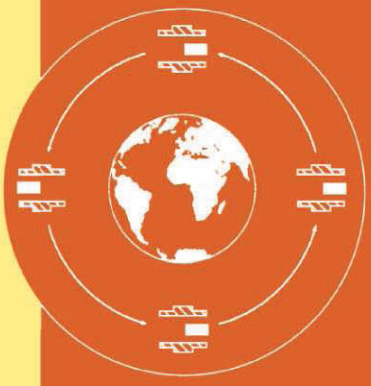
LECTURE NOTES IN PHYSICS

C. Lämmerzahl C.W.F. Everitt F.W. Hehl (Eds.)

Gyros, Clocks, Interferometers...: Testing Relativistic Gravity in Space



Springer

Lecture Notes in Physics

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Editorial Policy

The series *Lecture Notes in Physics* (LNP), founded in 1969, reports new developments in physics research and teaching -- quickly, informally but with a high quality. Manuscripts to be considered for publication are topical volumes consisting of a limited number of contributions, carefully edited and closely related to each other. Each contribution should contain at least partly original and previously unpublished material, be written in a clear, pedagogical style and aimed at a broader readership, especially graduate students and nonspecialist researchers wishing to familiarize themselves with the topic concerned. For this reason, traditional proceedings cannot be considered for this series though volumes to appear in this series are often based on material presented at conferences, workshops and schools (in exceptional cases the original papers and/or those not included in the printed book may be added on an accompanying CD ROM, together with the abstracts of posters and other material suitable for publication, e.g. large tables, colour pictures, program codes, etc.).

Acceptance

A project can only be accepted tentatively for publication, by both the editorial board and the publisher, following thorough examination of the material submitted. The book proposal sent to the publisher should consist at least of a preliminary table of contents outlining the structure of the book together with abstracts of all contributions to be included.
Final acceptance is issued by the series editor in charge, in consultation with the publisher, only after receiving the complete manuscript. Final acceptance, possibly requiring minor corrections, usually follows the tentative acceptance unless the final manuscript differs significantly from expectations (project outline). In particular, the series editors are entitled to reject individual contributions if they do not meet the high quality standards of this series. The final manuscript must be camera-ready, and should include both an informative introduction and a sufficiently detailed subject index.

Contractual Aspects

Publication in LNP is free of charge. There is no formal contract, no royalties are paid, and no bulk orders are required, although special discounts are offered in this case. The volume editors receive jointly 30 free copies for their personal use and are entitled, as are the contributing authors, to purchase Springer books at a reduced rate. The publisher secures the copyright for each volume. As a rule, no reprints of individual contributions can be supplied.

Manuscript Submission

The manuscript in its final and approved version must be submitted in camera-ready form. The corresponding electronic source files are also required for the production process, in particular the online version. Technical assistance in compiling the final manuscript can be provided by the publisher's production editor(s), especially with regard to the publisher's own Latex macro package which has been specially designed for this series.

Online Version/ LNP Homepage

LNP homepage (list of available titles, aims and scope, editorial contacts etc.):
http://www.springer.de/phys/books/lnpp/

LNP online (abstracts, full-texts, subscriptions etc.):
http://link.springer.de/series/lnpp/

C. Lämmerzahl C. W. F. Everitt F. W. Hehl (Eds.)

Gyros, Clocks, Interferometers...: Testing Relativistic Gravity in Space

Springer

Editors

Claus Lämmerzahl
Institute for Experimental Physics
Heinrich-Heine-University Düsseldorf
40225 Düsseldorf, German

C. W. Francis Everitt
W.W. Hansen Experimental Physics Laboratory
Stanford University
Stanford, CA 94305, USA

Friedrich W. Hehl
Institute for Theoretical Physics
University of Cologne
50923 Köln, Germany

Cover picture: see contribution by N. Lockerbie in this volume (Fig. 1, modified).

Library of Congress Cataloging-in-Publication Data applied for.

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Gyros, clocks, interferometers ... : testing relativistic gravity in space / C. Lämmerzahl ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 2001
(Lecture notes in physics ; Vol. 562)
(Physics and astronomy online library)
ISBN 3-540-41236-0

ISSN 0075-8450
ISBN 3-540-41236-0 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by the authors/editors
Camera-data conversion by Steingraeber Satztechnik GmbH Heidelberg
Cover design: *design & production*, Heidelberg

Printed on acid-free paper
SPIN: 10786632 55/3141/du - 5 4 3 2 1 0

Preface

Nowadays, Experimental Gravitation is one of the exciting areas in modern physics. Many important experiments have entered their final state of realization: In 2001 the interferometric gravitational wave detectors will start to collect data, in May 2002 Gravity Probe B (Stanford orbiting gyroscope experiment) will be launched, and, in 2005, STEP (the Satellite Test of the Equivalence Principle) is planned to be put in orbit. All these experiments will give new momentum to the field of experimental gravitation and, of course, to the development of theoretical concepts in gravity.

Besides this, the development of many new devices for carrying through high-precision measurements will open up new areas for testing the gravitational interaction. New atomic *clocks*, such as the atomic fountain clock, for example, or clocks which will be based on Bose–Einstein condensates, will lead to an increased precision, giving better results of tests of, for example, the gravitational red shift, or will enable the measurement of gravitomagnetic effects for clocks. Also, new quantum devices like atom *interferometers* or interferometers with Bose–Einstein condensates will give much better results while probing the gravitational field. And, last but not least, clocks will have an application in establishing a global reference system which is used for the GPS (Global Positioning System) and for telecommunication.

From the theoretical point of view all these experiments are highly interesting and the new results will stimulate the effort of unifying the four interactions and/or finding a consistent combination of gravitation theory and quantum theory, i.e., a quantum gravity theory. It is a general feature of such generalized theories of gravity that extra scalar interactions emerge accompanying the gravitational interaction. As one consequence they lead to a violation of the Equivalence Principle. Moreover, anomalous spin interactions may arise in that context. Therefore, tests of gravity may be at the same time tests of our microscopic view of the world.

Experiments are designed for testing the predictions of a theory as well as for testing the foundations of theories. In these proceedings both aspects are represented. The Lense–Thirring effect and gravitational waves are two of the main predictions of General Relativity which have not yet been confirmed directly or with convincing precision. As far as the foundations of General Relativity are concerned, we deal with the test of the Equivalence Principle which is at the very heart of the geometric nature of General Relativity. This principle is character-

istic for this kind of interaction. Other topics in this volume are experiments with clocks that measure the universality of the gravitational red shift. A violation of this universality implies the existence of more than one gravitational interaction. In addition, since the most precise clocks are atomic clocks, this class of experiments also yields information about the action of gravity in the quantum domain. Therefore, the application of new quantum-based devices, like atomic interferometers, has been examined for measurements of gravitational interaction. These and other devices seem to be very promising for increasing the experimental precision considerably. Thus, the effects of gravitation on quantum systems with spin are part of these proceedings. All these attempts will provide a better experimental foundation for the theory of gravity.

Accordingly, our meeting about *Gyros, Clocks, Interferometers, ... : Testing General Relativity in Space* held from 21 to 27 August 1999 in Bad Honnef and these proceedings are devoted to these topics:

- The Lense–Thirring effect
- The detection of gravitational waves
- Testing the Equivalence Principle
- Clocks and rods in gravitational fields
- Quantum tests of gravity
- Electromagnetic field and gravity

We want to present a review of the status of experimental gravity at the beginning of the 21st century. In doing so, we first tried to give a theoretical understanding of the various effects followed by reports about the status of the various experimental projects. We got the impression that this combination of theoretical and experimental talks gave a more complete picture of what is going on and contributed to the mutual understanding of theoreticians and experimentalists.

Finally, we would like to express our gratefulness to the *WE–Heraeus Foundation* for financing this meeting and the *Deutsche Forschungsgemeinschaft* (DFG) for giving support to many of our colleagues from Eastern Europe.

Konstanz and Düsseldorf, Stanford, Cologne
October 2000

Claus Lämmerzahl
Francis Everitt
Friedrich W. Hehl

Contents

List of Abbreviations

ALLEGRO	Gravitational bar detector, USA
ASTRE	Accelerometre Spatial TRiaxial Electrostatique
ASTROD	Astrodynamical Space Test of Relativity using Optical Devices
AURIGA	Gravitational bar detector, Italy
AXEL	AXial Experiment at Low temperature
CHAMP	CHAllenging Micro–satellite Payload for geophysical research and application
CODATA	Committee on Data for Science and Technology
CWDB	Close White Dwarf Binaries
EEP	Einstein Equivalence Principle
EP	Equivalence Principle
EXPLORER	A gravitational wave antenna located at CERN in Geneva
GAIA	Global Astrometric Interferometer for Astrophysics
GEO 600	German–British gravitational wave interferometer in Hannover
GOCE	Gravity and Ocean Circulation Explorer
GP-B	Gravity Probe B
GRACE	Gravity Recovery And Climate Experiment
HIPPARCOS	HIgh Precision PARallax COllecting Satellite
HYPER	HYPER sensitive cold atom interferometry in space
IPLR	InterPlanetary Laser Ranging
LAGEOS	LAser GEOdynamics Sattelite
LIGO	Laser Interferometer Gravitational wave Observatory
LISA	Laser Interferometer Space Antenna
LLI	Local Lorentz Invariance
LLR	Lunar Laser Ranging
MOT	Magneto–Optical Trap
MICROSCOPE	MICRO–Satellite à trainée Compensée pour l'Observation du Principle d'Equivalence
NAUTILUS	Gravitational bar detector at Laboratori Nazionali di Frascati, Italy
PPN	Parametrized Post–Newtonian approximation
SORT	Solar Orbit Relativity Test
SQUID	Super conducting QUantum Interference Device
STAR	Space Three–axis Accelerometer for Research
STEP	Satellite Test of the Equivalence Principle
TAMA300	Japanese gravitational wave interferometer
VIRGO	French–Italian gravitational wave interferometer in Pisa
VLBI	Very Long Baseline Interferometry
WEP	Weak Equivalence Principle
rf	radio frequency
rms	root mean square

List of Contributors

Luc Blanchet
Department of Relativistic Astrophysics
and Cosmology
Section de Meudon
Place Jules Janssen
92905 Meudon Cedex
France
blanchet@obspm.fr

Christian Bordé
Laboratoire de Physique des Lasers
Institut Galilée
UMR – CNRS 7538
Université Paris 13
Avenue Jean–Baptiste Clément
93430 Villetaneuse
France
chbo@ccr.jussieu.fr

Saps Buchman
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
sasha@relgyro.stanford.edu

Karsten Danzmann
Institut für Atom- und Molekülphysik
Universität Hannover
and
MPI für Quantenoptik
Außenstelle Hannover
Callinstraße 38
30167 Hannover
Germany
kvd@mpq.mpg.de

Daniel B. DeBra
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
ddebra@sun-valley.stanford.edu

Hansjörg Dittus
ZARM
Universität Bremen
Am Fallturm
28359 Bremen
Germany
dittus@zarm.uni-bremen.de

C.W. Francis Everitt
W.W. Hansen Experimental Physics
Laboratory
Stanford University
Stanford, CA 94305-4085
USA
francis@relgyro.stanford.edu

Ephraim Fischbach
Purdue University
Department of Physics
West Lafayette, IN 47907-1396
USA
ephraim@physics.purdue.edu

Leonid Grishchuk
Department of Physics and Astronomy
University of Wales, Cardiff
P.O. Box 913
Cardiff, Wales CF2 3YB
UK
Leonid.Grishchuk@astro.cf.ac.uk

Frank Gronwald
Universität Magdeburg
Institut für Allgemeine Elektrotechnik
und Leistungselektronik
Postfach 4120
39016 Magdeburg
Germany
frank.gronwald@e-technik.
uni-magdeburg.de

Mark Haugan
Purdue University
Department of Physics
1396 Physics Building
West Lafayette, Indiana 47907-1396
USA
mph@physics.purdue.edu

Friedrich W. Hehl
Inst. Theor. Physics
University of Cologne
50923 Köln
Germany
hehl@thp.uni-koeln.de

Eugen Holzschuh
Physik–Institut
Universität Zürich
Winterthurer Str. 190
8057 Zürich
Switzerland
holzschu@physik.unizh.ch

George M. Keiser
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
mac@relgyro.stanford.edu

Iosip Khriplovich
Butker Institute of Nuclear Physics
630090 Novosibirsk
Russia
I.B.Khriplovich@inp.nsk.su
kraused@wabash.edu

Helmar Koch
Institute of Solid State Physics
Universität Jena
Helmholtzweg 5
07743 Jena
Germany
Helmar.Koch@uni-jena.de

Dennis Krause
Department of Physics
Wabash College
Crawfordsville, IN 47933
USA
kraused@wabash.edu

Walter Kündig
Physik–Institut
Universität Zürich
Winterthurer Str. 190
8057 Zürich
Switzerland
kuendig@physik.unizh.ch

Claus Lämmerzahl
Faculty of Physics
University of Konstanz
Fach M 677
78457 Konstanz
Germany
claus.laemmerzahl@uni-konstanz.de
and
Institute for Experimental Physics
Heinrich–Heine–University Düsseldorf
40225 Düsseldorf
Germany
claus.laemmerzahl@uni-duesseldorf.de

Herbert I.M. Lichtenegger
Institut für Weltraumforschung
Österreichische Akademie der Wissenschaften
8010 Graz
Austria
him@iwf.tu-graz.ac.at

Stefan Lochmann
ZARM
Universität Bremen
Am Fallturm
28359 Bremen

Germany
lochmann@zarm.uni-bremen.de

Nick Lockerbie
University of Strathclyde
Department of Physics and Applied Physics
Glasgow G4 0NG
UK
nick@tesla.phys.strath.ac.uk

Jim M. Lockhart
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
jim@relgyro.stanford.edu

Dirk Lockowandt
ZARM
Universität Bremen
Am Fallturm
28359 Bremen
Germany

Lute Maleki
Jet Propulsion Laboratory (JPL)
4800 Oak Grove Drive M/S 298-100
Pasadena, CA 91109-8099
USA
Lute.Maleki@jpl.nasa.gov

Bahram Mashhoon
Department of Physics and Astronomy
University of Missouri – Columbia
Columbia, Missouri 65211
USA
physgrav@everest.cclabs.missouri.edu

Carsten Mehls
ZARM
Universität Bremen
Am Fallturm
28359 Bremen
Germany
mehls@zarm.uni-bremen.de

John Mester
W.W. Hansen Experim. Physics Labs
Satellite Test of the Equivalence Principle (STEP)
Stanford University
Stanford, CA 94305-4085
USA
mester@relgyro.stanford.edu

Barry F. Muhlfelder
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
barry@relgyro.stanford.edu

Ralf Neubert
Institute of Solid State Physics
Universität Jena
Helmholtzweg 5
07743 Jena
Germany
Ralf.Neubert@uni-jena.de

Gernot Neugebauer
Theoretisch–Physikalisches Institut
Universität Jena
Max–Wien–Platz 1
07743 Jena
Germany
G.Neugebauer@tpi.uni-jena.de

Wei–Tou Ni
National Tsing Hua University
Department of Physics No. 101, Sec. 2
Kung Fu Road
Hsinchu, Taiwan 300
Republic of China
wtni@phys.nthu.edu.tw

Sandor Nietzsche
Institute of Solid State Physics
Universität Jena
Helmholtzweg 5
07743 Jena
Germany
Sandor.Nietzsche@uni-jena.de

Frithjof Nolting
Lawrence Berkeley National Laboratory
1 Cyclotron Road, MS 7-222
Berkeley, CA. 94720
USA
FNolting@lbl.gov

Kenneth Nordtvedt
Northwest Analysis
118 Sourdough Ridge
Bozeman, MT 59715
USA
kennordtvedt@one800.net

Bradford W. Parkinson
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
brad@relgyro.stanford.edu

Achim Peters
Fakultät für Physik
Universität Konstanz
78457 Konstanz
Germany
achim.peters@uni-konstanz.de

Fritz Riehle
Physikalisch–Technische Bundesanstalt (PTB)
Section 4.31: "Unit of Length"
Bundesallee 100
38116 Braunschweig
Germany
fritz.riehle@ptb.de

Lewis Ryder
School of Physical Sciences
University of Kent
Canterbury CT2 7NR
UK
l.h.ryder@ukc.ac.uk

Gerhard Schäfer
Theoretisch–Physikalisches Institut
Universität Jena
Max–Wien–Platz 1
07743 Jena
Germany
Schaefer@tpi.uni-jena.de

Stephan Schlamminger
Physik–Institut
Universität Zürich
Winterthurer Str. 190
8057 Zürich
Switzerland
schlammi@physik.unizh.ch

Jürgen Schurr
Physikalisch Technische Bundesanstalt (PTB)
Bundesallee 100
38116 Braunschweig
Germany
Juergen.Schurr@ptb.de

Dietmar Theiss
Institute for Theoretical Physics
University of Cologne
50923 Köln
Germany
Dietmar.Theiss@mediaways.net

Rodney Torii
W.W. Hansen Experim. Physics Labs
Satellite Test of the Equivalence Principle (STEP)
Stanford University
Stanford, CA 94305-4085
USA
torii@step.Stanford.edu

Pierre Touboul
Physics, Instrumentation and Sensing Department
ONERA
29 Avenue de la division Leclerc
BP 72
92322 Châtillon Cedex
France
touboul@onera.fr

John P. Turneaure
W.W. Hansen Experim. Physics Labs
Stanford University
Stanford, CA 94305
USA
john@relgyro.stanford.edu

Stefano Vitale
Department of Physics
University of Trento
38050 Povo, Trento
Italy
vitale@alpha.science.unitn.it

Wolfgang Vodel
Institute of Solid State Physics
Universität Jena
Helmholtzweg 5
07743 Jena
Germany
Wolfgang.Vodel@uni-jena.de

Norbert Wex
Max–Planck–Institut für Radioastronomie
Auf dem Hügel 69
53121 Bonn
Germany
wex@MPIfR-Bonn.MPG.de

Paul W. Worden
W.W. Hansen Experim. Physics Labs
Satellite Test of the Equivalence Principle (STEP)
Stanford University
Stanford, CA 94305-4085
USA
worden@step.Stanford.EDU

J. von Zameck Glyscinski
Institute of Solid State Physics
Universität Jena
Helmholtzweg 5
07743 Jena
Germany
zameck@dpserv1.ifk.uni-jena.de

An Overview of Solar System Gravitational Physics: The Theory–Experiment Interface

Kenneth Nordtvedt

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

Abstract. If the gravitational metric tensor field of Einstein's General Relativity is supplemented by other long range, very weakly coupled interaction fields, then General Relativity's Equivalence Principle foundations are violated, or its post-Newtonian ($1/c^2$ order) structure is altered, or both. Space experiments test for and measure such possibilities; presently universality of free fall is confirmed to about a part in 10^{12}, and no deviations of post–Newtonian metric gravity from general relativity are seen down to the few parts in 10^4 level. Future experiments in space can significantly increase the precision to which fundamental physical law is probed. In particular, transponded interplanetary laser ranging can measure presence of metrically coupled scalar fields in gravity with two or three orders of magnitude higher precision than past experiments, and can begin to measure the second post–Newtonian ($1/c^4$) structure of gravity. A space–based experiment of the universality of free–fall (STEP) can detect additional interactions of a non–metric nature with five or six orders of magnitude higher precision than today's experiments.

1 Introduction

In General Relativity theory the gravitational interaction between matter is transmitted via the ten components of a symmetric tensor field $g_{\mu\nu}(\boldsymbol{r}, t)$. While the space and time gradients of these ten potentials determine the local gravitational equations of motion of matter, the comparative values of the potentials between different locations also establish global relationships between the rates of clocks and spans of rulers which are substantially separated from each other. This creates, not simply deforms, the metrical properties of space and time — the cosmic arena — which are not *a priori* or by default Euclidean, but rather are dynamical and dependent on the distribution of matter throughout the cosmos and the gravitational field for their existence [1]. This historical progression in our understanding of the the cosmic arena's nature — from Newton's separate and rigid *absolute time* and *absolute Euclidean space*, to Minkowski's united, but still globally rig id space–time arena, to the metric gravity field's creation of a dynamic arena shaped by its material content — is illustrated symbolically by

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 3–14, 2001.

the sequence of matrices

$$\mathbf{1} \quad + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \Longrightarrow \begin{bmatrix} g_{tt} & g_{tx} & g_{ty} & g_{tz} \\ g_{xt} & g_{xx} & g_{xy} & g_{xz} \\ g_{yt} & g_{yx} & g_{yy} & g_{yz} \\ g_{zt} & g_{zx} & g_{zy} & g_{zz} \end{bmatrix}$$

‘rigid’ time + ‘rigid’ space ‘rigid’ space–time space–time arena field (1)

Various experiments in space have contributed to the systematic mapping out of the long range gravitational interaction in the presence of source matter distributions of ever–increasing generality and complexity. The experimental goals for the future are not only to continue improving the precision of this mapping, but also to deliberately search for any additional long range force fields which may supplement Einstein’s tensor gravity by acting between bodies and thereby alter the details of the total in teraction.

