limit on Planck Scale Lorentz Violation from Gamma-ray Burst Polarization. Astropart Phys. 2011;35:95. doi:10.1016/j.astropartphys.2011.06.007.
112. Laurent P, Gotz D, Binetruy P, Covino S, Fernandez-Soto A. Constraints on Lorentz invariance violation using INTEGRAL/IBIS observations of GRB041219A. 2011.
113. Altschul B. Lorentz violation and synchrotron radiation. Phys Rev. 2005;D72:085003. doi:10.1103/PhysRevD.72.085003.
114. Jacobson T, Liberati S, Mattingly D. Lorentz violation and Crab synchrotron emission: a new constraint far beyond the Planck scale. Nature. 2003;424:1019–21.
115. Maccione L, Liberati S, Celotti A, Kirk J. New constraints on Planck-scale Lorentz violation in QED from the Crab Nebula. J Cosmol Astropart Phys. 2007;2007(10):013.
116. Kirk JG, Lyubarsky Y, Petri J. The Theory of Pulsar Winds and Nebulae. In: Becker W, editor. Astrophysics and Space Science Library. vol. 357 of Astrophysics and Space Science Library; 2009. p. 421.
117. Aharonian F, et al. The Crab nebula and pulsar between 500-GeV and 80-TeV: observations with the HEGRA stereoscopic air Cherenkov telescopes. Astrophys J. 2004;614:897–913.
118. Gelmini G, Nussinov S, Yaguna CE. On photon splitting in theories with Lorentz invariance violation. JCAP. 2005;0506:012. doi:10.1088/1475-7516/2005/06/012.
119. Ward B. Estimates of radiation by Superluminal neutrinos. Phys Rev. 2012;D85:073007.
120. Yao WM, et al. Review of particle physics. J Phys. 2006;G33:1–1232. doi:10.1088/0954-3899/33/1/001.
121. Heckel BR, Adelberger E, Cramer C, Cook T, Schlamminger S, et al. Preferred-frame and CP-violation tests with polarized electrons. Phys Rev. 2008;D78:092006. doi:10.1103/PhysRevD.78.092006.
122. Kostelecky AV, Tasson JD. Matter-gravity couplings and Lorentz violation. Phys Rev. 2011;D83:016013. doi:10.1103/PhysRevD.83.016013.
123. Altschul B. Bounds on spin-dependent Lorentz violation from inverse Compton observations. Phys Rev. 2007;D75:041301. doi:10.1103/PhysRevD.75.041301.
124. Klinkhamer F, Schreck M. New two-sided bound on the isotropic Lorentz-violating parameter of modified-Maxwell theory. Phys Rev. 2008;D78:085026. doi:10.1103/PhysRevD.78.085026.
125. Arkani-Hamed N, Cheng HC, Luty MA, Mukohyama S. Ghost condensation and a consistent infrared modification of gravity. JHEP. 2004;05:074.
126. Jacobson T, Mattingly D. Gravity with a dynamical preferred frame. Phys Rev. 2001;D64:024028. doi:10.1103/PhysRevD.64.024028.
127. Bluhm R, Gagne NL, Potting R, Vrublevskis A. Constraints and stability in vector theories with spontaneous Lorentz violation. Phys Rev. 2008;D77:125007. doi:10.1103/PhysRevD.77.125007, 10.1103/PhysRevD.79.029902.
128. Kostelecky VA, Tasson J. Prospects for large relativity violations in matter-gravity couplings. Phys Rev Lett. 2009;102:010402. doi:10.1103/Phys Rev Lett.102:010402.
129. Foster BZ, Jacobson T. Post-Newtonian parameters and constraints on Einstein-aether theory. Phys Rev. 2006;D73:064015. doi:10.1103/PhysRevD.73.064015.
130. Will CM. The confrontation between general relativity and experiment. Living Rev Rel. 2005;9:3.
131. Jacobson T, Mattingly D. Einstein-aether waves. Phys Rev. 2004;D70:024003. doi:10.1103/PhysRevD.70.024003.
132. Chatziioannou K, Yunes N, Cornish N. Model-independent test of general relativity: an extended post-Einsteinian framework with complete polarization content. Phys Rev. 2012;D86:022004. doi:10.1103/PhysRevD.86.022004.

133. Elliott JW, Moore GD, Stoica H. Constraining the new aether: gravitational Cherenkov radiation. JHEP. 2005;08:066.
134. Hohensee MA, Chu S, Peters A, Muller H. Equivalence principle and gravitational redshift. Phys Rev Lett. 2011;106:151102. doi 10.1103/PhysRevLett.106.151102.
135. Carroll SM, Lim EA. Lorentz-violating vector fields slow the universe down. Phys Rev. 2004;D70:123525. doi:10.1103/PhysRevD.70.123525.
136. Muller H, Peters A, Chu S. A precision measurement of the gravitational redshift by the interference of matter waves. Nature. 2010;463:926–29. doi:10.1038/nature08776.
137. Aguilar-Arevalo AA, et al. Test of Lorentz and CPT violation with short baseline neutrino oscillation excesses. Phys Lett. 2013;B718:1303. doi:10.1016/j.physletb.2012.12.020.
138. Kelley J. Searching for quantum gravity with high energy atmospheric neutrinos and AMANDA-II. Nucl Phys. 2009;A827:507C–9C. doi:10.1016/j.nuclphysa.2009.05.110.
139. Gonzalez-Garcia M, Maltoni M. Atmospheric neutrino oscillations and new physics. Phys Rev. 2004;D70:033010. doi:10.1103/PhysRevD.70.033010.
140. Huelsnitz W, Kelley J. Search for quantum gravity with IceCube and high energy atmospheric neutrinos. Proceedings of the 31st ICRC, LODZ 2009.
141. Abbasi R, et al. Search for a Lorentz-violating sidereal signal with atmospheric neutrinos in IceCube. Phys Rev. 2010;D82:112003. doi:10.1103/PhysRevD.82.112003.
142. Diaz JS, Kostelecky A. Phys Rev. 2012;D85:016013.
143. Stodolsky L. Lorentz- and CPT-violating models for neutrino oscillations. Phys Lett. 1988;B201:353. doi:10.1016/0370-2693(88)91154-9.
144. Longo MJ. New precision tests of the Einstein equivalence principle from SN1987A. Phys Rev Lett. 1988;60:173. doi:10.1103/PhysRevLett.60.173.
145. Ellis JR, Harries N, Meregaglia A, Rubbia A, Sakharov A. Probes of Lorentz violation in neutrino propagation. Phys Rev. 2008;D78:033013. doi:10.1103/PhysRevD.78.033013.
146. Sakharov A, Ellis J, Harries N, Meregaglia A, Rubbia A. Exploration of possible quantum gravity effects with neutrinos II: Lorentz violation in neutrino propagation. J Phys Conf Ser. 2009;171:012039. doi 10.1088/1742-6596/171/1/012039
147. Adamson P, et al. Measurement of neutrino velocity with the MINOS detectors and NuMI neutrino beam. Phys Rev. 2007;D76:072005. doi:10.1103/PhysRevD.76.072005.
148. Antonello M, et al. A search for the analogue to Cherenkov radiation by high energy neutrinos at superluminal speeds in ICARUS. Phys Lett. 2012;B711:270–5.
149. Mattingly DM, Maccione L, Galaverni M, Liberati S, Sigl G. Possible cosmogenic neutrino constraints on Planck-scale Lorentz violation. JCAP. 2010;1002:007. doi:10.1088/1475-7516/2010/02/007.
150. Barwick SW. ARIANNA: a new concept for UHE neutrino detection. J Phys Conf Ser. 2007;60:276–83. doi:10.1088/1742-6596/60/1/060.
151. Shomer A. A pedagogical explanation for the non-renormalizability of gravity. ArXiv e-prints: hep-th/0709.3555 2007.
152. Ellis JR, Mavromatos NE, Nanopoulos DV. String theory modifies quantum mechanics. Phys Lett. 1992;B293:37–48. doi:10.1016/0370-2693(92)91478-R.
153. Polchinski J. TASI lectures on D-branes. ArXiv High Energy Physics—Theory e-prints: hep-th/9611050 1996.
154. Ellis JR, Mavromatos NE, Nanopoulos DV, Sakharov AS. Space-time foam may violate the principle of equivalence. Int J Mod Phys. 2004;A19:4413–30. doi:10.1142/S0217751X04019780.
155. von Ignatowsky W. Einige allgemeine Bemerkungen zum Relativitätsprinzip. Verh Deutsch Phys Ges. 1910;12:788–96.
156. von Ignatowsky W. Einige allgemeine Bemerkungen zum Relativitätsprinzip. Phys Zeitsch. 1910;11:972–6.
157. von Ignatowsky W. Das Relativitätsprinzip. Arch Math Phys. 1911;3(17):1–24.
158. von Ignatowsky W. Das Relativitätsprinzip. Arch Math Phys. 1911;3(18):17–41.
159. von Ignatowsky W. Eine Bemerkung zu meiner Arbeit ‘Einige allgemeine Bemerkungen zum Relativitätsprinzip’. Phys Zeitsch. 1911;12:779.

160. Liberati S, Sonego S, Visser M. Faster-than-c signals, special relativity, and causality. Annals Phys. 2002;298:167–85. doi:10.1006/aphy.2002.6233.
161. Sonego S, Pin M. Foundations of anisotropic relativistic mechanics. J Math Phys. 2009;50:042902. doi:10.1063/1.3104065.
162. Baccetti V, Tate K, Visser M. Inertial frames without the relativity principle: breaking Lorentz symmetry. ArXiv e-prints: gr-qc/1302.5989 2013.
163. Cohen AG, Glashow SL. Very special relativity. Phys Rev Lett. 2006;97:021601. doi:10.1103/
PhysRevLett.97.021601.
164. Bogoslovsky G. Subgroups of the group of generalized Lorentz transformations and their geometric invariants. SIGMA. 2005;1:017. doi:10.3842/SIGMA.2005.017.
165. Bogoslovsky GY. Lorentz symmetry violation without violation of relativistic symmetry. Phys Lett. 2006;A350:5–10. doi:10.1016/j.physleta.2005.11.007.
166. Gibbons GW, Gomis J, Pope CN. General very special relativity is Finsler geometry. Phys Rev D. 2007;76(8):081701. doi:10.1103/PhysRevD.76.081701. http://link.aps.org/abstract/PRD/v76/e081701.
167. Amelino-Camelia G, Freidel L, Kowalski-Glikman J, Smolin L. The principle of relative locality. Phys Rev. 2011;D84:084010. doi:10.1103/PhysRevD.84.084010.
168. Smolin L. Could deformed special relativity naturally arise from the semiclassical limit of quantum gravity? ArXiv e-prints: hep-th/0808.3765 2008.
169. Rovelli C. A note on DSR. ArXiv e-prints: gr-qc/0808.3505 2008.
170. Amelino-Camelia G, Freidel L, Kowalski-Glikman J, Smolin L. Relative locality and the soccer ball problem. Phys Rev. 2011;D84:087702. doi:10.1103/PhysRevD.84.087702.
171. Hossenfelder S. Comment on arXiv:1104.2019, 'Relative locality and the soccer ball problem,' by Amelino-Camelia et al. Phys Rev. 2013;D88:028701. doi:10.1103/PhysRevD.88.028701.
172. Abraham J, et al. Measurement of the depth of maximum of extensive air showers above 10^{18} eV. Phys Rev Lett. 2010;104:091101. doi:10.1103/PhysRevLett.104.091101.
173. Glushkov A, Makarov I, Pravdin M, Sleptsov I, Gorbunov D, et al. Muon content of ultrahigh-energy air showers: Yakutsk data versus simulations. JETP Lett. 2008;87:190–4. doi:10.1134/S0021364008040024. (The Yakutsk EAS Array).
174. Hooper D, Taylor AM, Sarkar S. Cosmogenic photons as a test of ultra-high energy cosmic ray composition. Astropart Phys. 2011;34:340–3. doi:10.1016/j.astropartphys.2010.09.002.
175. Abraham J, et al. Measurement of the energy spectrum of cosmic rays above 10^{18} eV using the Pierre Auger Observatory. Phys Lett. 2010;B685:239–46. doi 10.1016/j.physletb.2010.02.013.
176. Saveliev A, Maccione L, Sigl G. Lorentz invariance violation and chemical composition of ultra high energy cosmic rays. JCAP. 2011;1103:046. doi:10.1088/1475-7516/2011/03/046.
177. Nicolis A, Rattazzi R, Trincherini E. The Galileon as a local modification of gravity. Phys Rev. 2009;D79:064036. doi:10.1103/PhysRevD.79.064036.

Possible Low-Energy Manifestations of Strings and Gravity

Ignatios Antoniadis

Abstract Despite the important experimental success of general relativity, there are several theoretical reasons indicating that gravitational phenomena may change radically from the predictions of Einstein's theory at very short distances. A main motivation comes from studies of unifying all fundamental forces in the framework of a consistent quantum theory, called string theory. In the first part of my lectures (Sects. 1–6), I discuss the main motivations and directions for physics beyond the Standard Model of elementary particles and I give a short introduction on perturbative string theory. This theory introduces a new physical constant, the string length, under which a new fundamental structure appears, changing drastically the laws of nature. In particular, lowering the string scale in the TeV region provides a theoretical framework for solving the so-called mass hierarchy problem: the apparent weakness of gravity can be accounted for by the existence of large internal dimensions, in the submillimeter region, and transverse to a braneworld where our observed universe is confined. In the second part of my lectures (Sects. 7–11), I describe the main properties of this scenario and its implications for new gravitational phenomena that can be observed at both particle colliders, and in non-accelerator experiments searching for new short-range forces.

1 Departure from the Standard Model

1.1 Why Beyond the Standard Model?

Almost all research in theoretical high-energy physics of the past 30 years has been concentrated on the quest for the theory that will replace the Standard Model (SM) as the proper description of nature at the energies beyond the TeV scale. Thus, it is important to know if this quest is justified or not, given that the SM is a very successful theory that has offered us some of the most striking agreements between experimental data and theoretical predictions. We will start by briefly examining

I. Antoniadis (✉)
Department of Physics, CERN-Theory Division, 1211 Geneva 23, Switzerland
e-mail: ignatios.antoniadis@cern.ch

CPHT (UMR CNRS 7644) Ecole Polytechnique, 91128 Palaiseau, France

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2_12

the reasons for leaving behind such a successful SM. Actually there is a variety of such arguments, both theoretical and experimental, that lead to the undoubtable conclusion that the SM should be only an effective theory of a more fundamental one.

We briefly mention some of the most important arguments. First, the SM does not include gravity. It says absolutely nothing about one of the four fundamental forces of nature. Another is the popular "mass hierarchy" problem. Within the SM, the mass of the Higgs particle is extremely sensitive to any new physics at higher energies and its natural value is of the order of the Planck mass, if the SM is valid up to that scale. This is several orders of magnitude higher than the electroweak scale implied by experiment. The way to cure this discrepancy within the SM requires an incredible fine-tuning of parameters. Also, the SM does not explain why the charges of elementary particles are quantized. Other equally important reasons include that the SM does not describe the dark matter or the dark energy of the universe. It does not explain the observed neutrino masses and oscillations, and does not predict any gauge coupling unification, suggested by experiments.

Here, we describe some of these arguments in more detail, starting from gravity. If we follow the intuition that the fundamental theory should describe within the same quantum framework all forces of nature, we need to extend the SM in a way that it consistently includes this force. However, this is particularly difficult for a variety of reasons. The most serious is that gravity cannot be quantized consistently as a field theory. Another one is that the new theory has to provide a natural explanation for the apparent weakness of gravity compared to the other three fundamental interactions.

The gravitational force between two particles of mass m is

$$F_{\rm grav} = G_N \frac{m^2}{r^2}, \qquad G_N^{-1/2} = M_{\rm Planck} = 10^{19}\ {\rm GeV}. \tag{1}$$

If we compare this with the electric force

$$F_{\rm el} = \frac{e^2}{r^2}, \tag{2}$$

for, say, the proton, we get

$$\frac{F_{\rm grav}}{F_{\rm el}} = \frac{G_N m_{\rm proton}^2}{e^2} \simeq 10^{-40}. \tag{3}$$

So, gravity is an extremely weak force at low energies (or equivalently large distances) even compared to the weak force as is shown in Table 1.

The distance at which gravitation becomes comparable to the other interactions is 10^{-33} cm, the Planck length. The energy that corresponds to this length is of the order of the Planck mass and is $\sim 10^{15}$ times the Large Hadron Collider (LHC) energy! Thus, another great difficulty is that even if we manage to construct such a model, its phenomenological verification will be extremely difficult.

Table 1 Fundamental forces

Force	Range	Intensity for two protons	Intensity at 10^{-16} cm
Gravitation	∞	10^{-38}	10^{-30}
Electromagnetic	∞	10^{-2}	10^{-2}
Weak (radioactivity β)	10^{-15} cm	10^{-5}	10^{-2}
Strong (nuclear forces)	10^{-12} cm	1	10^{-1}

Next, we describe the mass hierarchy problem. The mass of the Higgs particle is very sensitive to high-energy physics. For example, the main one-loop radiative corrections from the virtual exchange of top quarks and the Higgs particle itself are

$$\mu_{\text{eff}}^2 = \mu_{\text{bare}}^2 + \left(\frac{\lambda}{8\pi^2} - \frac{3\lambda_t^2}{8\pi^2}\right)\Lambda^2 + \cdots, \tag{4}$$

where, λ is the Higgs quartic coupling, λ_t is the top Yukawa coupling, and Λ is the ultraviolet (UV) cutoff set by the scale where new physics appears. From the masses of all other particles and for reasons that we will explain later, we know that the Higgs mass should be at the weak scale. Requiring the validity of the SM at energies $\Lambda \gg \mathcal{O}(100)$ GeV imposes an "unnatural" order by order fine-tuning between the bare mass parameter μ_{bare}^2 in the Lagrangian and the radiative corrections.

For example, for $\Lambda \sim \mathcal{O}(M_{\text{Planck}}) \sim 10^{19}$ GeV and a loop factor of the order of 10^{-2}, we get $\mu_{1-\text{loop}}^2 \sim 10^{-2} \times 10^{38} = \pm 10^{36}$ (GeV)^2. Thus, we need $\mu_{\text{bare}}^2 \sim \mp 10^{36}$ (GeV)$^2 + 10^4$ (GeV)^2. An incredible adjustment is then required, at the level of 1 part per 10^{32}: $\mu_{\text{bare}}^2/\mu_{1-\text{loop}}^2 = -1 \pm 10^{-32}$. Even more, at the next and all higher orders, a new adjustment is required. The correction will be of the order of $\mathcal{O}(100)$ GeV only beyond 17 loops: $\left(10^{-2}\right)^N \times 10^{38} \leq 10^4 \Rightarrow N \geq 17$ loops. If we want to avoid these unnatural adjustments we need $10^{-2}\Lambda^2 \leq 10^4$ (GeV)^2 $\Rightarrow \Lambda \leq 1$ TeV. The resolution of the mass hierarchy problem should reside within the energy reach of LHC!

Another open question is the charge quantization: All observed color singlet states have integer electric charges. The symmetry group of the SM is $SU(3) \times SU(2)_L \times U(1)_Y$ and the electric charge is given by: $Q = T_3 + Y$. The representations of the symmetry group of all fields are:

$$q = (3,2)_{1/6} \qquad q = \begin{pmatrix} u_{2/3} \\ d_{-1/3} \end{pmatrix}$$

$$u^c = (\bar{3},1)_{-2/3}$$

$$d^c = (\bar{3},1)_{1/3}$$

$$\ell = (1,2)_{-1/2} \qquad \ell = \begin{pmatrix} \nu_0 \\ e_{-1} \end{pmatrix}$$

$$e^c = (1,1)_1,$$

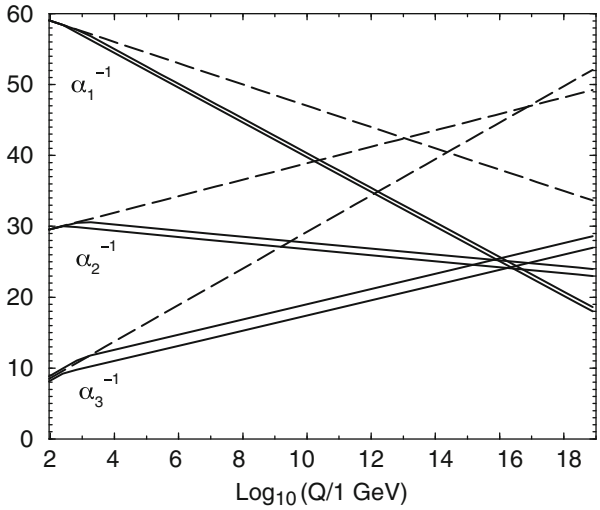


Fig. 1 Gauge coupling evolution of SM *(dashed lines)* versus the supersymmetric SM *(solid lines)*

where the subscripts denote the hypercharges in the left column and the electric charges in the right one. However, the SM does not tell us anything about this choice of the hypercharges that guarantees the observed electric charge quantization.

Also, the experimental indication of gauge coupling unification is a feature of nature that asks for an explanation. The renormalization group evolution of the three SM gauge couplings $\alpha_i = \frac{g_i^2}{4\pi}$ is given by:

$$\frac{d\alpha_i}{d\ln Q} = -\frac{b_i}{2\pi}\alpha_i^2 \quad \Rightarrow \quad \alpha_i^{-1}(Q) = \alpha_i^{-1}(Q_0) - \frac{b_i}{2\pi}\ln\frac{Q}{Q_0}, \tag{5}$$

where Q is the energy scale and b_i are the one-loop beta-function coefficients. The extrapolation of the low-energy experimental data at high energies under the "desert" assumption indicates an approximate unification of all couplings at energies of order $M_{GUT} \simeq 10^{15} \div 10^{16}$ GeV. By doing a more precise analysis however, using the current experimental precision, one finds that the SM fails to predict such a unification (see Fig. 1).

One more reason for going beyond the SM is the existence of dark matter. Astronomical observations tell us that the ordinary baryonic matter is only a tiny fraction of the energy of the universe. There are observations that point toward the fact that a kind of nonluminous matter is out there and that is actually much more abundant than baryonic matter, consisting of around 25 % of the total energy density of the universe. A natural explanation is that dark matter consists of a new kind of particles that are stable, massive at the electroweak scale, and weakly interacting.

We could mention many other important reasons that drive us beyond the SM, like neutrino masses, dark energy (see Fig. 2), and so on, but we will stop here.

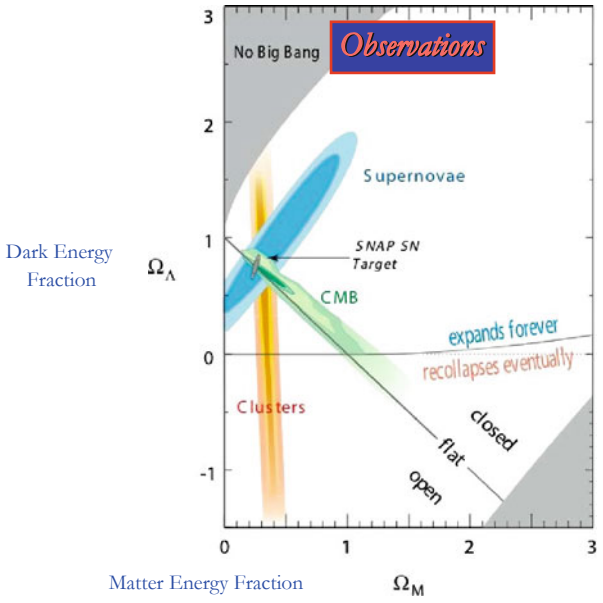


Fig. 2 Observational constraints on dark matter and dark energy [1]

Even this brief and incomplete presentation shows that there are a lot of fundamental questions that the SM cannot address so the search for a new fundamental theory is definitely justified.

Modifying the SM though is not as easy as it might seem. If we try to identify the scale Λ of new physics by looking at what energies there are deviations from the SM predictions by observations, we will be disappointed. With the exception of dark matter and the Higgs mass, all experimental indications suggest that Λ is very high. For example, the "see-saw" mechanism for the neutrino masses gives a very large scale: $m_\nu \sim \frac{v^2}{M} \sim 10^{-2}$ eV $\Rightarrow M \sim 10^{13}$ GeV, where v is the vacuum expectation value (VEV) of the SM Higgs. Similarly, the scale of gauge coupling unification is even higher: $M_{\rm GUT} \sim 10^{16}$ GeV.

Even more, modifying the SM in a way that solves the hierarchy problem is highly restricted by numerous experiments. Any new interactions could violate various SM predictions that have been experimentally tested with good accuracy such as lepton and baryon number conservation, absence of flavor changing neutral currents (FCNC), charge parity (CP) violation etc. For example, theories with baryon number violating operators like $\frac{1}{\Lambda_B^2}\left(\bar{q}\gamma^\mu u^c\right)\left(\bar{\ell}\gamma_\mu d^c\right)$ can lead to proton decay: $p = [uud] \rightarrow (e^+[\bar{d}d]) = e^+\pi^0, \quad uu \rightarrow e^+\bar{d}$. The bound from the Superkamiokande experiment on the proton lifetime of $\tau_p \geq 2.6 \times 10^{33}$ years highly restricts the allowed scale: $\Lambda_B \geq 10^{15}$ GeV. Similarly, for the lepton number, in this case, the lowest dimensional operator that violates lepton number has dimension

5: $\frac{1}{\Lambda_L}(\ell H^*)^2$, where H is the Higgs field. This operator leads in particular to Majorana neutrino masses. However, as we said before, neutrino masses are very low, $m_\nu = \frac{v^2}{\Lambda_L} \leq 0.1$ eV. With a VEV for the Higgs at $\mathcal{O}(100)$ GeV, the scale Λ_L is also constrained to very high values: $\Lambda_L \geq 10^{14}$ GeV. Finally, FCNC operators like $\frac{1}{\Lambda_F^2}\left(\bar{d}_L\gamma_\mu s_L\right)^2$ are restricted by the measured value of $K_0 - \overline{K}_0$ mixing to be $\Lambda_F \geq 10^6$ GeV. It is very challenging to construct a model that provides us with a solution to the theoretical problems mentioned above and at the same time to comply with all experimental requirements.

1.2 Directions Beyond the Standard Model

One way to encapsulate in the low-energy regime the effects of a more fundamental theory is to use effective operators. These can be produced by integrating out heavy states in the fundamental theory. However, we can always write down generic effective operators even when that theory is unknown. We just require Lorentz symmetry and gauge invariance of the low-energy effective action. These terms need not be renormalizable because by definition, the validity of an effective theory holds only up to the scale where the high-energy physics appears.

Here we show an example of integrating out a heavy gauge boson of mass M

$$-\frac{1}{4}F_{\mu\nu}^2 - \frac{1}{2}M^2A_\mu^2 + g\bar{\psi}\not{A}\psi \quad \Rightarrow \quad \frac{g^2}{M^2}\left(\bar{\psi}\gamma^\mu\psi\right)^2. \tag{6}$$

It gives rise to a six-dimensional (6D) non-renormalizable four-fermion effective operator. This is, for instance, the case in the popular Fermi theory obtained by integrating out at low energies $E \ll m_W$ the heavy $SU(2)$ gauge bosons $W^\pm$:

$$\mathcal{L}_{\text{eff}} = G_F\left(\bar{q}_L\gamma_\mu q_L + \bar{\ell}_L\gamma_\mu\ell_L\right)^2 + \cdots, \tag{7}$$

with $G_F = g_2^2/m_W^2$.

Generally, effective field theory parametrizes our ignorance of the high-energy physics by introducing local effective operators O_n^i of dim $(4+n)$

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_n^i}{M^n}O_n^i, \tag{8}$$

for $E \ll M$; M is the new physics scale where the heavy states become dynamical. If this is not too far away from the electroweak scale, the lowest dimensional (in powers of M) operators O_n^i can significantly affect the low-energy physics. They can shift the mass of the Higgs or violate symmetries of the SM such as baryon number, lepton number, induce FCNC and important CP violation, etc. as we mentioned previously.

Effective field theory offers a systematic but blind way to explore the physics beyond the SM. The alternative is to propose a totally new theory and study its consequences. There are several such proposals. They all cure some major weaknesses of the SM but they also have their own drawbacks. Keeping the mass hierarchy as guidance, the following proposals exist so far:

1. Compositeness
2. Symmetry:
   - Supersymmetry (SUSY)
   - Higgs as pseudo-Goldstone boson, little Higgs
   - Conformal
   - Higgs as component of a higher-dimensional gauge field
3. Low-UV cutoff:
   - Low-scale gravity $\Rightarrow$ · large extra dimensions
     · warped extra dimensions
   - Low-string scale $\Rightarrow$ · low-scale gravity
     · ultra-weak string coupling
   - Large N degrees of freedom
   - Higgsless models
4. Live with the hierarchy:
   - Landscape of vacua, environmental selection
   - Split SUSY

We will briefly describe only the first two, and later we will focus on strings and extra dimensions.

The first proposal for physics beyond the SM is "compositeness." This idea postulates the existence of new strong dynamics at the energy range of TeV. The Higgs field which breaks the electroweak symmetry is then a bound state of fermion bilinears just like the pions that break chiral symmetry in quantum chromodynamics (QCD). A concrete model realizing this idea is called "technicolor." However, most technicolor models have conflicts with electroweak precision measurements, leading in particular to the appearance of FCNCs, and are generally disfavored.

Another idea is to introduce a new symmetry. This provides a natural way to keep the mass of the Higgs small and solve the mass hierarchy. After all, there are various examples where symmetries have been used to naturally obtain small masses for fermions and bosons. In the case of fermions, it is the chiral symmetry as it can be seen in the following simple Lagrangian

$$\mathcal{L}_F = i\bar{\psi}_L \not{D}\psi_L + i\bar{\psi}_R \not{D}\psi_R + m\left(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L\right). \tag{9}$$

When $m = 0$ there is an enhanced symmetry $\psi_L \to e^{i\theta_1}\psi_L$; $\psi_R \to e^{i\theta_2}\psi_R$ which breaks when the mass terms are included. Thus, imposing this symmetry on our model, one forbids any mass terms in the Lagrangian. Moreover, in their

presence, since chiral symmetry is "softly" broken, radiative corrections behave like $\delta m \propto g^2 m$ where g is the gauge coupling (e.g., in quantum electrodynamics (QED) with $\psi \equiv$ electron and $g \equiv e$). An alternative reason to understand why there is no $g^2\Lambda$ term is because in a relativistic quantum field theory, any linear divergence cancels between the electron and positron contributions.

The corresponding symmetry for vector bosons is the gauge invariance. Consider, for instance, a massive abelian vector field:

$$\mathcal{L}_V = -\frac{1}{4}F_{\mu\nu}^2 - \frac{1}{2}m^2 A_\mu^2. \tag{10}$$

We see again that $m = 0$ respects $A_\mu \to A_\mu + \partial_\mu \omega$. This is, for instance, how the gauge invariance of QED forbids any mass for the photon.

Is there a symmetry that can protect scalar masses, too? We can introduce, for example, a shift symmetry, implying that the Lagrangian should depend only on derivatives of the scalar field:

$$\phi \to \phi + c \quad \Rightarrow \quad \mathcal{L}\left(\partial_\mu \phi\right) = \frac{1}{2}(\partial_\mu \phi)^2 + \frac{1}{\Lambda^4}\left[(\partial_\mu \phi)^2\right]^2 + \cdots . \tag{11}$$

This is the case for a Goldstone boson without potential and in particular it does not hold when a quartic coupling is included. Scenarios like the "little Higgs" take advantage of this property and treat the Higgs as a pseudo-Goldstone boson. In this case, the shift symmetry is broken by new gauge interactions that generate quartic Higgs interactions at higher orders. Another possibility is to use scale invariance:

$$x^\mu \to a x^\mu \quad \varphi_d \to a^{-d}\varphi(ax), \tag{12}$$

where d is the conformal dimension of the field φ_d and $d = 1$ for a scalar field. This symmetry allows a quartic coupling but in a renormalizable theory is hardly broken by radiative corrections. In order to take advantage of this symmetry, the SM should be embedded in a conformally invariant theory at the TeV scale.

Finally, we can use a different strategy by postulating a new symmetry that connects scalars to fermions or gauge fields:

$$\delta\phi = \xi\psi \quad \text{or} \quad \delta\phi = \varepsilon_\mu A^\mu. \tag{13}$$

Then the chiral or gauge invariance that protects the fermion or the gauge boson masses should also protect the masses of their scalar partners. This idea leads us to SUSY or to theories with extra dimensions, correspondingly. Indeed, the second case leads to the identification of the Higgs field with an internal component of a higher-dimensional gauge field.

1.3 Advantages and Problems of Supersymmetry

We would like to conclude this introductory section by summarizing the main advantages and drawbacks of SUSY. First of all, SUSY offers a solution to the

hierarchy problem predicting a Higgs mass that can be reached by colliders like LHC. Moreover, SUSY offers an excellent candidate for dark matter, the lightest supersymmetric particle. Another attractive feature is the gauge coupling unification, a property that indicates that the theory might stay perturbative up to the Grand Unified Theory (GUT) scale. On the conceptual level, we are given a framework where elementary scalars can be naturally included and treated on the same footing as fermions. Also, SUSY expanded our notion of space-time including new dimensions of Grassmannian nature. Last but not least, SUSY is a theory that makes specific predictions and can be verified or disproved experimentally in the real feature. If SUSY is true, there will be a rich spectrum of new particles within LHC reach.

On the other hand, one of the weak points is the fact that the Minimal Supersymmetric Standard Model (MSSM), without imposing phenomenological restrictions, has far too many free parameters. Also, global symmetries like lepton and baryon number conservation are not automatic like in the SM but need to be imposed, for instance, by R-parity. Restrictions by hand also need to be imposed on the soft terms for suppressing dangerous FCNC processes. In addition there is a problem with the parameter μ. This is a SUSY mass parameter and a priori has nothing to do with SUSY breaking but its phenomenologically acceptable value is of the order of the soft breaking masses. Finally, there is some tension from experimental results in colliders such as LEP2 and TEVATRON. The fact that they have not seen any SUSY signatures yet has introduced already a fine-tuning in the theory of the order of a few percent.

Because of these problems, several alternative theories have been proposed that can also offer an explanation to the mass hierarchy, predicting new phenomena at the TeV scale. One radical idea, motivated by string theory, is based on extra dimensions (ordinary bosonic and not fermionic as in SUSY).

2 Strings and Extra Dimensions

In all physical theories, the number of dimensions is a free parameter fixed to three by observation, with one exception: string theory, which predicts the existence of six new spatial dimensions. This is the only known theory today that unifies the two great discoveries of twentieth century: quantum mechanics, describing the behavior of elementary particles, and Einstein's general relativity, describing gravitational phenomena in our universe [2].

String theory replaces all elementary point particles that form matter and its interactions with a single extended object of vanishing width: a tiny string. Thus, every known elementary particle, such as the electron, quark, photon, or neutrino, corresponds to a particular vibration mode of the string (see Fig. 3).

The diversity of these particles is due to the different properties of the corresponding string vibrations.

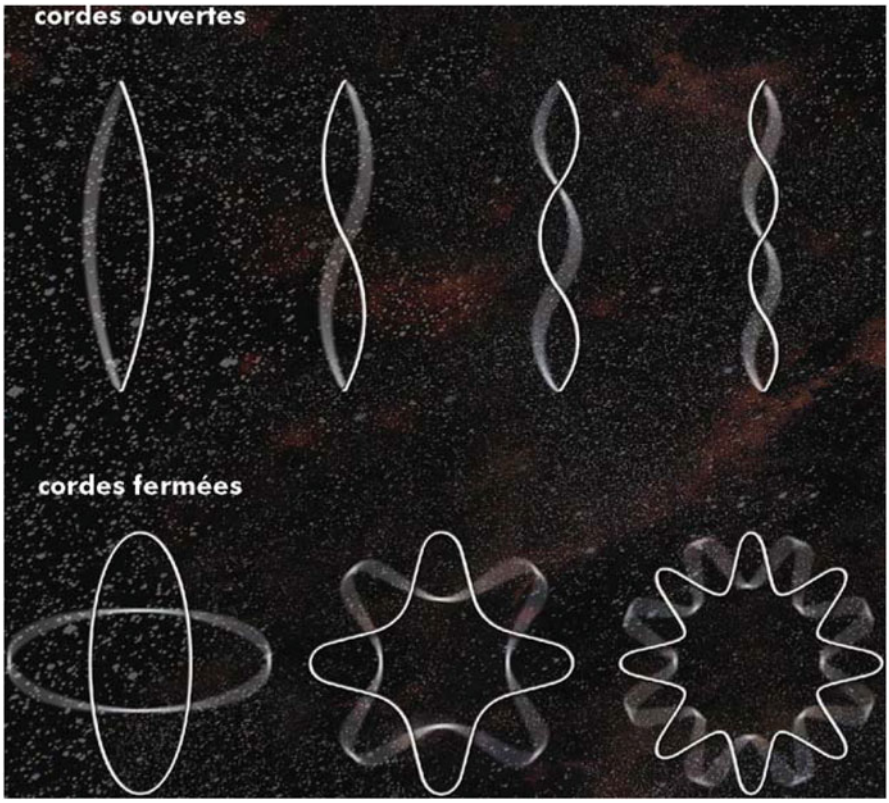


Fig. 3 In string theory, the elementary constituent of matter is a miniscule string, having vanishing width but finite size. It can be open with free ends *(upper part)*, or closed *(lower part)*. Its vibration modes, like the ones shown above, correspond to various elementary particles

Until now, there is no experimental confirmation of string theory. No one has ever observed strings, not even indirectly, neither the space of extra dimensions where they live. The main arguments in its favor are theoretical, because it provides a coherent framework for unification of all fundamental interactions. For a long time, string physicists thought that strings were extremely thin, having the smallest possible size of physics, associated to the Planck length $\sim 10^{-35}$ m.

However, the situation changed drastically over the last decade. It has been realized that the "hidden" dimensions of string theory may be much larger than what we thought in the past and they become within experimental reach in the near future, together with the strings themselves [3,4]. These ideas lead in particular to proposals of experimental tests of string theory that can be performed at TEVATRON and LHC.

The main motivation for these ideas came from considerations of the hierarchy problem that was described in the previous section. As mentioned already, a possible solution is provided by the introduction of SUSY. The appropriate and most

convenient framework for low-energy SUSY and grand unification is the perturbative heterotic string. Indeed, in this theory, gravity and gauge interactions have the same origin, as massless modes of the closed heterotic string, and they are unified at the string scale M_s. As a result, the Planck mass M_P is predicted to be proportional to M_s:

$$M_P = M_s/g\,, \tag{14}$$

where g is the gauge coupling. In the simplest constructions, all gauge couplings are the same at the string scale, given by the four-dimensional (4D) string coupling, and thus no grand unified group is needed for unification. In our conventions $\alpha_{\rm GUT} = g^2 \simeq 0.04$, leading to a discrepancy between the string and grand unification scale $M_{\rm GUT}$ by almost two orders of magnitude. Explaining this gap introduces in general new parameters or a new scale, and the predictive power is essentially lost. This is the main defect of this framework, which remains though an open and interesting possibility [5].

On the other hand, a new idea has been proposed that solves the problem if the fundamental string length is fixed to $10^{-18} \div 10^{-19}$ m [4]. In this case, quantum corrections are controlled by the string scale, which is in the TeV region, and do not destabilize the masses of elementary particles. Moreover, it offers the remarkable possibility that string physics may be testable soon in particle colliders.

How is it possible to lower the string scale from the traditional quantum gravity Planck scale to the TeV region without contradicting observations? In particular, what happens to the extra dimensions of string theory and why they have not been observed? A crucial role is played by the discovery of p-branes which are higher-dimensional objects extended in p spatial dimensions, generalizing the notion of point particle ($p = 0$) and string ($p = 1$). A main consequence of this discovery is that the string scale is in general a free parameter that can be dissociated from the Planck mass, if the universe is localized on a p-brane and does not feel all extra dimensions of string theory. The braneworld description of our universe separates the dimensions of space in two groups: those extended along our p-braneworld, called parallel, and those transverse to it.

If our universe has additional dimensions, we should observe new phenomena related to their existence. Why nobody has detected them until now? A possible answer was given in the beginning of twenteith century by Kaluza and Klein [6]: because the size of the new dimensions is very small, in contrast to the size of the other three that we know, which is infinitely large. An infinite and narrow cylinder, for example, is a two-dimensional (2D) space, with one dimension forming a very small cycle: one can move infinitely far away along the axis, while one returns back at the same point when moving along the orthogonal direction (see Fig. 4).

If one of the three known dimensions of space was small, say of millimeter size, we would be flat and, while we could move freely toward left or right, forward or backward, it would be impossible to do more than a few millimeters up or down where space ends. It follows that extra dimensions along our universe escape observation if their size is less than about 10^{-18} m which is the smallest distance scale that can be probed in present high-energy experiments [3, 7].

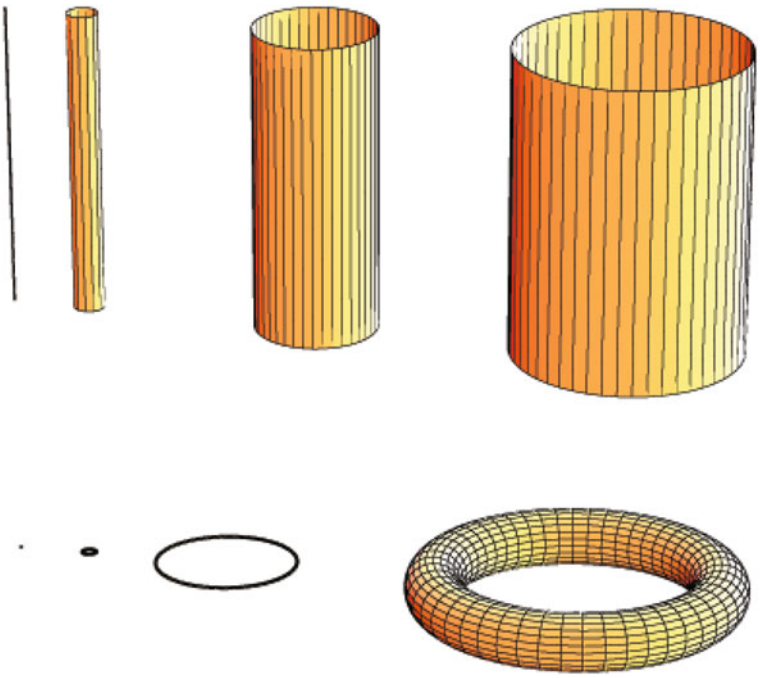


Fig. 4 Possible forms of small extra dimensions of space. Far away they are unobservable, but at distances comparable to their size we start feeling their existence and exploring their shapes

The next question is how we can detect their existence if we dispose sufficient energy to probe their size? The minimum energy required is given by their inverse size and is called compactification scale. The answer was given again by Kaluza and Klein: The motion of a particle in extra dimensions of finite size manifests to us as a tower of massive particles, called Kaluza–Klein excitations. If, for instance, the photon propagates along an extra compact dimension, one should observe a tower of massive particles with the same properties as the photon but with a mass that becomes higher as the size of the extra dimension is getting smaller. It follows, that for a size of order 10^{-18} m, an energy of order of a few TeV would be sufficient to produce them.

The above analysis and bound does not apply though for transverse dimensions to our universe, since in this case we cannot send light or any form of observable matter to probe their existence. The only way to communicate is through gravity which couples to any kind of energy density. However, our knowledge of gravity at short distances is much weaker than the other interactions, allowing for sizes of such "hidden" dimensions as large as almost a millimeter. This is roughly the shortest distance where Newton's law is tested in the laboratory.