The major achievements of physics in the 20th century depend not only on the two major revolutions of relativity and quantum physics and their applications, but also on the elimination of action–at–a–distance from physical law and its replacement with the field theory of interaction — that matter at one location acts as source of certain dynamical fields, which once produced then propagate in space and time, and then interact with other matter located elsewhere. Maxwell’s equations for the electromagnetic field in presence of sources, plus the force law of Lorentz which gives the response of charges (sources) to those fields, formed the breakthrough model for this paradigm over a century ago

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= 4\pi\, \rho(\boldsymbol{r}, t)\,, & \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0\,, \\ \boldsymbol{\nabla} \times \boldsymbol{B} &= \frac{4\pi}{c}\, \boldsymbol{j}(\boldsymbol{r}, t) + \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t}\,, & \boldsymbol{\nabla} \times \boldsymbol{E} &= -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t}\,, \\ \frac{d\boldsymbol{p}}{dt} &= e\left(\boldsymbol{E}(\boldsymbol{r}, t) + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{B}(\boldsymbol{r}, t)\right)\,, \end{aligned} \tag{2}$$

and includes all the basic ingredients for this new conceptual framework. A compact way to formulate the electromagnetic interaction is in terms of a *vector* field of potentials $A_\mu(\boldsymbol{r}, t),\ \mu = 0, x, y, z$

$$E_i = -\frac{\partial A_o}{\partial x^i} + \frac{\partial A_i}{c\partial t}, \qquad B_i = \sum_{j,k} e_{ijk}\, \frac{\partial A_k}{\partial x^j}\,, \tag{3}$$

with indices i, j, k ranging over the three spatial dimensions, and e_{ijk} being the anti–symmetric permutation matrix. Today’s so–called *Standard Model* for the electromagnetic, weak, and strong nuclear forces in matter is built upon the repeated use of vector fields of interaction, with there being three additional fields for the weak force, $W_\mu^{(3)}$, and eight ’gluon’ fields for the strong force, $G_\mu^{(8)}$. These vector fields have been successfully integra ted into quantum physics (involving

the careful handling of still not fully understood renormalizable infinities), but the metric gravitational field of General Relativity has not been so integrated, which is one hint that something is still missing in our theory of gravity. (It should be noted, however, that the post–Newtonian 'near fields' of gravity which accompany source bodies contribute to matter's quantum mechanical Hamiltonian; otherwise body center–of–mass wave packets would not follow the post –Newtonian classical general relativistic trajectories which we observe in various experiments.)

The simplest type of field, the scalar field, presently plays no role in modeling observed phenomena, yet theoretical speculations beyond the Standard Model suggest it should be a natural partner of the metric tensor gravitational field. A scalar field can, in fact, combine with the tensor field to jointly produce a hybrid metric field for the space–time arena. Because of an interesting equality between static–limit tensor and scalar source strengths for bodies of negligible external world stresses, gravi tational self–energy, and deviation from internal equilibrium,

$$\int T_0^0 \, d^3x \; = \; \int T \, d^3x \tag{4}$$

(T_ν^μ is matter's stress–energy tensor, T is its trace), such bodies fulfill composition–independence of free fall rates and clock rate shifts in presence of other bodies. Presence of the scalar field in this case only manifests itself in post–Newtonian phenomena. The search for a 'metric' scalar field in gravity has therefore been, and continues to be, one of the main themes of experimental gravity in space.

A scalar field can not only alter the phenomena of metric gravity if coupled to matter in the special manner indicated above, this field also has the possibility (indeed the probability, according to string theorists!) of coupling differently or selectively to the various (electromagnetic, weak, or nuclear) scalar attributes of matter, and consequently violating Einstein's Equivalence Principle by producing composition–dependence of both gravitational free fall rates and clock rate shifts in proximity to m atter (gravitational 'red–shift').

Other very weakly coupled, long range force fields may also be present, waiting for discovery in more precise space experiments. For instance, a vector field B_μ coupled to some combination of baryon and lepton number would also generate composition–dependent accelerations of bodies.

From what has been experimentally confirmed to date, it is not likely we'll find a breakdown (in long range phenomena) of Einstein's general relativistic equations for the dominant tensor field gravitational interaction

$$G_\nu^\mu \; = \; \kappa \, T_\nu^\mu \, , \tag{5}$$

but this still leaves open the question of whether there are any supplementary long range and very weak interactions between bodies?

2 Testable Models for Experiments

The ingredients for modeling the space experiments consist primarily of the equations of propagation of electromagnetic signals, the rate equations of clocks, and the equations of motion for bodies. The totality of all the long range interactions in physical law will contribute to one or more of these dynamical equations. Clocks and bodies exist with a wide variety of internal compositions and structure. This affects how they can couple to external force fields, and our models must take this into account. Some useful representations of these key dynamical equations are as follows:

$$c(\boldsymbol{r},t) = c_\infty \left(1 - (1+\gamma)\frac{U(\boldsymbol{r},t)}{c^2} + \dots\right), \tag{6}$$

$$d\tau_c = \left(1 - \frac{1}{2}\frac{v^2}{c^2} - \alpha_c \frac{U(\boldsymbol{r},t)}{c^2} + \dots\right) dt\,, \tag{7}$$

in which $U(\boldsymbol{r},t)$ is the Newtonian potential function of proximate matter

$$U(\boldsymbol{r},t) = G\sum_i \frac{M_i}{|\boldsymbol{r}-\boldsymbol{r}_i|}\,. \tag{8}$$

The gravitational clock rate shift parameter α_c has the universal, composition–independent value of one in metric theories, but more generally it may depend on the nature of the clock if other interaction fields supplement the tensor metric field of general relativity. The *Eddington* parameter γ which appears in the light propagation equation equals one in general relativity but has smaller values in scalar–tensor metric gravity, for example. The light equat ion leads to two measurable phenomena: the deflection of light's direction of travel through angle

$$\delta\theta \;=\; 2(1+\gamma)\,\frac{GM}{c^2 D} \tag{9}$$

when making a complete passage by a central body M with distance of closest approach D, and a delay in integrated time of propagation

$$\delta t \;\cong\; (1+\gamma)\,\frac{GM}{c^3}\ln\left(\frac{4R_1R_2}{D^2}\right), \tag{10}$$

when a signal travels between two sites at distances R_1 and R_2 from the central body and passes at distance of closest approach D from that body. In the last year or so, analysis of very long baseline interferometry (VLBI) observations of deflections of light from very many astronomical sources has produced the most precise measurement of the light propagation equation,

$$|\gamma - 1|_{\rm VLBI} \;\le\; 6\times 10^{-4} \qquad (2\,\sigma)\,, \tag{11}$$

whereas the time delay effect had for the previous couple decades supplied the premier experimental data for this measurement.

Accepting that local gravitational physics is Lorentz–invariant and Lagrangeian–based with energy, momentum, and angular momentum conservation laws, foundations empirically established by the very space experiments under discussion in this report, the gravitational equations of motion for the general N–body solar system, valid not only in Einstein's pure tensor gravity, but also generalized to represent plausible alternative metric theory possibilities by introduction of the two *Eddington* para meters γ and β, is

$$
\begin{aligned}
A \qquad \boldsymbol{a}_i = & \left(1 + \frac{\dot{G}}{G}(t - t_0)\right)\left(\frac{M(G)}{M(I)}\right)_i \sum_j \boldsymbol{g}_{ij} \\
B \qquad & - (2\beta - 1)\sum_{j,k}\left(\frac{\mu_k}{r_{ik}} + \frac{\mu_k}{r_{jk}}\right)\boldsymbol{g}_{ij} \\
C \qquad & + \sum_j \left((2\gamma + 2)\ \boldsymbol{v}_i \times (\boldsymbol{g}_{ij} \times \boldsymbol{v}_j) + (2\gamma + 1)\ \boldsymbol{g}_{ij} \cdot \boldsymbol{v}_j\ \boldsymbol{v}_i\right) \\
D \qquad & + \sum_j \Big(\frac{1}{2}\left((2\gamma + 1)v_i^2 + (2\gamma + 2)v_j^2 - 3(\boldsymbol{v}_j \cdot \hat{\boldsymbol{r}}_{ij})^2\right)\boldsymbol{g}_{ij} \\
& \qquad\qquad - (2\gamma + 1)\left((\boldsymbol{g}_{ij} \cdot \boldsymbol{v}_j)\boldsymbol{v}_j + (\boldsymbol{g}_{ij} \cdot \boldsymbol{v}_i)\boldsymbol{v}_i\right)\Big) \\
E \qquad & + \frac{1}{2}\sum_j \frac{\mu_j}{r_{ij}}\left((4\gamma + 3)\boldsymbol{a}_j + (\boldsymbol{a}_j \cdot \hat{\boldsymbol{r}}_{ij})\hat{\boldsymbol{r}}_{ij}\right) \\
F \qquad & - \frac{1}{2}v_i^2\ \boldsymbol{a}_i - (\boldsymbol{a}_i \cdot \boldsymbol{v}_i)\boldsymbol{v}_i - (2\gamma + 1)\sum_j \frac{\mu_j}{r_{ij}}\boldsymbol{a}_i\,, \qquad (12)
\end{aligned}
$$

with $i, j, k = 1, \ldots, N$, $\boldsymbol{v}_i = d\boldsymbol{r}_i/dt$, $\boldsymbol{a}_i = d\boldsymbol{v}_i/dt$, and $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$. A factor $1/c^2$ multiplies all the post–Newtonian lines B through F, but has been set equal to 1 to simplify presentation. This total equation is very close to Eq. (1) of reference [4]; I have simply organized the terms of the equation of motion by type on separate named lines $A, \ldots, F$ for purposes of discussion (see page 321). And on line F, for purposes of stressing the inertial nature of these forces, I have retained the proportionality of these forces to the body acceleration $\boldsymbol{a}_i$ which in reference [4] are set equal to their Newtonian values. $\mu_j = GM_j$ are the body gravitational mass parameters, and the Newtonian acceleration functions are indicated

$$\boldsymbol{g}_{ij} = \frac{\mu_j}{r_{ij}^3}\,\boldsymbol{r}_{ji}\,. \qquad (13)$$

Setting $\gamma = \beta = 1$, all body factors $M(G)/M(I) = 1$, and G equal to a constant as in general relativity, this becomes the equation of motion which JPL's computer ephemeris uses to integrate and produce the default orbits of the planets and satellites in the solar system. Deviations from this nominal general relativistic case then produce experimental signals proportional to scientific parameters $\gamma - 1$, $\beta - 1$, $M(G)/M(I) - 1$, dG/Gdt, etc., which can then be measured in various experim ents by least squares fit type data analysis techniques.

Two necessary modifications of the Newtonian acceleration appear in line A of Eq.(12). In alternative theories of gravity the local Newtonian coupling strength G will generally be changing in some proportion to the universe's Hubble expansion rate H [2]

$$G(t) \cong G_0 + \dot{G}\,(t - t_0) + \ldots\,, \qquad \text{with} \qquad \dot{G}/G \;\sim\; (4\beta - 3 - \gamma)\,H\,. \tag{14}$$

In such theories the necessary and consistent application of the total equation of motion to the internal gravitational structures of the bodies and their interactions with the gravity of the external world also alters the gravitational to inertial mass ratio of each celestial body in proportion to its internal gravitational binding energy [3]:

$$\left(\frac{M(G)}{M(I)}\right)_i = 1 - (4\beta - 3 - \gamma)\,\frac{1}{2M_i c^2}\int \frac{G\rho(\boldsymbol{x}_i)\rho(\boldsymbol{y}_i)}{|\boldsymbol{x}_i - \boldsymbol{y}_i|} d^3x_i\, d^3y_i\,. \tag{15}$$

Celestial bodies, in this general situation, will therefore not fall in gravitational fields at test body rates. Measuring this ratio for the Earth by fitting the lunar laser ranging (LLR) data has been one of the key methods for measuring the metric theory, *Eddington* coupling parameters γ and (especially) β which both take values different from their general relativity values of one in scalar–tensor theories [4,5].

Different terms of the general equation of motion given by Eq.(12), either individually or collectively, play roles in determining the different so–called 'effects' of post–Newtonian gravity — perihelion precession, *deSitter* precession, *Lense–Thirring* precession, etc. But it must be stressed that all effects must be calculable from a common underlying equation of motion valid for the general case. The equation of motion can not be altered to 'fit' one effect without considering how that affects the 'fit' to other observed effects.

It is worthwhile to exhibit the static limit force produced by an additional long range, very weakly coupled force

$$\boldsymbol{F}_i \;=\; K_i\,\boldsymbol{\nabla}_i \sum_j \frac{K_j}{R_{ij}}\,e^{-\mu R_{ij}}\,. \tag{16}$$

The coupling strengths of this interaction, K_i, generally will not be proportional to the mass–energy contents of the interacting bodies, causing violations of key predictions of the Equivalence Principle. For example, two bodies A, B will now accelerate at different rates toward another source body S located at distance R away

$$\frac{|\boldsymbol{a}_A - \boldsymbol{a}_B|}{|\boldsymbol{g}_S|} = \frac{K_S}{GM_S}\left(\frac{K_A}{M_A} - \frac{K_B}{M_B}\right)\,(1 + \mu R)\,e^{-\mu R}\,. \tag{17}$$

And atomic clocks whose frequencies are determined by the energy differences between two quantum states m, n,

$$f_{nm} = \frac{E_n - E_m}{2\pi\hbar} \tag{18}$$

will now generally acquire anomalous, clock–dependent frequency shifts when in proximity to a source body, thereby modifying the universal rate shifts predicted by the Equivalence Principle and metric theories of gravity

$$\frac{\delta f_{nm}}{f_{nm}} \;=\; -\,\frac{GM_S}{c^2R}\,\left(1-\frac{K_S}{GM_S}\left(\frac{K_n-K_m}{M_n-M_m}\right)\,e^{-\mu R}\right)\,. \tag{19}$$

These expressions for the two types of Equivalence Principle violations allow comparisons to be made between experiments involving clocks and those measuring gravitational free–fall rate differences among assorted materials, with the purpose of noting their relative abilities to detect any new interaction with coupling strengths K_i. These expressions can also guide choice of experimental materials and the design of special clocks in order to maximize experimental sensitivities to specific models for ne w interactions.

Even if the two quantum levels of an atomic clock have common coupling strengths $K_n = K_m$, *absolute* clock rate shift experiments could measure the presence of such an additional interaction by comparing the strength of the clock shifts now proportional to only GM_S to the effective Newtonian coupling strength of the same central body with some planet P, which has now become

$$\Gamma_S = GM_S \;+\; K_S\,\frac{K_P}{M_P}\,(1+\mu R)e^{-\mu R}\,. \tag{20}$$

If the additional interaction given by Eq. (16) has finite range, then its contribution to the periastron precession rate of orbits of size comparable to the interaction range $1/\mu$ provides a very sensitive means for its detection. For orbits of small eccentricity, the extra acceleration term from Eq. (16) generates an orbital precession rate of

$$\frac{\delta\omega(\mathrm{per})}{\omega(\mathrm{orb})} = \frac{1}{2}\,\frac{K_iK_j}{GM_iM_j}(\mu R)^2\,e^{-\mu R}\,. \tag{21}$$

The orbits of satellites and celestial bodies of size from about an Earth radius and up to the orbit of Mars have been measured with high precision using laser or radar ranging techniques. The absences of any anomalies in the precession rates of the various orbits fit to the data give very strong constraints on the presence of finite range interactions. These constraints are shown in Fig. 1 by the four labeled curves, and they apply for the supplementary interaction being either metric or non–metric. The s cientific significance of orbital precession rate measurements has not been fully appreciated in the past, and their future improvement over the total range of available orbits can be an effective way to enhance the search for new interactions.

The full content of the equation of motion given by Eq. (12) includes the grouped contributions to the equation given in lines marked A through F. A rich variety of *post–Newtonian* phenomena results, including

- rescaling of the inertia of a body or its mass elements in proportion to the strength of the gravitational potential produced by the other bodies and mass elements of the system (line F),

- accelerations of a body due to motion of either the acted–upon body — v_i^2 — or the acting body — v_j^2 (line D),
- non–linear accelerations proportional to the second power of source body masses or to the product of source (acting) bodies and acted–upon body masses (line B),
- *gravito–magnetic* accelerations proportional to the velocities of the acted–upon as well as the acting bodies — $\boldsymbol{v}_i\boldsymbol{v}_j$ (line C),
- inductive accelerations of a body i in proportion to the acceleration $\boldsymbol{a}_j$ of other nearby bodies j (line E).

From the entire collection of experimental observations, these different parts of the total equation of motion are all presently confirmed and measured to better than a part in 10^3 precision, and as well the *local Lorentz invariance* and *conservation laws* foundations are empirically established to even higher precision. The LLR experiment has proven to be a particularly comprehensive probe of the equation of motion; there are measured lunar orbit perturbations affected by every part of the general equation $(A, \ldots, F)$. By contrast, the initial historic success of general relativity to explain the anomalous precession of perihelion for Mercury's orbit results from the special case of the general Eq. (12) in which a single test body (Mercury) moves in the static, spherically symmetric gravity of the Sun

$$\left(1 + \frac{1}{2}v^2 + (2\gamma+1)\frac{GM}{r}\right)\boldsymbol{a} + \boldsymbol{v}\,(\boldsymbol{v}\cdot\boldsymbol{a}) \tag{22}$$
$$= \left(1 - (2\beta-1)\frac{GM}{r} + (2\gamma+1)v^2\right)\boldsymbol{g} - (2\gamma+1)\,\boldsymbol{v}\,(\boldsymbol{v}\cdot\boldsymbol{g})$$

(here again, $1/c^2$ has been set equal to one in the post–Newtonian terms), and

$$\boldsymbol{g} = -\frac{GM}{r^3}\,\boldsymbol{r}\,. \tag{23}$$

3 Lunar Laser Ranging

Modeling of the LLR experiment, which includes a moving and accelerating Moon orbitting a moving and accelerating Earth, both bodies in the combined gravitational field of the Sun, of each other, and of themselves, results in LLR being a comprehensive laboratory for testing the total structure of the gravitational equation of motion for bodies (LLR also tests the more general parameterized metric field equation of motion developed by Will and Nordtvedt [8,9]). The measured Earth–Moon range is sensitive to literally every one of the phenomena in the body equation of motion described above. From the fit of the range data, several types of terms in the general relativistic equation of motion are today confirmed with precisions of at least a few parts in 10^4. But there are three LLR measurements in gravitational physics which are of particular scientific interest.

1. Earth and Moon are found to fall toward the Sun at equal rates to three parts in 10^{13} precision. This confirms both the non–linear self–coupling of gravity to itself,

$$|\beta - 1|_{\rm LLR} \ \leq \ 2 \times 10^{-4} \tag{24}$$

   and the composition–independence of gravity's coupling to matter (Earth and Moon have somewhat different average chemical compositions).
2. Inertial space in the locality of the Earth–Moon system is found to rotate relative to distant inertial space at the rate predicted in General Relativity — *deSitter's geodetic precession* — to precision of three or four parts in a thousand.
3. Newton's G shows no time variation at the few parts in 10^{12} *per year* level.

For more detail on the LLR experiment and theory, see the other contribution to these proceedings.

4 Geodetic Precession and NASA's Gravity Probe–B

When a system moves non–radially at velocity $\boldsymbol{V}$ through the gravitational field of an external body M, the accelerations $\boldsymbol{a}_i$ of the system's individual bodies moving in that system at the additional velocities $\boldsymbol{v}_i$ experience Coriolis–like acceleration terms equal to

$$\delta \boldsymbol{a}_i = 2 \left(\gamma + \frac{1}{2} \right) \frac{GM}{c^2 R^3} \ (\boldsymbol{R} \times \boldsymbol{V}) \times \boldsymbol{v}_i \, . \tag{25}$$

This indicates that the local inertial frame moving with that system rotates with respect to the distant inertial frame at the *deSitter geodetic precession* rate (about 19.2 mas/year)

$$\boldsymbol{\Omega}_{\rm dS} = \left(\gamma + \frac{1}{2} \right) \frac{GM}{c^2 R^3} \ \boldsymbol{R} \times \boldsymbol{V} \, . \tag{26}$$

LLR presently measures this rotation of local inertial space to good precision through its .07 mas/year fit of the perigee precession of the lunar orbit. NASA's GP–B mission of Earth–orbitting gyroscopes, scheduled to be launched early 2002, is designed to measure this precession due to motion through Earth's gravity with two orders of magnitude more precision; if successful this will yield a very significant improvement in our knowledge of the fundamental metric field *Eddington* p arameter γ which enters Eq.(26).

5 Interplanetary Laser Ranging

Two decades ago NASA's *Viking* mission to Mars included transponded radar ranging to both a spacecraft orbitting that planet and a lander. From the individual range measurements of about 7 meters precision, and relying especially on

those observations made when the ranging signals passed close to the Sun, the propagation time delays given by Eq. 8 were fitted, and the *Eddington* γ parameter was measured to precision of 2×10^{-3}. Transponded laser ranging between Earth and Mars co uld be performed in a future mission with precision comparable to that of LLR — of order centimeters. By fitting the laser pulse propagation times, and also taking into account the perturbations of the Mars and Earth orbits which emerge from the relativistic equation of motion, both parameters γ and β can be measured to a part in 10^6. For this purpose, a key perturbation results from the 4–body problem involving Sun, Jupiter, Earth and Mars. If the Sun accelerates anomalously toward Jup iter because its gravitational to inertial mass ratio differs from one at either the first or second post–Newtonian order (χ is a collection of second post–Newtonian order metric field expansion coefficients which sum to zero in general relativity),

$$\begin{aligned}\frac{M(G)}{M(I)} &= 1 - (4\beta - 3 - \gamma)\frac{G}{2Mc^2}\int \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}d^3x\, d^3x' \\ &\quad + \chi\frac{G^2}{Mc^4}\int \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{x}')\rho(\boldsymbol{x}'')}{|\boldsymbol{x}-\boldsymbol{x}'||\boldsymbol{x}-\boldsymbol{x}''|}d^3x\, d^3x'\, d^3x'' + ... \\ &\cong 1 - 4\times 10^{-6}\,(4\beta - 3 - \gamma) + 7\times 10^{-11}\,\chi\,, \end{aligned} \tag{27}$$

then the orbit of an inner planet is polarized toward Jupiter by amount

$$\delta r_p \;\cong\; \left(\frac{M(G)}{M(I)} - 1\right)_S \frac{M_J}{M_S}\,\frac{3\Omega_P - \Omega_J}{2\Omega_P - \Omega_J}\,\frac{\Omega_J}{\Omega_P - \Omega_J}\,R_J\,, \tag{28}$$

in which Ω_P and Ω_S refer to the orbital frequencies of the planet and Jupiter, respectively, and R_J is Jupiter's orbital radius. The Earth and Mars polarizations then produce a unique time–dependent Earth–Mars range signal. As seen in Eq.(28), replacing Mars by a more distant asteroid as the ranging target would enhance the effect and should be considered as a mission option. Laser ranging to a space–based gravity wave observatory like *LISA* will also show this orbit al perturbation. Although this may be the only relativistic effect in the solar system involving second post–Newtonian order ($1/c^4$) features of the gravitational field which can be measured in the near future, theoretical work suggests that the precision measurement of the first post–Newtonian parameters γ and β which will result is of more importance than the novel but low precision measurement of the second post–Newtonian order parameters. An experiment at this level could be a decisive type of solar system experiment for probing metric gravity. It has been shown that scalar–tensor metric gravity, which during the early universe may have contained comparable coupling strengths of these two interactions with matter, would within some general scenarios find the cosmic background scalar field dynamically evolving as the universe expands toward a field value where the scalar field's coupling to matter measured by $1-\gamma$ and $1-\beta$ has naturally become very small, but not exactly zero, in our era [6]. (A similar scenario has been studied in non–metrically coupled scalar–tensor theories, giving a motivation for testing, for example, the composition–independence of gravitational free fall beyond the present–day part in 10^{12} confirmation [7].)