The appropriate string perturbative framework that provides such a description, and has been studied extensively in the more recent years is type I string theory

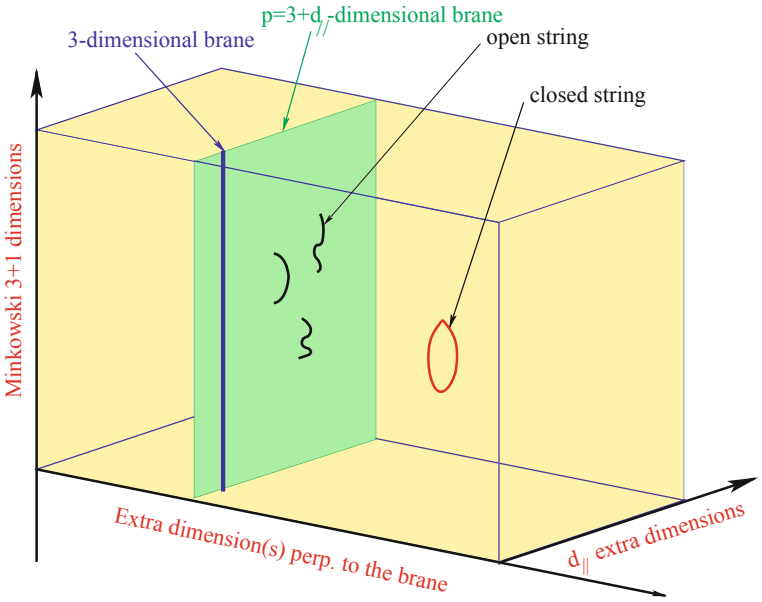


Fig. 5 In the type I string framework, our universe contains, besides the three known spatial dimensions (denoted by a single *blue line*), some extra dimensions ($d_{\parallel} = p - 3$) parallel to our world p-brane (*green plane*) where endpoints of open strings are confined, as well as some transverse dimensions (*yellow space*) where only gravity described by closed strings can propagate

with D-branes. Unlike in the heterotic string, gauge and gravitational interactions have now different origin. The latter are described again by closed strings, while the former emerge as excitations of open strings with endpoints confined on D-branes [8]. This leads to a braneworld description of our universe, which should be localized on a hypersurface, that is, a p-brane extended in p spatial dimensions (see Fig. 5). Closed strings propagate in all nine dimensions of string theory: in those extended along the p-brane (parallel) as well as in the transverse ones. On the contrary, open strings are attached on the p-brane.

Obviously, our p-brane world must have at least the three known dimensions of space. But it may contain more: the extra $d_{\parallel} = p - 3$ parallel dimensions must have a finite size, in order to be unobservable at present energies, and can be as large as $\text{TeV}^{-1} \sim 10^{-18}$ m. On the other hand, transverse dimensions interact with us only gravitationally and experimental bounds are much weaker: Their size should be less than about 0.1 mm [9]. In the following, I review the main properties and experimental signatures of low-string scale models [4, 10].

3 A Brief Introduction to Perturbative String Theory

Before proceeding, in this section, we give a short introduction of perturbative string theory [2, 11–13].

3.1 From Point Particle to Extended Objects

A p-brane is a p-dimensional spatially extended object, generalizing the notion of point particles ($p=0$) to strings ($p=1$), membranes ($p=2$), etc. We will discuss the dynamics of a free p-brane, propagating in D space-time dimensions ($p \leq D-1$), using the first quantized approach in close analogy with the case of point particles. Indeed, the propagation of a point leads to a world line. The corresponding action that describes the dynamics of a free particle is proportional to the length of this line. The trajectory which minimizes the action in flat space is then a straight line. Similarly, the propagation of a p-brane leads to a $(p+1)$-dimensional world-volume. The action describing the dynamics of a free p-brane is then proportional to the area of the world-volume and its minimization implies that the classically prefered motion is the one of minimal volume.

More precisely, in order to describe the dynamics of a p-brane in the embedding D-dimensional space-time, we introduce space-time coordinates $X^\mu(\xi_\alpha)$ ($\mu = 0, 1, \ldots, D-1$) depending on the world-volume coordinates ξ_α ($\alpha = 0, 1, \ldots, p$). The Nambu–Gotto action in a flat space-time is then given by

$$S = -T \int \sqrt{-\det h}\, d^{p+1}\xi\,, \tag{15}$$

where, T is the brane tension of dimensionality (mass)$^{p+1}$, and

$$h_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X_\mu \equiv \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu} \tag{16}$$

is the induced metric on the brane.

An equivalent but more convenient description is given by the covariant Polyakov action:

$$S = -\frac{T}{2} \int \sqrt{-\det h}\, \{h^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X_\mu - (p-1)\}\, d^{p+1}\xi\,, \tag{17}$$

where the intrinsic metric of the brane world-volume $h_{\alpha\beta}$ is introduced as an independent variable. Solving the equations of motion for X^μ and $h_{\alpha\beta}$

$$\partial_\alpha(\sqrt{-\det h}\, h^{\alpha\beta}\partial_\beta X_\mu) = 0\,, \tag{18}$$

$$\partial_\alpha X^\mu \partial_\beta X_\mu - \frac{1}{2} h_{\alpha\beta}(\partial X)^2 + \frac{1}{2}(p-1)h_{\alpha\beta} = 0\,, \tag{19}$$

with $(\partial X)^2 \equiv h^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X_\mu$, one finds

$$h_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X_\mu\,; \qquad p \neq 1. \tag{20}$$

$h_{\alpha\beta}$ is thus identified with the induced metric (16), and after substitution into the Polyakov action (17) we obtain back the Nambu–Gotto form (15).

The symmetries of the Polyakov action are (i) global space-time Lorentz invariance, and (ii) local world-volume reparameterization invariance under an arbitrary change of coordinates $\xi^\alpha \to \xi^{\alpha'}(\xi^\beta)$ with X^μ and $h_{\alpha\beta}$ transforming as a $(p+1)$-dimensional scalar and symmetric tensor, respectively. The local symmetry needs a gauge fixing: We should impose $p+1$ gauge conditions, so that only the $D-(p+1)$ ones transverse to the brane oscillations are physical.

The equations of motion (18, 19) are generally complicated and simplify only for the cases of particle ($p = 0$) and string ($p = 1$). Indeed, to solve the equations, we apply the gauge-fixing conditions on the elements of the $p \times p$ symmetric matrix $h_{\alpha\beta}$, so that there remain $\frac{1}{2}p(p+1)$ free components.

For $p = 0$, one can fix the metric to a constant, $h_{\alpha\beta} = -m^2$. Furthermore, there is only one world line coordinate, the time $\xi^\alpha = \tau$, and the equation of motion (18) becomes $\ddot{X}^\mu = 0$. The general solution is

$$X^\mu(\tau) = X_0^\mu + p^\mu \tau \,, \tag{21}$$

which is a straight line. On the other hand, the second equation of motion (19) leads to the constraint

$$\dot{X}^2 = -m^2 \,, \tag{22}$$

which is the on-shell condition $p^2 = -m^2$. The quantization can be done following the canonical procedure, that is, by applying the (equal-time) commutators

$$[X^\mu, \dot{X}^\nu] = i\eta^{\mu\nu} \,, \tag{23}$$

leading to the usual relation $[X_0^\mu, p^\nu] = i\eta^{\mu\nu}$.

For $p = 1$, we have two world-sheet coordinates $\xi^\alpha = (\tau, \sigma)$, where σ is a parameter along the string. Using the gauge freedom, one can fix two components of the metric and bring it into the conformally flat form

$$h_{\alpha\beta}(\sigma, \tau) = e^{\Phi(\sigma,\tau)}\eta_{\alpha\beta} \,, \tag{24}$$

where the scale factor Φ is the only remaining degree of freedom. However for $p = 1$, the action (17) has an additional symmetry under local rescalings of the metric (local conformal invariance), so that Φ decouples from both the action and the equations of motion (18, 19). As a result, there is an additional gauge freedom and one can again fix the metric to a constant, $h_{\alpha\beta} = \eta_{\alpha\beta}$, as in the case $p = 0$. The equations of motion for X^μ and constraints then read:

$$\partial_\alpha \partial^\alpha X^\mu = 0 \,, \tag{25}$$

$$\partial_\alpha X^\mu \partial_\beta X_\mu = \frac{1}{2}\eta_{\alpha\beta}(\partial X)^2 \,. \tag{26}$$

The above equations are further simplified by defining light-cone variables $\tau, \sigma \to \sigma_\pm = \tau \pm \sigma$. The equations of motion (25) then become

$$\partial_+ \partial_- X^\mu = 0 \,, \tag{27}$$

and the general solution for $X^\mu(\sigma_\pm)$ is separated into left- and right-movers:

$$X^\mu(\sigma_\pm) = X^\mu_L(\sigma_+) + X^\mu_R(\sigma_-)\,. \tag{28}$$

These functions are subject to the constraints (26):

$$(\partial_+ X_L)^2 = (\partial_- X_R)^2 = 0\,. \tag{29}$$

In order to write explicit solutions, we need to impose boundary conditions that correspond in general to two kind of strings: Closed and open.

Closed strings satisfy the periodicity condition

$$X^\mu(\tau,\sigma) = X^\mu(\tau,\sigma+2\pi)\,, \tag{30}$$

leading to the general solution

$$X^\mu = X^\mu_0 + P^\mu\tau + \frac{i}{\sqrt{2}}\sum_{n\neq 0}\frac{1}{n}\{a^\mu_n e^{-in(\tau+\sigma)} + \tilde{a}^\mu_n e^{-in(\tau-\sigma)}\}\,,$$

in mass units of $T = \frac{1}{2\pi}$.[1] The first two terms on the right-hand side describe the motion of the closed string center of mass, while the remaining terms in the sum correspond to the string oscillations of the left- and right-movers which are subject to the constraints (29). Reality of the coordinates X^μ imply $(a^\mu_n)^\dagger = a^\mu_{-n}$ and similar for the right-movers. Applying the canonical quantization procedure:

$$[X^\mu(\tau,\sigma), \dot{X}^\nu(\tau,\sigma')] = 2i\pi\delta(\sigma-\sigma')\eta^{\mu\nu}\,, \tag{31}$$

one obtains the commutation relations

$$[a^\mu_m, a^\nu_n] = [\tilde{a}^\mu_m, \tilde{a}^\nu_n] = m\delta_{m+n,0}\eta^{\mu\nu}\,, \tag{32}$$

$$[\tilde{a}^\mu_m, a^\nu_n] = 0 \quad , \quad [X^\mu_0, P^\nu] = i\eta^{\mu\nu}\,. \tag{33}$$

For open strings, the only Lorentz-invariant boundary condition is the Neumann one:

$$\partial_\sigma X^\mu|_{\sigma=0,\pi} = 0\,, \tag{34}$$

implying that the ends of the string propagate with the speed of light. The general solution (28) becomes in this case

$$X^\mu(\sigma,\tau) = X^\mu_0 + 2P^\mu\tau + i\sqrt{2}\sum_{n\neq 0}\frac{1}{n}a^\mu_n e^{-in\tau}\cos(n\sigma)\,,$$

and can be obtained from the closed string expression (28) by identifying left- and right-movers ($a \equiv \tilde{a}$).[2]

[1] This convention corresponds to fixing the Regge slope $\alpha' \equiv \frac{1}{2\pi T} = 1$.

[2] Note though the factor 2 in P^μ due to the change in the range of σ.

3.2 Free Closed and Open String Spectrum

By inspection of the commutation relations (32, 33), one can identify a_n^μ, $\tilde{a}_n^\mu$ with $n > 0$ ($a_{-n}^\mu, \tilde{a}_{-n}^\mu$) with the annihilation (creation) operators. Without loss of generality, let us first restrict to the holomorphic (left-moving) part. As usual, we can define a vacuum $|p>$ of momentum p annihilated by a_n^μ:

$$a_n^\mu |p>= 0 \quad ; \qquad P^\mu |p>= p^\mu |p> \ . \tag{35}$$

The physical states are then created by the action of a_{-n}^μ's on this vacuum after imposing the constraints (29). Expanding the latter into Fourier decomposition, one should require:

$$L_m \equiv \frac{1}{2} \sum_{n \in \mathbf{Z}} : a_{m-n} \cdot a_n := 0 \tag{36}$$

for $m > 0$, and

$$\begin{aligned} L_0 &\equiv \frac{1}{2\pi} \int_0^{2\pi} d\sigma (\partial_+ X)^2 \\ &= \frac{1}{4} p^2 + \frac{1}{2} \sum_{n \neq 0} : a_{-n} \cdot a_n := \frac{1}{4} p^2 + N = c \,, \end{aligned} \tag{37}$$

where N is the number operator and the constant c appears due to the normal ordering. These constraints are in fact necessary to eliminate the negative norm states generated by the action of the time component of creation operators due to the commutator (32); for instance, $||a_{-m}^\mu |p> || = m\eta^{\mu\mu} < 0$ for $\mu = 0$.

Unitarity is manifest in the light-cone gauge which fixes the residual symmetry of the covariant gauge-fixed Eqs. (27) and (29) under $\sigma_+ \rightarrow \sigma'_+(\sigma_+)$ and $\sigma_- \rightarrow \sigma'_-(\sigma_-)$. Defining

$$X^\pm = X^0 \pm X^{D-1} \,, \tag{38}$$

this invariance can be used to set:

$$X^+ = X_0^+ + p^+ \tau \,, \tag{39}$$

while the constraints can be solved for X^- in terms of the transverse coordinates X^T:

$$\partial_+ X^- = \frac{2}{p^+} (\partial_+ X^T)^2 \,. \tag{40}$$

Thus, we are left only with the transverse a_{-n}^T that are the independent physical oscillations.

We can now derive the spectrum. Note that the zero-mode of Eq. (40) reproduces the global Hamiltonian constraint (37) that gives the mass formula:

$$-\frac{1}{4} p^2 = N - c \,. \tag{41}$$

The first excited state $a_{-1}^{T}|p>$ with $N = 1$ is a space-time vector with $D-2$ transverse independent components. Therefore, by Lorentz invariance, it should be massless, implying $c = 1$. Furthermore, the quantum algebra of Lorentz generators—or equivalently the absence of conformal anomaly—fixes the space-time dimensionality $D = 26$. As a result, one obtains:

$$\begin{array}{lll} N=0 & |p> & -\frac{1}{4}p^2=-1 \quad \text{tachyon} \\ N=1 & a_{-1}^{\mu}|p> & -\frac{1}{4}p^2=0 \quad \text{massless vector} \\ N=2 \left\{ \begin{array}{c} a_{-1}^{\mu}a_{-1}^{\nu}|p> \\ a_{-2}^{\mu}|p> \end{array} \right. & & -\frac{1}{4}p^2=1 \quad \text{massive spin 2}\,, \end{array} \tag{42}$$

and so on. Note that the states of highest spin J at each mass level form a Regge trajectory:

$$\frac{1}{4}M^2 = \frac{(J-1)}{\alpha'}\,. \tag{43}$$

The spectrum of a closed string is obtained by the direct product of states from left- and right-movers with the condition $M_L^2 = M_R^2$. At the massless level, one thus has:

$$a_{-1}^{\mu}\tilde{a}_{-1}^{\nu}|p>\,, \tag{44}$$

that can be decomposed into a spin-2 graviton (symmetric traceless part), a scalar dilaton (trace), and a 2-index antisymmetric tensor (2-form). The automatic and unavoidable occurrence of the graviton in the massless spectrum is of course welcome and constitutes a strong motivation for string theory as quantum theory of gravity.

For open strings, there are no separate left- and right-movers, but one has the freedom to introduce additional quantum numbers associated to their ends. As a result, the vacuum of oscillators becomes $|p, ij>$, where the indices i, j are Chan–Paton charges "living" at the two ends of the open string. Their transformation properties define a (non-abelian) gauge group, and the massless states are now gauge fields:

$$a_{-1}^{\nu}|p, ij>. \tag{45}$$

They can form antisymmetric, symmetric, or complex representations corresponding to orthogonal, symplectic, or unitary gauge groups, respectively. Obviously, open strings cannot be a theory of gravitation as they do not contain a massless graviton in the spectrum. However, in the presence of interactions, closed strings appear by unitarity making string theory a candidate for unification of gauge with gravitational forces.

String interactions of splitting and joining can be introduced in analogy with the first quantized approach of point particles, where the world line can split using n-point vertices. However, unlike the particle case, the string interaction vertex is unique modulo world-sheet reparameterizations. As a result, interactions correspond to world-sheets with nontrivial topology and string perturbation theory becomes a topological expansion in the number of handles (for oriented closed strings), as

well as holes and crosscaps (for open strings and orientation flips of closed strings, respectively). As we will see in Sect. 3.7, string diagrams are weighted by powers of a coupling constant λ^{2g-2}, with $g = n + (h + c)/2$ the genus of the surface given in terms of the number of handles n, holes h, and crosscaps c. An important property of string interactions is that (modulo reparametrizations) there is no local notion of interaction point, which makes string perturbation theory free of UV divergences, that occur in point particles by products of distributions at the same point. Thus, string theory provides a unique mathematical framework of describing nontrivial particle interactions with no UV divergences.

3.3 Compactification on a Circle and T-duality

Since the bosonic string lives in 26 dimensions, one has to compactify 22 of those on some internal manifold of small (presently unobserved) size. The simplest example of compactification is given by one dimension X on a circle of radius R. X must then satisfy the periodicity condition:[3]

$$X = X + 2\pi R\,, \tag{46}$$

so that the solution to the equations of motion (27) is

$$\begin{aligned} X &= X_0 + p\tau + w\sigma + \frac{i}{\sqrt{2}}\sum_{n\neq 0}\frac{1}{n}\{a_n e^{-in(\tau+\sigma)} + \tilde{a}_n e^{-in(\tau-\sigma)}\} \\ &= X_0 + \frac{p+w}{2}\sigma_+ + \frac{p-w}{2}\sigma_- + \text{oscillators.} \end{aligned}$$

Note the appearance of a new term linear in σ, which is allowed to be nonvanishing due to the periodicity condition (46). Its coefficient is quantized in units of R, $w = nR$, since for $\sigma \to \sigma + 2\pi$ we must have $X \to X + 2n\pi R$, where n is the (integer) winding number of the string around the circle. On the other hand, the internal momentum is also quantized in units of $1/R$, $p = m/R$ for $m \in \mathbf{Z}$, which follows from the requirement that plane waves e^{ipX} must be single-valued around the circle: $X \to X + 2\pi R$.

For closed strings, it is convenient to define left and right momenta:

$$p_{L,R} = \frac{m}{R} \pm nR\,, \tag{47}$$

in terms of which the mass of physical states becomes (in 25 dimensions):

$$-\frac{1}{4}p^2 = \frac{1}{4}p_L^2 + N_L - 1 = \frac{1}{4}p_R^2 + N_R - 1\,. \tag{48}$$

[3] Here we define $X \equiv X^{25}$.

The above spectrum is invariant under the T-duality symmetry $R \to 1/R$ with the simultaneous interchange of momenta and windings:

$$R \to \frac{1}{R}\,,\ m \leftrightarrow n \quad \text{or} \quad X_L \to X_L\,,\ X_R \to -X_R\,. \tag{49}$$

It can be shown that T-duality is also an exact symmetry of the string interactions to all orders of perturbation theory and it is conjectured to hold at the non-perturbative level, as well. Note that the string coupling $\lambda \equiv \lambda_{26}$ should transform, so that the lower-dimensional coupling of the compactified theory $\lambda_{25} = \lambda_{26}/\sqrt{R}$ remains inert:

$$\lambda \to \frac{\lambda}{R}\,. \tag{50}$$

An important consequence of the winding modes in closed strings is the appearance of enhanced non-abelian gauge symmetries at special values of the compactification radii. For instance, the generic gauge group in the case of one dimension is the Kaluza–Klein $U(1)_L \times U(1)_R$:

$$a^\mu_{-1}|p>_L \otimes \tilde{a}^{25}_{-1}|p>_R\,,\ a^{25}_{-1}|p>_L \otimes \tilde{a}^\mu_{-1}|p>_R \text{ with } p^{25}_L = p^{25}_R = 0\,.$$

However, for the special value of the radius $R = 1$, there is an enhanced gauge symmetry $SU(2)_L \times SU(2)_R$, due to the appearance of extra massless states:

$$\begin{aligned} m = n = \pm 1 \quad &\Rightarrow p_L = \pm 2,\, p_R = 0 \ : |\pm 2>_L \otimes \tilde{a}^\mu_{-1}|p>_R \\ m = -n = \pm 1 &\Rightarrow p_L = 0,\, p_R = \pm 2 \ : a^\mu_{-1}|p>_L \otimes |\pm 2>_R\,. \end{aligned} \tag{51}$$

This property will be generalized later on to non-perturbative states.

For open strings, the Neumann boundary condition (34) imposes the windings to vanish, $w = 0$. Thus, a T-duality is not anymore a symmetry of the spectrum. Under T-duality, $p \to w$ ($\sigma \leftrightarrow \tau$), and the Neumann becomes a Dirichlet boundary condition:

$$\partial_\tau X|_{\sigma=0,\pi} = 0\,, \tag{52}$$

implying the ends of the string to be fixed at particular points of the circle. In fact, a coordinate (35) satisfying the Neumann (N) condition can be written in the form:

$$X = X_L(\sigma_+) + X_R(\sigma_-), \tag{53}$$

with left- and right-moving oscillators identified, $a_n = \tilde{a}_n$. After the T-duality transformation (49), it becomes

$$\begin{aligned} \tilde{X} &= X_L - X_R \\ &= \tilde{X}_0 + 2w\sigma + i\sqrt{2}\sum_{n\neq 0}\frac{1}{n}a^\mu_n e^{-in\tau}\sin(n\sigma)\,, \end{aligned}$$

which is the general solution of the wave Eq. (27) satisfying the Dirichlet (D) condition (52), implying that the endpoints of the string are fixed at $\tilde{X}|_{\sigma=0,\pi} = \tilde{X}_0$.