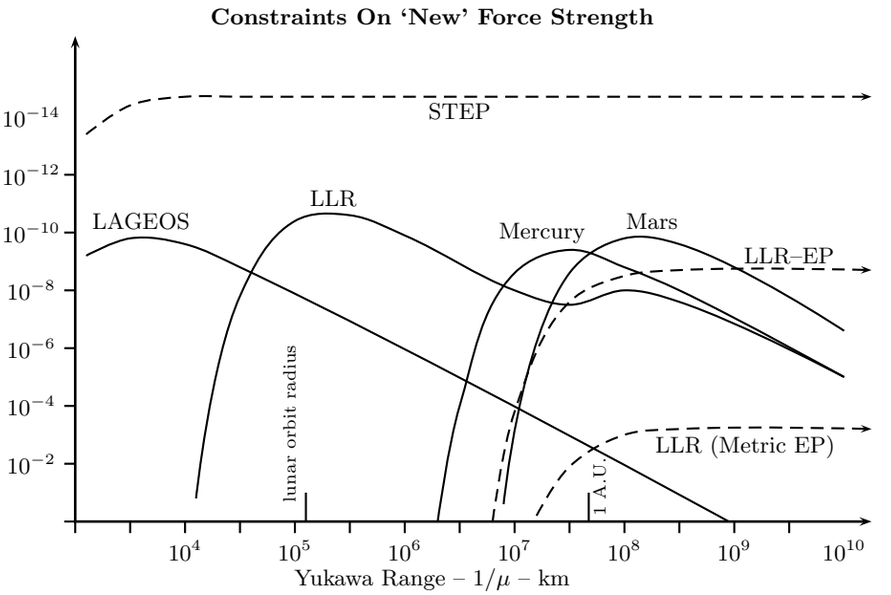


Fig. 1. The solid curves marked "LAGEOS", "LLR", "Mercury", and "Mars" show constraints on the strength of any metric or non–metric force which may supplement general relativistic gravity, plotted as function of Yukawa range. These constraints result from measurements of the periastron precession rates of the corresponding orbits. The LLR curve is double humped, reflecting the roles of both Earth and Sun in determining the lunar orbit. The dashed curves marked "LLR–EP" and "L LR (Metric EP)" are constraints from the measurement of the equality of the Earth's and Moon's acceleration rates toward the Sun. This result has separate interpretations for the cases of the supplementary force being metric or non–metric, and these constraints extend to 'infinite' range, i.e. inverse–square. The dashed curve marked "STEP" shows the level of very weak non–metric force which can be detected by that future space experiment.

6 A STEP Experiment

From laboratory experiments and lunar laser ranging, we presently know that differently composed bodies accelerate in a gravitational field at rates identical to about a part in 10^{12}. From Eq.(12), and noting that various attributes per unit mass of atoms K_i/M_i fractionally differ by about a part in 10^3 as one selects from different regions of the periodic table of elements (baryon number, nuclear electrostatic energy, etc.), it can be concluded that most any supplementary long range interaction must be at least a factor of about 10^9 weaker than the strength of gravity (for certain types of coupling of the supplementary field, the constraint would be stronger)

$$\left| \frac{K_i K_j}{G M_i M_j} \right| \; \leq \; 10^{-9} \, . \tag{29}$$

This experimental constraint is shown by one of the curves in Fig, 1. Exploiting several favorable features of the experimental environment in an orbitting spacecraft, a STEP (Satellite Test of the Equivalence Principle) mission is being designed to measure differences in the gravitational free fall rates of a variety of elements toward Earth with precision of a part in 10^{18}. This will extend the search for new long range interactions by five or six orders of magnitude higher precision for interaction ranges from an earth radius up to infinity (inverse square). This STEP goal is shown by another curve in Fig.1. Present–day investigations of unified theories beyond the Standard Model, such as *string theories*, suggest that such additional interactions may exist, especially a scalar field coupled selectively to the different scalar attributes of matter. A STEP–like experiment is one of the very few ways we may be able to search for and possibly find consequences of these under lying theories in our 'low energy' realms of experience. Even a discovery by STEP of no composition–dependent accelerations at these incredible levels of precision would place strong constraints on and guidance for the theoretical investigations working toward unification of physical law.

Acknowledgement

This work supported by NASA contracts NASW-97008 and NASW-98006.

References

1. That the metric gravitational field *creates, not simply deforms,* space–time geometry means that there is no default Euclidean geometry for the cosmos which is then altered by gravity. That the space–time arena locally possesses the relatively simple structure of Minkowski follows from a mathematical theorem that local coordinates can always be found which convert the ten–component metric field of gravity into the Minkowski matrix plus forty vanishing first derivatives with respect to the space–time coordinates. That the cosmos–at–large seems to be quasi–Euclidean or quasi–Minkowskian (after taking the Hubble expansion into account) follows from the approximate large scale uniformity of the cosmic matter density which taken collectively dominates determination of the metric field throughout most of space–time.
2. C. Brans and R.H. Dicke: *Phys. Rev.* **124**, 925 (1961).
3. K. Nordtvedt: *Phys. Rev.* **169**, 1017 (1968).
4. J.G. Williams, X.X. Newhall, and J.O. Dickey: *Phys. Rev.* **D 53**, 6730 (1996).
5. J. Müller and K. Nordtvedt: *Phys. Rev.* **D 58**, 2001 (1998).
6. T. Damour and K. Nordtvedt: *Phys. Rev.* **D 48**, 3436 (1993).
7. T. Damour and A.M. Polyakov: *Nucl. Phys.* **B 423**, 229 (1995).
8. C. Will and K. Nordtvedt: *Astrophys. J.* **177**, 757 (1972).
9. K. Nordtvedt: *Phys. Rev.* **D 7**, 2347 (1973).

Determination of the Gravitational Constant

Stephan Schlamminger[1], Eugen Holzschuh[1], Walter Kündig[1],
Frithjof Nolting[2], and Jürgen Schurr[3]

[1] Physik–Institut, Universität Zürich, Winterthurer Str. 190, 8057 Zürich, Switzerland
[2] present adress: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] present adress: Physikalisch Technische Bundesanstalt, 38023 Braunschweig, Germany

Abstract. The Newtonian gravitational constant G was the first known fundamental constant of physics. Nevertheless, the measurement of its value still seem to be in a rather sad shape. Recently, the CODATA Task Group on Fundamental Constants recommended a preliminary value of G with a relative uncertainty of 0.15 %. This is more than ten times larger as the previous recommendation!

In the first part of this lecture, a brief summary is given of recent experimental efforts to determine G. The second part is a description of our experiment at the University of Zürich.

We use a beam balance to measure the gravitational forces of large field masses (13.5×10^3 kg mercury) on 1 kg test masses. A first result with an uncertainty of 220×10^{-6} has been published recently. Presently we are working to improve the experiment.

1 Introduction

Henry Cavendish published in 1798 – more than 200 years ago – his famous paper titled "Experiments to determine the density of the earth" [1]. This experiment is the first laboratory experiment to determine the Newtonian gravitational constant G. Cavendish employed a torsion balance, a device which, up to now, is widely used to measure G. Although the measurement techniques improved considerably during the last 200 years, CODATA[1] in 1998 recommended an uncertainty of 1500 ppm and a value of $G = 6.673(10) \times 10^{-11}\,\mathrm{m^3 kg^{-1} s^{-2}}$ [2,3].

Compared to the previous CODATA recommendation of 1986 [4], the uncertainty was increased by more than a factor of ten. The reason for this surprising step can be seen in Fig.1, showing the results of recent measurements. There are large and apparently not well understood discrepancies. The CODATA task group on Fundamental Constants decided to recommend the old 1996 value of G, but to drastically enlarge the recommended uncertainty.

On the positive side, these discrepancies provided motivation for many new experiments. Several of these new results were presented at a conference held in London 1998 to celebrate the 200th anniversary of the Cavendish experiment.

[1] The Committee on Data for Science and Technology of the International Council of Scientific Unions.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 15–28, 2001.

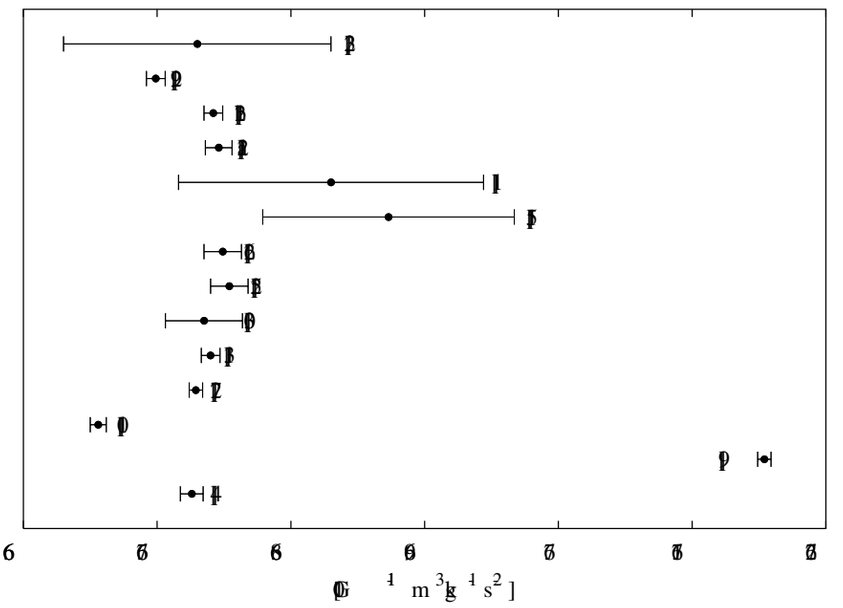


Fig. 1. Latest results of published G experiments and the recommended CODATA values of 1986 [4] and 1998 [2,3].

2 Currently Running Experiments

We will restrict our discussion to earth bound experiments. There are also intentions to perform measurements of G in space [5]. However, these efforts seem to be in an early stage and are not yet fully funded.

The present experiments are conveniently divided into two groups, those using a torsion balance and other experiments using a variety of techniques. In all experiments there are movable field masses producing an attractive force on relatively small test masses. Many techniques have been developed to sense the effect of these small forces. We will first discuss techniques employing a torsion balance.

Cavendish operated his torsion balance in the static deflection mode. A dumbbell was suspended by a torsion wire and the angle of deflection was measured, when two field masses were brought nearby. This method is not used anymore. Instead three other techniques are employed today.

First, there is the time of swing method. The angular frequency of a torsion pendulum is measured with and without field masses. The torque caused by the gravitational attraction of the field masses increases the angular frequency when compared with the free pendulum. The difference of the squared angular frequencies is proportional to the second derivative of the gravitational potential

Table 1. Summary of the presently running experiments employing a torsion balance to determine the Newtonian gravitational constant.

group	method, features	status	value for G [10^{-11} m^3 kg^{-1}s^{-2}]	rel. uncertainty [ppm]
M. P. Fitzgerald, T. R. Armstrong [12], New Zealand	force compensation, electrometer compensated	working on improvements	6.6746 [12]a 6.6742 [12]b	150 105
J. H. Gundlach[13], Washington	acceleration, optimized mass arrangement	under construction		
O. V. Karagioz, V. P. Izmailov [27], Moscow	time of swing	published results	6.6729	75
J. Luo et. al [29] People's Rep. of China	time of swing	published results	6.6699	105
G. G. Luther, C. H. Bagley Los Alamos	time of swing	old published [31] new under constr. [32]	6.6740	105
R. D. Newman, M. K. Bantel [8], Irvine CA	time of swing, cryogenic, optimized mass arrangement	under construction		
S. J. Richman et al. [11], BIPM[a]	time of swing and force compensation in one apparatus	first results	6.683	1700

[a] Bureau International des Poids et Mesures

U of the field masses [6]:

$$G = \frac{(\omega_{f+g}^2 - \omega_f^2)I}{k_g}\,, \qquad \text{where} \qquad k_g = \frac{1}{G}\,\frac{\mathrm{d}^2U}{\mathrm{d}\theta^2}\,.$$

The angle θ is formed by the field masses and the axis of the pendulum and I denotes the moment of inertia of the dumbbell. This equation is valid under the assumption that the elastic coefficients of the torsion wire is independent of the angular frequency. However, in 1995 Kuroda [7] pointed out that inelasticity in the fiber could produce a systematic shift of the frequency. Newman and Bantel showed [8], assuming a certain model for the inelasticity, that an upper limit for the induced relative error of G is given by $\frac{1}{2}Q^{-1} + O(Q^{-2})$, where Q is the quality factor of the torsion pendulum. It is thus advantageous to have a Q–value as large as possible. As Q can be increased by lowering the temperature, Newmann and Bantel intend to operate their pendulum at the temperature of

liquid helium. As shown in Table 1 there are five groups employing the time of swing technique.

Table 2. Summary of the present running experiments to determine G without using a torsion balance.

group	method, features	status	value $[10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^2]$	rel. uncertainty [ppm]
U. Kleinevoß *et al.* [30] Wuppertal	Fabry-Pérot resonator (microwaves)	working on improvements	6.6735	432
A. De Marchi [17] *et al.* Torino	dynamic free pendulum	first results		
Wei-Tou Ni *et al.* [33] Taiwan	Fabry–Pérot resonator (optical)	under construction		
F. Nolting *et al.* [26] Zürich	beam balance	working on improvements	6.6749	217
H. Parks *et al.* Colorado	Fabry–Pérot resonator (optical)	under construction		
J. P. Schwarz *et al.* [14,15] Colorado	free fall method	published results	6.6873	1400

The second technique is called force compensation method. A electromagnetic force is applied to the torsion pendulum in order to prevent the torsion fiber from twisting. The electromagnetic quantity, for example the applied voltage on an electrometer, is measured. Thus a calibration has to be done, i. e. the measured electromagnetic quantity must be converted into a mechanical force or torque. On the other hand, these experiments are static in principle and therefore should not be affected by inelasticity effects. Two groups have exclusively used this methods and both results showed large discrepancies [9,10] when compared to the CODATA value. Although unlikely, one may suspect that this could be an indication for some fundamental problem. For this reason the group at the Bureau International des Poids et Mesures (BIPM) [11] built a torsion balance where both the time of swing and the compensation method can be used. At present only preliminary measurements with the first method have been performed. The group is working on improvements and it is hoped that a accurate comparison of the two methods should soon be possible. The group in New Zealand has discovered a fault in the analysis of their 1995 measurement [10]. In 1999 they reported a corrected value [12] and a new and more precise result. Both are in agreement with most other results. Unfortunately the experiment of Michaelis *et al.* [9] has been discontinued some time ago.

The third important technique is the so–called acceleration method where the torsion balance is mounted on a turntable. This turntable is accelerated and controlled by a feedback loop in such a way that the torsion fiber does not twist. The angular acceleration of the turntable can be measured directly and the torque produced by the field masses be calculated. Furthermore, the measurement is static again and the experiments should not be affected by inelasticity. Obviously, this technique requires high precision turntables and angular encoders. One such experiment is presently set up by Gundlach [13] at Washington.

Altogether there are at present seven torsion balance experiments, which are listed in Table 1. Five groups have published a result.

There are presently six experiments measuring G without a torsion balance. In 1998, J.P. Schwarz published a interesting free–fall experiment [14,15]. The test mass was a corner cube reflector falling vertically through the center of a doughnut shaped field mass. The acceleration of the test mass was determined by tracking its position using an optical interferometer. The force on the test mass was modulated by periodically moving the field mass in vertical direction. The field mass was made by tungsten because of its high density. The published result is given in Table 2.

Three experiments use a Fabry–Pérot resonator. The resonator is formed by two pendulums which act as test masses. The field masses are moved horizontally. Their gravitational force changes the relative distance of the two pendulums, proportional to the horizontal force gradient. The distance is measured by the eigenfrequency of the resonator. This technique was pioneered by the Wuppertal group [16]. They are using microwaves with typically 20 GHz, i.e. a rather large wavelength compared to the distance change of the two pendulums, which is about 10 nm. The measurements reported are very precise and the latest published result (see Table 2) is dominated by systematic uncertainties. There are two other experiments using this Fabry–Pérot technique. They intend to use visible light, that is a much shorter wavelength. Both experiments are presently under construction.

Another interesting experiment is performed by De Marchi *et al.* in Italy [17]. They exploit the fact that frequencies can easily be measured with very high precision. A simple pendulum, mounted in vacuum, is used and the frequency change is measured when a field mass is brought nearby.

The experiment at the University of Zürich uses a beam balance and will be described in more detail below.

All present experiments not using a torsion balance are listed in Table 2. A detailed review of the experiments to measure the Newtonian gravitational constant can be found in [18].

3 The Principle of Our Experiment

The first experiment to measure the gravitational constant with the help of a beam balance was published in 1879 by J.H. Poynting [19]. Subsequently several similar measurements were performed but soon this technique was given up in

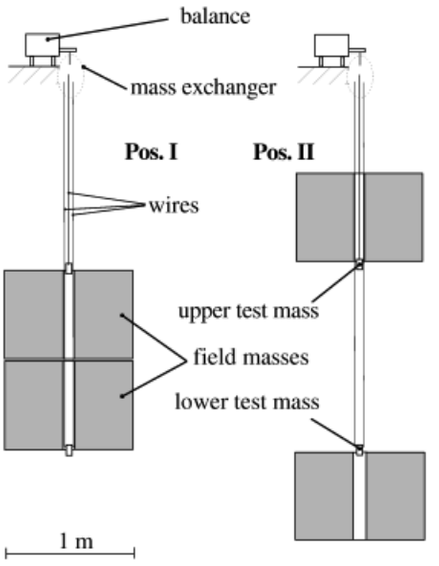


Fig. 2. The principle of the experiment. The two field masses are shown in the two positions as used for measurements.

favor of the torsion balance. In 1983 the first modern beam balance experiment was done by C. Speake et al. [20] to search for the "fifth force". The University of Zürich group used a beam balance in a experiment performed at the Gigerwald storage lake [21,22]. There, the gravitational constant was measured at an effective distance of approximately 100 m. Since 1997 we are using a beam balance to measure G in a laboratory experiment.

The principle of our experiment is shown in Fig.2. There are two cylindrical field masses and two test masses. The field masses can be moved vertically and have a central hole such that the test masses can pass through. In position one the field masses are close together, while in position two they are apart. The two test masses remain at the same position but can be connected alternately to the balance.

First the weight difference between the test masses in position one is measured. Obviously, the weight of the upper test mass is increased by the gravitational force of both field masses, while the weight of the lower test mass is decreased. The weight difference, defined as the difference in the readings of the balance, weighing first the upper test mass and then the lower test mass, is increased, when compared with weights of the test masses in absence of the field masses.

In position two the situation is reversed. The weight of the upper test mass is decreased, while the weight of the lower test mass is increased.

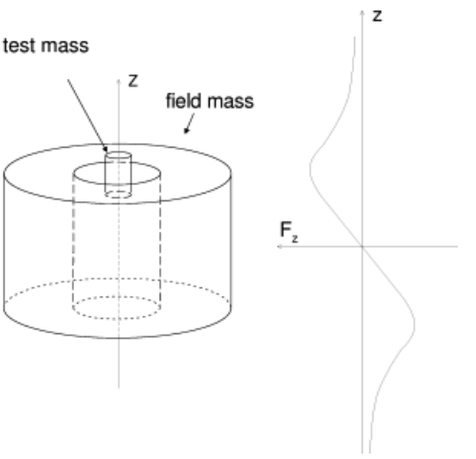


Fig. 3. Schematic illustration of the force acting on a test mass on the axis of a field mass with the form of a hollow cylinder.

The second difference, that is the difference of the weight differences for field mass positions I and II, is independent of small but unavoidable mass differences of the two test masses and directly proportional to G. All effects which are the same for the two test masses, such as tidal forces and zero–point drifts of the balance, cancel already in the first difference.

The force field of a cylinder with a central hole has an advantageous feature which is exploited in this experiment. The z–component of the force F_z, produced by one cylinder and acting on a small test mass on the symmetry axis, is shown in Fig.3. There are two local extrema near both ends of the cylinder, which in fact are saddle points. This is shown in more detail in Fig.4, where F_z is plotted as a function of the radial distance and the z–coordinate. The plot was generated assuming the dimensions of one field mass as used in the experiment. Clearly the accurate position of a test mass is quite uncritical if placed near the saddle point. For an assumed position error of 1 mm, the force error would be of order 10 ppm.

4 The Experimental Setup

The experiment is located in a 4.5 m deep pit inside an experimental building at the Paul–Scherrer–Institut (PSI). The necessary thermal and mechanical stability is provided by the thick concrete walls of the pit. The experiment is divided vertically in two parts separated by a working platform. As shown in Fig.5, the field and test masses are located in the lower part. The upper part of the experiment houses the balance, the electronics and the vacuum pumps.

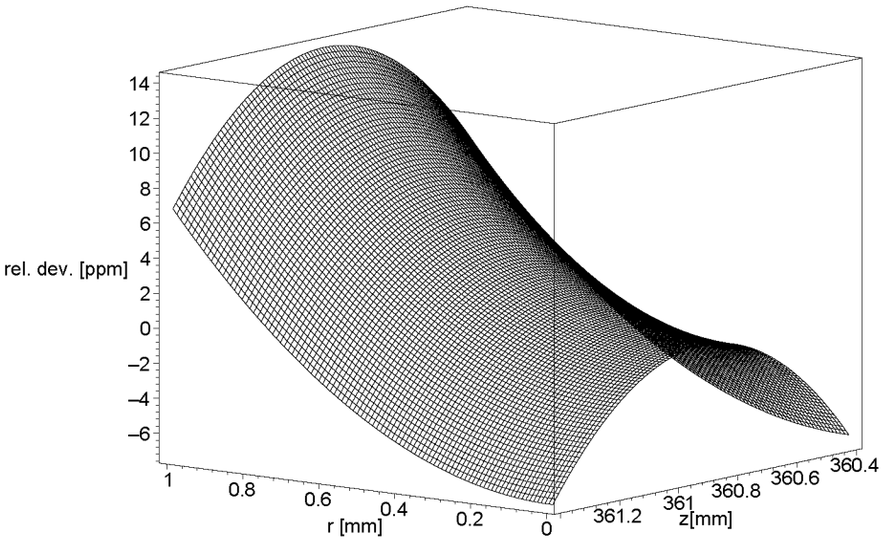


Fig. 4. The relative deviation of the z–component of the force with respect to its value at the saddle point. The force is generated by one homogeneous cylindrical field mass (see text).

Both parts are mechanically and thermally insulated from each other and each room has its own temperature stabilization system.

The balance is placed on a massive granite plate, supported by two steel girders. They are attached to the wall of the pit and have no direct mechanical contact to the working platform. The balance and the two test masses are inside a vacuum system consisting of a chamber and a long tube passing through the central hole of the field masses. At present the pressure in the system is typically 10^{-4} Pa. In that way problems due to convection and buoyancy are avoided.

The most important part of the experiment is the balance. It is a modified Mettler–Toledo AT 1006 comparator balance. The commercial version of this balance was designed for a high precision comparison of 1 kg standard masses. A schematic view is shown in Fig.6. It is a single–pan, flexure–strip balance with a servo–controlled beam. A weight change on the cantilever is compensated by a magnetic system, consisting of a permanent magnet and a coil mounted on the beam. The current through the coil is measured and converted into a weight. A more detailed description may be found in [23]. The original resolution of the balance was 1 µg and has been changed to a present value of 12.5 ng. Currently we achieve a reproducibility of 200 ng which we define to be the standard deviation of 10 weight comparisons. The balance is calibrated by putting small and precisely known standard masses (presently two 0.5 g) on the cantilever. Finally the balance reading is converted into a force using the measured value of the local gravity, $g = 9.8072335(6)\,\mathrm{ms}^{-2}$.

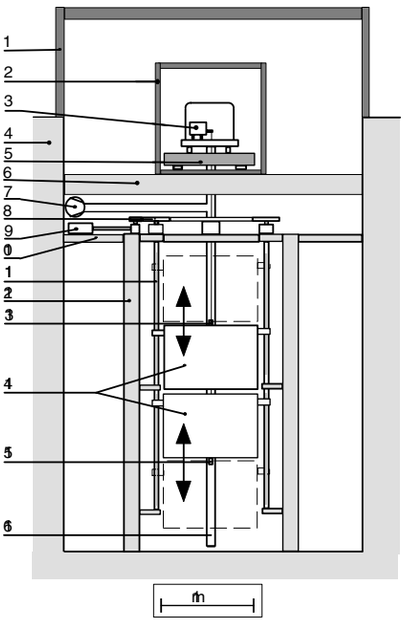


Fig. 5. The experimental setup. Legend: 1=chamber, 2=thermally insulated chamber, 3=the balance inside of the vacuum chamber, 4=concrete walls of the pit, 5=granite plate, 6=steel girders which support the balance, 7=vacuum pumps, 8=gear assembly, 9=motor, 10=working platform, 11=spindle, 12=steel girder of the main support, 13=upper test mass, 14=field masses, 15=lower test mass, 16=vacuum tube.

The field masses are vessels made of stainless steel and can be filled with a liquid. A liquid is used, because of its high homogeneity. Each vessel has a volume of 500 l and a mass of 800 kg when unfilled. The outer diameter of the field mass is 1046 mm, the inner diameter 100 mm and the height 700 mm. Measurements with empty and with water filled vessels have been performed and first results were reported in [24,25].

At present the vessels are filled with mercury. The total filled in mass (13.5×10^3 kg) was determined with a relative uncertainty of 2 ppm. A detailed description of the filling procedure may be found in [23].

The 1 kg test masses were made from copper and are gold plated. Small samples were taken from the same piece of material as the test masses. The magnetic susceptibility of the samples was measured and found to be sufficiently small such that magnetic forces should be negligible. Also the density homogeneity was checked and at the level of 2×10^{-4} no differences were found.

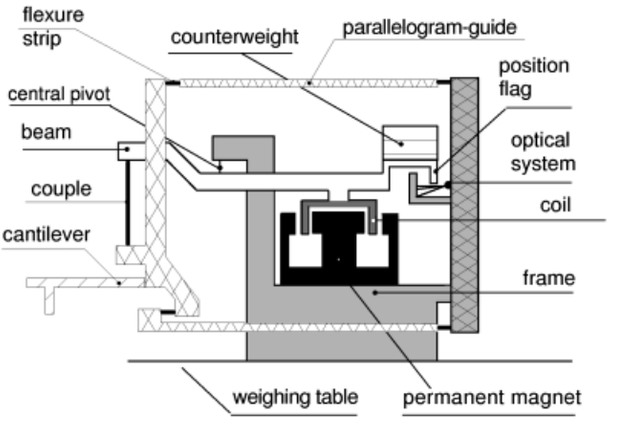


Fig. 6. Schematic view of the balance.

5 The Measuring Procedure and First Results

The whole measuring procedure is fully automated and computer controlled. Initially the field masses are in position one, that is together. One test mass is suspended on the balance and weighed. This lasts typically 2 minutes. Then the test mass is lifted from the balance and at the same time the other test mass is connected to the balance. This is done, using a special exchange device, in such a way that the total load on the balance does not change by more than typically a few grams. The second test mass is also weighed and then the sequence above is repeated typically 20 times. Next a standard mass is put on the balance for calibration. Then the vessels are moved in position two, that is apart and the complete weighing procedure above is done again. All taken together this forms one cycle of the measurement (see Fig.7), which is then repeated indefinitely. Typical data are shown in Fig.8. Plotted is the first difference as a function of time. The amplitude of the rectangular signal is the second difference and has a approximate value of 0.8 mg for mercury filled vessels. The inset shows the data of the last half–cycle with the vertical scale expanded by a factor 500.

Measurements with empty and water filled vessels were carried out during 1997 and the results were published in [24,25]. From these data the gravitational constant was found to be

$$G = (6.6754 \pm 0.0005 \pm 0.0013) \times 10^{-11}\,\mathrm{m^3 kg^{-1} s^{-2}} \qquad \text{(water)}.$$

The statistical uncertainty is 75 ppm and the systematic uncertainty is 209 ppm. The detailed error budget is reproduced in Table 3.

Since the beginning of 1998 measurements with mercury filled vessels are in progress. A preliminary result of these measurements was published in [26]. During the mercury runs a systematic variation of the measured signal as a

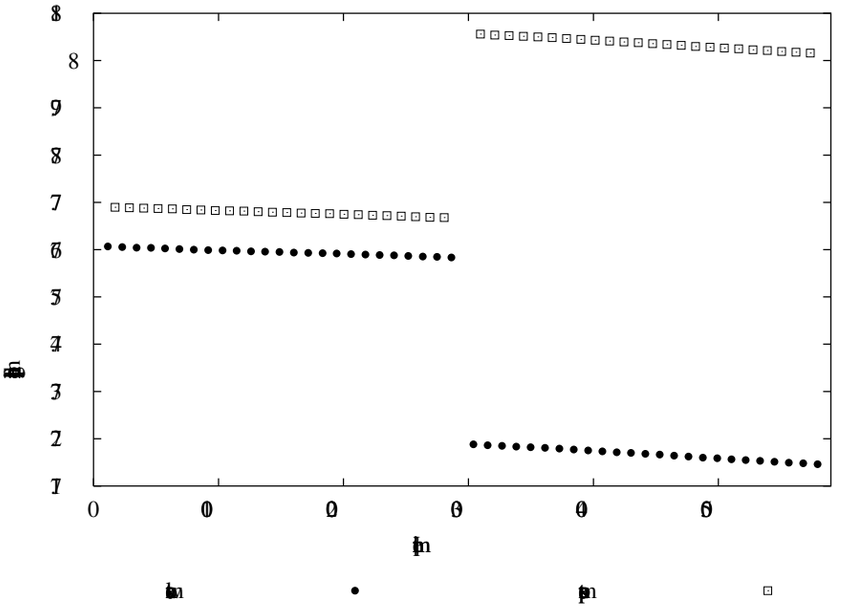


Fig. 7. A typical example of a measurement cycle. Plotted is the reading of the balance for the two field mass positions. Initially – in the left part of the graph – the vessels are apart and the weight of the lower test mass is increased, the weight of the upper test mass is decreased. At the right, corresponding to the first field mass position, the situation is reversed. The smooth variation with time is caused by the drift of the balance.

function of time was observed. In February of 1998 the results were slightly larger than in later months until November. Surprisingly, a similar effect was also observed in 1999. For this year we have data until May. Up to now, no fully satisfactory explanation has been found. To take this into account, an extra uncertainty of 80 ppm, which is one half of the observed variation, was included into the error budget. As a consequence, the result from the mercury runs is not more accurate than the water results, despite the larger signal. The published value is

$$G = (6.6749 \pm 0.0014) \times 10^{-11}\,\mathrm{m^3 kg^{-1} s^{-2}} \qquad \text{(mercury)}.$$

Here only the systematic uncertainty is given and the statistical uncertainty is comparatively negligible.

6 Outlook

The results of the measurements with the water filled vessels and the mercury filled vessels agree within their errors (see Fig.1). Our results are 420 ppm or ap-

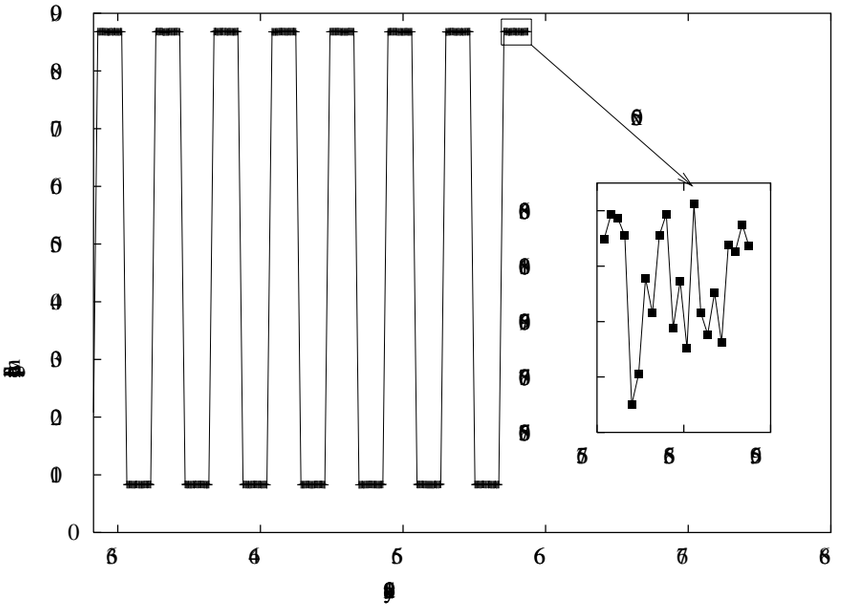


Fig. 8. Measured weight difference from three days of data during 1999. The inset shows the difference of the last half–cylce expanded vertically by a factor 500.

proximately two of our standard deviations larger than the CODATA value from 1986. The error budget is dominated by a possible, but not yet fully investigated, nonlinearity of the balance. This large uncertainty arose, because the balance was calibrated with a 1 g standard weight, whereas the measured signal was only 0.8 mg (mercury). We hope that with future work a significant improvement is possible. The next largest error was assigned because of the observed systematic variation mentioned above. This effect is difficult to understand, because most conceivable causes should precisely cancel in the result. In fact, we have not yet been able to identify a unique cause but believe that it is related to some instabilities in the experiment. At present we are working on a greatly improved temperature stabilization. Also sensitive tilt meters will be installed to monitor possible tilt changes of the balance. We hope to reach the design accuracy of the experiment of 10×10^{-6} in the near future.

Acknowledgment

We thank the Paul–Scherrer–Institut for the helpful support. We wish to thank the companies Mettler–Toledo and metrotec for providing and modifying the balance. We gratefully acknowledge the work of the machine shop of our institute. Furthermore, we want to thank all the metrological institutes that have helped

Table 3. Systematic uncertainties of the gravitational constant, measured with water–filled and with mercury–filled vessels as published in [26]. The uncertainties are grouped according to the contributions of the test masses, the vessels, the liquid and further effects. They are estimated at a confidence level of 68.3% and the total uncertainty is the root–mean–square value

	ΔG/G[10^-6]	
Source of uncertainty	Water	Mercury
test mass position	11	10
test mass dimension	2.7	2.0
density inhomogeneity	≤ 1	≤ 2.2
masses of test masses	0.27	0.27
joints and spindle drive	16	2.2
shape and volume	14	2
dimensions of the vessels	11	1.6
density inhomogeneity	≤ 5	≤ 0.06
masses of the vessels	4	0.5
air displaced by the vessels	7	2
density of the liquid	15	18
mass of the liquid	8	0.7
nonlinearity of the balance	≤ 200	≤ 200
systematic variation	-	80
sorption effect	45	6.4
integration	≤ 13	≤ 5
tilt effect	≤ 18	≤ 4
calibration	5	8
local gravity	0.06	0.06
total	209	217

us. This experiment is supported by the Swiss National Science Foundation and the Dr. Tomalla Foundation.

References

1. H. Cavendish: *Phil. Trans. R. Soc.* (London) **88**, 469 (1798).
2. P.J. Mohr and B.N. Taylor: *Rev. Mod. Phys.* **72**, 351 (2000).
3. ibid, see also http://physics.nist.gov/cuu/Constants/index.html
4. E.R. Cohen and B.N. Taylor: Rev. Mod. Phys. **59**, 1121 (1987).
5. A.J. Sanders and G.T. Gillies: *Riv. Nuovo Cim.* **19**, 1 (1996).
6. G. G. Luther and W. R. Towler: *Phys. Rev. Lett.* **48**, 121 (1982).
7. K. Kuroda: *Phys. Rev. Lett.* **75**, 2796 (1995).
8. R.D. Newman and M.K. Bantel: *Meas. Sci. Technol.* **10**, 445 (1999).
9. W. Michaelis *et al.*: *Metrologia* **32**, 267 (1996).
10. M.P. Fitzgerald and T.R. Armstrong: *IEEE Trans. Instrumen. Meas.* **44**, 494 (1995).
11. S.J. Richman *et al.*: *Meas. Sci. Technol.* **10**, 460 (1999).
12. M.P. Fitzgerald and T.R. Armstrong: *Meas. Sci. Technol.* **10**, 439 (1999).

13. J.H. Gundlach: *Meas. Sci. Technol.* **10**, 454 (1999).
14. J.P. Schwarz *et al.*: *Science* **282**, 2230 (1998).
15. J.P. Schwarz *et al*: *Meas. Sci. Technol.* **10**, 478 (1999).
16. J. Schurr *et al.*: *Metrologia* **28**, 397 (1991).
17. A. De Marchi *et al.*: Conference of Precision Electromagnetic Measurements Conferrence Digest 1998, p. 147.
18. G.T. Gillies: *Rep. Prog. Phys.* **60**, 151 (1997).
19. J.H. Poynting: *Proc. R. Soc.* (London) **28**, 2 (1879).
20. C.C. Speake and G.T. Gillies: *Proc. R. Soc. London* **A 414**, 315 (1983).
21. A. Cornaz *et al.*: *Phys. Rev. Lett.* **72**, 1152 (1994).
22. B. Hubler *et al.*: *Phys. Rev.* **D 51**, 4005 (1995).
23. F. Nolting: *Determination of the Gravitational Constant by Means of a Beam Balance*, PhD Thesis (Universität Zürich, Zürich 1998).
24. J. Schurr *et al.*: *Phys. Rev. Lett.* **80**, 1142 (1998).
25. J. Schurr *et al.*: *Phys. Lett.* **A 248**, 295 (1998).
26. F. Nolting *et al.*: *Meas. Sci. Technol.* **10**, 487 (1999).
27. O.V. Karagioz and V.P. Izmailov: *Meas. Tech.* **39**, 979 (1996).
28. J. Luo *et al.*: *Science in China Ser.* **A 41** 1289 (1998).
29. J. Luo *et al.*: *Phys. Rev.* **D 59**, 042001 (1999).
30. U. Kleinevoß *et al.*: *Meas. Sci. Technol.* **10**, 492 (1999).
31. C.H. Bagley and G. G. Luther: *Phy. Rev. Lett.* **78**, 3047 (1997).
32. G. Luther: *Meas. Sci. Technol.* **10**, 426 (1999).
33. W.-T. Ni *et al.*: *Meas. Sci. Technol.* **10**, 495 (1999).

The Lense–Thirring Effect: From the Basic Notions to the Observed Effects

Claus Lämmerzahl[1] and Gernot Neugebauer[2]

[1] Fakultät für Physik, Universität Konstanz, 78457 Konstanz, Germany
[2] Istitut für Theoretische Physik, Friedrich–Schiller–Universität Jena, Max–Wien–Platz 1, 07743 Jena, Germany

Abstract. A pedagogical derivation is given of the Lense–Thirring effect using basic notions from the motion of point particles and light rays. First, the notion of rotation is introduced using the properties of light rays only. Second, two realizations for a non–rotating propagation of space–like directions are presented: the gyroscope and the spin of elementary particles. Then the gravitational field around a rotating body is specified which is taken for determining the various effects connected with a point particle or a gyroscope: the deSitter precession (geodesic precession) and the Lense–Thirring effect ('frame dragging'). The results are applied to the precession of gyroscopes and to the motion of satellites around the earth.

1 Introduction

In the Lense–Thirring effect three rotations are involved: the rotation of the gravitating body, the rotation of the test body around its own axis, and the rotation of the axis of rotation of the test body. The properties of the gravitating body are prescribed, the rotating test body can be shown to move approximately on a geodesics around the gravitating body, and its axis of rotation can be shown to be Fermi propagated along the path of the test body resulting in a precession of the axis of rotation with respect to distant stars. In this note we want to derive all these notions and equations of motion from scratch in order to indicate clearly that everything is provided by General Relativity: Everything follows from the geodesic equation for point particles and the validity of Einstein's equations; we don't have to use additional assumptions. At the end we will discuss several experimental approaches to test the various effects related to rotating bodies.

In the following we (i) introduce the notion of rotation, (ii) derive the equation of motion of the spinning axis of a rotating test body, (iii) derive the gravitational field of a rotating gravitating body, and (iv) use these results for discussing and analyzing the equation of motion of the test body and of its spinning axis.

All of what we assume is that gravity is described by means of a Riemannian geometry endowed with a space–time metric g and that light rays and freely falling point particles move along geodesics of that metric,

$$D_v v = \alpha v \,, \tag{1}$$

where D is the unique metric compatible, $Dg = 0$, torsion–free, $D_u v - D_v u - [u, v] = 0$, covariant derivative. In components, $(D_w v)^\nu = w^\mu(\partial_\mu v^\nu + \{\,{}^{\nu}_{\mu\sigma}\,\}v^\sigma)$,

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 31–51, 2001.

with the Christoffel symbol $\{ {\nu \atop \mu\sigma} \} := \frac{1}{2} g^{\nu\rho} \left(\partial_\mu g_{\rho\sigma} + \partial_\sigma g_{\rho\mu} - \partial_\rho g_{\mu\sigma} \right)$. We do not assume any normalization condition of the 4–velocity. Light rays with tangents l obey the same equation of motion (1), but with the additional condition $g(l, l) = 0$. – In addition, we will assume Einstein's field equations.

2 Rotation

In order to define the notion of rotation, we use the so–called *zig–zag construction*, or the *bouncing photon*, as introduced by Pirani [1]. This construction uses a central point particle which moves along an arbitrary path, see Fig.1. In the neighborhood of that central particle there are two other point particles equipped with a mirror. At first, the central point particle emits a flash of light which hits the auxilary point particles. These auxilary point particles reflect this flash of light in such a way that it again meets the central point particle and, in addition, the other auxilary particle positioned appropriately behind the central particle. Then the auxilarty particles again reflect the flash of light so that it again meets the central particle and the other auxilary particle, and so on, see Fig. 1. For this construction we assume that the two satellites are "near" to the central particle which means that no curvature effects should be involved. This is a condition which can always be fulfilled.

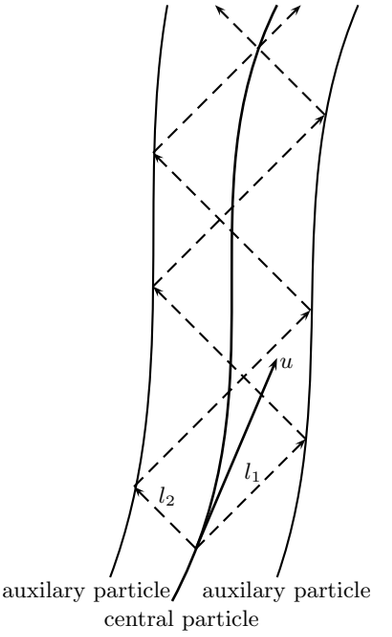


Fig. 1. Geometry of the *bouncing photon*. Two auxilary particles communicate via light rays in such a way that all light rays (dashed lines) have to meet the central particle. No particle is assumed to be in geodesic motion.