Given the two ends of the open string, one can choose independently the N or D condition for each endpoint, implying three kinds of boundary conditions: NN, DD, and ND (for which $p = w = 0$). The introduction of D conditions along the internal directions leads to the notion of D-branes which are subsurfaces where open strings can end. A Dp-brane is then defined by imposing N conditions along its longitudinal directions $X^{\mu=0,\ldots,p}$, and D conditions along its transverse ones $X^{I=p+1,\ldots,25}$. The endpoints are thus fixed in the X^I transverse space, while they can move freely in a (p+1)-dimensional space-time spanned by the world-volume of the p-brane.

It is now easy to see that performing a T-duality along a compact transverse (longitudinal) direction, a Dp-brane is transformed into a D(p+1)-brane (D-(p-1)-brane). Moreover, in this picture, the Chan–Paton multiplicity corresponds to the multiplicity of D-branes. Thus, D-branes on top of each other, labeled by a Chan–Paton index, give a gauge group enhancement.

3.4 The Superstring: Type IIA and IIB

Two immediate problems with the bosonic string are the presence of tachyon (signaling a vacuum instability) and the absence of fermions in the spectrum. Both problems can be solved in the context of superstring, obtained by adding fermions on the world-sheet.

The fermionic coordinates are Weyl–Majorana 2D fermions $\psi^\mu_{L,R}(\sigma, \tau)$, which carry a space-time index μ. The 2D Dirac equation implies that the left- (right-) handed fermions depend only on σ^+ (σ^-). As in the case of bosonic coordinates, the timelike components $\psi^0_{L,R}$ generate extra negative norm states that need a new local symmetry to be removed. This is indeed the super-reparameterization invariance, or equivalently local SUSY on the world-sheet. In the superconformal gauge (24), the SUSY transformations read (for the left-movers):

$$\begin{aligned} \delta X^\mu_L &= -i\varepsilon_R \psi^\mu_L \\ \delta \psi^\mu_L &= \varepsilon_R \partial_+ X^\mu_L, \end{aligned} \tag{54}$$

and emerge from the 2D supercurrent

$$T_F = \psi^\mu_L \partial_+ X_\mu \,. \tag{55}$$

Similar transformations hold for the right-movers by exchanging $L \leftrightarrow R$ and $+ \leftrightarrow -$. The cancellation of conformal anomalies implies, in this case, that the superstring must live in $D = 10$ space-time dimensions.

In order to obtain solutions to the equations of motion, we need to discuss boundary conditions. For both closed and open strings, there are two possible conditions compatible with space-time Lorentz invariance:

$$\psi^\mu(\tau, \sigma) = \pm \psi^\mu(\tau, \sigma + 2\pi)\,, \tag{56}$$

corresponding to the Ramond (R) and Neveu–Schwarz (NS) ones. The general solution can then be expanded as (for instance for the left-movers):

$$\psi_L^\mu = \sum_r b_r^\mu e^{-ir(\tau+\sigma)} \quad ; \quad (b_r^\mu)^\dagger = b_{-r}^\mu \,, \tag{57}$$

where r is half-integer (integer) for NS (R) boundary conditions, and upon quantization one has the usual canonical anticommutator relations:

$$\{b_r^\mu, b_s^\nu\} = \eta_{\mu\nu}\delta_{r+s,0} \,. \tag{58}$$

The mass formula (41) now becomes:

$$-\frac{1}{4}p^2 = N + N_\psi - c \,, \tag{59}$$

where $N_\psi = \frac{1}{2}\sum_r r : b_r \cdot b_{-r} :$ is a sum of half-integer (integer) frequencies for NS (R) fermions, and $c = 1/2$ ($c = 0$) in the NS (R) sector.

As in the bosonic string, unitarity is manifest in the light-cone gauge, where only the transverse oscillators create physical states. In the superstring, there are however two sectors in the spectrum: The NS and the R sectors, corresponding to antiperiodic and periodic world-sheet supercurrent (55) and giving rise to space-time bosons and fermions, respectively. In the NS sector, the ground state is still a tachyon $|p>$ of momentum p and mass given by $-(1/4)p^2 = -1/2$, while at the massless level there is a vector:

$$|\mu; p> = b_{-1/2}^\mu |p> \ . \tag{60}$$

In the R sector, the fermionic coordinates (57) have zero-modes which satisfy the anticommutator relations (58), generating a ten-dimensional (10D) Clifford algebra:

$$\{b_0^\mu, b_0^\nu\} = \eta^{\mu\nu} \,. \tag{61}$$

As a result, the oscillator vacuum forms a representation of this algebra and corresponds to a space-time spinor $|p, \alpha>$ of dimension $2^{d/2} = 32$, on which b_0^μ act as the 10D gamma matrices:

$$i\sqrt{2}\, b_0^\mu |p, \alpha> = \gamma_{\alpha\beta}^\mu |p, \beta> \ . \tag{62}$$

Moreover, in the absence of oscillators, the constraint corresponding to the zero-mode of the 2D supercurrent (55) generates the 10D massless Dirac equation which reduces the dimensionality of the lowest lying state to 16, corresponding to a massless 10D real (Majorana) fermion.

At this point, the problem of the tachyon remains. However, it turns out that the above spectrum is not consistent with world-sheet modular invariance, which guarantees the absence of global anomalies under 2D diffeomorphisms disconnected from the identity that can be performed in topologically nontrivial surfaces. Consistency of the theory already at the one-loop level (torus topology) implies that one should impose the Gliozzi–Scherk–Olive (GSO) projection:

$$(-)^F = -1 \,, \tag{63}$$

where F is the fermion number operator. This projection eliminates the tachyon from the NS sector, while in the R sector acts as space-time chirality. In fact, from the anticommutator

$$\{(-)^F, b_0^\mu\} = 0\,, \tag{64}$$

$(-)^F$ can be identified with the 10D γ^{11} matrix, in the absence of oscillators:

$$(-)^F = \gamma^{11}(-)^{\sum_{n=1}^{\infty} b_{-n}^\mu b_n^\mu}\,. \tag{65}$$

It follows that at the massless level there is one massless vector and one Weyl–Majorana spinor, having eight bosonic and eight fermionic degrees of freedom. This fermion–boson degeneracy holds to all massive levels and is a consequence of a resulting space-time SUSY.

We can now discuss the spectrum of closed superstrings. There are two different theories depending on the relative (space-time) chirality between left- and right-movers: The type IIA or type IIB corresponding to opposite or same chirality, respectively. The massless spectrum is obtained by the tensor product:

$$(|\mu>, |\alpha>_L) \otimes (|\nu>, |\beta>_{R,L})\,, \tag{66}$$

where L, R denotes left, right 10D chiralities, and $|\beta>_R$ ($|\beta>_L$) stands for type IIA (type IIB). Decomposing the spectrum into representations of the 10D little group $SO(8)$, one finds the bosons

$$\begin{aligned} |\mu> \otimes |\nu> &= 1 + 35_S + 28_A \\ |\alpha> \otimes |\beta> &\overset{\text{IIA}}{=} 8_{(1-\text{form})} + 56_{(3-\text{form})} \\ &\overset{\text{IIB}}{=} 1_{(0-\text{form})} + 28_{(2-\text{form})} + 35^+_{(4-\text{form})}, \end{aligned} \tag{67}$$

and the fermions

$$\begin{aligned} |\alpha> \otimes |\nu> &= 8_L + 56_L \\ |\mu> \otimes |\beta> &\overset{\text{IIA}}{=} 8_R + 56_R \\ &\overset{\text{IIB}}{=} 8_L + 56_L\ . \end{aligned} \tag{68}$$

The NS–NS states 1, 35_S, and 28_A in Eq. (67) denote the trace, the 2-index symmetric traceless, and antisymmetric combinations corresponding to the dilaton, graviton, and antisymmetric tensor, respectively, that are part of the massless spectrum of any consistent string theory. The above massless spectra coincide with those of IIA and IIB $N = 2$ supergravities in ten dimensions; note the two gravitini 56s in Eq. (68) with opposite or same chirality.

Upon compactification in nine dimensions on a circle of radius R, the two type II theories are equivalent under T-duality:

$$R \to \frac{1}{R} \quad : \qquad \text{IIA} \leftrightarrow \text{IIB}\,. \tag{69}$$

This can be easily seen from the transformation (49), whose action on the fermions is $\psi_L \to \psi_L$ and $\psi_R \to -\psi_R$, in order to preserve the form of the 2D supercurrent (55). As a result, $(-)^{F_R} = \gamma_R^{11} \to -\gamma_R^{11}$, implying a flip of the fermion chirality in the right-movers.

3.5 Heterotic String and Orbifold Compactifications

The heterotic string is a closed string obtained by the tensor product of the superstring (for the left-movers) and the bosonic string (for the right-movers). Since the left-moving coordinates live in ten dimensions while the right-moving ones in 26, 16 of the latter are compactified on an internal momentum lattice. One-loop modular invariance then implies that the corresponding momenta p_R^I $(I = 1, \ldots, 16)$ belong to an even self-dual lattice ($\mathbf{p}_R \cdot \mathbf{p}_R^{'} \in Z$ and $\mathbf{p}_R^2 \in 2Z$). There are only two such lattices in 16 dimensions generated by the roots of $SO(32)$ and $E_8 \times E_8$ groups.

The massless spectrum of the heterotic string is given by the tensor product

$$(|\mu>, |\alpha>) \otimes |\nu> = 1 + 35_S + 28_A + 8_L + 56_L, \tag{70}$$

which forms the particle content of the $N = 1$ supergravity multiplet in ten dimensions, and by

$$(|\mu>, |\alpha>_L) \otimes |p_R^I; \mathbf{p}_R^2 = 2>, \tag{71}$$

which forms a 10D $N = 1$ gauge supermultiplet of $SO(32)$ or $E_8 \times E_8$.

Upon compactification in nine dimensions on a circle, the two heterotic theories are equivalent under T-duality:

$$R \to \frac{1}{R} \quad : \qquad \text{Het } SO(32) \leftrightarrow \text{Het } E_8 \times E_8. \tag{72}$$

The continuous connection of the two theories can be easily seen using the freedom of gauge symmetry breaking by turning on Wilson lines (flux lines in the compact direction). These correspond to the constant values of the internal (10th) component of the gauge fields A_9^I along the 16 Cartan generators, and break generically the gauge group to its maximal abelian subgroup $U(1)^{16}$. One can then show that Eq. (72) is valid at the point with $SO(16) \times SO(16)$ gauge symmetry.

The heterotic string appears to be the only perturbative closed string theory that can describe our observable world. One of the main problems is however to get a 4D superstring which is phenomenologically viable. In particular, in order to be chiral, the spectrum should be at most $N = 1$ supersymmetric. For a toroidal compactification, the 10D $N = 1$ spectrum is converted into a non-chiral $N = 4$ supersymmetric spectrum in four dimensions.

A solution to this problem is provided by orbifold compactifications. These are obtained from toroidal compactifications by identifying points under some discrete subgroup of the internal rotations that remains an exact symmetry of the compactified theory. The resulting compact spaces are not smooth manifolds because there are singularities associated to the fixed points. String propagation, however, is made consistent because modular invariance requires the presence of a new (twisted) sector corresponding to strings with center of mass localized at the orbifold fixed points. In four dimensions, the internal rotations form an $SO(6) \equiv SU(4)$ and the condition of unbroken $N = 1$ SUSY amounts to divide the torus T^6 with a discrete subgroup of $SU(3)$ that leaves one of the four gravitini invariant.

The simplest orbifold example reduces the number of supersymmetries by half and can be studied in six dimensions. It is defined by $T^4/\mathbf{Z}_2$, where the $\mathbf{Z}_2$ inverts the sign of the four internal coordinates $X^i \to -X^i$ $(i = 1, \ldots, 4)$ and of their fermionic superpartners. Since in the heterotic string there are no right-moving superpartners, $\mathbf{Z}_2$ should also act nontrivially on the gauge degrees of freedom breaking partly the gauge symmetry together with the $N = 4$ SUSY. To see the reduction of SUSY in the massless sector, consider the (left-moving) Ramond vacuum that forms a 10D Weyl–Majorana spinor with $(-)^F = -1$. Its decomposition under $SO(4) \times SO(4)$, with the first factor corresponding to the 6D little group and the second to the internal rotations, reads

$$\begin{array}{cc} \psi_r{}^{\mu} & \psi_r{}^{i} \\ + & - \\ - & + \end{array}$$

where the signs denote the chiralities under the two $SO(4)$. The $\mathbf{Z}_2$ orbifold action on this spinor is identical to the chirality projection in the internal part, and thus eliminates half of the gravitini.

The Hilbert space of the theory consists in two sectors:

- The untwisted sector which is obtained from the Hilbert space of the toroidal compactification on T^4 projected into the $\mathbf{Z}_2$ invariant states:

$$\begin{aligned} &|p^i = 0>_+ \,, \quad |p^i> + |-p^i> \\ &\qquad \text{with even number of oscillators} \\ &|p^i = 0>_- \,, \quad |p^i> - |-p^i> \\ &\qquad \text{with odd number of oscillators} \end{aligned} \tag{73}$$

  where $|p^i = 0>_{\pm}$ denote the $\mathbf{Z}_2$ even $(+)$ and odd $(-)$ states with vanishing internal momentum p^i in which the $\mathbf{Z}_2$ action is nontrivial.
- The twisted sector which contains states localized at the $2^4 = 16$ fixed points and, thus, are confined to live on five-dimensional (5D) subspaces (in the large-volume limit).

3.6 Type I String Theory

Up to this point, we have seen four consistent closed superstring theories in ten dimensions: Type IIA and IIB with two space-time supersymmetries, and heterotic $SO(32)$ and $E_8 \times E_8$ with one SUSY. Moreover, the two type II theories and the two heterotic ones are connected by T-duality upon compactification to nine dimensions. Actually, there is a fifth consistent superstring theory in 10D with $N = 1$ SUSY, the

type I theory of open and closed strings; open strings provide the gauge sector, whereas closed strings provide gravity needed for unitarity.

A consistent algorithm to construct type I theory is to "orbifolding" type IIB by the world-sheet involution Ω that exchanges left- and right-movers and is a symmetry of the theory:

$$\Omega \quad : \quad \sigma \to -\sigma \qquad (L \leftrightarrow R). \tag{74}$$

We thus obtain type I theory = IIB/Ω. As in ordinary orbifolds, the spectrum consists in an untwisted and a twisted sector.

The untwisted sector contains closed strings projected by Ω (unoriented closed strings). This turns to symmetrize the NS–NS sector and to antisymmetrize the R–R one. As a result, the two-index NS–NS antisymmetric tensor is projected out, together with the R–R scalar and 4-form (see Eq. (68)), and we are left with the dilaton scalar ϕ and the symmetric tensor (graviton) $G_{\mu\nu}$ from NS–NS, and a 2-form $B_{\mu\nu}$ from R–R.

The twisted sector corresponds to the "fixed points" $X(-\sigma, \tau) = X(\sigma, \tau)$ which are equivalent to the (Neumann) boundary conditions for open strings: $\partial_\sigma X|_{\sigma=0,\pi} = 0$. Moreover, the "fixed-point multiplicity" N corresponds to the Chan–Paton charges. It is determined by the tadpole cancellation condition that plays the role of modular invariance for open and closed unoriented strings and guarantees the absence of potential gauge and gravitational anomalies. One finds N=32 which can also be interpreted as the multiplicity of D9-branes, leading to an $SO(32)$ gauge group.

Under a T-duality along one compact direction, we get

$$X = X_L + X_R \to \tilde{X} = X_L - X_R \tag{75}$$

and the action of $\Omega(L \leftrightarrow R)$ becomes

$$\tilde{X} \to -\tilde{X}\,. \tag{76}$$

So the effect of a T-duality on Ω is

$$\Omega \to \Omega\mathcal{R}\,, \tag{77}$$

where $\mathcal{R}$ is defined as $\mathcal{R} : X \to -X$. Therefore, T-duality gives

$$\text{type I} = \text{IIB}/\Omega \to \text{type I}' = \text{IIA}/\Omega\mathcal{R} \tag{78}$$

$$\text{D9} - \text{branes} \to \text{D8} - \text{branes} \tag{79}$$

3.7 *Effective Field Theories*

At low energies, lower than the string scale $\alpha'^{-1/2}$, one can integrate out all massive string modes to obtain an effective field theory for the massless excitations of the

string. There are two ways to obtain this effective action. Either by computing the string scattering amplitudes or by considering string propagation in the presence of nontrivial background fields. The latter is described by the world-sheet action:

$$S = -\frac{1}{4\pi}\int d^2\xi\{G_{\mu\nu}(X)\partial_\alpha X^\mu \partial^\alpha X^\nu + B_{\mu\nu}(X)\varepsilon^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X^\nu - \phi(X)\mathcal{R}^{(2)} + A^a_\mu(X)J^\mu_a + ...\} \quad (80)$$

where $G_{\mu\nu}$, $B_{\mu\nu}$, ϕ, A^a_μ, ... are backgrounds for the massless fields (metric, 2-index antisymmetric tensor, dilaton, gauge fields, etc), and $\mathcal{R}^{(2)}$ denotes the 2D scalar curvature. Note that this is the most general nonlinear sigma model which is renormalizable in two dimensions. Conformal invariance implies the vanishing of all beta functions which reproduce the space-time equations of motion for the background fields.

A particular property of string theories, which can be seen from the action (80), is that the constant dilaton background e^ϕ plays the role of the string coupling. Indeed, a shift $\phi \to \phi + c$ amounts to multiply the path integral by a factor

$$e^{-S} \to e^{2c(g-1)}e^{-S}, \quad (81)$$

where g is the genus of the world-sheet, and we used the Euler integral

$$\frac{1}{4\pi}\int d^2\xi\mathcal{R}^{(2)} = 2(g-1). \quad (82)$$

It follows that the dilaton shift can be absorbed in a rescaling of the string coupling $\lambda \to e^c\lambda$. The 10D effective action can therefore by expanded in powers of e^ϕ corresponding to the perturbative topological string expansion:

$$S_{eff} = \int d^{10}X\{e^{-2\phi}[\mathcal{R}^{(10)} + \cdots] + e^{-\phi}[\cdots] + 1[\cdots] + \cdots\}.$$

The first term proportional to $e^{-2\phi}$ corresponds to the tree-level contribution associated to spherical world-sheet topology ($g = 0$), the second term multiplying $e^{-\phi}$ denotes the disk contribution ($g = 1/2$), the third term proportional to the identity corresponds to the one-loop toroidal topology ($g = 1$), and so on. Closed oriented string diagrams give rise to even powers $e^{2(g-1)\phi}$ with g integer, whereas closed unoriented and open string diagrams introduce boundaries and crosscaps having g half-integer and can lead to odd powers of e^ϕ, as well.

4 Framework of Low-Scale Strings

In type I theory, the different origin of gauge and gravitational interactions implies that the relation between the Planck and string scales is not linear as (14) of the heterotic string. The requirement that string theory should be weakly coupled, constrains the size of all parallel dimensions to be of order of the string length, whereas

transverse dimensions remain unrestricted. Assuming an isotropic transverse space of $n = 9 - p$ compact dimensions of common radius $R_\perp$, one finds:

$$M_P^2 = \frac{1}{g_s^2} M_s^{2+n} R_\perp^n \,, \qquad g_s \simeq g^2 \quad \text{for D}-\text{branes}\,, \tag{83}$$

where g_s is the string coupling. It follows that the type I string scale can be chosen hierarchically smaller than the Planck mass [4, 14] at the expense of introducing extra-large transverse dimensions felt only by gravity, while keeping the string coupling small [4]. The weakness of 4D gravity compared to gauge interactions (ratio M_W/M_P) is then attributed to the largeness of the transverse space $R_\perp$ compared to the string length $l_s = M_s^{-1}$.

An important property of these models is that gravity becomes effectively $(4+n)$-dimensional with a strength comparable to those of gauge interactions at the string scale. The first relation of Eq. (83) can be understood as a consequence of the $(4 + n)$-dimensional Gauss law for gravity, with

$$M_*^{(4+n)} = M_s^{2+n}/g^4 \tag{84}$$

the effective scale of gravity in $4 + n$ dimensions. Taking $M_s \simeq 1$ TeV, one finds a size for the extra dimensions $R_\perp$ varying from 10^8 km, 0.1 mm, down to the fermi scale for $n = 1, 2$, or 6 large dimensions, respectively. This shows that while $n = 1$ is excluded, $n \geq 2$ is allowed by present experimental bounds on gravitational forces [9,15]. Thus, in these models, gravity appears to us very weak at macroscopic scales because its intensity is spread in the "hidden" extra dimensions. At distances shorter than $R_\perp$, it should deviate from Newton's law, which may be possible to explore in laboratory experiments (see Fig. 6).

The main experimental implications of TeV scale strings in particle accelerators are of three types, in correspondence with the three different sectors that are generally present: (i) new compactified parallel dimensions, (ii) new extra-large transverse dimensions and low-scale quantum gravity, and (iii) genuine string and quantum gravity effects. On the other hand, there exist interesting implications in non-accelerator table-top experiments due to the exchange of gravitons or other possible states living in the bulk.

5 Large Number of Species

Here, we open a parenthesis to describe that low-scale gravity with large extra dimensions is actually a particular case of a more general framework, where the UV cutoff is lower than the Planck scale due to the existence of a large number of particle species coupled to gravity [16]. Indeed, it was shown that the effective UV cutoff M_* is given by

$$M_*^2 = M_P^2/N \,, \tag{85}$$

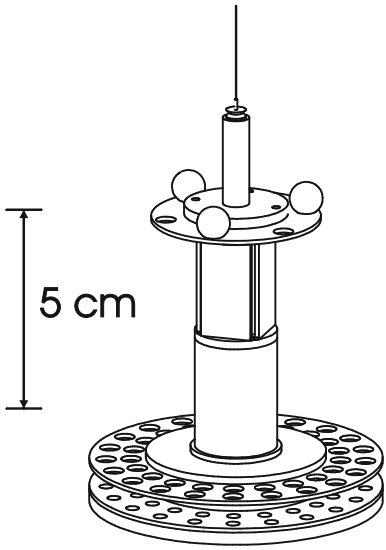


Fig. 6 Torsion pendulum that tested Newton's law at 55 μm [9]

where the counting of independent species N takes into account all particles which are not broad resonances, having a width less than their mass. The derivation is based on black hole evaporation, but here we present a shorter argument using quantum information storage [17]. Consider a pixel of size L containing N species storing information. The energy required to localize N wave functions is then given by N/L, associated to a Schwarzschild radius $R_s = N/LM_P^2$. The latter must be less than the pixel size in order to avoid the collapse of such a system to a black hole, $R_s \leq L$, implying a minimum size $L \geq L_{min}$ with $L_{min} = \sqrt{N}/M_P$ associated precisely to the effective UV cutoff $M_* = L_{min}$ given in Eq. (85). Imposing $M_* \simeq 1$ TeV, one should then have $N \sim 10^{32}$ particle species below about the TeV scale!