It is not assumed that the central particle moves along a geodesic. And even if the path of this particle is geodesic, then the two auxilary particles in general are not geodesic because they always have to be re–positioned in order to meet the above construction.

In this way the particles define a time–like 2–surface, or, after projection into the rest space of the central particle, a direction in the rest space propagating along the path of the central particle. This construction defines the propagation of the direction of the light rays in the rest space of the central particle. It turns out that this propagation can be used to define the notion of a "non–rotating" propagation of a vector. This notion coincides with the notion

of Fermi–displacement which usually is used for the description of non–rotating propagation. However, in our case we get this notion from an *operational* procedure.

We now turn to the mathematical description of this procedure. We make use of the equation of motion (1). The condition that the light rays with tangent l_1 and l_2 lie in the same plane with the central particle u is

$$u = \sigma_1 l_1 + \sigma_2 l_2 \,, \qquad \text{for some } \sigma_1, \sigma_2 \in \mathbb{R} \,. \tag{2}$$

The condition that, after reflection, the light rays will cross the central worldline again, is secured by

$$\mathcal{L}_{l_1} l_2 = \epsilon_1 l_2 + \epsilon_2 l_1 \,, \qquad \text{for some } \epsilon_1, \epsilon_2 \in \mathbb{R} \,. \tag{3}$$

Now we derive the equation which governs the transport of the directions

$$V_1 := P_u l_1 \quad \text{or} \quad V_2 := P_u l_2 \,, \tag{4}$$

along u, where

$$P_u A := A - \frac{g(A,u)}{g(u,u)} u \tag{5}$$

is the projection operator onto the rest space of the world line of the central particle.

Using $\sigma_1 V_1 = -\sigma_2 V_2$ we get

$$P_{V_1} P_u D_u V_1 = \frac{1}{\sigma_1} P_{V_1} P_u D_u (\sigma_1 V_1) = \frac{1}{2\sigma_1} P_{V_1} P_u D_u (\sigma_1 V_1 - \sigma_2 V_2) \,. \tag{6}$$

Inserting (2, 3, 4) and the equation of motion (1) for the light rays l_1 and l_2, we finally get

$$P_{V_1} P_u D_u V_1 = 0 \,. \tag{7}$$

The expression $F_u V := P_V P_u D_u V$ is the so–called *Fermi–derivative* of the vector V along u. Eq.(7) is invariant against reparametrization of the paths, so that it indeed describes the propagation of the direction V. This is precisely the characterization of the bouncing photon which we take as definition for a non–rotating propagation of a vector along a given path [1]. If for another vector W defined along P the above expression does not vanish, then the operator Ω, which is a $\binom{1}{1}$-tensor defined by $P_W P_u D_u W = \Omega(P_u W)$, is called the *rotation* of W. This is the characterization of the notion "rotation" we announced.

Now we turn to the question whether, beside this bouncing photon, there are other physical realizations of a non–rotating propagation.

3 Equation of Motion for Angular Momentum

There are indeed two further realizations for a non–rotating propagation which are experimentally easier to handle with than the bouncing photon: One realization is given by a rotating gyroscope possessing orbital angular momentum, the other is the elementary particle with spin.

3.1 The Motion of Gyroscopes

The metrical energy–momentum tensor $T^{\mu\nu}$ which, within Einstein's theory, is the source of the gravitational field, is symmetric and divergence–free

$$D_\mu T^{\mu\nu} = 0\,. \tag{8}$$

This is the equation of motion for matter in General Relativity.

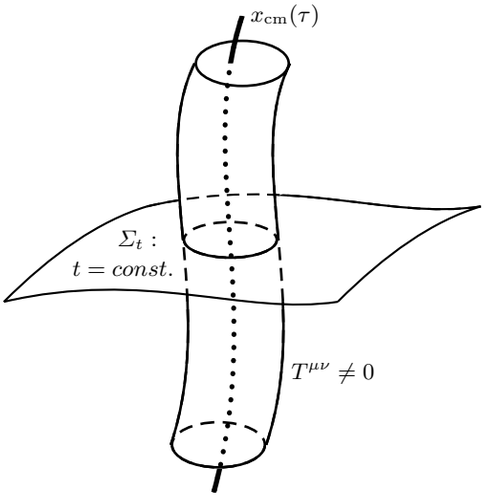


Fig. 2. The world–tube of space–time points for which the energy–momentum tensor in non–vanishing. The center–of–mass wordline is dotted.

In the case that matter is fairly localized (that is, $T^{\mu\nu} \neq 0$ for a compact space–like region only, within which the gravitational field in terms of the curvature does not vary too much), it is possible to extract from this tensor a center–of–mass and correspondingly the equation of motion of this center–of–mass. We assume that this center–of–mass $x_{\rm cm}(\tau)$, where τ is the proper time, lies within the body, that is, $\{x_{\rm cm}(\tau) | \tau \in \mathbb{R}\} \subset \operatorname{supp} T^{\mu\nu}$, see Fig.2. With this center–of–mass worldline there is connected the 4–velocity $v = \frac{d}{d\tau} x_{\rm cm}$.

In addition, one can define an angular momentum with respect to this distinguished point is space–time and derive its temporal evolution. In doing so we follow the procedure originated by Papapetrou [2,3] and developed further by Dixon and [4] and Ehlers and Rudolph [5].

We first introduce a 3+1–slicing of the space–time by introducing hypersurfaces Σ_t with normal n_μ, see Fig. 2. From the energy momentum tensor we can define various moments ($\mathfrak{T}^{\mu\nu} = \sqrt{-g}\, T^{\mu\nu}$)

$$P^{\mu_1\mu_2\dots\mu_n\nu} := \int_{\Sigma_t} \delta x^{\mu_1} \delta x^{\mu_2} \cdots \delta x^{\mu_n} \mathfrak{T}^{\nu 0} d^3x\,, \tag{9}$$

$$M^{\rho_1\rho_2\dots\rho_n\mu\nu} := \int_{\Sigma_t} \delta x^{\rho_1} \delta x^{\rho_2} \cdots \delta x^{\rho_n} \mathfrak{T}^{\mu\nu} d^3x\,, \tag{10}$$

where $\delta x^\mu = x^\mu - x^\mu_{\rm cm}$ with $x, x_{\rm cm} \in \Sigma_t$. We use coordinates so that $t = const.$ over Σ_t. Consequently, $\delta x^0 = 0$ in (9) and $n_\mu = \delta^0_\mu$.

We can distinguish between various types of particles: If all momenta (9) but the symmetrical

$$M^{\mu\nu} = \int_{\Sigma_t} \mathfrak{T}^{\mu\nu} d^3x \tag{11}$$

vanish, then we have a particle which possesses no internal structure and thus is called a *point particle*. It will become clear in the following that in this case

$$P^\mu = \int_{\Sigma_t} \mathfrak{T}^{\mu 0} d^3x \tag{12}$$

can be identified with the momentum of that point particle. — If all momenta but $M^{\mu\nu}$ and

$$M^{\mu\nu\rho} = \int_{\Sigma_t} \delta x^\mu \mathfrak{T}^{\nu\rho} d^3x \tag{13}$$

vanish, then we have a particle with mass and, in addition, an orbital angular momentum which in this case is given by

$$L^{\mu\nu} := 2P^{[\mu\nu]} = 2\int_{\Sigma_t} \delta x^{[\mu} \mathfrak{T}^{\nu]0} d^3x\,. \tag{14}$$

That this quantity indeed describes the angular momentum will become clear later, too. This quantity characterizes a special type of internal motion, so that we call this type of matter a *spinning particle* (please note, that this notion "spin" does not mean the elementary particle spin), or spinning top. It can be shown that the order of the highest moment is an invariant [2]. That means that a spinning particle cannot become a point particle by a coordinate transformation.

We assume in the following that all other moments but (11) and (13) vanish. From these definitions and the basic equation of motion (8) we first derive the equations of motion for a point particle, and second for a spinning particle.

The point particle. From (8) we have $0 = \partial_0 \mathfrak{T}^{\mu 0} + \partial_i \mathfrak{T}^{\mu i} + \{\,{}^{\mu}_{\nu\sigma}\,\}\mathfrak{T}^{\nu\sigma}$ from which we get by integration

$$\frac{d}{dt} P^\mu = \int_{\Sigma_t} \partial_0 \mathfrak{T}^{\mu 0} d^3x = -\{\,{}^{\mu}_{\nu\sigma}\,\} M^{\nu\sigma}\,. \tag{15}$$

In an analogous way we analyze the quantity $\int x^\rho \mathfrak{T}^{\mu\nu} d^3x$. With (8) we get

$$\partial_0(x^\rho \mathfrak{T}^{\mu 0}) + \partial_i(x^\rho \mathfrak{T}^{\mu i}) = \mathfrak{T}^{\mu\rho} - x^\rho \{\,{}^{\mu}_{\nu\sigma}\,\}\mathfrak{T}^{\nu\sigma} \tag{16}$$

and thus, by integration,

$$M^{\mu\rho} = \frac{d}{dt}\int_{\Sigma_t} x^\rho \mathfrak{T}^{\mu 0} d^3x + \int_{\Sigma_t} x^\rho \{\,{}^{\mu}_{\nu\sigma}\,\}\mathfrak{T}^{\nu\sigma} d^3x\,. \tag{17}$$

We expand x and $\{\,{}^{\mu}_{\nu\sigma}\,\}$ around the coordinate of the worldline x_{cm}

$$x = x_{\mathrm{cm}} + \delta x\,, \qquad \{\,{}^{\mu}_{\nu\sigma}\,\}(x) = \{\,{}^{\mu}_{\nu\sigma}\,\}(x_{\mathrm{cm}}) + \delta x^\kappa \partial_\kappa \{\,{}^{\mu}_{\nu\sigma}\,\}(x_{\mathrm{cm}}) \tag{18}$$

and get, taking into account the condition for a point particle, $\int \delta x \mathfrak{T} d^3x = 0$,

$$M^{\mu\rho} = v^\rho P^\mu\,. \tag{19}$$

Since $M^{\mu\nu}$ is symmetric, we must have $P^\mu = mv^\mu$ with $m := P^0/v^0$. Therefore $M^{\mu\nu} = m\ v^\mu v^\nu$, and, consequently, we get from (15) $D_v(mv) = 0$. From this and the fact that v is a normalized 4–velocity, $g(v,v) = -1$, we get $D_v m = 0$ which means that m is the mass of the particle which is constant. Then we also have the geodesic equation for the center–of–mass trajectory:

$$D_v v = 0\,. \tag{20}$$

The spinning particle. This kind of particle is defined by $\int \mathfrak{T}^{\mu\rho} d^3x \neq 0$ and $\int \delta x^\mu \mathfrak{T}^{\nu\rho} d^3x \neq 0$; all other moments vanish. We consider the divergences $\partial_\rho \mathfrak{T}^{\mu\rho}$, $\partial_\rho(x^\mu \mathfrak{T}^{\nu\rho})$, and $\partial_\rho(x^\mu x^\nu \mathfrak{T}^{\sigma\rho})$ and get in a way analogous to above

$$\frac{d}{dt} M^{\mu 0} = -\{\,{}^{\mu}_{\rho\sigma}\,\} M^{\rho\sigma} - \partial_\kappa \{\,{}^{\mu}_{\rho\sigma}\,\} M^{\kappa\rho\sigma}\,, \tag{21}$$

$$\frac{d}{dt} M^{\mu\nu 0} = M^{\mu\nu} - v^\mu M^{\nu 0} - \{\,{}^{\nu}_{\rho\sigma}\,\} M^{\mu\rho\sigma}\,, \tag{22}$$

$$v^\mu M^{\nu\sigma 0} + v^\nu M^{\mu\sigma 0} = M^{\mu\nu\sigma} + M^{\nu\mu\sigma}\,, \tag{23}$$

where all Christoffel symbols are evaluated at the center–of–mass position $x_{\rm cm}$ and where, in addition to (11), we used (13). With definition (13) we also have (note $\delta x^0 = 0$)

$$L^{\mu\nu} = M^{\mu\nu 0} - M^{\nu\mu 0}\,, \tag{24}$$

$$L^{\mu 0} = M^{\mu 00}\,. \tag{25}$$

Now we first express, using (23), $M^{\mu\nu\rho}$ in terms of v and L, and second, using (22), the propagation of L, and, at last, with (21) the equation for the center–of–mass motion.

Cyclic permutation of the three indices in (23) and adding two and subtracting the third relation gives

$$2M^{\mu\nu\rho} = v^\mu (M^{\nu\rho 0} + M^{\rho\nu 0}) + v^\nu L^{\mu\rho} + v^\rho L^{\mu\nu}\,, \tag{26}$$

where we used (24). By specifying $\sigma = 0$ in (23), we can express the first part also in terms of the angular momentum, $M^{\mu\nu 0} + M^{\nu\mu 0} = v^\mu L^{\nu 0} + v^\nu L^{\mu 0}$, so that we finally find

$$2M^{\mu\nu\rho} = v^\mu (v^\nu L^{\rho 0} + v^\rho L^{\nu 0}) + v^\nu L^{\mu\rho} + v^\rho L^{\mu\nu}\,. \tag{27}$$

Choosing $\nu = 0$ in (22) and reinserting this into (22) gives

$$0 = \frac{d}{dt} M^{\mu\nu 0} + \{\,{}^{\nu}_{\rho\sigma}\,\} M^{\mu\rho\sigma} - M^{\mu\nu} + v^\mu \left(v^\nu M^{00} + \frac{d}{dt} M^{\nu 00} + \{\,{}^{0}_{\rho\sigma}\,\} M^{\nu\rho\sigma} \right). \tag{28}$$

Antisymmetrization leads to an equation of motion for L:

$$0 = \frac{d}{dt} L^{\mu\nu} - \{\,{}^{\mu}_{\rho\sigma}\,\} M^{\nu\rho\sigma} + \{\,{}^{\nu}_{\rho\sigma}\,\} M^{\mu\rho\sigma} + \left[v^\mu \left(\frac{d}{dt} L^{\nu 0} + \{\,{}^{0}_{\rho\sigma}\,\} M^{\nu\rho\sigma} \right) - (\mu \leftrightarrow \nu) \right]. \tag{29}$$

Inserting $M^{\mu\rho\sigma}$ in terms of the center–of–mass velocity and the angular momentum, Eq.(27), we obtain a covariant equation of motion for the angular momentum L:

$$0 = D_v L^{\mu\nu} + v^\mu D_v L^{\nu 0} - v^\nu D_v L^{\mu 0} \,. \tag{30}$$

Multiplication with v_ν gives $D_v L^{\mu 0} = -v_\nu D_v L^{\mu\nu} - v_\nu v^\mu D_v L^{\nu 0}$ which can be inserted into (30)

$$0 = D_v L^{\mu\nu} + v^\mu v_\rho D_v L^{\rho\nu} + v^\nu v_\rho D_v L^{\mu\rho} = P_v D_v L^{\mu\nu} \,. \tag{31}$$

This is the equation of motion for the angular momentum.

In a similar fashion [2] we derive from (21) the equation of motion for the path. We get

$$D_v \left(m v^\mu - v_\rho D_v L^{\mu\rho} \right) = \frac{1}{2} R^\mu{}_{\nu\rho\sigma} v^\nu L^{\rho\sigma} \,. \tag{32}$$

By counting the degrees of freedom, it is clear that (31) and (32) are 6 equations for 3 components of v^μ and 6 components of $L^{\mu\nu}$. Therefore we have to reduce the numer of unknown components in the angular momentum. What is still unspecified in our approach is the center–of–mass coordinate. The center–of–mass coordinate can be determined by the so–called *Frenkel condition*

$$0 = L^{\mu\nu} v_\nu = v_\nu \int \delta x^{[\mu} \mathfrak{T}^{\nu]0} d^3x \,, \tag{33}$$

which leads to an expression of the form $0 = \int \rho \delta x^\mu d^3x +$ relativistic corrections, where $\rho = \mathfrak{T}^{00}$ is the energy density, see [5], e.g., for a detailed treatment of the center–of–mass problem.

If the Frenkel condition is valid, then it makes sense to introduce a vector for the angular momentum $L^\mu := \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} v_\nu L_{\rho\sigma}$. In terms of this vector, Eq.(30) means that the angular momentum vector is Fermi propagated,

$$F_v L = P_v D_v L = 0 \tag{34}$$

and thus is non–rotating.

3.2 Motion of an Elementary Particle with Spin $\frac{1}{2}$

A spin–$\frac{1}{2}$–particle ψ is assumed to obey an equation of motion which can be derived from the minimally coupled Lagrangian for the Dirac field in a Riemannian geometry:

$$\mathcal{L} = \sqrt{-g} \left[\frac{i\hbar}{2} \left(\bar{\psi} \gamma^\mu D_\mu \psi - (D_\mu \bar{\psi}) \gamma^\mu \psi \right) - m \bar{\psi} \psi \right] \,. \tag{35}$$

The parameter m is the mass of the Dirac particle. The matrices γ^μ are given by $\gamma^\mu = h^\mu_a \gamma^a$ where the γ^a are the special relativistic Dirac matrices obeying the Clifford algebra $\gamma^a \gamma^b + \gamma^b \gamma^a = 2\eta^{ab}$ ($\eta^{ab} = \mathrm{diag}(-1, +1, +1, +1)$). The tetrads h^μ_a are defined by $g_{\mu\nu} h^\mu_a h^\nu_b = \eta_{ab}$. Therefore

$$\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2 g^{\mu\nu} \,. \tag{36}$$

D_μ is the covariant spinorial derivative $D_\mu = \partial_\mu + \Gamma_\mu$ with the spinorial connection $\Gamma_\mu := -\frac{1}{2}(D_\mu h^\nu_a) h_{b\nu} G^{ab}$ (the covariant derivative D_μ acts on the vectorial index μ in $D_\mu h^\nu_a$ only). The $G^{ab} := \frac{1}{4}(\gamma^a\gamma^b - \gamma^b\gamma^a)$ are the generators of the Lorentz–group. The *adjoint* spinor is defined by $\bar\psi := \psi^+\gamma^{(0)}$ (in this case $\bar\psi\psi$ transforms as a scalar). Here $\gamma^{(0)}$ is the zeroth special relativistic Dirac–matrix.

Variation of the above Lagrangian with respect to $\bar\psi$ gives the field equation

$$0 = i\hbar\gamma^\mu D_\mu\psi - m\psi\,. \tag{37}$$

This is the Dirac equation in curved space–time. Here we use $c = 1$.

Now we describe a particle in a quasiclassical approximation. That means, we look for a solution of the Dirac equation (37) which locally has the form of a plane wave:

$$\psi = e^{\frac{i}{\hbar}S(x)} a(x)\,. \tag{38}$$

Inserting this ansatz into the field equations (37) gives

$$0 = -(\gamma^\mu\partial_\mu S + m)a + i\hbar\gamma^\mu D_\mu a\,. \tag{39}$$

The main step of the quasiclassical approximation consists in the assumption that the external fields are weak enough, so that, in first approximation, the derivatives of the amplitudes can be neglected. That means $(-\gamma^\mu\partial_\mu S + m)a + i\hbar\gamma^\mu D_\mu a \approx (-\gamma^\mu\partial_\mu S + m)a$, or $|\hbar\gamma^\mu D_\mu a| \ll |ma|$, where $|\cdot|$ denotes some norm on a complex vector space. If we use this condition, then we get from (39) with $p_\mu := -\partial_\mu S$

$$0 = (\gamma^\mu p_\mu - m)a\,. \tag{40}$$

Here p_μ is the momentum of the plane wave. Eq (40) is an algebraic condition which possesses a solution for p_μ if and only if the determinant of the coefficient matrix vanishes, $0 = \det(\gamma^\mu p_\mu - m)$. This leads to the condition

$$0 = \left(g^{\mu\nu}p_\mu p_\nu + m^2\right)^2\,. \tag{41}$$

(The exponent 2 characterizes the fact that we have for both spin states the same mass shell.) Eq (41) is a Hamilton–Jacobi partial differential equation for the phase $S(x,t)$ which always possess a solution.

From the plane wave ansatz (38) we can define a wave packet by superposition of plane waves from a continuous spectrum of momenta peaked around $\widehat{p}_\mu$. Then one can show that the tangent vector of the path of the peak of this wave packet is given by the group velocity $v^\mu := \frac{1}{m} g^{\mu\nu}p_\nu|_{p=\widehat{p}}$ which fulfills the normalization condition, $g(v,v) = -1$. Differentiating (41) once more yields immediately the geodesic equation for this group velocity

$$0 = D_v v\,. \tag{42}$$

The integral curves of this geodesic equation are the paths of the peaks of wave packets.

If we have a solution of the first order equation, then the first part in (39) vanishes and we get as equation for the next order of approximation

$$0 = \gamma^\mu D_\mu a \,. \tag{43}$$

What we are looking for is a propagation equation for the amplitude, that is, an equation of the form $D_v a = f(x,v)a$ which describes the evolution of a along the path given by v. For this we multiply (43) with $(\gamma^\nu p_\nu + m)/m$ and get

$$\begin{aligned} 0 &= \frac{1}{m}(p_\nu\gamma^\nu + m)\gamma^\mu D_\mu a \\ &= \frac{1}{m}\left(\gamma^\mu\left(D_\mu\left((p_\nu\gamma^\nu + m)a\right) - D_\mu(\gamma^\nu p_\nu + m)a\right) + p_\nu[\gamma^\nu,\gamma^\mu]D_\mu a\right) \\ &= -D_\mu v^\mu a - 2v^\mu D_\mu a \,. \end{aligned} \tag{44}$$

Here we used $\partial_{[\mu}p_{\nu]} = 0$. For obtaining this result, the existence of a Clifford algebra is important. With the definition for the expansion $\theta := D_\mu v^\mu$ we finally find

$$D_v a = -\frac{1}{2}\theta a \,. \tag{45}$$

The same holds true for the adjoint spinor: $D_v\bar{a} = -\frac{1}{2}\theta\bar{a}$. Within the frame of the theory of congruences [6] (see also the Appendix), θ is interpreted as the divergence of the trajectories given by the phase $S(x)$. If we define a normalized spinorial amplitude $b := a/\sqrt{\bar{a}a}$, then we get [7,8]

$$D_v b = 0 \,, \qquad D_v\bar{b} = 0 \,. \tag{46}$$

That means that the normalized spinors b and $\bar{b}$ are parallely propagated along the path of the center of the wave packet.

With these propagation equations for the spinors b and $\bar{b}$, we can calculate propagation equations for the bilinears [9] $S := \bar{b}b$, $P := \bar{b}i\gamma_5 b$, $j^\mu := \bar{b}\gamma^\mu b$, $S^\mu := \bar{\gamma}_5\gamma^\mu b$, and $S^{\mu\nu} := \bar{b}2iG^{\mu\nu}b$. Using (40) one can derive

$$Sp^\mu = mj^\mu \,, \tag{47}$$

$$S^\mu p_\mu = 0 \,, \tag{48}$$

$$P = 0 \,, \tag{49}$$

$$\epsilon^{\mu\nu\rho}{}_\sigma S^\sigma p_\rho = -mS^{\mu\nu} \,. \tag{50}$$

The last relation can be inverted:

$$v^a S^b - v^b S^a = \frac{1}{2}\epsilon^{ab}{}_{cd}S^{cd} \,. \tag{51}$$

From these identities we get an interpretation of these bilinear quantities. Eq.(47) yields $j^a = Sv^a$ so that S is the intensity of the Dirac field. Since S^a is an axial vector and, according to (48) a rest–frame quantity, it is identified with the spin of the Dirac particle.

Therefore, the only independent normalized quantities are the normalized current $\hat{j}^\mu = \bar{b}\gamma^\mu b = v^\mu$ and the normalized spin–vector $S^\mu = \bar{b}\gamma_5\gamma^\mu b$. The propagation equations (46) for b and $\bar{b}$ then give propagation equations for j^μ and S^μ:

$$D_v v = 0\,, \tag{52}$$

$$D_v S = 0\,. \tag{53}$$

Therefore the direction of the spin is parallely propagated along the path of the Dirac particle. The spin behaves in the same way as a spinning top. (For a gravitational theory with torsion it can be shown that the spin couples to torsion while the orbital angular momentum does not [10].)

We note without proof that in the next approximation it is possible to get an influence of the spin on the path of the wave packet [8]:

$$D_v v = \frac{1}{2}\lambda_{\mathrm{C}} R^*(\cdot, v, S, v)\,, \tag{54}$$

where R^* is the right–dual of the curvature tensor and $_{\mathrm{C}}$ the Compton wavelength of the Dirac particle.

4 Gravitational Field of a Rotating Body

In this section we want to derive the general features of a gravitational field which is created by a rotating body. The gravitational field, that is, the space–time metric $g_{\mu\nu}$, is given by Einstein's equations

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^2}\, T_{\mu\nu}\,, \tag{55}$$

where $G = 6.673(10) \times 10^{-11}\,\mathrm{m^3 kg^{-1} s^{-2}}$ is Newton's gravitational constant [11]. We now analyze two aspects of a rotation in the gravitational field: (i) we discuss the gravitational field of an arbitrary stationary situation and (ii) discuss the general structure of the gravitational field created from a rotating mass given in form of the energy momentum tensor.

4.1 Stationary Gravitational Field

A stationary gravitational field is characterized by a time–like Killing vector ξ, $g(\xi, \xi) < 0$, with $\pounds_\xi g = 0$, whereas a gravitational field with an axial symmetry is characterized by a space–like Killing vector η, $g(\eta, \eta) > 0$, with $\pounds_\eta g = 0$ and the integral curves of η are space–like closed curves. An axisymmetric space–time possess both Killing vectors ξ and η which, in addition, should commute $[\xi, \eta] = 0$. For any Killing vector field then there exists a coordinate system, which coordinate lines are the integral curves of the Killing vector field, so that the metric does not depend on the coordinates corresponding to the Killing field. For a stationary gravitational field this means $g(x) \stackrel{*}{=} g(x^1, x^2, x^3)$ and, if the

gravitational field possesses an additional axial symmetry, then $g(x) \stackrel{*}{=} g(x^1, x^3)$, where x^0 plays the role of the time coordinate and x^2 the role of the angle φ. Therefore $ds^2 = g_{\mu\nu}(x^1, x^3)dx^\mu dx^\nu$. If we choose $x^1 = \rho$ and $x^3 = \vartheta$, then

$$ds^2 \stackrel{*}{=} g_{00}(\rho, \vartheta)\, dt^2 + 2g_{0i}(\rho, \vartheta)\, dt\, dx^i + g_{ij}(\rho, \vartheta)\, dx^i\, dx^j\,. \tag{56}$$

We call a gravitational field static if the rotation of the time–like Killing congruence vanishes (cf. Appendix): $\omega = 0$ or $\epsilon^{\mu\nu\rho\sigma}\xi_\nu\partial_\rho\xi_\sigma = 0$. The gravitational field is stationary, if the Killing congruence rotates: $\omega \neq 0$ or $\epsilon^{\mu\nu\rho\sigma}\xi_\nu\partial_\rho\xi_\sigma \neq 0$.

The standard example for a static space–time is given by the Schwarzschild solution, and an example for a stationary space–time is given by the Kerr solution, see [12], e.g., or the space–time determined from a thin rotating disk [13].

4.2 Gravitational Field of a Rotating Source

It is intuitively clear what a rotating source is: The source consists of a set of point particles (a gas, or a rigid body, for example), which form a rotating congruence. The particles of the source may interact with one another. Therefore, the source is a conguence of point particles moving on trajectories with 4–velocity u. This congruence may possess rotation, acceleration, expansion, and shear. If we have, as a very simple example, a perfect fluid, then we have as source of the gravitational field the energy–momentum tensor $T^{\mu\nu} = (\rho + p)u^\mu u^\nu + pg^{\mu\nu}$, where ρ is the energy density and p the pressure. For $p = 0$ (dust) one can show that the geodesic equation (1) follows from (8). If the vector field u belongs to a rotating congruence, then this energy–momentum tensor describes a rotating source.

We now calculate the gravitational field which is created by such a rotating source. For this purpose, we split the metric into two parts, $g = g_0 + g_1$. The curvature associated with g can be split into a term corresponding to g_0 and terms depending on g_1 [12]: $R(g) = R(g_0) + \delta R(g_0, g_1)$. We assume that the curvature associated with g_0 vanishes, $R(g_0) = 0$. From the Einstein equations (55) we finally get a differential equation for the part $\widetilde{g}_1 := g_1 - \frac{1}{2}\bar{g}_1 g_0$ with $\bar{g} := g_0^{\mu\nu} g_{1\mu\nu}$:

$$\Box\, \widetilde{g}_1 = \kappa T\,, \tag{57}$$

where $\Box$ is the d'Alambertian with respect to the metric g_0 and where we have chosen a coordinate system such that $\partial_\nu g_1^{\mu\nu} = 0$, with $g_1^{\mu\nu} := g_0^{\mu\rho} g_0^{\nu\sigma} g_{1\rho\sigma}$. Because the curvature associated with g_0 vanishes, it is possible to introcduce a global coordinate system such that $g_{0\mu\nu} = \eta_{\mu\nu}$. If the source is stationary, then g_1 does not depend on the time and (57) reduces to the Poisson equation $\Delta\, \widetilde{g}_1 = \kappa T$ which can be integrated,

$$\widetilde{g}_1 = \kappa \int \frac{T(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d^3x'\,, \tag{58}$$

provided T falls off appropriately at spatial infinity.

The component $T_{00} =: \rho$ of the energy–momentum tensor is interpreted as the energy density, the components $T_{i0} = T_{i0} =: \rho v_i$ as the energy flux, and the T_{ij} as the stress tensor ($i, j, \ldots = 1, 2, 3$). Therefore,

$$(\widetilde{g}_1)_{00} = G \int \frac{\rho(t, \boldsymbol{r}')}{|\boldsymbol{x} - \boldsymbol{r}'|} d^3x' \,, \tag{59}$$

$$(\widetilde{g}_1)_{i0} = G \int \frac{\rho(t, \boldsymbol{r}') v_i(t, \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d^3x' \,, \tag{60}$$

$$(\widetilde{g}_1)_{ij} = G \int \frac{T_{ij}(t, \boldsymbol{r}')}{|\boldsymbol{x} - \boldsymbol{r}'|} d^3x' \,. \tag{61}$$

It is clear that $(\widetilde{g}_1)_{i0}$ is smaller than $(\widetilde{g}_1)_{00}$ by a factor v/c and $(\widetilde{g}_1)_{ij}$ by a factor $(v/c)^2$, cf. [12]. In the case of an isolated body and large distances, we have

$$U := (\widetilde{g}_1)_{00} = G\,\frac{M}{r} \,, \qquad h_i := (\widetilde{g}_1)_{i0} = -\frac{G}{2}\,\frac{(\boldsymbol{r} \times \boldsymbol{L})_i}{r^3} \,, \tag{62}$$

where M is the total mass and $\boldsymbol{L}$ the angular momentum of the gravitating body.

In a coordinate system where the components of the metric are isotropic, we have as line element

$$ds^2 = -(1 - 2U + 2U^2)\, dt^2 + (1 + 2U)(dx^2 + dy^2 + dz^2) - 4h_i\, dx^i\, dt \,, \tag{63}$$

or

$$g_{\mu\nu} = \begin{pmatrix} -1 + 2U - U^2 & -2h_i \\ -2h_i & (1 + 2U)\delta_{ij} \end{pmatrix} . \tag{64}$$

This metric is time–independent. Thus a time–like Killing vector $\xi = \partial_t$ exists, in components $\xi^\mu = \delta^\mu_0$. The different components of this vector ξ are given by

$$\xi^\mu = \delta^\mu_0 \,, \qquad \xi_0 = g_{0\nu}\xi^\nu = g_{00} \,, \qquad \xi_i = g_{i\mu}\xi^\mu = g_{0i} = -2h_i \,. \tag{65}$$

Hence the curl of the Killing vector field (see Appendix) is connected with $\partial_{[i} h_{j]}$.

The tetrads ϑ^a (one–forms, $a = 0, \ldots 3$) connected with the Killing vector field ξ are given by

$$\vartheta^{(0)}_0 = -1 + U - \frac{1}{2}U^2 \,, \qquad \vartheta^{(0)}_i = -h_i \tag{66}$$

$$\vartheta^{\hat{a}}_0 = -h_{\hat{a}} \,, \qquad \vartheta^{\hat{a}}_i = (1 + U)\delta^{\hat{a}}_i \,. \tag{67}$$

Later we need to boost this tetrad to a comoving (with the gyroscope) tetrad $\bar{\vartheta}^a = L^a{}_b(\dot{x})\vartheta^b$ where $\dot{x}$ is the relative velocity of the gyroscope with respect to the tetrad ϑ^a. Since for a pure boost with small velocity v we have $L = \begin{pmatrix} 1 - \frac{1}{2}\dot{x}^2 & \dot{x} \\ \dot{x} & \delta^i_j + \frac{1}{2}\dot{x}^i\dot{x}_j \end{pmatrix}$, we find for the new tetrad

$$\bar{\vartheta}^{(0)}_0 = -1 + U - \frac{1}{2}U^2 - \frac{1}{2}\dot{x}^2 \,, \qquad \bar{\vartheta}^{(0)}_i = -h_i + \dot{x}_i \tag{68}$$

$$\bar{\vartheta}^{\hat{a}}_0 = -h_{\hat{a}} + \dot{x}_{\hat{a}} \,, \qquad \bar{\vartheta}^{\hat{a}}_i = (1 + U)\, \delta^{\hat{a}}_i + \dot{x}^{\hat{a}}\dot{x}_i \,. \tag{69}$$

5 Lense–Thirring Effect

Let us combine the results and notions derived above in order to describe the dynamics of a rotating test body in the neighborhood of a gravitating rotating body. We consider a stationary situation: The gravitating body rotates with a constant angular velocity. In the field of such a body we will consider the motion of a point particle as well as the motion of a gyroscope.

5.1 Motion of a Point Particle

The equation of motion for a point particle is the geodesic equation $D_v v = 0$. The four components of this equation can be evaluated by using the solution for a static spherically symmetric mass distribution, that is the Schwarzschild solution, or the solution for rotating bodies, like the Kerr metric or the metric for a rotating disk of dust [13]. However, in our approach, we restrict ourselves to the case of a weak stationary gravitational field. In this case, for a point particle without spin and in 3–notation, the geodesic equation reads

$$\frac{d^2\boldsymbol{r}}{dt^2} = \boldsymbol{\nabla} U + \boldsymbol{F} - 2\boldsymbol{v} \times (\boldsymbol{\nabla} \times \boldsymbol{h}) \,. \tag{70}$$

The first term is the Newtonian part whereas $\boldsymbol{F}$ symbolizes nonlinear contributions of the gravitostatic field U which are responsible for the perihelion shift, for example. The last term is the gravitomagnetic part due to the rotation of the gravitating body. Since the motion of a satellite around the earth represents a gyroscope, too, this interaction results in a precession of the angular momentum of the satellite around the earth. Therefore, the plane of the path of the satellite is no longer fixed, as it is in the Schwarzschild case, but starts to precess instead. Thus the pericenter (perihelion) or the nodes (intersections of the paths of the satellite with the equatorial plane of the earth) move. This should be observable in an experiment proposed by Ciufolini [14] according to which two excentric satellites orbit around the earth.

5.2 Motion of a Gyroscope

For the description of the motion of a direction attached to a gyroscope we use Eqs.(34) and (32) or (53) and (54). For simplicity we assume that the path is geodesic which is very well fulfilled because, according to Eqs.(32) and (54), all non–geodesic terms can be neglected for weak gravitational fields.

The non–rotating frame defined by the gyroscope will be compared with a direction given by a fixed star far away from the gravitating body. The light from that star comes from a fixed direction. Thus the tangent vector l of light rays of this star is stationary: $\pounds_\xi l = 0$. Accordingly, we can introduce a stationary space–like unit vector: $e_{(1)} = P_u l/(g(P_u l, P_u l))^{1/2}$, which again is stationary $\pounds_\xi e_{(1)} = 0$. We can complete this unit vector to give a 3–bein by adding two more spatial unit vectors which are orthogonal to $e_{(1)}$ and to each other and

which are stationary, too: $g(e_{\hat{a}}, e_{\hat{b}}) = \delta_{\hat{a}\hat{b}}$, $\pounds_\xi e_{\hat{a}} = 0$. If we take $e_{(0)} = u$, then $e_a = (u, e_{\hat{a}})$ defines a tetrad with $\pounds_\xi e_a = 0$. Accordingly, also for the dual basis ϑ^a we have $\pounds_\xi \vartheta = 0$.

In order to determine the behaviour of the gyroscope with respect to the direction given by the fixed star, we project the spin of the gyroscope on the comoving basis connected with the fixed star: $\bar{S}^a := \bar{\vartheta}^a(S)$. This projection is the quantity observed. We calculate the dynamics of this projection $\dot{\bar{S}}^a = \partial_v \bar{S}^a$, where S is parallely displaced along the path of the gyroscope, while the basis ϑ^a is Lie–displaced along ξ:

$$\begin{array}{ccc} \bar{\vartheta} & \xrightarrow{\text{Lorentz–transf. } L^a{}_b} & \vartheta \\ \text{comoving with gyro} & & \text{attached to distant stars} \\ \updownarrow & & \updownarrow \\ D_v S = 0 & & \pounds_\xi \vartheta = 0 \end{array}$$

The four–velocity of the gyroscope is related to the fourth leg u via $v = \gamma u + \gamma \dot{x}^{\hat{a}} e_{\hat{a}}$ where γ is the Lorentz factor $(1 - \dot{x}^2)^{-1/2}$ and $\dot{x}^{\hat{a}}$ is the relative velocity measured between u and v. Moreover, because of $\bar{\vartheta}^{(0)}(S) = 0$, we have $0 = L^{(0)}{}_b \vartheta^b(S) = L^{(0)}{}_{(0)} \vartheta^{(0)}(S) + L^{(0)}{}_{\hat{a}} \vartheta^{\hat{a}}(S)$ so that

$$\vartheta^{(0)}(S) = -\dot{x}_{\hat{a}} \vartheta^{\hat{a}}(S)\,. \tag{71}$$

We calculate, using (34) and (53), respectively, ($\hat{a} = 1, 2, 3$),

$$\begin{aligned} \dot{\bar{S}}^{\hat{a}} &= D_v(\bar{\vartheta}^{\hat{a}}(S)) \\ &= (D_v \bar{\vartheta}^{\hat{a}})(S) + \bar{\vartheta}^{\hat{a}}(D_v S) \\ &= (D_v \bar{\vartheta}^{\hat{a}})(S) \\ &= (D_v(L^{\hat{a}}{}_b \vartheta^b))(S) \\ &= D_v L^{\hat{a}}{}_b \vartheta^b(S) + L^{\hat{a}}{}_b (D_v \vartheta^b)(S)\,. \end{aligned} \tag{72}$$

The first term can be evaluated by using $L^a{}_b = \begin{pmatrix} 1 + \frac{1}{2}\dot{x}^2 & \dot{x}_{\hat{b}} \\ \dot{x}^{\hat{a}} & \delta^{\hat{a}}_{\hat{b}} + \frac{1}{2}\dot{x}^{\hat{a}}\dot{x}_{\hat{b}} \end{pmatrix}$ and $D_v L^a{}_b = \begin{pmatrix} \dot{x}_{\hat{c}}\ddot{x}^{\hat{c}} & \ddot{x}_{\hat{b}} \\ \ddot{x}^{\hat{a}} & \frac{1}{2}(\ddot{x}^{\hat{a}}\dot{x}_{\hat{b}} + \dot{x}^{\hat{a}}\ddot{x}_{\hat{b}}) \end{pmatrix}$. With (71) this yields

$$D_v L^{\hat{a}}{}_b \vartheta^b(S) = D_v L^{\hat{a}}{}_{(0)} \vartheta^{(0)}(S) + D_v L^{\hat{a}}{}_{\hat{b}} \vartheta^{\hat{b}}(S) \approx -\frac{1}{2}(\dot{v}^{\hat{a}} v_{\hat{b}} - v^{\hat{a}} \dot{v}_{\hat{b}}) \bar{\vartheta}^{\hat{b}}(S)\,. \tag{73}$$

For the seond term

$$L^{\hat{a}}{}_b (D_v \vartheta^b)(S) = L^{\hat{a}}{}_b (D_{\gamma u + \gamma \dot{x}} \vartheta^b)(S) \approx L^{\hat{a}}{}_b (D_u \vartheta^b)(S) = L^{\hat{a}}{}_b e^{-U} (D_\xi \vartheta^b)(S) \tag{74}$$

we use the fact that the frame ϑ is stationary: $\pounds_\xi \vartheta^a = 0$. In components: $0 = \xi^\nu D_\nu \vartheta^a_\mu + \vartheta^a_\nu D_\mu \xi^\nu$, so that $(D_\xi \vartheta^a)_\mu = -\vartheta^a_\nu D_\mu \xi^\nu$. Thus

$$(D_\xi \vartheta^a)_\mu = -\vartheta^a_\nu D_\mu \xi^\nu = -\vartheta^a_\nu (e^U \omega_\mu{}^\nu + \overset{\xi}{a}_\mu \xi^\nu - \overset{\xi}{a}{}^\nu \xi_\mu)\,. \tag{75}$$

Therefore we find for the total precession of the spin

$$\begin{aligned}\dot{\bar{S}}^{\hat{a}} &= -\frac{1}{2}(\ddot{x}^{\hat{a}}\dot{x}_{\hat{b}} - \dot{x}^{\hat{a}}\ddot{x}_{\hat{b}})\;\bar{\vartheta}^{\hat{b}}(S) - L^{\hat{a}}{}_{b}e^{-U}\vartheta^{b}_{\nu}\;(e^{U}\omega^{\nu}{}_{\mu} + \overset{\xi}{\hat{a}}_{\mu}\xi^{\nu} - \overset{\xi}{\hat{a}}{}^{\nu}\xi_{\mu})\;S^{\mu} \\ &= -\omega^{a}{}_{b}\bar{S}^{b} - \left[\left(\overset{\xi}{\hat{a}}{}^{\hat{a}} + \frac{1}{2}\ddot{x}^{\hat{a}}\right)\dot{x}_{\hat{b}} - \dot{x}^{\hat{a}}\left(\overset{\xi}{\hat{a}}_{\hat{b}} + \frac{1}{2}\ddot{x}_{\hat{b}}\right)\right]\bar{S}^{\hat{b}}\,, \end{aligned} \tag{76}$$

with $u^{\hat{a}} = 0$ and where we neglected terms with "velocity $\times$ gravitomagnetic field" and terms of the order $\dot{x}^2$. Using $\ddot{x}^{\hat{a}} = a^{\hat{a}} + \overset{\xi}{\hat{a}}{}^{\hat{a}}$, where a is any non–gravitational acceleration, we finally arrive in 3–notation at

$$\frac{d}{d\tau}\boldsymbol{S} = \boldsymbol{\Omega}\times\boldsymbol{S}\,, \tag{77}$$

with

$$\boldsymbol{\Omega} = \boldsymbol{v}\times\left(-\frac{1}{2}\boldsymbol{a} + \frac{3}{2}\boldsymbol{\nabla}U\right) + \boldsymbol{\nabla}\times\boldsymbol{h}\,. \tag{78}$$

The first term $\boldsymbol{v}\times\boldsymbol{a}$ is a special relativistic term, called the *Thomas precession* which is known from atomic physics. It describes the precession of the spin due to inertial forces. Thus, the second term, $\boldsymbol{v}\times\boldsymbol{\nabla}U$, is a gravity–induced Thomas precession, the so–called *de Sitter precession* or *geodetic precession*. Note that only the Newtonian potential enters this term. The last term is purely post–Newtonian and describes the Lense–Thirring effect. This is the rotation of the locally non–rotating frame with respect to distant fixed stars due to the rotation of a nearby rotating gravitating body ('frame dragging').