In the string theory context, there are two ways of realizing such a large number of particle species by lowering the string scale at a TeV:

1. In large-volume compactifications, with the SM localized on D-brane stacks, as described in the previous section. The particle species are then the Kaluza–Klein (KK) excitations of the graviton (and other possible bulk modes) associated to the large extra dimensions, given by $N = R_\perp^n l_s^n$, up to energies of order $M_* \simeq M_s$.
2. By introducing an infinitesimal string coupling $g_s \simeq 10^{-16}$ with the SM localized on Neveu–Schwarz NS5-branes in the framework of little strings [18]. In this case, the particle species are the effective number of string modes that contribute to the black hole bound [19]: $N = 1/g_s^2$ and gravity does not become strong at $M_s \sim \mathcal{O}$(TeV).

Note that both TeV string realizations above are compatible with the general expression (83), but in the second case there is no relation between the string and gauge couplings.

6 Warped Spaces

In these models, space-time is a slice of anti-de Sitter space (AdS) in $d = 5$ dimensions whereas our universe forms a 4D flat boundary [20]. The corresponding line element is:

$$ds^2 = e^{-2k|y|}\eta_{\mu\nu}dx^\mu dx^\nu + dy^2 \quad ; \quad \Lambda = -24M^3k^2\,, \tag{86}$$

where M, Λ are the 5D Planck mass and cosmological constant, respectively, and the parameter k is the curvature of AdS_5. The fifth coordinate y is restricted on the interval $[0, \pi r_c]$. Thus, this model requires two "branes," a UV and an infrared (IR), located at the two endpoints of the interval, $y = 0$ and $y = \pi r_c$, respectively. The vanishing of the 4D cosmological constant requires to fine-tune the two tensions: $T = -T' = 24M^3k^2$. The 4D Planck mass is given by:

$$M_P^2 = \frac{1}{k}(1 - e^{-2\pi kr_c})M^3\,. \tag{87}$$

On the other hand, there is an exponential hierarchy of the mass scales in the IR relative to the UV-brane due to the warped factor in the metric (86). One can then identify the electroweak scale $M_W \simeq M_P e^{-2\pi kr_c}$.

By studying the corresponding wave functions, one can show that 4D gravity is localized on the UV-brane, whereas KK gravitons on the IR. It follows that the main prediction of this model is the existence of spin-2 resonances at the TeV scale coupled with (TeV)-1 strength. Their masses are given by:

$$m_n = c_n k e^{-2\pi kr_c} \quad ; \quad c_n \simeq \left(n + \frac{1}{4}\right) \text{ for large } n \tag{88}$$

and they are weakly coupled for $m_n < Me^{-2\pi kr_c}$, that is, for $k < M$. Viable models can be constructed when SM gauge bosons propagate in the 5D bulk, fermions are localized near the UV-brane, whereas the Higgs near the IR [21]. Using AdS/CFT correspondence, they are dual to strongly coupled 4D field theories with composite Higgs; their gauge coupling $g_{YM} \sim M/k$ is strong when spin-2 KK modes are weakly coupled in the gravity dual description.

7 Experimental Implications in Accelerators

7.1 World-Brane Extra Dimensions

In this case $RM_s \gtrsim 1$, and the associated compactification scale $R_\parallel^{-1}$ would be the first scale of new physics that should be found increasing the beam energy [3, 7, 22]. There are several reasons for the existence of such dimensions. It is a logical possibility, since out of the six extra dimensions of string theory only two are needed for lowering the string scale, and thus the effective p-brane of our world has in general $d_\parallel \equiv p - 3 \leq 4$. Moreover, they can be used to address several physical problems in braneworld models, such as obtaining different SM gauge couplings, explaining fermion mass hierarchies due to different localization points of quarks and leptons in the extra dimensions, providing calculable mechanisms of SUSY breaking, etc.

The main consequence is the existence of KK excitations for all SM particles that propagate along the extra parallel dimensions. Their masses are given by:

$$M_m^2 = M_0^2 + \frac{m^2}{R_\parallel^2} \quad ; \quad m = 0, \pm 1, \pm 2, \ldots \tag{89}$$

where we used $d_\parallel = 1$, and M_0 is the higher-dimensional mass. The zero-mode $m = 0$ is identified with the 4D state, whereas the higher modes have the same quantum numbers as the lowest one, except for their mass given in (89). There are two types of experimental signatures of such dimensions [7, 23, 24]: (i) virtual exchange of KK excitations, leading to deviations in cross sections compared to the SM prediction, that can be used to extract bounds on the compactification scale; (ii) direct production of KK modes.

On general grounds, there can be two different kinds of models with qualitatively different signatures depending on the localization properties of matter fermion fields. If the latter are localized in 3D-brane intersections, they do not have excitations and KK momentum is not conserved because of the breaking of translation invariance in the extra dimension(s). KK modes of gauge bosons are then singly produced giving rise to generally strong bounds on the compactification scale and new resonances that can be observed in experiments. Otherwise, they can be produced only in pairs due to the KK momentum conservation, making the bounds weaker but the resonances difficult to observe.

When the internal momentum is conserved, the interaction vertex involving KK modes has the same 4D tree-level gauge coupling. On the other hand, their couplings to localized matter have an exponential form factor suppressing the interactions of heavy modes. This form factor can be viewed as the fact that the branes intersection has a finite thickness. For instance, the coupling of the KK excitations of gauge fields $A^\mu(x, y) = \sum_m A_m^\mu \exp i\frac{my}{R_\parallel}$ to the charge density $j_\mu(x)$ of massless localized

fermions is described by the effective action [25]:

$$\int d^4x \sum_m e^{-\ln 16 \frac{m^2 l_s^2}{2R_\parallel^2}} j_\mu(x) A_m^\mu(x). \tag{90}$$

After Fourier transform in position space, it becomes:

$$\int d^4x\, dy \frac{1}{(2\pi \ln 16)^2} e^{-\frac{y^2 M_s^2}{2 \ln 16}} j_\mu(x) A^\mu(x, y), \tag{91}$$

from which we see that localized fermions form a Gaussian distribution of charge with a width $\sigma = \sqrt{\ln 16}\, l_s \sim 1.66\, l_s$.

To simplify the analysis, let us consider first the case $d_\parallel = 1$ where some of the gauge fields arise from an effective 4-brane, while fermions are localized states on brane intersections. Since the corresponding gauge couplings are reduced by the size of the large dimension $R_\parallel M_s$ compared to the others, one can account for the ratio of the weak to strong interactions strengths if the $SU(2)$-brane extends along the extra dimension, whereas $SU(3)$ does not. As a result, there are three distinct cases to study [24], denoted by (t, l, l), (t, l, t), and (t, t, l), where the three positions in the brackets correspond to the three SM gauge group factors $SU(3) \times SU(2) \times U(1)$ and those with l (longitudinal) feel the extra dimension, whereas those with t (transverse) do not.

In the (t, l, l) case, there are KK excitations of $SU(2) \times U(1)$ gauge bosons: $W_\pm^{(m)}$, $\gamma^{(m)}$, and $Z^{(m)}$. Performing a χ^2 fit of the electroweak observables, one finds that if the Higgs is a bulk state (l), $R_\parallel^{-1} \gtrsim 3.5$ TeV [26]. This implies that LHC can produce at most the first KK mode. Different choices for localization of matter and Higgs fields lead to bounds, lying in the range $1 \div 5$ TeV [26].

In addition to virtual effects, KK excitations can be produced on-shell at LHC as new resonances [23] (see Fig. 7).

There are two different channels, neutral Drell–Yan processes $pp \to l^+ l^- X$ and the charged channel $l^\pm \nu$, corresponding to the production of the KK modes $\gamma^{(1)}$, $Z^{(1)}$, and $W_\pm^{(1)}$, respectively. The discovery limits are about 6 TeV, while the exclusion bounds 15 TeV. An interesting observation in the case of $\gamma^{(1)} + Z^{(1)}$ is that interferences can lead to a "dip" just before the resonance. There are some ways to distinguish the corresponding signals from other possible origin of new physics such as models with new gauge bosons. In fact, in the (t, l, l) and (t, l, t) cases, one expects two resonances located practically at the same mass value. This property is not shared by most of other new gauge boson models. Moreover, the heights and widths of the resonances are directly related to those of SM gauge bosons in the corresponding channels.

In the (t, l, t) case, only the $SU(2)$ factor feels the extra dimension and the limits set by the KK states of $W^\pm$ remain the same. On the other hand, in the (t, t, l) case where only $U(1)_Y$ feels the extra dimension, the limits are weaker and the exclusion bound is around 8 TeV. In addition to these simple possibilities, brane constructions lead often to cases where part of $U(1)_Y$ is t and part is l. If $SU(2)$ is l the limits

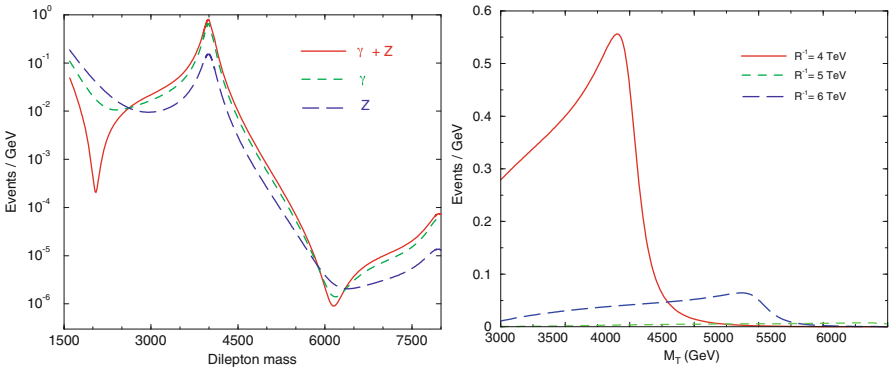


Fig. 7 Production at LHC of the first KK modes of the photon and of the Z boson decaying to electron–positron pairs (*upper panel*), as well as of the W boson decaying to lepton–neutrino (*lower panel*). The number of expected events is plotted as a function of the energy of the pair in GeV (transverse energy for W excitations). From highest to lowest on the *upper panel*: excitation of $\gamma + Z$, γ, and Z

come again from $W^{\pm}$, while if it is t then it will be difficult to distinguish this case from a generic extra $U(1)'$. A good statistics would be needed to see the deviation in the tail of the resonance as being due to effects additional to those of a generic $U(1)'$ resonance. Finally, in the case of two or more parallel dimensions, the sum in the exchange of the KK modes diverges in the limit $R_{\parallel} M_s \gg 1$ and needs to be regularized using the form factor (90). Cross sections become bigger yielding stronger bounds, while resonances are closer implying that more of them could be reached by LHC.

On the other hand, if all SM particles propagate in the extra dimension (called universal)[4], KK modes can only be produced in pairs and the lower bound on the compactification scale becomes weaker, of the order of 300–500 GeV. Moreover, no resonances can be observed at LHC, so that this scenario appears very similar to low-energy SUSY. In fact, KK parity can even play the role of R-parity, implying that the lightest KK mode is stable and can be a dark matter candidate in analogy to the (lightest supersymmetric particle, LSP) [27].

7.2 *Extra Large Transverse Dimensions*

The main experimental signal is gravitational radiation in the bulk from any physical process on the world-brane. In fact, the very existence of branes breaks translation invariance in the transverse dimensions and gravitons can be emitted from the brane

[4] Although interesting, this scenario seems difficult to be realized, since 4D chirality requires nontrivial action of orbifold twists with localized chiral states at the fixed points.

Table 2 Limits on $R_\perp$ in mm [10]

Experiment	$n=2$	$n=4$	$n=6$
Collider bounds			
LEP 2	5×10^{-1}	2×10^{-8}	7×10^{-11}
TEVATRON	5×10^{-1}	10^{-8}	4×10^{-11}
LHC	4×10^{-3}	6×10^{-10}	3×10^{-12}
NLC	10^{-2}	10^{-9}	6×10^{-12}
Present non-collider bounds			
SN1987A	3×10^{-4}	10^{-8}	6×10^{-10}
COMPTEL	5×10^{-5}	–	–

LHC Large Hadron Collider, *NLC* Next Linear Collider

into the bulk. During a collision of center of mass energy $\sqrt{s}$, there are $\sim(\sqrt{s}R_\perp)^n$ KK excitations of gravitons with tiny masses, that can be emitted. Each of these states looks from the 4D point of view as a massive, quasi-stable, extremely weakly coupled (s/M_P^2 suppressed) particle that escapes from the detector. The total effect is a missing-energy cross section roughly of order:

$$\frac{(\sqrt{s}R_\perp)^n}{M_P^2}\sim\frac{1}{s}\left(\frac{\sqrt{s}}{M_s}\right)^{n+2}. \tag{92}$$

Explicit computation of these effects leads to the bounds given in Table 2.

However, larger radii are allowed if one relaxes the assumption of isotropy, by taking, for instance, two large dimensions with different radii.

Figure 8 shows the cross section for graviton emission in the bulk, corresponding to the process $pp\to jet+graviton$ at LHC, together with the SM background [28].

For a given value of M_s, the cross section for graviton emission decreases with the number of large transverse dimensions, in contrast to the case of parallel dimensions. The reason is that gravity becomes weaker if there are more dimensions because there is more space for the gravitational field to escape. It turns out that here is a particular energy and angular distribution of the produced gravitons that arise from the distribution in mass of KK states of spin-2. This can be contrasted to other sources of missing energy and might be a smoking gun for the extra-dimensional nature of such a signal.

Table 2 also included astrophysical and cosmological bounds. Astrophysical bounds [29,30] arise from the requirement that the radiation of gravitons should not carry on too much of the gravitational binding energy released during core collapse of supernovae. In fact, the measurements of Kamiokande and IMB for SN1987A suggest that the main channel consists of neutrino fluxes. The best cosmological bound [31] is obtained from requiring that decay of bulk gravitons to photons does not generate a spike in the energy spectrum of the photon background measured by the COMPTEL instrument. Bulk gravitons are expected to be produced just before nucleosynthesis due to thermal radiation from the brane. The limits assume that the

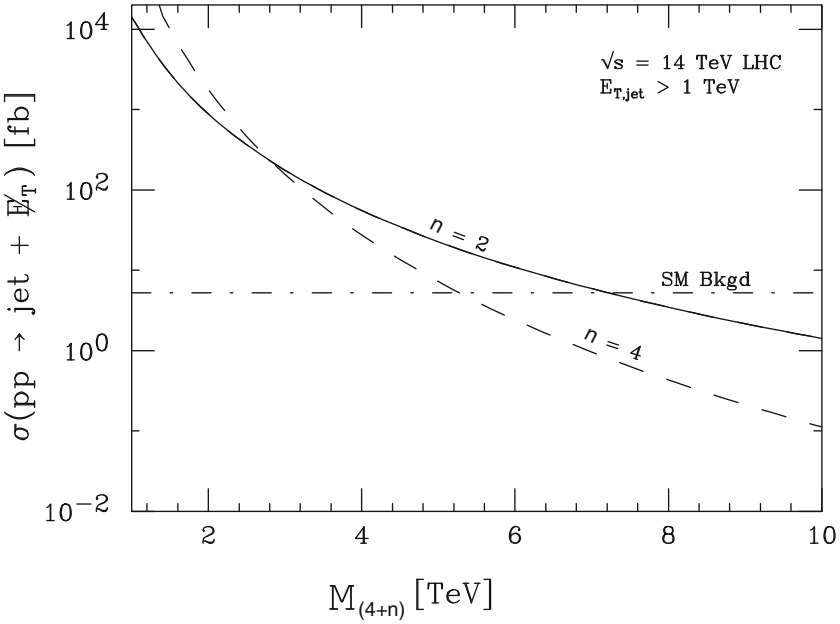


Fig. 8 Missing energy due to graviton emission at LHC, as a function of the higher-dimensional gravity scale M_*, produced together with a hadronic jet. The expected cross section is shown for $n = 2$ and $n = 4$ extra dimensions, together with the SM background [28]

temperature was at most 1 MeV as nucleosynthesis begins, and become stronger if temperature is increased.

7.3 String Effects

At low energies, the interaction of light (string) states is described by an effective field theory. Their exchange generates in particular four-fermion operators that can be used to extract independent bounds on the string scale. In analogy with the bounds on longitudinal extra dimensions, there are two cases depending on the localization properties of matter fermions. If they come from open strings with both ends on the same stack of branes, exchange of massive open string modes gives rise to dimension-eight effective operators, involving four fermions and two space-time derivatives [25, 32]. The corresponding bounds on the string scale are then around 500 GeV. On the other hand, if matter fermions are localized on nontrivial brane intersections, one obtains dimension-six four-fermion operators and the bounds become stronger: $M_s \gtrsim 2 \div 3$ TeV [10,25]. At energies higher than the string scale, new spectacular phenomena are expected to occur, related to string physics and quantum gravity effects, such as possible micro-black hole production [33–35].

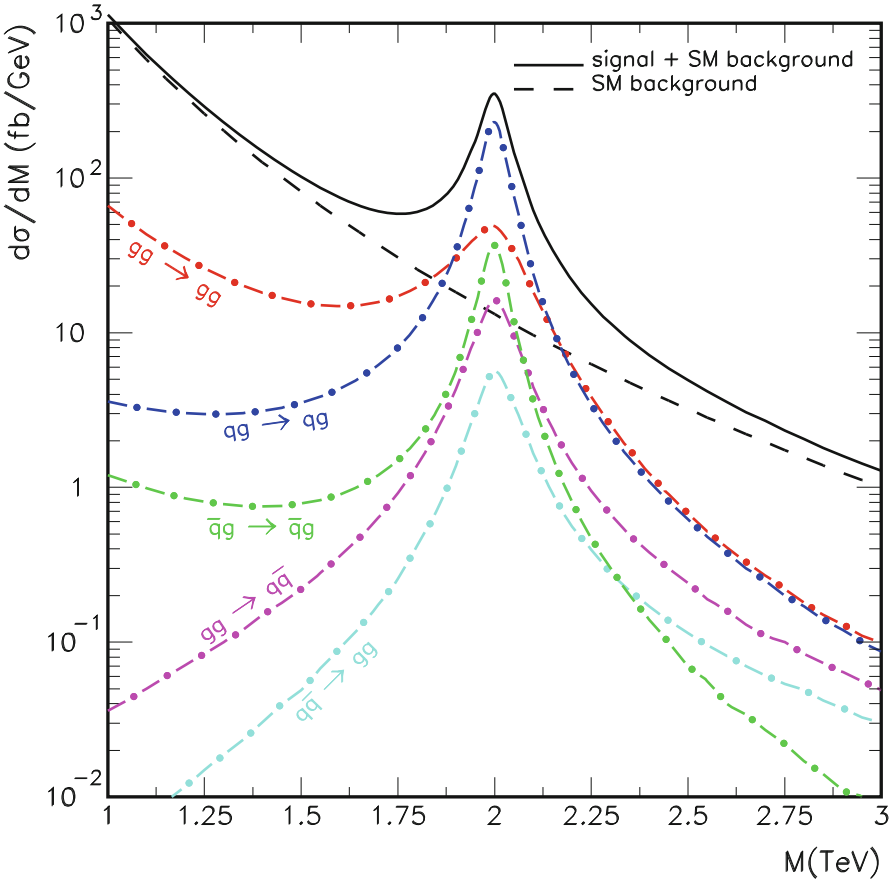


Fig. 9 Production of the first Regge excitations at LHC in the dijet channel, for a string scale $M_s = 2$ TeV. The cross-section is plotted as a function of the dijet invariant mass M [36, 37]

Particle accelerators would then become the best tools for studying quantum gravity and string theory.

Direct production of string resonances in hadron colliders leads generically to a universal deviation from SM in jet distribution [36]. In particular, the first Regge excitation of the gluon has spin-2 and a width an order of magnitude lower than the string scale, leading to a characteristic peak in dijet production; similarly, the first excitations of quarks have spin 3/2. The dijet cross section is shown in Fig. 9 for LHC energies.

The reason for the universal behavior is that tree N-point open superstring amplitudes involving at most two fermions and gluons are completely model independent from the details of the compactification, including the number of supersymmetries that are left unbroken in four dimensions (even if all are broken). Such tree-level amplitudes do not receive contributions from KK, string winding or closed string

graviton modes, but are given as a universal sum over exchanges of open string Regge excitations lying in Regge trajectories with masses

$$M_n^2 = M_s^2 n \quad ; \quad n = 0, 1, \ldots \tag{93}$$

and maximal spin $n + 1$.