6 On the Observation of Gravitomagnetic Effects

The systematic analysis of relativistic effects of planetary motion and motion of the moons of the planets dates back to the first years after the publication of Einstein's theory in 1915. In 1916, W. de Sitter [15] predicted a geodetic precession of the rotating earth-moon system ('earth–moon gyroscope') in the gravitational field of the sun. (The de Sitter term in (78) provides a simple model of the dynamics of that system.) The effect has finally been detected in the late 80's by means of an elaborate combination of lunar ranging and radio interferometry data [16]; refined data can be found in [17]. The accuracy of this verification is of the order of 1%. Three years after de Sitter, J. Lense and H. Thirring published their pioneering work "about the influence of the proper rotation of the central bodies on the motion of the planets and moons according to Einstein's gravitational theory" [18]. Analyzing the equations of motion (70), they excluded measurable effects for the moon orbit as well as for the orbits of the planets, but found considerable secular relativistic perturbations of the orbital parameters of the moons of the outer planets. In particular, Jupiter V evidenced a gravitomagnetic shift of its pericenter of 2.26 arcsec/yr. However, a confirmation of their prediction by observation was not possible at that time.

In the next two sections, we will discuss the theoretical fundamentals of recent satellite experiments. In this context, the earth will be considered to be a sphere (radius R) rotating with a constant angular velocity about its axis, which has a fixed orientation in an inertial system connected with the (distant) stars.

6.1 Lense–Thirring Effect for Point Particles

According to (70) and (62), the equation of motion for a spinless particle (satellite) in the gravitational field of the rotating earth is given by[1]

$$\frac{d^2\boldsymbol{r}}{dt^2} = -\frac{1}{r^3}GM\mathbf{r} + \frac{2G}{c^2r^3}\frac{d\boldsymbol{r}}{dt} \times \left(\boldsymbol{L} - \frac{3\,(\boldsymbol{L}\cdot\boldsymbol{r})}{r^2}\boldsymbol{r}\right) , \tag{79}$$

where M is the mass of the earth and $\boldsymbol{L}$ is its angular momentum. A detailed discussion of these equations and explicit expressions for the relativistic perturbations of the particle orbit may be found in the original paper of Lense and Thirring [18].

According to the proposal of Ciufolini, it should be possible, with present-day technology, to measure the advance of the pericenter and the nodes of highly eccentric satellites. A first attempt to do this using the LAGEOS I and LAGEOS II satellites, has been carried through in [19–21]. It was possible to verify the gravitomagnetic effect with an precision of about 10%. This poor precision is a result of the low eccentricity of the orbits of the satellites and the difficulties in eliminating the multipoles of the earth which give rise to contributions of comparable order. The idea of a further experiment [14] is to orbit a new LAGEOS satellite with the same orbital parameters as those of an existing LAGEOS, but with supplementary inclinations, and to observe the bisector of the angle between the nodal lines which defines a kind of gyroscope. The expected precision of the verification of the gravitomagnetic effect is of the order of 3% after 3 years of Laser measurements (see also [22]).

6.2 Lense–Thirring Effect for Gyroscopes

Since the early 60's, the Stanford orbiting gyroscope experiment, Gravity Probe B, has been under development [23]. The experimental construction has been completed and the experiment should be performed in 2000/2001. The idea is to put a spacecraft in a polar orbit equipped with four gyroscopes (see Fig.1 on page 53 of Everitts's paper in this volume) and to measure the gravitomagnetic precession of the spins of the gyroscopes. To calculate the expected numerical values for the de Sitter and Lense–Thirring effects, we consider a single gyroscope with spin $\boldsymbol{S}$ at a circular polar orbit. We introduce a co–moving but non–rotating orthogonal coordinate system Σ the z–axis of which is parallel to the earth's angular momentum $\boldsymbol{L}$ and the x–axis of which lies in the orbital plane ($y = 0$), which has a fixed position with respect to the distant stars. The orientation of the co–moving frame can be maintained by two telescopes on board the satellite each of which points at a particular fixed star.

[1] For simplicity, the nonlinear term of (70) has been omitted.

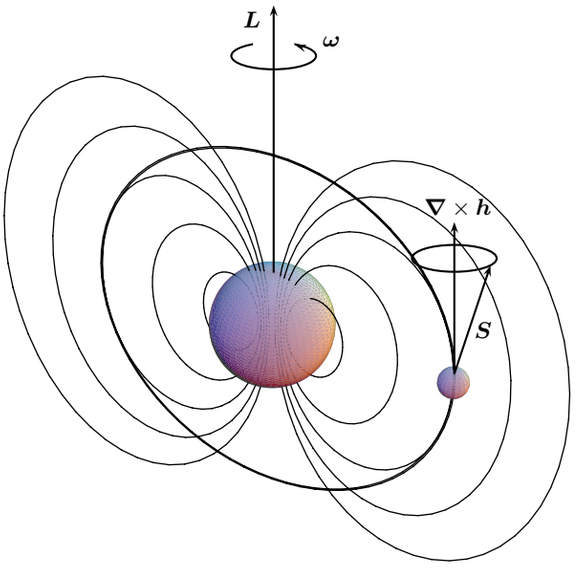


Fig. 3. The Lense–Thirring effect for gyroscopes: The earth rotating with angular velocity $\boldsymbol{\omega}$ and angular momentum $\boldsymbol{L}$ creates a gravitomagnetic field with the shape of a magnetic dipole. A gyroscope with angular momentum or spin $\boldsymbol{S}$ moves around the earth along a geodesic circular polar orbit (thick solid line). The Lense–Thirring effect consists in the precession of $\boldsymbol{S}$ around the direction given by the field lines of $\boldsymbol{\nabla} \times \boldsymbol{h}$.

In order to apply Eqs. (77,78), we have to specify the position vector $\boldsymbol{r}$ of the gyroscope (origin of Σ) and the angular velocity $\boldsymbol{\Omega}$. Obviously,

$$\boldsymbol{r} = (r \cos \omega_0 t, 0, r \sin \omega_0 t) \,. \tag{80}$$

Here r is the constant distance of the gyroscope from the center of the earth and

$$\omega_0 = \frac{1}{r} \sqrt{\frac{GM}{r}} \tag{81}$$

is the orbital angular velocity of the satellite. Then,

$$\boldsymbol{\Omega} = \left(\frac{3GL}{2r^3c^2} \sin 2\omega_0 t, \quad \frac{3}{2} \sqrt{\frac{GM}{r}} \frac{GM}{r^2c^2}, \quad \frac{GL}{2r^3c^2} \left(1 - 3 \cos 2\omega_0 t\right) \right) . \tag{82}$$

A good approximation for the angular momentum of the earth is

$$L = 0.3306 \cdot MR^2\omega, \tag{83}$$

where $M = (5.974 \pm 0.004)\, 10^{27}$ g is the mass, $R = (6378140 \pm 5)$ m the equatorial radius, and ω the angular velocity of the earth.

Note that the choice of Σ has separated the de Sitter and the Lense–Thirring contributions to $\boldsymbol{\Omega}$: Ω_x and Ω_z are pure Lense–Thirring terms whereas Ω_y is of geodetic origin. After a decomposition of the spin vector $\boldsymbol{S}$ in spherical polar coordinates,

$$S_x = S\,\cos\varphi\,\sin\vartheta\,, \quad S_y = S\,\sin\varphi\,\sin\vartheta\,, \quad S_z = S\cos\vartheta\,, \quad S = |\boldsymbol{S}|\,, \tag{84}$$

Eqs.(77,78) take the form

$$\dot{\varphi} = -\Omega_x\,\cos\varphi\,\cot\vartheta - \Omega_y\,\cot\vartheta\,\sin\varphi + \Omega_z\,, \tag{85}$$

$$\dot{\vartheta} = -\Omega_x \sin\varphi + \Omega_y\,\cos\varphi\,. \tag{86}$$

In order to keep the two effects separate, we may start from an equatorial position ($t = 0; \boldsymbol{r} = (r, 0, 0)$) and choose the spin vector $\boldsymbol{S}$ to be perpendicular to the angular momentum $\boldsymbol{L}$ of the earth ($t = 0 : \varphi = 0, \vartheta = \frac{\pi}{2}$). From the linearized Eqs.(85,86)

$$\dot{z} = i\Omega_x z + (\Omega_y + i\Omega_z)\,, \tag{87}$$

where $z = \vartheta - \frac{\pi}{2} + i\varphi$, we finally obtain the desired secular angular precessions

$$\Delta\vartheta = \Omega_y \Delta t = \frac{3}{2}\sqrt{\frac{GM}{r}}\frac{GM}{r^2c^2}\Delta t \tag{88}$$

$$\Delta\varphi = \overline{\Omega_z}^t \Delta t = \frac{GL}{2r^3c^2}\Delta t\,. \tag{89}$$

Here $\overline{\Omega_z}^t$ is the time–averaged Ω_z as experienced by the gyroscope. For a polar orbit at about 650 km altitude ($r = (6371 + 650)$ km) this leads (note also (83)) to a rate of

$$\frac{\Delta\vartheta}{\Delta t} = 6.6 \text{ arcsec/yr} \tag{90}$$

for the geodetic precession and

$$\frac{\Delta\varphi}{\Delta t} = 0.041 \text{ arcsec/yr} \tag{91}$$

for the Lense–Thirring precession.

The goal for the precision of the Gravity Probe B experiment is about 0.01% for the de Sitter effect and about 1% for the Lense–Thirring effect (in contrast to the Lense–Thirring effect for orbiting point particles the multipole moments of the earth play no role here [24]). The measurement of these effects is based on SQUIDs. The numerical values (90) are illustrated in Fig.1 of Everitt's talk (see p. 53), which, moreover, describes the technicalities o f the sophisticated equipment.

6.3 Lense–Thirring Effect in Quantum Physics

Due to huge improvements in the accuracy of devices based on the quantum properties of matter, it may be useful to estimate the effect of a rotating gravitating body on quantum particles. Two types of effects can be imagined: The effect on a matter wave interferometer and the effect on the spectrum of atoms.

In the first case, the field $\boldsymbol{\Omega}$ acts like a rotation of the interferometer if the interferometer is attached to the fixed stars. The effect of rotation of the interferometer on the phase of the quantum interference is the famous Sagnac effect $\delta\phi_{\text{Sagnac}} = \frac{m}{\hbar}\boldsymbol{\omega}\cdot\boldsymbol{A}$, where $\boldsymbol{\omega}$ is the angular velocity of the interferometer and $\boldsymbol{A}$ its area. Due to great success, e.g., in the cooling of atoms which makes it possible to prepare interfering atoms which stay for a long time inside the interferometer and thus possess a long interaction time, it may be possible to detect the Lense–Thirring effect with atomic interferometry, see [25] for a recent account on the sensitivity of atomic interferometers on rotation.

A realization of this effect is attempted within the HYPER project which is planned to put atomic interferometers in space and to measure, beside the fine structure constant and the quantum gravity induced foam structure of space, the Lense–Thirring effect. For this purpose, two atomic interferometers based on Mg and two based on Cs will be placed in two orthogonal planes. The resolution of rotation rates aimed at is 10^{-14} rad/s for an integration time of 1000 s. Note that, contrary to the GP-B approach where the cumulative effect over apoproximately one year is read out, in this case the *angular velocity* $\boldsymbol{\nabla}\times\boldsymbol{h}$ is measured *locally.* No integration over many days is carried through. The integration takes place for a few minutes only, that is, for a duration, over which the curl $\boldsymbol{\nabla}\times\boldsymbol{h}$ is approximately constant. HYPER is planned to be put into orbit within the next 10 years.

It has been shown that the rotation of the earth has an influence on the spectrum of atoms: While searching for anomalous spin–couplings in atoms [26], one has to compensate for the influence of the earth's rotation on the spin. This in fact establishes [27] an experimental verification of the coupling between spin and rotation, see [28]. However, since the accuracy of this result is not very high, and since the rotation caused by $\boldsymbol{\Omega}$ is about 9 orders of magnitude smaller than the earth's rotation, there is no hope in near future to use this approach for a verification of the Lense–Thirring effect.

Appendix: Theory of Congruences

For a time–like vector field u, $g(u,u) = -1$, which may be interpreted as a field of four velocities being tangents at a set of point–like particles, like dust, we have [6]

$$D_\mu u_\nu = \omega_{\nu\mu} + \sigma_{\nu\mu} + \frac{1}{3}\theta P_{\mu\nu} - u_\mu a_\nu\,, \tag{92}$$

where

$$\omega_{\nu\mu} := (P_u)^\rho_\nu (P_u)^\sigma_\mu D_{[\sigma}u_{\rho]}\,, \tag{93}$$

$$\theta := P_u^{\mu\nu} D_{(\mu}u_{\nu)}\,, \tag{94}$$

$$\sigma_{\nu\mu} := (P_u)^\rho_\nu (P_u)^\sigma_\mu D_{(\sigma}u_{\rho)} - \frac{1}{3}\theta (P_u)_{\mu\nu}\,, \tag{95}$$

$$a := D_u u\,, \tag{96}$$

with the projection operator

$$(P_u)^\nu_\mu := \delta^\nu_\mu + u^\nu u_\mu \,. \tag{97}$$

For a time–like Killing congruence, that is, a congruence the four–velocity of which is proportional to a Killing vector field,

$$\xi = e^U u \,, \qquad \pounds_\xi g = 0 \,, \qquad g(u,u) = -1 \,, \tag{98}$$

we have $\sigma_{\mu\nu} = 0$ and $\theta = 0$. A Killing congruence possesses only rotation and acceleration. (One can show, that ω is, indeed, a rotation as defined in section 2.) Then the acceleration of u is given by

$$\overset{\xi}{a} := D_u u = dU \,, \qquad D_u U = 0 \,, \tag{99}$$

which we get from projecting $0 = D_\mu \xi_\nu + D_\nu \xi_\mu$ onto u^ν and $u^\mu u^\nu$. With this result we get

$$D_\mu \xi_\nu = D_\mu (e^U u_\nu) = e^U \omega_{\nu\mu} + \overset{\xi}{a}_\mu \xi_\nu - \overset{\xi}{a}_\nu \xi_\mu \,. \tag{100}$$

References

1. F.A.E. Pirani: A note on bouncing photons, *Bull. Acad. Polon. Sci., Ser. Sci. Math. Astr. Phys.* **13**, 239 (1965).
2. A. Papapetrou: Spinning test–particles in general relativity I, *Proc. Roy. Soc.* (London) **A** , 248 (1951).
3. E. Corinaldesi and A. Papapetrou: Spinning test–particles in general relativity II, *Proc. Roy. Soc.* (London) **A**, 259 (1951).
4. W.G. Dixon: Dynamics of extended bodies in General Relativity III: Equations of motion, *Phil. Trans. R. Soc. London* **A 277**, 59 (1974).
5. J. Ehlers and E. Rudolph: Dynamics of extended bodies in general relativity: center-of-mass description and quasirigidity, *Gen. Rel. Grav.* **8** 197 (1977).
6. J. Ehlers: Beiträge zur relativistischen Mechanik kontinuierlicher Medien, *Akad. Wiss. Lit. Mainz Abh. Math.-Nat. Kl.*, Seite 793 (1961).
7. J. Stachel und J. Plebanski: Classical particles with spin I: The WKBJ approximation, *J. Math. Phys.* **18**, 2368 (1977).
8. J. Audretsch: Trajectories and Spin Motion of Massive Spin $\frac{1}{2}$ Particles in Gravitational Fields, *J. Phys. A: Math. Gen.* **14**, 411 (1981).
9. J.D. Bjorken and S.D. Drell: *Relativistic Quantum Mechanics*, McGraw-Hill, New York (1964).
10. P.B. Yasskin and W.R. Stoeger: Propagation equations for test bodies with spin and rotation in theories of gravity with torsion, *Phys. Rev.* **D 21**, 2081 (1980).
11. S. Schlamminger, E. Holzschuh, W. Kündig, F. Nolting, and J. Schurr: Determination of the Gravitational Constant, this volume p. 15.
12. C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation*, Freeman, San Francisco 1973.
13. G. Neugebauer and R. Meinel: General Relativistic Gravitational Field of a Rigidly Rotating Disk of Dust: Solution in Terms of Ultraelliptic Functions, *Phys. Rev. Lett.* **75**, 3046 (1995).

14. I. Ciufolini and J.A. Wheeler: *Gravitation and Inertia*, Princeton Series in Physics, Princeton University Press, Princeton 1995.
15. W. De Sitter: On Einstein's theory of gravitation and its astronomical consequences, *Mon. Not. Roy. Astron. Soc.* **77**, 155 and 481 (1916).
16. I.I. Shapiro, R.D. Reasenberg, J.F. Chandler, R.W. Babcock: Measurement of the deSitter perecession of the Moon: A relativistic three–body effect, *Phys. Rev. Lett.* **61**, 2643 (1988).
17. J.G. Williams, and X.X. Newhall, J.O. Dickey: Relativity parameters determined from lunar laser ranging, *Phys. Rev.* **D 53**, 6730 (1995).
18. J. Lense and H. Thirring: Über den Einfluß der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie, *Physik. Zeitschr.* **19**, 156 (1918).
19. I. Ciufolini, D. Lucchesi, F. Vespe, and F. Chieppa: Measurement of gravitomagnetism, *Europhys. Lett.* **39**, 359 (1997).
20. I. Ciufolini, F. Chieppa, D. Luccesi, and F. Vespe: Test of Lense–Thirring orbital effect due to spin, *Class. Quantum Grav.* **14**, 2701 (1997).
21. I. Ciufolini, E. Pavlis, F. Chieppa, E. Fernandes–Vieira, and J. Pérez–Mercader: Test of General Relativity and measurement of the Lense–Thirring effect with two Earth satellites, *Science* **279**, 2100 (1998).
22. M. Schneider: *Himmelsmechanik (Band 4: Theorie der Satellitenbewegung, Bahnbestimmung)* (*Celestial Mechanics, Volume 4: Theory of Satellite Motion, Determination of Paths*, in German), Spektrum Akademischer Verlag, Heidelberg, Berlin 1999.
23. C.W.F. Everitt *et al.*: this volume, p. 52.
24. R.J. Adler, A.S. Silbergleit: A General Treatment of Orbiting Gyroscope Precession, http://arXiv.org/abs/gr-qc/9909054.
25. T.L. Gustavson, P. Boyer, M. Kasevich: Precision rotation measurements with an atom interferometer gyroscope, *Phys. Rev. Lett.* **78**, 2046 (1997).
26. B.J. Venema, P.K. Majumder, S.K. Lamoreaux, B.R. Heckel, and E.N. Fortson: Search for a Coupling of the Earth's Gravitational Field to Nuclear Spins in Atomic Mercury, *Phys. Rev. Lett.* **68**, 135 (1992).
27. B. Mashhoon: On the coupling of intrinsic spin with the rotation of the earth, *Phys. Lett.* **A 198**, 9 (1995).
28. B. Mashhoon: Neutron Interferometry in a Rotating Frame of Reference, *Phys. Rev. Lett.* **61**, 2639 (1988).

Gravity Probe B: Countdown to Launch

C.W.F. Everitt[1], S. Buchman[1], D.B. DeBra[1], G.M. Keiser[1], J.M. Lockhart[1 2], B. Muhlfelder[1], B.W. Parkinson[1], J.P. Turneaure[1], and other members of the Gravity Probe B team

[1] W.W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA

[2] Physics Department, San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132, USA

Abstract. NASA's Gravity Probe B Mission is a test of two predictions of Einstein's General Theory of Relativity based on observations on very precise cryogenic gyroscopes in a satellite in a 650 km polar orbit about the Earth. Construction and the first round of testing of the flight payload was completed in December 1999. Of the 32 planned qualification tests 28 were passed with complete success, meeting or in several instances surpassing the program requirements. However, one test very unexpectedly revealed a p roblem in the thermal performance of the Dewar/Probe system which has required a significant redesigin and rework, now successfully completed. Gravity Probe B is scheduled for launch on April 1, 2002. This article reviews from the physicist's viewpoint the experience of living through a space flight program.

1 Gravity Probe B: An Experiment in Physics – and Management

To physicists trained in ground–based research the thought of carrying out an experiment in space is a daunting prospect. The laboratory experiments we are accustomed to seldom work at a first attempt: many stages of redesign and reconstruction are necessary before the desired performance can be met. Space is different. Space experiments of their very nature have to be designed and built so that once on orbit they will work first time. Furthermore the design has to be 'robust'. Robustness means introducing redundancy into areas that might fail. Redundancy, however, adds complexity. Striking the right balance between these competing requirements is critical to any successful space mission. By now, after forty years of space flight, considerable experience exists about how to approach such issues (not least how to bring together the right mix of science, engineering and management skills to do so).

The NASA/Stanford Gravity Probe B Mission (GP-B) is a Fundamental Physics experiment designed to provide two extremely precise tests of Einstein's theory of gravitation, General Relativity, based on observations of electrically suspended gyroscopes in a satellite in a 640 km circular polar orbit around the Earth, see Fig.1. It will measure the geodetic effect due to the curvature of space–time by the Earth to approximately 2 parts in 10^5 and the frame–dragging effect, with its subtle connections to gravitomagnetism and Mach's principle, to an accuracy approaching 0.3%.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 52–82, 2001.