The relevant partonic cross sections for dijet production, involving at most two quarks are $|\mathcal{M}(gg \to gg)|^2$, $|\mathcal{M}(gg \to q\bar{q})|^2$, $|\mathcal{M}(q\bar{q} \to gg)|^2$, and $|\mathcal{M}(qg \to qg)|^2$, where g and q denote gluons and quarks, respectively. They can be obtained from the following two independent processes, up to crossing symmetries [37]:

$$\begin{aligned}
|\mathcal{M}(gg \to gg)|^2 &= g^4 \left(\frac{1}{s^2} + \frac{1}{t^2} + \frac{1}{u^2}\right) \\
&\quad \times \left[\frac{9}{4}\left(s^2 V_s^2 + t^2 V_t^2 + u^2 V_u^2\right) - \frac{1}{3}\left(s V_s + t V_t + u V_u\right)^2\right] \\
|\mathcal{M}(gg \to q\bar{q})|^2 &= g^4 \frac{t^2 + u^2}{s^2}\left[\frac{1}{6}\frac{1}{tu}(t V_t + u V_u)^2 - \frac{3}{8} V_t V_u\right],
\end{aligned} \tag{94}$$

where s, t, u are the usual Mandelstam variables in string mass units and

$$V_s = -\frac{tu}{s}\, B(t,u) = 1 - \frac{2}{3}\pi^2 tu + \ldots \qquad V_t : \; s \leftrightarrow t \qquad V_u : \; s \leftrightarrow u\,. \tag{95}$$

The low-energy expansion of these amplitudes reproduces the usual QCD expressions (upon inserting the color factors), while higher-order terms describe the string corrections due to the exchange of Regge excitations. Besides these amplitudes, there are also those involving four quarks such as $|\mathcal{M}(q\bar{q} \to q\bar{q})|^2$ and $|\mathcal{M}(qq \to qq)|^2$. These are model dependent, because the details of the compactification do not decouple. The reason is that they involve four vertices containing twist fields that describe quark states arising from open strings stretched in brane intersections. However, taking into account QCD color factors, their contribution is suppressed because parton luminosities in proton–proton collisions at TeV energies favor at least one gluon in the initial state. As a result, the dominant contribution comes from the model independent cross sections (94) leading to the effect of Fig. 9.

We finish this section with some comments related to the possible micro-black hole production. Independently on the unresolved issue of the convergence of string perturbation theory in the kinematic region relevant to micro-black hole formation, there is a simple argument showing that at least within the perturbative TeV string framework, the energy threshold for black hole production is far above the LHC reach. Indeed, a string size black hole has a horizon radius $r_H \sim 1$ in string units, while the d-dimensional Newton's constant behaves as $G_N \sim g_s^2$. It follows that the mass of a d-dimensional black hole is [38]:

$$M_{\rm BH} \sim r_H^{d/2-1}/G_N \simeq 1/g_s^2\,. \tag{96}$$

Thus, for a weakly coupled theory, this energy threshold is much higher than the string and the higher-dimensional Planck scales M_s and M_* of Eq. (84). Comparing this energy threshold with the mass of Regge excitations (93), one finds $n \sim 1/g_s^4$ which is actually compatible with the relation one obtains by identifying the black hole entropy $S_{\rm BH} \sim 1/G_N \sim 1/g_s^2$ with the perturbative string entropy $S_{\rm string} \sim \sqrt{n}$. Using now relation (83), and the value of the SM gauge couplings $g_s \simeq g^2 \sim 0.1$, one finds that the energy threshold $M_{\rm BH}$ of micro-black hole production is about four orders of magnitude higher than the string scale, implying that one would produce 10^4 string states before reaching $M_{\rm BH}$.

8 Supersymmetry in the Bulk and Short Range Forces

8.1 Submillimeter Forces

Besides the spectacular predictions in accelerators, there are also modifications of gravitation in the submillimeter range, which can be tested in "table-top" experiments that measure gravity at short distances. There are three categories of such predictions:
(i) Deviations from the Newton's law $1/r^2$ behavior to $1/r^{2+n}$, which can be observable for $n = 2$ large transverse dimensions of submillimeter size. This case is particularly attractive on theoretical grounds because of the logarithmic sensitivity of SM couplings on the size of transverse space [39], that allows to determine the hierarchy [40].
(ii) New scalar forces in the submillimeter range, related to the mechanism of SUSY breaking, and mediated by light scalar fields φ with masses [4,41]:

$$m_\varphi \simeq \frac{m_{susy}^2}{M_P} \simeq 10^{-4} \div 10^{-6}\ {\rm eV}\,, \tag{97}$$

for a SUSY breaking scale $m_{susy} \simeq 1 \div 10$ TeV. They correspond to Compton wavelengths of 1 mm to 10 μm. m_{susy} can be either $1/R_\parallel$ if supersymmetry is broken by compactification [41], or the string scale if it is broken "maximally" on our world-brane [4]. A universal attractive scalar force is mediated by the so-called radion modulus, whose expectation value determines the size of the extra dimension, namely $\varphi \equiv M_P \ln R$, with R the radius of the longitudinal or transverse dimension(s). In the former case, the result (97) follows from the behavior of the vacuum energy density $\Lambda \sim 1/R_\parallel^4$ for large $R_\parallel$ (up to logarithmic corrections). In the latter, SUSY is broken primarily on the brane, and thus its transmission to the bulk is gravitationally suppressed, leading to (97). For $n = 2$, there may be an enhancement factor of the radion mass by $\ln R_\perp M_s \simeq 30$ decreasing its wavelength by an order of magnitude [40].

The coupling of the radius modulus to matter relative to gravity can be easily computed and is given by:

$$\sqrt{\alpha_\varphi} = \frac{1}{M}\frac{\partial M}{\partial \varphi}\ ;\ \alpha_\varphi = \begin{cases} \frac{\partial \ln \Lambda_{\rm QCD}}{\partial \ln R} \simeq \frac{1}{3} \ \text{ for } R_\parallel \\ \\ \frac{2n}{n+2} = 1 \div 1.5 \text{ for } R_\perp , \end{cases} \tag{98}$$

where M denotes a generic physical mass. In the longitudinal case, the coupling arises dominantly through the radius dependence of the QCD gauge coupling [41], while in the case of transverse dimension, it can be deduced from the rescaling of the metric which changes the string to the Einstein's frame and depends slightly on the bulk dimensionality ($\alpha = 1 \div 1.5$ for $n = 2 \div 6$) [40]. Such a force can be tested in microgravity experiments and should be contrasted with the change of Newton's law due to the presence of extra dimensions that is observable only for $n = 2$ [9,15]. The resulting bounds from an analysis of the radion effects are [42]:

$$M_* \gtrsim 6\,\text{TeV}\,. \tag{99}$$

In principle, there can be other light moduli which couple with even larger strengths. For example the dilaton, whose VEV determines the string coupling, if it does not acquire large mass from some dynamical supersymmetric mechanism, can lead to a force of strength 2000 times bigger than gravity [43].

(iii) Nonuniversal repulsive forces much stronger than gravity, mediated by possible abelian gauge fields in the bulk [29,44]. Such fields acquire tiny masses of the order of M_s^2/M_P, as in (97), due to brane localized anomalies [44]. Although their gauge coupling is infinitesimally small, $g_A \sim M_s/M_P \simeq 10^{-16}$, it is still bigger than the gravitational coupling E/M_P for typical energies $E \sim 1$ GeV (corresponding to the nucleon mass), and the strength of the new force would be $10^6 \div 10^8$ stronger than gravity. This is an interesting region which will be soon explored in microgravity experiments (see Fig. 10). Note that in this case supernova constraints impose that there should be at least four large extra dimensions in the bulk [29].

In Fig. 10 we depict the actual information from previous, present, and upcoming experiments [15,40].

The solid lines indicate the present limits from the experiments indicated. The excluded regions lie above these solid lines. Measuring gravitational strength forces at short distances is challenging. The horizontal lines correspond to theoretical predictions, in particular, for the graviton in the case $n = 2$ and for the radion in the transverse case. These limits are compared to those obtained from particle accelerator experiments in Table 2. Finally, in Figs. 11 and 12, we display some improved bounds for new forces at very short distances by focusing on the left-hand side of Fig. 10, near the origin [15].

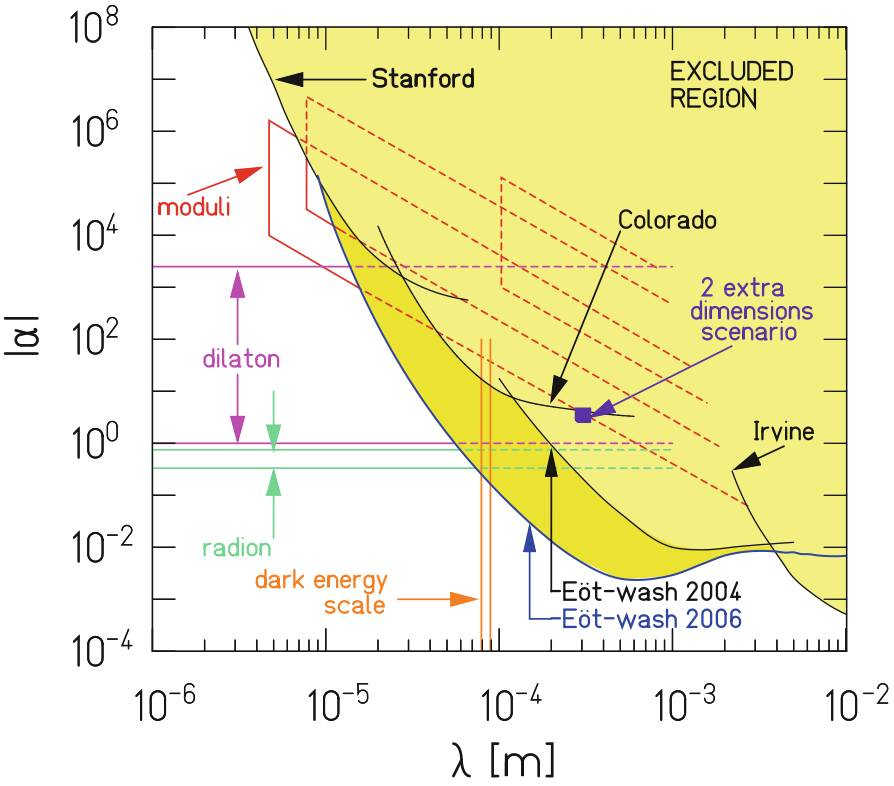


Fig. 10 Present limits on new short-range forces (*yellow regions*), as a function of their range λ and their strength relative to gravity α [9, 15]. The limits are compared to new forces mediated by the graviton in the case of two large extra dimensions, and by the radion

8.2 *Brane Nonlinear Supersymmetry*

When the closed string sector is supersymmetric, SUSY on a generic brane configuration is nonlinearly realized even if the spectrum is not supersymmetric and brane fields have no superpartners. The reason is that the gravitino must couple to a conserved current locally, implying the existence of a goldstino on the brane world-volume [45]. The goldstino is exactly massless in the infinite (transverse) volume limit and is expected to acquire a small mass suppressed by the volume, of order (97). In the standard realization, its coupling to matter is given via the energy momentum tensor [46], whereas in general there are more terms invariant under nonlinear SUSY that have been classified, up to dimension eight [47, 48].

An explicit computation was performed for a generic intersection of two-brane stacks, leading to three irreducible couplings, besides the standard one [48]: two of dimension six involving the goldstino, a matter fermion and a scalar or gauge field,

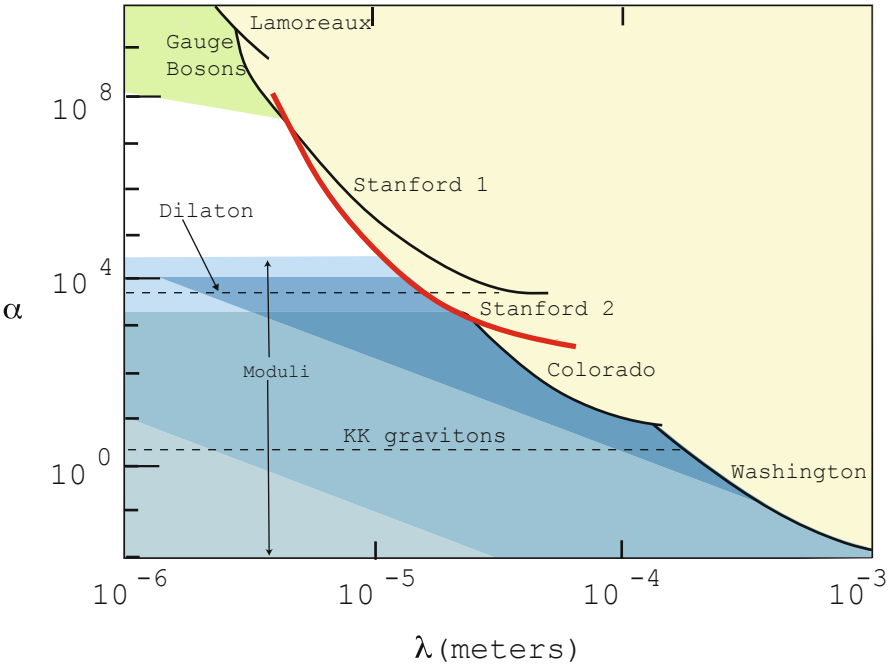


Fig. 11 Bounds on non-Newtonian forces in the range $6 \div 20$ μm (see Smullin et al. [15])

and one four-fermion operator of dimension eight. Their strength is set by the goldstino decay constant κ, up to model-independent numerical coefficients which are independent of the brane angles. Obviously, at low energies the dominant operators are those of dimension six. In the minimal case of (non-supersymmetric) SM, only one of these two operators may exist, that couples the goldstino χ with the Higgs H and a lepton doublet L:

$$\mathcal{L}_{\chi}^{int} = 2\kappa (D_{\mu} H)(L D^{\mu} \chi) + \text{ complex conjugate}\,, \tag{100}$$

where the goldstino decay constant is given by the total brane tension

$$\frac{1}{2\,\kappa^2} = N_1\, T_1 + N_2\, T_2\,; \quad T_i = \frac{M_s^4}{4\pi^2 g_i^2}\,, \tag{101}$$

with N_i the number of branes in each stack. It is important to notice that the effective interaction (100) conserves the total lepton number L, as long as we assign to the goldstino a total lepton number $L(\chi) = -1$ [49]. To simplify the analysis, we will consider the simplest case where (100) exists only for the first generation and L is the electron doublet [49].

The effective interaction (100) gives rise mainly to the decays $W^{\pm} \to e^{\pm}\chi$ and $Z, H \to \nu\chi$. It turns out that the invisible Z width gives the strongest limit on κ

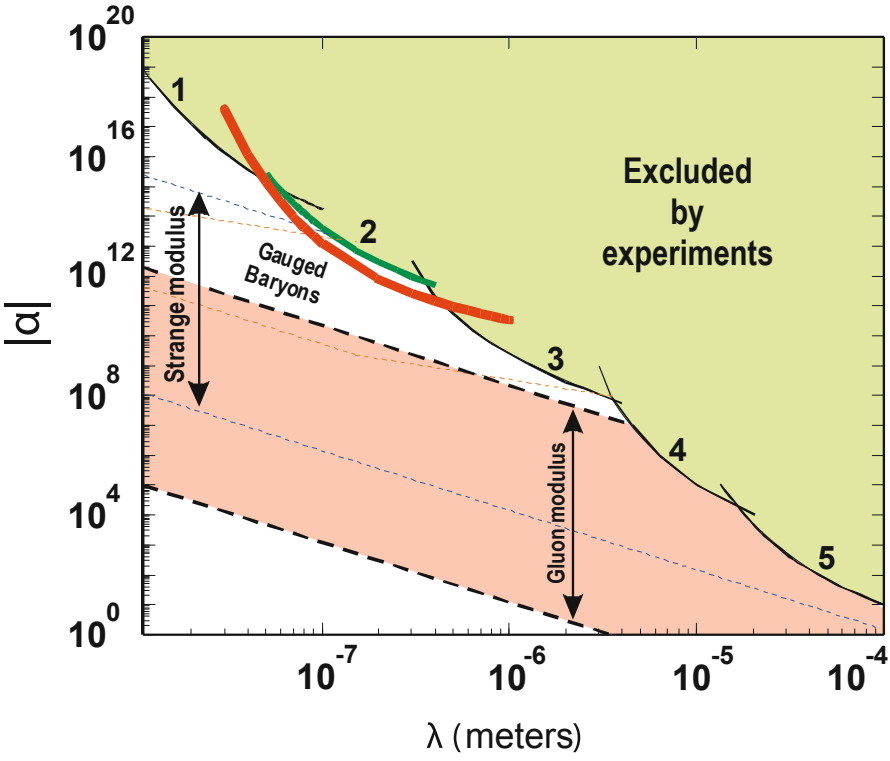


Fig. 12 Bounds on non-Newtonian forces in the range $10 \div 200$ nm (see Decca et al. in Ref. [15]). Curves 4 and 5 correspond to Stanford and Colorado experiments, respectively, of Fig. 11 (see also Long and Price of Ref. [15])

which can be translated to a bound on the string scale $M_s \gtrsim 500$ GeV, comparable to other collider bounds. This allows for the striking possibility of a Higgs boson decaying dominantly, or at least with a sizable branching ratio, via such an invisible mode, for a wide range of the parameter space (M_s, m_H), as seen in Fig. 13.

9 Electroweak Symmetry Breaking

Non-supersymmetric TeV strings also offer a framework to realize gauge symmetry breaking radiatively. Indeed, from the effective field theory point of view, one expects quadratically divergent one-loop contributions to the masses of scalar fields. The divergences are cut off by M_s and if the corrections are negative, they can induce electroweak symmetry breaking and explain the mild hierarchy between the weak and a string scale at a few TeV, in terms of a loop factor [50]. More precisely, in the

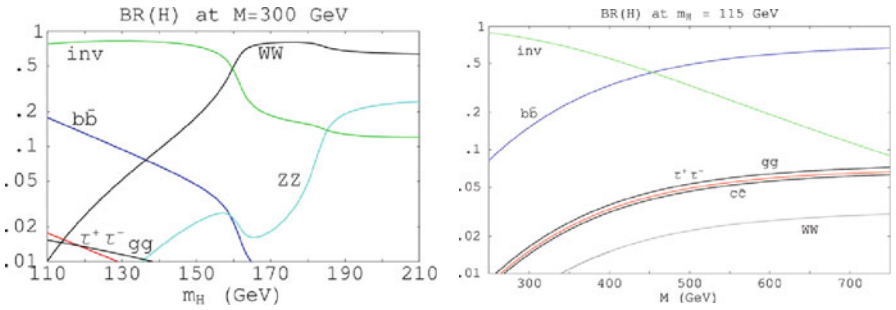


Fig. 13 Higgs branching rations, as functions either of the Higgs mass m_H for a fixed value of the string scale $M_s \simeq 2M = 600$ GeV, or of $M \simeq M_s/2$ for $m_H = 115$ GeV

minimal case of one Higgs doublet H, the scalar potential is:

$$V = \lambda(H^\dagger H)^2 + \mu^2(H^\dagger H), \tag{102}$$

where λ arises at tree level. Moreover, in any model where the Higgs field comes from an open string with both ends fixed on the same brane stack, it is given by an appropriate truncation of a supersymmetric theory. Within the minimal spectrum of the SM, $\lambda = (g_2^2 + g'^2)/8$, with g_2 and g' the $SU(2)$ and $U(1)_Y$ gauge couplings. On the other hand, μ^2 is generated at one loop:

$$\mu^2 = -\varepsilon^2\, g^2\, M_s^2, \tag{103}$$

where ε is a loop factor that can be estimated from a toy model computation and varies in the region $\varepsilon \sim 10^{-1} \div 10^{-3}$.

Indeed, consider for illustration a simple case where the whole one-loop effective potential of a scalar field can be computed. We assume, for instance, one extra dimension compactified on a circle of radius $R > 1$ (in string units). An interesting situation is provided by a class of models where a nonvanishing VEV for a scalar (Higgs) field ϕ results in shifting the mass of each KK excitation by a constant $a(\phi)$:

$$M_m^2 = \left(\frac{m + a(\phi)}{R}\right)^2, \tag{104}$$

with m the KK integer momentum number. Such mass shifts arise, for instance, in the presence of a Wilson line, $a = q \oint \frac{dy}{2\pi} g A$, where A is the internal component of a gauge field with gauge coupling g, and q is the charge of a given state under the corresponding generator. A straightforward computation shows that the ϕ-dependent part of the one-loop effective potential is given by [51]:

$$V_{eff} = -Tr(-)^F \frac{R}{32\,\pi^{3/2}} \sum_n e^{2\pi i n a} \int_0^\infty dl\; l^{3/2}\, f_s(l)\; e^{-\pi^2 n^2 R^2 l}, \tag{105}$$

where $F = 0, 1$ for bosons and fermions, respectively. We have included a regulating function $f_s(l)$ which contains, for example, the effects of string oscillators. To

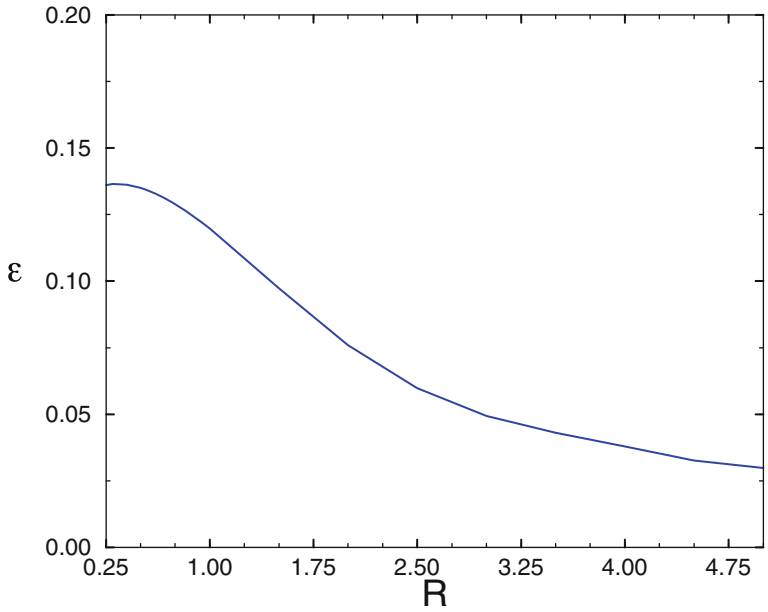


Fig. 14 The coefficient ε of the one-loop Higgs mass (103)

understand its role we will consider the two limits $R \gg 1$ and $R \ll 1$. In the first case, only the $l \to 0$ region contributes to the integral. This means that the effective potential receives sizable contributions only from the infrared (field theory) degrees of freedom. In this limit we would have $f_s(l) \to 1$. For example, in the string model considered in [50]:

$$f_s(l) = \left[\frac{1}{4l}\frac{\theta_2}{\eta^3}(il + \frac{1}{2})\right]^4 \to 1 \qquad \text{for} \qquad l \to 0, \tag{106}$$

and the field theory result is finite and can be explicitly computed. As a result of the Taylor expansion around $a = 0$, we are able to extract the one-loop contribution to the coefficient of the term of the potential quadratic in the Higgs field. It is given by a loop factor times the compactification scale [51]. One thus obtains $\mu^2 \sim g^2/R^2$ up to a proportionality constant which is calculable in the effective field theory. On the other hand, if we consider $R \to 0$, which by T-duality corresponds to taking the extra dimension as transverse and very large, the one-loop effective potential receives contributions from the whole tower of string oscillators as appearing in $f_s(l)$, leading to squared masses given by a loop factor times M_s^2, according to Eq. (103).

More precisely, from the expression (105), one finds:

$$\varepsilon^2(R) = \frac{1}{2\pi^2}\int_0^\infty \frac{dl}{(2\,l)^{5/2}}\frac{\theta_2^4}{4\eta^{12}}\left(il + \frac{1}{2}\right)R^3\sum_n n^2 e^{-2\pi n^2 R^2 l}\,, \tag{107}$$

which is plotted in Fig. 14.

For the asymptotic value $R \to 0$ (corresponding upon T-duality to a large-transverse dimension of radius $1/R$), $\varepsilon(0) \simeq 0.14$, and the effective cutoff for the mass term is M_s, as can be seen from Eq. (103). At large R, $\mu^2(R)$ falls off as $1/R^2$, which is the effective cutoff in the limit $R \to \infty$, as we argued above, in agreement with field theory results in the presence of a compactified extra dimension [41,52]. In fact, in the limit $R \to \infty$, an analytic approximation to $\varepsilon(R)$ gives:

$$\varepsilon(R) \simeq \frac{\varepsilon_\infty}{M_s\, R}\,, \qquad \varepsilon_\infty^2 = \frac{3\,\zeta(5)}{4\,\pi^4} \simeq 0.008\,. \tag{108}$$

The potential (102) has the usual minimum, given by the VEV of the neutral component of the Higgs doublet $v = \sqrt{-\mu^2/\lambda}$. Furthermore, from (103), one can compute M_s in terms of the Higgs mass $m_H^2 = -2\mu^2$:

$$M_s = \frac{m_H}{\sqrt{2}\, g\varepsilon}\,, \tag{109}$$

yielding naturally values in the TeV range.

10 Standard Model on D-bratnes

The gauge group closest to the SM one can easily obtain with D-branes is $U(3) \times U(2) \times U(1)$. The first factor arises from three coincident "color" D-branes. An open string with one end on them is a triplet under $SU(3)$ and carries the same $U(1)$ charge for all three components. Thus, the $U(1)$ factor of $U(3)$ has to be identified with *gauged* baryon number. Similarly, $U(2)$ arises from two coincident "weak" D-branes and the corresponding abelian factor is identified with *gauged* weak-doublet number. Finally, an extra $U(1)$ D-brane is necessary in order to accommodate the SM without breaking the baryon number [53]. In principle this $U(1)$-brane can be chosen to be independent of the other two collections with its own gauge coupling. To improve the predictability of the model, we choose to put it on top of either the color or the weak D-branes [54]. In either case, the model has two independent gauge couplings g_3 and g_2 corresponding, respectively, to the gauge groups $U(3)$ and $U(2)$. The $U(1)$ gauge coupling g_1 is equal to either g_3 or g_2.

Let us denote by Q_3, Q_2 and Q_1 the three $U(1)$ charges of $U(3) \times U(2) \times U(1)$, in a self explanatory notation. Under $SU(3) \times SU(2) \times U(1)_3 \times U(1)_2 \times U(1)_1$, the members of a family of quarks and leptons have the following quantum numbers:

$$\begin{aligned}
&Q\ \ (\mathbf{3}, \mathbf{2}; 1, w, 0)_{1/6} \\
&u^c\ (\bar{\mathbf{3}}, \mathbf{1}; -1, 0, x)_{-2/3} \\
&d^c\ (\bar{\mathbf{3}}, \mathbf{1}; -1, 0, y)_{1/3} \\
&L\ \ (\mathbf{1}, \mathbf{2}; 0, 1, z)_{-1/2} \\
&l^c\ \ (\mathbf{1}, \mathbf{1}; 0, 0, 1)_1\,.
\end{aligned} \tag{110}$$

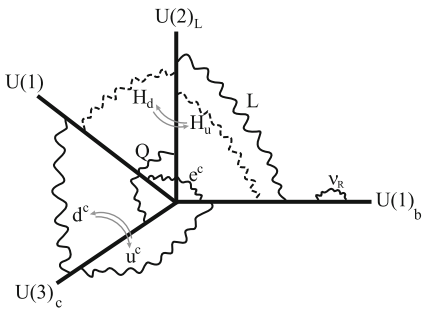


Fig. 15 A minimal Standard Model embedding on D-branes

The values of the $U(1)$ charges x, y, z, w will be fixed below so that they lead to the right hypercharges, shown for completeness as subscripts.

It turns out that there are two possible ways of embedding the SM particle spectrum on these stacks of branes [53], which are shown pictorially in Fig. 15.

The quark doublet Q corresponds necessarily to a massless excitation of an open string with its two ends on the two different collections of branes (color and weak). As seen from the figure, a fourth brane stack is needed for a complete embedding, which is chosen to be a $U(1)_b$ extended in the bulk. This is welcome since one can accommodate right-handed neutrinos as open string states on the bulk with sufficiently small Yukawa couplings suppressed by the large volume of the bulk [55]. The two models are obtained by an exchange of the up and down antiquarks, u^c and d^c, which correspond to open strings with one end on the color branes and the other one either on the $U(1)$-brane, or on the $U(1)_b$ in the bulk. The lepton doublet L arises from an open string stretched between the weak branes and $U(1)_b$, while the antilepton l^c corresponds to a string with one end on the $U(1)$-brane and the other in the bulk. For completeness, we also show the two possible Higgs states H_u and H_d that are both necessary in order to give tree-level masses to all quarks and leptons of the heaviest generation.

10.1 Hypercharge Embedding and the Weak Angle

The weak hypercharge Y is a linear combination of the three $U(1)$'s:

$$Y = Q_1 + \frac{1}{2}Q_2 + c_3 Q_3 \quad ; \quad c_3 = -1/3 \text{ or } 2/3\,, \tag{111}$$

where Q_N denotes the $U(1)$ generator of $U(N)$ normalized so that the fundamental representation of $SU(N)$ has unit charge. The corresponding $U(1)$ charges appearing in Eq. (110) are $x = -1$ or 0, $y = 0$ or 1, $z = -1$, and $w = 1$ or -1, for

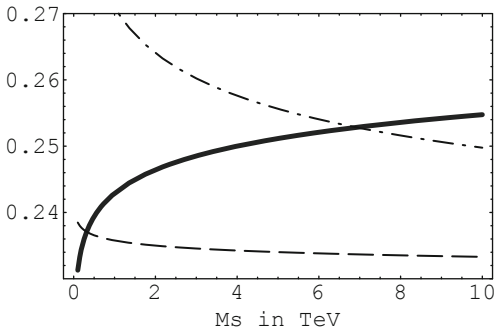


Fig. 16 The experimental value of $\sin^2\theta_W$ (*thick curve*), and the theoretical predictions (113) [53]

$c_3 = -1/3$ or $2/3$, respectively. The hypercharge coupling g_Y is given by [5]:

$$\frac{1}{g_Y^2} = \frac{2}{g_1^2} + \frac{4c_2^2}{g_2^2} + \frac{6c_3^2}{g_3^2}\,. \tag{112}$$

It follows that the weak angle $\sin^2\theta_W$, is given by:

$$\sin^2\theta_W \equiv \frac{g_Y^2}{g_2^2 + g_Y^2} = \frac{1}{2 + 2g_2^2/g_1^2 + 6c_3^2 g_2^2/g_3^2}\,, \tag{113}$$

where g_N is the gauge coupling of $SU(N)$ and $g_1 = g_2$ or $g_1 = g_3$ at the string scale. In order to compare the theoretical predictions with the experimental value of $\sin^2\theta_W$ at M_s, we plot in Fig. 16 the corresponding curves as functions of M_s. The solid line is the experimental curve. The dashed line is the plot of the function (113) for $g_1 = g_2$ with $c_3 = -1/3$ whereas the dotted-dashed line corresponds to $g_1 = g_3$ with $c_3 = 2/3$. The other two possibilities are not shown because they lead to a value of M_s which is too high to protect the hierarchy. Thus, the second case, where the $U(1)$-brane is on top of the color branes, is compatible with low-energy data for $M_s \sim 6\ div8$ TeV and $g_s \simeq 0.9$.

From Eq. (113) and Fig. 16, we find the ratio of the $SU(2)$ and $SU(3)$ gauge couplings at the string scale to be $\alpha_2/\alpha_3 \sim 0.4$. This ratio can be arranged by an appropriate choice of the relevant moduli. For instance, one may choose the color and $U(1)$-branes to be D3-branes while the weak branes to be D7-branes. Then, the ratio of couplings above can be explained by choosing the volume of the four compact dimensions of the seven branes to be $V_4 = 2.5$ in string units. This being larger than one is consistent with the picture above. Moreover, it predicts an interesting spectrum of KK states for the SM, different from the naive choices that have appeared hitherto: The only SM particles that have KK descendants are the W

[5] The gauge couplings $g_{2,3}$ are determined at the tree-level by the string coupling and other moduli, like radii of longitudinal dimensions. In higher orders, they also receive string threshold corrections.

bosons as well as the hypercharge gauge boson. However, since the hypercharge is a linear combination of the three $U(1)$'s, the massive $U(1)$ KK gauge bosons do not couple to the hypercharge but to the weak doublet number.

10.2 The Fate of U(1)'s, Proton Stability and Neutrino Masses

It is easy to see that the remaining three $U(1)$ combinations orthogonal to Y are anomalous. In particular there are mixed anomalies with the $SU(2)$ and $SU(3)$ gauge groups of the SM. These anomalies are cancelled by three axions coming from the closed string RR (Ramond) sector, via the standard Green–Schwarz mechanism [56]. The mixed anomalies with the non-anomalous hypercharge are also cancelled by dimension-five Chern–Simons type of interactions [53]. An important property of the above Green–Schwarz anomaly cancellation mechanism is that the anomalous $U(1)$ gauge bosons acquire masses leaving behind the corresponding global symmetries. This is in contrast to what would have happened in the case of an ordinary Higgs mechanism. These global symmetries remain exact to all orders in type I string perturbation theory around the orientifold vacuum. This follows from the topological nature of Chan–Paton charges in all string amplitudes. On the other hand, one expects non-perturbative violation of global symmetries and consequently exponentially small in the string coupling, as long as the vacuum stays at the orientifold point. Thus, all $U(1)$ charges are conserved and since Q_3 is the baryon number, proton stability is guaranteed.

Another linear combination of the $U(1)$'s is the lepton number. Lepton number conservation is important for the extra-dimensional neutrino mass suppression mechanism described above, that can be destabilized by the presence of a large Majorana neutrino mass term. Such a term can be generated by the lepton-number violating dimension-five effective operator $LLHH$ that leads, in the case of TeV string scale models, to a Majorana mass of the order of a few GeV. Even if we manage to eliminate this operator in some particular model, higher-order operators would also give unacceptably large contributions, as we focus on models in which the ratio between the Higgs VEV and the string scale is just of order $\mathcal{O}(1/10)$. The best way to protect tiny neutrino masses from such contributions is to impose lepton number conservation.

A bulk neutrino propagating in $4+n$ dimensions can be decomposed in a series of 4D KK excitations denoted collectively by $\{m\}$:

$$S_{kin} = R_\perp^n \int d^4x \sum_{\{m\}} \left\{ \bar{\nu}_{Rm} \partial\!\!\!/ \nu_{Rm} + \bar{\nu}^c_{Rm} \partial\!\!\!/ \nu^c_{Rm} + \frac{m}{R_\perp} \nu_{Rm} \nu^c_{Rm} + \text{complex conjugate} \right\}, \tag{114}$$

where ν_R and ν_R^c are the two Weyl components of the Dirac spinor and for simplicity we considered a common compactification radius $R_\perp$. On the other hand, there is a localized interaction of ν_R with the Higgs field and the lepton doublet, which leads

to mass terms between the left-handed neutrino and the KK states ν_{Rm}, upon the Higgs VEV v:

$$S_{int} = g_s \int d^4x H(x) L(x) \nu_R(x, y = 0) \quad \rightarrow \quad \frac{g_s v}{R_\perp^{n/2}} \sum_m \nu_L \nu_{Rm} \,, \tag{115}$$

in string units. Since the mass mixing $g_s v / R_\perp^{n/2}$ is much smaller than the KK mass $1/R_\perp$, it can be neglected for all the excitations except for the zero-mode ν_{R0}, which gets a Dirac mass with the left-handed neutrino

$$m_\nu \simeq \frac{g_s v}{R_\perp^{n/2}} \simeq v \frac{M_s}{M_p} \simeq 10^{-3} \div 10^{-2} \text{ eV} \,, \tag{116}$$

for $M_s \simeq 1 \div 10$ TeV, where the relation (83) was used. In principle, with one bulk neutrino, one could try to explain both solar and atmospheric neutrino oscillations using also its first KK excitation. However, the later behaves like a sterile neutrino which is now excluded experimentally. Therefore, one has to introduce three bulk species (at least two) ν_R^i in order to explain neutrino oscillations in a "traditional way", using their zero-modes ν_{R0}^i [57]. The main difference with the usual seesaw mechanism is the Dirac nature of neutrino masses, which remains an open possibility to be tested experimentally.

11 Non-compact Extra Dimensions and Localized Gravity

There are several motivations to study localization of gravity in non-compact extra dimensions: (i) it avoids the problem of fixing the moduli associated to the size of the compactification manifold; (ii) it provides a new approach to the mass hierarchy problem; (iii) there are modifications of gravity at large distances that may have interesting observational consequences. Two types of models have been studied: Warped metrics in curved space discussed in Sect. 6 and infinite size extra dimensions in flat space [58]. The former, although largely inspired by stringy developments and having used many string-theoretic techniques, have not yet a clear and calculable string theory realization [59]. In any case, since curved space is always difficult to handle in string theory, in the following we concentrate mainly on the latter, formulated in flat space with gravity localized on a subspace of the bulk. It turns out that these models of induced gravity have an interesting string theory realization [60] that we describe below, after presenting first a brief overview of the warped case.

11.1 The Induced Gravity Godel

The DGP model and its generalizations are specified by a bulk Einstein–Hilbert (EH) term and a 4D EH term [58]:

$$M^{2+n}\int_{\mathcal{M}_{4+n}} d^4x d^n y\,\sqrt{G}\,\mathcal{R}_{(4+n)} + M_P^2\int_{\mathcal{M}_4} d^4x\sqrt{g}\,\mathcal{R}_{(4)}\,;\quad M_P^2\equiv r_c^n M^{2+n} \tag{117}$$

with M and M_P the (possibly independent) respective Planck scales. The scale $M \geq 1$ TeV would be related to the short-distance scale below which UV quantum gravity or stringy effects are important. The 4D metric is the restriction of the bulk metric $g_{\mu\nu} = G_{\mu\nu}\big|$ and we assume the WORLD[6] rigid, allowing the gauge $G_{i\mu}\big| = 0$ with $i \geq 5$.

11.1.1 Co-dimension One

In the case of co-dimension one bulk ($n = 1$) and δ-function localization, it is easy to see that r_c is a crossover scale where gravity changes behavior on the WORLD. Indeed, by Fourier transform the quadratic part of the action (117) with respect to the 4D position x, at the WORLD position $y = 0$, one obtains $M^{2+n}(p^{2-n}+r_c^n p^2)$, where p is the 4D momentum. It follows that for distances smaller than r_c (large momenta), the first term becomes irrelevant and the graviton propagator on the "brane" exhibits 4D behavior $(1/p^2)$ with Planck constant $M_P = M^3 r_c$. On the contrary, at large distances, the first term becomes dominant and the graviton propagator acquires a 5D fall-off $(1/p)$ with Planck constant M.

An important consequence of this model is the existence of self-accelerating solutions for the universe at late times [61]. Indeed, the Friedmann equations become:

$$H^2 \pm \frac{H}{r_c} = \frac{8\pi G_N}{3}\rho(t)\,, \tag{118}$$

where H is the Hubble constant and ρ the 4D energy density. For the upper sign (+), one finds usual Robertson–Walker cosmologies extrapolating between 4D and 5D at large scales, while for the lower (negative) sign, one finds a de Sitter solution at late times. This can account for the present dark energy if $r_c \simeq 10^{28}$ cm. On the other hand, this model suffers from the van Dam–Veltman–Zakharov discontinuity of massive gravity, implying a strong coupling regime at distances shorter than the Vainshtein radius $r_V = (r_c^2 w)^{1/3}$, where w is the UV cutoff, corresponding to the "brane" thickness [62]. Moreover, imposing in general r_c to be larger than the size of the universe, $r_c \gtrsim 10^{28}$ cm, one finds $M \lesssim 100$ MeV, which is in conflict with experimental bounds.

[6] We avoid calling $\mathcal{M}_4$ a brane because, as we will see below, gravity localizes on singularities of the internal manifold, such as orbifold fixed points. Branes with localized matter can be introduced independently of gravity localization.

11.1.2 Higher Co-dimension

The situation changes drastically for more than one non-compact bulk dimensions, $n > 1$, due to the UV properties of the higher-dimensional theories. Indeed, from the action (117), the effective potential between two test masses in four dimensions

$$\int [d^3x]\, e^{-ip\cdot x}\, V(x) = \frac{D(p)}{1 + r_c^n\, p^2\, D(p)} \left[\tilde{T}_{\mu\nu} T^{\mu\nu} - \frac{1}{2+n}\, \tilde{T}^{\mu}_{\mu}\, T^{\nu}_{\nu}\right] \quad (119)$$

$$D(p) = \int [d^n q]\, \frac{f_w(q)}{p^2 + q^2} \quad (120)$$

is a function of the bulk graviton retarded Green's function $G(x, 0;0, 0) = \int [d^4 p]\, e^{ip\cdot x}\, D(p)$ evaluated for two points localized on the WORLD ($y = y' = 0$). The integral (120) is UV-divergent for $n > 1$ unless there is a nontrivial brane thickness profile $f_w(q)$ of width w. If the 4D WORLD has zero thickness, $f_w(q) \sim 1$, the bulk graviton does not have a normalizable wave function. It therefore cannot contribute to the induced potential, which always takes the form $V(p) \sim 1/p^2$ and Newton's law remains 4D at all distances.

For a nonzero thickness w, there is only one crossover length scale, R_c:

$$R_c = w \left(\frac{r_c}{w}\right)^{\frac{n}{2}}, \quad (121)$$

above which one obtains a higher-dimensional behavior [63]. Therefore, the effective potential presents two regimes: (i) at short distances ($w \ll r \ll R_c$) the gravitational interactions are mediated by the localized 4D graviton and Newton's potential on the WORLD is given by $V(r) \sim 1/r$ and, (ii) at large distances ($r \gg R_c$) the modes of the bulk graviton dominate, changing the potential. Note that for $n = 1$ the expressions (119) and (120) are finite and unambiguously give $V(r) \sim 1/r$ for $r \gg r_c$. For a co-dimension bigger than 1, the precise behavior for large-distance interactions depends *crucially* on the UV completion of the theory. At this point we stress a fundamental difference with the *finite extra dimensions* scenarios. In these cases Newton's law gets higher-dimensional at distances smaller than the characteristic size of the extra dimensions. This is precisely the opposite of the case of infinite volume extra dimensions that we discuss here.

As mentioned above, for higher co-dimension, there is an interplay between UV regularization and IR behavior of the theory. Indeed, several works in the literature raised unitarity [64] and strong coupling problems [65] which depend crucially on the UV completion of the theory. A unitary UV regularization for the higher co-dimension version of the model has been proposed in [66]. It would be interesting to address these questions in a precise string theory context. Actually, using for UV cutoff on the "brane" the 4D Planck length $w \sim l_P$, one gets for the crossover scale (121): $R_c \sim M^{-1}(M_P/M)^{n/2}$. Putting $M \gtrsim 1$ TeV leads to $R_c \lesssim 10^{8(n-2)}$ cm. Imposing $R_c \gtrsim 10^{28}$ cm, one then finds that the number of extra dimensions must be at least six, $n \geq 6$, which is realized nicely in string theory and provides an additional motivation for studying possible string theory realizations.

11.2 String Theory Realization

The gravity induced model (117) with $n = 6$ is realized as the low-energy effective action of type II string theory on a non-compact 6D manifold $\mathcal{M}_6$ [60], preserving $\mathcal{N} = 2$ SUSY in four dimensions. The resulting expressions for the Planck masses M and M_p are:

$$M^2 \sim M_s^2/g_s^{1/2} \quad ; \quad M_P^2 \sim \chi \left(\frac{c_0}{g_s^2} + c_1\right) M_s^2 \,, \tag{122}$$

with $c_0 = -2\zeta(3)$, $c_1 = \pm 4\zeta(2) = \pm 2\pi^2/3$ and χ is the Euler number of $\mathcal{M}_6$.

It is interesting that the appearance of the induced 4D localized term preserves $\mathcal{N} = 2$ SUSY and is independent of the localization mechanism of matter fields (for instance on D-branes). Localization requires the internal space M_6 to have a nonzero Euler characteristic $\chi \neq 0$. Actually, in type IIA/B compactified on a Calabi–Yau manifold, χ counts the difference between the numbers of $\mathcal{N} = 2$ vector multiplets and hypermultiplets: $\chi = \pm 4(n_V - n_H)$ (where the graviton multiplet counts as one vector). Moreover, in the non-compact limit, the Euler number can in general split in different singular points of the internal space, $\chi = \sum_I \chi_I$, giving rise to different localized terms at various points y_I of the internal space. A number of conclusions (confirmed by string calculations in [60]) can be reached by looking closely at (122):

▷ $M_p \gg M$ requires a large nonzero Euler characteristic for M_6, and/or a weak string coupling constant $g_s \to 0$.

▷ Since χ is a topological invariant the localized $\mathcal{R}_{(4)}$ term coming from the closed string sector is universal, independent of the background geometry and dependent only on the internal topology. It is a matter of simple inspection to see that if one wants to have a localized EH term in less than ten dimensions, namely something linear in curvature, with non-compact internal space in all directions, *the only possible dimension is four* (or five in the strong coupling M-theory limit).

▷ In order to find the width w of the localized term, one has to do a separate analysis. On general grounds, using dimensional analysis in the limit $M_P \to \infty$, one expects the effective width to vanish as a power of $l_P \equiv M_P^{-1}$: $w \sim l_P^\nu / l_s^{\nu-1}$ with $\nu > 0$. The computation of ν for a general Calabi–Yau space, besides its technical difficulty, presents an additional important complication: from the expression (122), $l_P \sim g_s l_s$ in the weak coupling limit. Thus, w vanishes in perturbation theory and one has to perform a non-perturbative analysis to extract its behavior. Alternatively, one can examine the case of orbifolds. In this limit, $c_0 = 0$, $l_P \sim l_s$, and the hierarchy $M_P > M$ is achieved only in the limit of large χ. In this case, one finds the power $\nu = 1$.

11.2.1 Summary of the Results

Using $w \sim l_P$ and the relations (122) in the weak coupling limit (with $c_0 \neq 0$), the crossover radius of Eq. (121) is given by the string parameters ($n = 6$)

$$R_c = \frac{r_c^3}{w^2} \sim g_s \frac{l_s^4}{l_P^3} \simeq g_s \times 10^{32}\ \text{cm}\,, \tag{123}$$

for $M_s \simeq 1$ TeV. Because R_c has to be of cosmological size, the string coupling can be relatively small, and the Euler number $|\chi| \simeq g_s^2\, l_P \sim g_s^2 \times 10^{32}$ must be very large. The hierarchy is obtained, mainly thanks to the large value of χ, so that lowering the bound on R_c lowers the value of χ. Our actual knowledge of gravity at very large distances indicates [67] that R_c should be of the order of the Hubble radius $R_c \simeq 10^{28}$ cm, which implies $g_s \geq 10^{-4}$ and $|\chi| \gtrsim 10^{24}$. A large Euler number implies only a large number of closed string massless particles with no a priori constraint on the observable gauge and matter sectors, which can be introduced for instance on D3-branes placed at the position where gravity localization occurs. All these particles are localized at the orbifold fixed points (or where the Euler number is concentrated in the general case), and should have sufficiently suppressed gravitational-type couplings, so that their presence with such a huge multiplicity does not contradict observations.

The explicit string realization of localized induced gravity models offers a consistent framework that allows to address a certain number of interesting physics problems. In particular, the effective UV cutoff and the study of the gravity force among matter sources localized on D-branes. It would be also interesting to perform explicit model building and study in detail the phenomenological consequences of these models and compare to other realizations of TeV strings with compact dimensions.

Acknowledgment This work was supported in part by the European Commission under the ERC Advanced Grant 226371 and the contract PITN-GA-2009-237920.

References

1. Liddle AR, Lyth DH. Cosmological inflation and large-scale structure. Cambridge: Cambridge University Press; 2000.
2. Green M, Schwarz J, Witten E. Superstring Theory, Vols. I and II, Cambridge: Cambridge University Press; 1987.
3. I. Antoniadis. A possible new dimension at a few TeV. Phys Lett B. 1990;246:377.
4. Arkani-Hamed N, Dimopoulos S, Dvali GR. The hierarchy problem and new dimensions at a millimeter. Phys Lett B. 1998;429:263 [arXiv:hep-ph/9803315]; Antoniadis I, Arkani-Hamed N, Dimopoulos S, Dvali GR. New dimensions at a millimeter to a Fermi and superstrings at a TeV. Phys Lett B. 1998;436:257 [arXiv:hep-ph/9804398].
5. Dienes KR. String Theory and the Path to Unification: A Review of Recent Developments. Phys Rept. 1997;287:447 [arXiv:hep-th/9602045]; and references therein.

6. Kaluza T. On the problem of unity in physics. Preuss Akad Wiss. 1921;966; Klein O. Z Phys. 1926;37:895.
7. Antoniadis I, Benakli K. Limits on extra dimensions in orbifold compactifications of superstrings. Phys Lett B. 1994;326:69.
8. Angelantonj C, Sagnotti A. Open strings. Phys Rept. 2002;371:1 [Erratum-ibid. 2003;376:339] [arXiv:hep-th/0204089].
9. Kapner DJ, Cook TS, Adelberger EG, Gundlach JH, Heckel BR, Hoyle CD, Swanson HE. Tests of the gravitational inverse-square law below the dark-energy length scale. Phys Rev Lett. 2007;98:021101. Hoyle CD, Kapner DJ, Heckel BR, Adelberger EG, Gundlach JH, Schmidt U, Swanson HE. Sub-millimeter tests of the gravitational inverse-square law. Phys Rev D. 2004;70:042004.
10. Antoniadis I. Physics of low scale string models. Prepared for NATO Advanced Study Institute and EC Summer School on Progress in String, Field and Particle Theory, Cargese, Corsica, France; 2002; and references therein.
11. Lüst D, Theisen S. Lectures in string theory, lecture notes in physics, 346. Berlin: Springer Verlag; 1989.
12. Polchinski J. String theory, Vols. I and II. Cambridge: Cambridge University Press; 1998.
13. Kiritsis E. Introduction to superstring theory. hep-th/9709062.
14. Lykken JD. Weak scale superstrings. Phys Rev D. 1996;54:3693 [arXiv:hep-th/9603133].
15. Long JC, Price JC. Current short-range tests of the gravitational inverse square law. Comptes Rendus Phys. 2003;4:337; Decca RS, Lopez D, Chan HB, Fischbach E, Krause DE, Jamell CR. Constraining new forces in the Casimir regime using the isoelectronic technique. Phys Rev Lett. 2005;94:240401; Decca RS, Lopez D, Fischbach E, Klimchitskaya GL, Krause DE, Mostepanenko VM. Novel constraints on light elementary particles and extra-dimensional physics from the Casimir effect. arXiv:0706.3283 [hep-ph]; Smullin SJ, Geraci AA, Weld DM, Chiaverini J, Holmes S, Kapitulnik A. New constraints on Yukawa-type deviations from Newtonian gravity at 20-microns. arXiv:hep-ph/0508204; Abele H, Haeßler S, Westphal A. In 271th WE-Heraeus-Seminar. Bad Honnef; 2002.
16. Dvali G. Black holes and large N Species Solution to the Hierarchy Problem. arXiv:0706.2050 [hep-th]; Dvali G. Nature of microscopic black holes and gravity in theories with particle Species. Int J Mod Phys A. 2010;25:602 [arXiv:0806.3801 [hep-th]]; Dvali G, Redi M. Black hole bound on the number of species and quantum gravity at LHC. Phys Rev D. 2008;77:045027 [arXiv:0710.4344 [hep-th]]; Brustein R, Dvali G, Veneziano G. A bound on the effective gravitational coupling from semiclassical black holes. JHEP. 2009;0910:085 [arXiv:0907.5516 [hep-th]].
17. Dvali G, Gomez C. Quantum information and gravity cutoff in theories with species. Phys Lett B. 2009;674:303. [arXiv:0812.1940 [hep-th]].
18. Antoniadis I, Pioline B. Low scale closed strings and their duals. Nucl Phys B. 1999;550:41; Antoniadis I, Dimopoulos S, Giveon A. Little string theory at a TeV. JHEP. 2001;0105:055. [arXiv:hep-th/0103033]; Antoniadis I, Arvanitaki A, Dimopoulos S, Giveon A. Phenomenology of TeV little string theory from holography. arXiv:1102.4043.
19. Dvali G, Lust D. Evaporation of microscopic black holes in string theory and the bound on species. arXiv:0912.3167 [hep-th]; Dvali G, Gomez C. Species and strings. arXiv:1004.3744 [hep-th].
20. Randall L, Sundrum R. An alternative to compactification. Phys Rev Lett. 1999;83:4690 [arXiv:hep-th/9906064]; and Randall L, Sundrum S. A large mass hierarchy from a small extra dimension. Phys Rev Lett. 1999;83:3370. [arXiv:hep-ph/9905221].
21. Contino R, Pomarol A. The holographic composite Higgs. Comptes Rendus Phys. 2007;8:1058; and references therein.
22. Dienes KR, Dudas E, Gherghetta T. Extra spacetime dimensions and unification. Phys Lett B. 1998;436:55. [arXiv:hep-ph/9803466]; Dienes KR, Dudas E, Gherghetta T. Grand unification at intermediate mass scales through extra dimensions. Nucl Phys B. 1999;537:47 [arXiv:hep-ph/9806292].

23. Antoniadis I, Benakli K, Quirós M. Production of Kaluza-Klein states at future colliders. Phys Lett B. 1994;331:313 and Direct collider signatures of large extra dimensions. Phys Lett B. 1999;460:176; Nath P, Yamada Y, Yamaguchi M. Probing the nature of compactification with Kaluza-Klein excitations at the Large Hadron Collider. Phys Lett B. 1999;466:100; Rizzo TG, Wells JD. Electroweak precision measurements and collider probes of the standard model with large extra dimensions. Phys Rev D. 2000;61:016007; Rizzo TG. Testing the nature of Kaluza-Klein excitations at future lepton colliders. Phys Rev D. 2000;61:055005 [arXiv:hep-ph/9909232]. De Rujula A, Donini A, Gavela MB, Rigolin S. Fat brane phenomena. Phys Rev D. 2000;482:195;
24. Accomando E, Antoniadis I, Benakli K. Looking for TeV scale strings and extra dimensions. Nucl Phys B. 2000;579:3.
25. Antoniadis I, Benakli K, Laugier A. Contact interactions in D-brane models. JHEP. 2001;0105:044.
26. Nath P, Yamaguchi M. Effects of extra space-time dimensions on the Fermi constant. Phys Rev D. 1999;60:116004; Effects of Kaluza-Klein excitations on (g(mu)-2). Phys Rev D. 1999;60:116006; Masip M, Pomarol A. Effects of SM Kaluza-Klein excitations on electroweak observables. Phys Rev D. 1999;60:096005; Marciano WJ. Fermi constants and 'New Physics'. Phys Rev D. 600930061999; Strumia A. Bounds on Kaluza-Klein excitations of the SM vector bosons from electroweak tests. Phys Lett B. 1999;466:107; Casalbuoni R, De Curtis S, Dominici D, Gatto R. SM Kaluza-Klein excitations and electroweak precision tests. Phys Lett B. 1999;462:48; Carone CD. Electroweak constraints on extended models with extra dimensions. Phys Rev D. 2000;61:015008; Delgado A, Pomarol A, Quirós M. Electroweak and flavor physics in extensions of the standard model with large extra dimensions. JHEP 2000;1:30.
27. Servant G, Tait TMP. Is the lightest Kaluza-Klein particle a viable dark matter candidate? Nucl Phys B. 2003;650:391.
28. Giudice GF, Rattazzi R, Wells JD. Quantum gravity and extra dimensions at high-energy colliders. Nucl Phys B. 1999;544:3; Mirabelli EA, Perelstein M, Peskin ME. Collider signatures of new large space dimensions. Phys Rev Lett. 1999;82:2236; Han T, Lykken JD, Zhang R. On Kaluza-Klein states from large extra dimensions. Phys Rev D. 1999;59:105006; Cheung K, Keung W-Y. Direct signals of low scale gravity at $e^+ e^-$ colliders. Phys Rev D. 1999;60:112003; Balázs C, et al. Collider tests of compact space dimensions using weak gauge bosons. Phys Rev Lett. 1999;83:2112; L3 Collaboration (Acciarri M, et al.) Search for low scale gravity effects in $e^+ e^-$ collisions at LEP. Phys Lett B. 1999;464:135 and 1999;470:281; Hewett JL. Search for extra dimensions in boson and fermion pair production in $e^+ e^-$ interactions at LEP. Phys Rev Lett. 1999;82:4765.
29. Arkani-Hamed N, Dimopoulos S, Dvali G. Phenomenology, astrophysics and cosmology of theories with submillimeter dimensions and TeV scale quantum gravity. Phys Rev D. 1999;59:086004.
30. Cullen S, Perelstein M. SN1987A constraints on large compact dimensions. Phys Rev Lett. 1999;83:268; Barger V, Han T, Kao C, Zhang RJ. Astrophysical constraints on large extra dimensions. Phys Lett B. 1999;461:34.
31. Benakli K, Davidson S. Baryogenesis in models with a low quantum gravity scale. Phys Rev D. 1999;60:025004; Hall LJ, Smith D. Cosmological constraints on theories with large extra dimensions. Phys Rev D. 1999;60:085008.
32. Dudas E, Mourad J. String theory predictions for future accelerators. Nucl Phys B. 2000;575:3 [arXiv:hep-th/9911019]; Cullen S, Perelstein M, Peskin ME. Phys Rev D. 2000;62:055012; Bourilkov D. Phys Rev D. 2000:62:076005; L3 Collaboration (Acciarri M, et al.) Phys Lett B. 2000;489:81.
33. Argyres PC, Dimopoulos S, March-Russell J. Black holes and sub-millimeter dimensions. Phys Lett B. 1998;441:96 [arXiv:hep-th/9808138]; Banks T, Fischler W. A model for high energy scattering in quantum gravity. arXiv:hep-th/9906038.

34. Giddings SB, Thomas S. High energy colliders as black hole factories: The end of short distance physics. Phys Rev D. 2002;65:056010 [arXiv:hep-ph/0106219]; Dimopoulos S, Landsberg G. Black holes at the LHC. Phys Rev Lett. 2001:87;161602 [arXiv:hep-ph/0106295].
35. Meade P, Randall L. Black holes and quantum gravity at the LHC. arXiv:0708.3017 [hep-ph].
36. Anchordoqui LA, Goldberg H, Lust D, Nawata S, Stieberger S, Taylor TR. Dijet signals for low mass strings at the LHC. Phys Rev Lett. 2008:101:241803 [arXiv:0808.0497 [hep-ph]].
37. Lust D, Stieberger S, Taylor TR. The LHC string Hunter's companion. Nucl Phys B. 2009:808:1 [arXiv:0807.3333 [hep-th]].
38. Horowitz GT, Polchinski J. A correspondence principle for black holes and strings. Phys Rev D. 1997;55:6189 [arXiv:hep-th/9612146].
39. Antoniadis I, Bachas C. Branes and the gauge hierarchy. Phys Lett B. 1999;450:83.
40. Antoniadis I, Benakli K, Laugier A, Maillard T. Brane to bulk supersymmetry breaking and radion force at micron distances. Nucl Phys B. 2003;662:40 [arXiv:hep-ph/0211409].
41. Antoniadis I, Dimopoulos S, Dvali G. Millimeter range forces in superstring theories with weak scale compactification. Nucl Phys B. 1998;516:70; Ferrara S, Kounnas C, Zwirner F. Mass formulae and natural hierarchy in string effective supergravities. Nucl Phys B. 1994;429:589.
42. Adelberger EG, Heckel BR, Hoedl S, Hoyle CD, Kapner DJ, Upadhye A. Constraints on exotic interactions from a recent test of the gravitational inverse sqaure law. Phys Rev Lett. 2007;98:131104.
43. Taylor TR, Veneziano G. Dilaton couplings at large distances. Phys Lett B. 1988;213:450.
44. Antoniadis I, Kiritsis E, Rizos J. Anomalous U(1)'s in type I string vacua. Nucl Phys B. 2002;637:92; Antoniadis I, Benakli K, Laugier A. Nucl Phys B. 2002;631:3 [arXiv:hep-th/0111209].
45. Dudas E, Mourad J. Consistent gravitino couplings in non-supersymmetric strings. Phys Lett B. 2001;514:173 [arXiv:hep-th/0012071].
46. Volkov DV, Akulov VP. Possible universal neutrino interaction. JETP Lett. 1972;16:438 and Phys Lett B. 1973;46:109.
47. Brignole A, Feruglio F, Zwirner F. On the effective interactions of a light gravitino with matter fermions. JHEP. 1997:9711:001 [arXiv:hep-th/9709111]; Clark TE, Lee T, Love ST, Wu G. Phys Rev D. 1998;57:5912 [arXiv:hep-ph/9712353]; Luty MA, Ponton E. Phys Rev D. 1998;57:4167 [arXiv:hep-ph/9706268v3]; Antoniadis I, Benakli K, Laugier A. Nucl Phys B. 2002;631:3 [arXiv:hep-th/0111209].
48. Antoniadis I, Tuckmantel M. Non-linear supersymmetry and intersecting D-branes. Nucl Phys B. 2004;697:3 [arXiv:hep-th/0406010].
49. Antoniadis I, Tuckmantel M, Zwirner F. Phenomenology of a leptonic goldstino and invisible Higgs boson decays. Nucl Phys B. 2005;707:215 [arXiv:hep-ph/0410165].
50. Antoniadis I, Benakli K, Quirós M. Radiative symmetry breaking in brane models. Nucl Phys B. 2000;583:35.
51. Antoniadis I, Benakli K, Quiros M. Finite Higgs mass without supersymmetry. New J Phys. 2001;3:20 [arXiv:hep-th/0108005].
52. Antoniadis I, Muñoz C, Quirós M. Dynamical supersymmetry breaking with a large internal dimension. Nucl Phys B. 1993;397:515; Antoniadis I, Quirós M. Phys Lett B. 1997;392:61; Pomarol A, Quirós M. Phys Lett B. 1998;438:225; Antoniadis I, Dimopoulos S, Pomarol A, Quirós M. Nucl Phys B. 1999;544:503; Delgado A, Pomarol A, Quirós M. Phys Rev D. 1999;60:095008; Barbieri R, Hall LJ, Nomura Y, Phys Rev D. 2001;63:105007.
53. Antoniadis I, Kiritsis E, Tomaras TN. A D-brane alternative to unification. Phys Lett B. 2000;486:186; Antoniadis I, Kiritsis E, Rizos J, Tomaras TN. D-branes and the standard model. Nucl Phys B. 2003;660:81 [arXiv:hep-th/0210263].
54. Shiu G, Tye S-HH. TeV scale superstring and extra dimension. Phys Rev D. 1998;58:106007; hep-th/9805157; Kakushadze Z, Tye S-HH. Nucl Phys B. 1999;548:180 [arXiv:hep-th/9809147]; Ibáñez LE, Muñoz C, Rigolin S. Aspects of type I string phenomenology. Nucl Phys B. 1999;553:43 [arXiv:hep-ph/9812397].

55. Dienes KR, Dudas E, Gherghetta T. Light neutrinos without heavy mass scales: A higher-dimensional seesaw mechanism. Nucl Phys B. 1999;557:25 [arXiv:hep-ph/9811428]; Arkani-Hamed N, Dimopoulos S, Dvali GR, March-Russell J. Neutrino masses from large extra dimensions. Phys Rev D. 2002;65:024032 [arXiv:hep-ph/9811448]; Dvali GR, Smirnov AY. Probing large extra dimensions with neutrinos. Nucl Phys B. 1999;563:63 [arXiv:hep-ph/9904211].
56. Sagnotti A. A Note on the Green-Schwarz mechanism in open string theories. Phys Lett B. 1992;294:196; Ibáñez LE, Rabadán R, Uranga AM. Anomalous U(1)'s in type I and type IIB D = 4, N = 1 string vacua. Nucl Phys B. 1999;542:112; Poppitz E. On the one loop Fayet-Iliopoulos term in chiral four-dimensional type I orbifolds. Nucl Phys B. 1999;542:31.
57. Davoudiasl H, Langacker P, Perelstein M. Constraints on large extra dimensions from neutrino oscillation experiments. Phys Rev D. 2002;65:105015 [arXiv:hep-ph/0201128].
58. Dvali GR, Gabadadze G, Porrati M. 4D gravity on a brane in 5D Minkowski space. Phys Lett B. 2000;485:208 [arXiv:hep-th/0005016].
59. Verlinde H. Holography and compactification. Nucl Phys B. 2000;580:264 [arXiv:hep-th/9906182]; Giddings SB, Kachru S, Polchinski J. Hierarchies from fluxes in string compactifications. Phys Rev D. 2002;66:106006 [arXiv:hep-th/0105097].
60. Antoniadis I, Minasian R, Vanhove P. Non-compact Calabi-Yau manifolds and localized gravity. Nucl Phys B. 2003;648:69 [arXiv:hep-th/0209030].
61. Deffayet C, Dvali GR, Gabadadze G. Accelerated universe from gravity leaking to extra dimensions. Phys Rev D. 2002;65:044023 [arXiv:astro-ph/0105068].
62. Luty MA, Porrati M, Rattazzi R. Strong interactions and stability in the DGP model. arXiv:hep-th/0303116.
63. Dvali GR, Gabadadze G. Gravity on a brane in infinite-volume extra space. Phys Rev D. 2001;63:065007 [arXiv:hep-th/0008054]; Dvali GR, Gabadadze G, Kolanovic M, Nitti F. The power of brane-induced gravity. Phys Rev D. 2001;64:084004 [arXiv:hep-ph/0102216].
64. Dubovsky SL, Rubakov VA. Brane-induced gravity in more than one extra dimensions: Violation of equivalence principle and ghost. Phys Rev D. 2003;67:104014 [arXiv:hep-th/0212222].
65. Rubakov VA. Strong coupling in brane-induced gravity in five dimensions. arXiv:hep-th/0303125.
66. Kolanovic M, Porrati M, Rombouts JW. Regularization of brane induced gravity. Phys Rev D. 2003;68:064018 [arXiv:hep-th/0304148].
67. Lue A, Starkman G. Gravitational leakage into extra dimensions: Probing dark energy using local gravity. Phys Rev D. 2003;67:064002 [arXiv:astro-ph/0212083].

Index

R. Peron et al. (eds.), *Gravity: Where Do We Stand?*, DOI 10.1007/978-3-319-20224-2