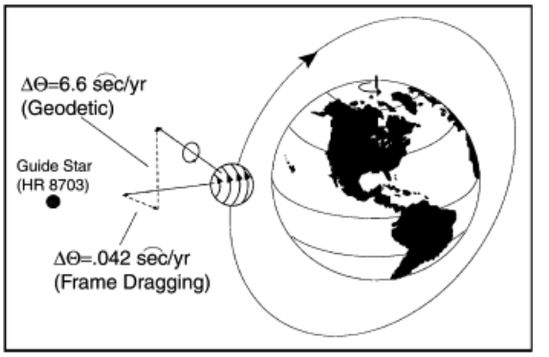


Fig. 1. The Two Relativity Effects Predicted by Schiff.

A crucial feature of GP-B is that it is 'a controlled physics experiment'. It is so in two senses. In the first place the two relativistic effects dominate the data: error terms such as the Newtonian drifts of gyroscopes are reduced to negligible values. Second, no less important, designed into the experiment is a rigorous program of on–orbit verification and calibration in which mission parameters are varied in a controlled way to check and, if necessary, remove specific sources of err or.

After many years of technology development, GP-B is now nearing completion. Final assembly of the payload with the spacecraft takes place in March 2001. Launch is scheduled for May 1, 2002.

It was not until late in 1959, 44 years after Einstein published his theory, that any conceptually new test of General Relativity was proposed. It was then that two men, Leonard Schiff and George Pugh, in complete independence of each other, were able to define one, or rather two, fundamentally new tests based on observations of non–Newtonian precessions of gyroscopes in Earth orbit. Remarkably, almost at the same time a number of other tests of the theory were conceived including the Shapiro time delay ex periment and redshift tests leading to the Gravity Probe A sub–orbital clock experiment.Then came various new or extended theoretical approaches, best known of which is the parameterized post–Newtonian (PPN) formalism due mainly to Kenneth Nordtvedt and Clifford Will. PPN provides a framework for comparing metric theories of gravitation; it has also suggested interesting new null experiments.

It is no accident that Schiff and Pugh came to their ideas just two years after Sputnik. Both recognized that an orbiting gyroscope would experience the geodetic and frame–dragging effects. They also saw that the performance of a gyroscope in space is potentially much better than on Earth because of the reduction in support force. Schiff's principal contribution was to provide the first elegant and correct derivation of the two effects. Pugh in a striking, too little known paper (November 1959) gave an imp ressively complete error analysis of a possible experiment. In particular he suggested for the first time the brilliant

and far–reaching concept of a drag–free or drag–compensated satellite. Gravity Probe B has evolved out of these early thoughts, together with discussions Schiff had at Stanford in 1959–60 with William Fairbank of the Physics Department and Robert Cannon of the Department of Aeronautics and Astronautics. From there GP-B became a cooperative effort between the two departments, a most for tunate occurrence.

Indeed, without a collaboration of this kind it is hard to imagine even attempting a program involving so many diverse technologies. In this test of Einstein, physics and engineering are inextricably linked. Important also has been the presence on the Stanford campus of an infrastructure capable of making the collaboration effective, provided by the interdisciplinary Hansen Experimental Physics Laboratory (HEPL) with its long history of developing medium–scale, long–term scientific programs. HEPL has prov ed to be an ideal administrative vehicle for a mission of this complexity. In addition, a rarity on university campuses, it offers an extended building with flexible high bay space and crane coverage. To have such a building within easy walking distance of the two departments so that physics and engineering students, academic staff and faculty can come together in a common enterprise is an extraordinary benefit. This happy combination of circumstances, not planned in advance, deserves reflection. Others int ending to perform fundamental space experiments may find it wise to think through not only the experiment itself but the structures and expertise that will be needed to make it happen.

No less important is to think through, in conjunction with NASA or the relevant space agency, the relationships between academia and aerospace industry. On GP-B with invaluable support from NASA Marshall Center, excellent working relations were established with Ball Aerospace in the early study phase of the program and subsequently in flight contracts with Lockheed Martin. In establishing the university–industry relation it is essential to formulate a clear plan of what should be done in academia and what in industry, and to devise a suitable contractual vehicle for managing the industrial effort. In facing, as one must, technical surprises, it is well to be prearmed as much as possible against the contractual surprises that too often accompany them.

2 Shape of the Experiment

The GP-B Science Instrument Assembly (SIA) is illustrated schematically in Fig.2. It comprises four gyroscopes with their spin axes aligned parallel to the line of sight to the guide star and mounted in line in a quartz block structure, to which is attached a reference telescope, all operating at a temperature of 1.8 K. The gyroscopes are fused quartz spheres 38 mm in diameter, coated with a 1.25 μm film of superconducting niobium, suspended electrically within fused quartz housings by voltages applied to three mutually perpendicular saucer–shaped electrodes, as illustrated in Fig.3. To support the rotor on Earth the voltage required over the 32 μm gap is 650 V. In space it is reduced to 0.1 V. The gyroscopes are spun up on orbit to about 100 Hz by means of helium gas at

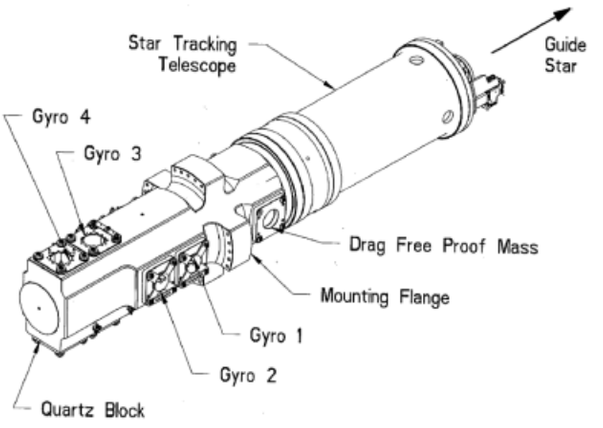


Fig. 2. The GP-B Science Instrument Assembly.

a temperature of 6.5 K running at sonic velocity through a differentially pumped channel in one half of the housing.

The gyro readout is based on a special property of superconductors known as the London moment. A spinning superconductor develops a magnetic field aligned with its instantaneous spin axis, and this serves as a magnetic 'pointer' even in a perfectly round, perfectly uniform sphere. The direction of the London moment, and therefore of the spin, is measured magnetically by a SQUID magnetometer connected to a four–turn superconducting loop sputtered on to the parting plane of the housing. Because the spacecraft rolls around the line of sight to the star the signal is chopped at roll rate, reducing limits on measurement from $1/f$ noise. The roll period is between one and three minutes (17 to 5.5 mHz). At 5.5 mHz the noise performance is about 7×10^{-29} J/Hz, equivalent to resolving 1 mas[1] in an integration time of 7 hours, meeting the GP-B mission goal of 1 mas in 10 hours.

More generally, rolling the spacecraft proves to be a very important overall symmetrizing principles, averaging out other potential sources of error in Gravity Probe B: drifts in the telescope readout, and many of the Newtonian drift torques on the gyroscope.

The SIA is enclosed in a cylindrical cryogenic vacuum 'probe', 0.3 m in diameter and 2.4 m long, inserted into the inner well of a dewar vessel containing 2400 l of superfluid helium (Fig.4). The dewar, which maintains cryogenic temperatures on orbit for 18 months, serves as the main structural element of the spacecraft. Boil–off gas from the dewar, vented through proportional thrusters, provides thrust authority for attitude, translational, and roll control. Also shown in Fig.4 is a sunshade that prevents scattered sunlight entering the telescope.

[1] 1 mas = 1 milli–arc–second.

Fig. 3. The GP-B Flight Gyroscope: Note the difference between the two halves of the gyro housing. Each has three circular support electrodes, but the one on the right has the differentially pumped spin–up channel, while the one on the left has the superconducting loop sputtered on to its parting–plane.

An essential feature of the experiment is the provision of both dc and ac magnetic shielding. For the London moment readout to work, the field trapped in the gyro rotor during cooling through its superconducting transition temperature must be below the extremely low value of 3×10^{-10} T – well below the limit achievable in practice by conventional ferromagnetic shields. Equally important is to eliminate ac disturbances from external magnetic fields such as the Earth's: the required ac shielding fa ctor is 10^{12}. These requirements are met by surrounding the instrument with a nested series of conventional and superconducting magnetic shields, and placing rigorous constraints on the magnetic properties of the probe and SIA. To produce the required ultralow magnetic field use is made of a technique developed originally by B. Cabrera, based on expanding a succession of superconducting lead shields in the inner well of the dewar. A conventional 'cryoperm' shield produces an initial ambient field in th e well of about 10^{-6} T. Because the quantity conserved in superconductors is magnetic flux (field times area), if a folded lead bag is cooled in this low field and then expanded, the new field level will be substantially lower. The process is then repeated with a second lead bag to obtain an even lower field. In practice, with three expansions, field levels as low as 10^{-11} T are regularly obtained.

In their earliest proposals for an orbiting gyroscope experiment Pugh, and independently Schiff and his colleagues, recognized that the performance of a gyroscope would be greatly improved by operating in the low g environment of space. The reason is simple. By far the most precise kind of gyroscope, then and now, is an electrically supported spinning sphere. The simplest of the torques

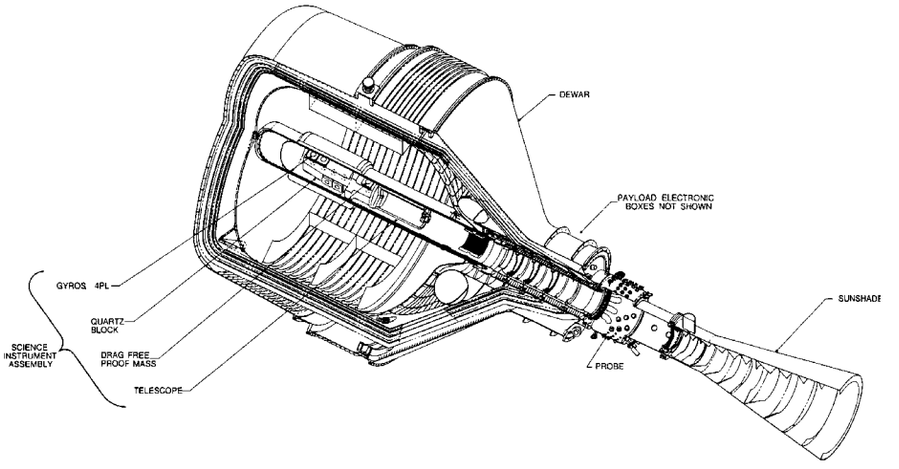


Fig. 4. The Science Mission Dewar and Probe-C: Cutaway view, with the Science Instrument Assembly in position with its separate vacuum 'probe'.

acting on such a suspended sphere is the 'mass–unbalance' torque which occurs when an acceleration acts on a sphere that is perfectly round but not quite perfectly homogeneous. The drift–rate Ω caused by an acceleration f at right angles to the spin axis is

$$\Omega = \frac{5}{2}\frac{f}{v_s}\frac{\delta r}{r} \tag{1}$$

where r is the radius and v_s the peripheral velocity of the spinning sphere, and δr is the distance between center of mass and center of support. For a sphere homogeneous to 1 part in 10^6 with a peripheral velocity of 20 m/s the resulting mean transverse acceleration required to reduce the drift–rate to 0.03 mas/yr is 4×10^{-12} g. This is an instructive number. It demonstrates the insight of George Pugh – and also of B.O. Lange, who independently advanced the same idea in August 1961 – in recognizing the need for drag compensation. The acceleration levels on satellites at the GP-B altitude with normal area/mass ratios are typically of the order 10^{-8} g. In Gravity Probe B as a *drag–free satellite* rolling about the common axis of the gyroscopes the mean cross–track acceleration will be below 10^{-12} g.

Arguments of the same kind developed in detail by G.M. Keiser and A. Silbergleit, and discussed further in Section 4 below, establish that torques due to the action of the suspension voltages on the out–of–roundness of the gyroscope are reduced to similarly acceptable levels. In all, three different classes of errors have to be considered in Gravity Probe B: gyro drift errors, gyro readout errors, uncertainty in the proper motion of the guide star. Fig.5 gives current estimates

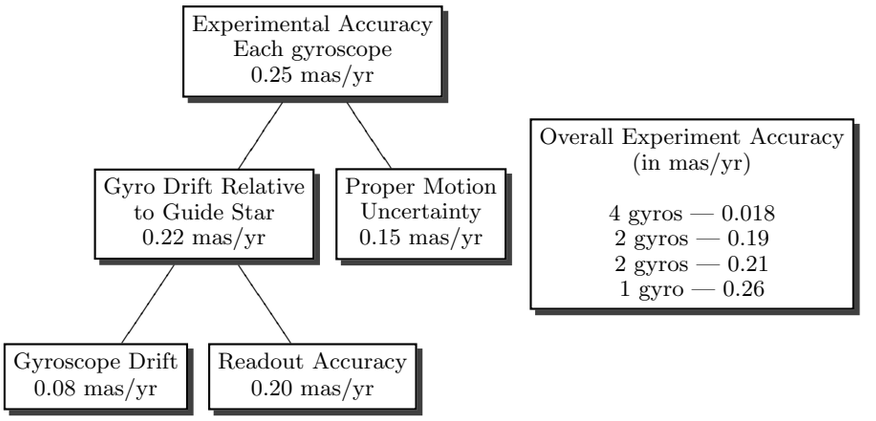


Fig. 5. Expected GP-B Experiment Accuracy.

of each of these terms and the result for an overall experiment with one, two, three, and four gyroscopes.

The spacecraft carrying electronics, telemetry, solar panels, mass trim mechanisms, a sunshield for the telescope, and other support equipment, fits around the dewar. Fig.6 is a general view of GP-B in orbit.

3 Incremental Prototyping

From 1964 to 1984 the main effort on Gravity Probe B went to advanced technology development, along with Mission Definition and Phase A studies and an in–house Phase B study performed at NASA Marshall Center. In US fiscal year 1985 NASA decided to proceed with an instrument development program, the Shuttle Test of the Relativity Experiment (STORE). As first conceived, the object of STORE was to fabricate the Science Payload, and then launch it on Shuttle for a seven–day rehearsal experiment that would tes t all systems in space and operate the gyros at Shuttle acceleration levels. The Payload would then be returned to Earth, mounted into the spacecraft, and launched once more by Shuttle from the SL-6 Western Test Range at Vandenberg Air Force Base in California for release into a free–flying polar orbit.

STORE provided the essential mechanism for developing the Gravity Probe B Payload. However, with the Challenger accident in 1986 and subsequent closure of the Western Test Range the planning had to be rethought. The new constraints on Shuttle payloads and the fact that it was no longer possible to launch Shuttle into a polar orbit were serious complications. A modified program was therefore devised with an increased emphasis on *incremental prototyping*. Although the main design principles of the Scienc e Instrument Assembly were

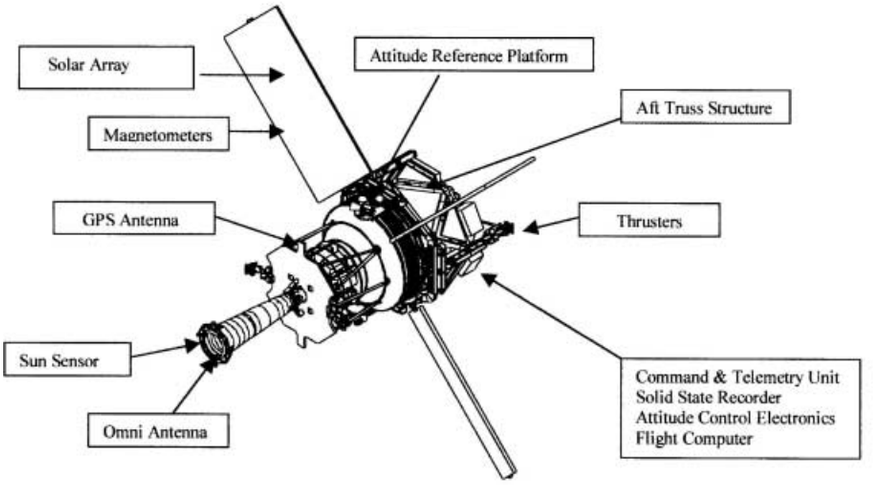


Fig. 6. Gravity Probe B Spacecraft.

well established by 1985, the engineering challenge was formidable. Equally challenging was the design of the cryogenic vacuum 'probe' containing the SIA which is complex and has many features unique to GP-B[2]. Thus the neck tube of the probe must contain cooled radiation windows to intercept infrared radiation that would otherwise be a severe heat load on the SIA and dewar. On the other hand, it also must serve as a high speed pumping line to exhaust gas at low pressure from differential pumping of the gyro spin system. It was by no means obvious in advance how these two requirements could be reconciled.

Nor could the sophistication of the GP-B dewar be underestimated. Though there was important engineering and flight heritage for superfluid helium dewars from the IRAS and COBE programs, some of GP-B's most critical features were unprecedented in space. A major issue was the ultralow magnetic field requirement, which demands having a permanently cold superconducting lead bag in the inner well of the dewar and therefore inserting the warm probe into the cold dewar. This is an intricate process in which it is necessary to mount a temporary airlock on top of the dewar to prevent solid air from condensing into the inner well. Once in place, the probe has to be locked down thermally and mechanically at appropriate poins in the dewar. Incremental prototyping has helped immensely in arriving at a sound design and establishing correct insertion procedures.

Incremental prototyping attacks the hardest engineering problems first by using full–size flight prototypical designs with flight interfaces, but with reduced

[2] Following standard cryogenic engineering practice, reference is made in this paper to three successive cryogenic 'probes' to be inserted into the dewars, Probe-A, Probe-B, Probe-C; this terminology is not to be confused with the use of the word 'probe' in Gravity Prob e B.

functionality and without costly flight build standards. Specifically, three full–sized payload system tests have been conducted with the three 'probes' of increasing fidelity just referred to (A, B, C), where Probe-A is a pre–prototype; Probe-B (designated as the backup flight probe) is essentially of flight–quality; Probe-C is the actual flight probe. These have been evaluated in the following successive ground–based tests:

A. First Integrated Systems Test (FIST) 1990–1991. FIST consisted of a demonstration test of Probe-A in a semi–prototypical Engineering Development Dewar (EDD) with science–mission–like interfaces. Probe-A incorporated a prototypical quartz block but no telescope and only two gyroscopes. It operates in a magnetic field of about 10^{-6} T obtained with conventional shields. The use of only two gyroscopes made additional electrical and other connections available for instrumentation to investigat e the thermal and vacuum performance of the probe.

B. Ground Test Unit GTU-0 1991–1994. GTU-0 was also performed with Probe-A and the EDD but with an ultra–low magnetic field shield (10^{-11} T) installed in the dewar. Cold insertion of the probe into the dewar by means of the airlock resulted in low trapped flux in the gyroscopes.

C. Ground Test Unit GTU-1 1994–1995. GTU-1 was the first use of Probe-B. It included: (a) operation with all four gyroscopes and a mass model of the telescope; (b) achieving a < 0.1 nT trapped dipole in the gyro rotor; (c) test of the fully coupled dc SQUID readout; (d) electromagnetic interference (EMI) tests; (e) operation at ultra high vacuum with a sintered titanium cryopump. Also performed at room temperature was a successful protoqual shake test of the probe with the te lescope mass model and caged gyroscopes, fully confirming the reliability of the design.

D. Ground Test Unit GTU-2 1997. GTU-2 was performed again with Probe-B but mounted in the Science Mission Dewar (SMD). This provided a verification of a complete payload integration and test procedures using the SMD together with a full functional test of all subsystems (SIA, Probe and Dewar) before and after an integrated system shake test at protoqual levels.

E. First Complete Payload Test August–December 1999. Results of this test are discussed in Section 5.

4 Risk Mitigation and Verification Matrix

Developing a flight mission compels the physicist to address two interrelated sets of problems very different from traditional laboratory experience. An apparatus has to be made that will survive launch and function in the unfamiliar environment of Earth orbit or outer space. It also has to work even though not all aspects of its performance can be verified experimentally on the ground. Addressing these two issues requires, in the first place, systematic thought about principles of risk mitigation and the c oncept of a verification matrix. Secondly it requires a well–ordered test plan to assure that the risks and uncertainties have indeed been properly accounted for.

From an engineering standpoint 'risk mitigation' means first that the experiment has to be designed conservatively to withstand the vibration of launch and the temperature conditions and charged particle environment encountered on orbit. In each case it is necessary to think not only of the design but of the testing necessary to verify that the completed payload meets the engineering and scientific requirements. Some of the following observations should be regarded as general space practice; others are spec ialized to Gravity Probe B.

A. Surviving Launch: The Tradeoff Between Strength and Weight. The recipe for surviving the hostile launch environment might seem simple – extreme over–design so that the apparatus cannot possibly break. The flaw in this simple nostrum is that there is another constraint, weight. It seems almost a law of human nature that during the development of any spacecraft its weight will grow beyond the capability of the launch vehicle. Always at some point a strenuous program of weight reduction becomes necessary. In Gravity Probe B the largest contributing factor was the hea vy outer shell of the dewar, especially the two end–domes. After investigating several options, including the use of special low–density alloys we chose the well established but sophisticated method in which the inner surface is relieved with suitably–patterned rib structures (an 'isogrid' design). The final effect was a reduction of dewar weight of about 18 percent. Another frequently encountered weight penalty, surprising to the laboratory physicist, is from the electrical harnesses. Long experience ha s shown that these are always underestimated. In Gravity Probe B we were fortunate to begin harness layout early and were thus subject to less surprise than some programs have been.

More generally, the physicist engaged in a space program is impressed to observe that through all the development a constant weekly watch has to be kept on weight and power budgets.

B. Temperatures of Electronics Boxes: Two Issues and Their Resolution. The conditions under which electronics boxes have to function in space are very different from those typical of ground–based laboratories. In Gravity Probe B the surface temperature of the spacecraft is around 220 K ($-50°$C). Also as it rolls around the line of sight to the star the electronic boxes will heat and cool in the presence of the Sun. This is a critical point in the design because it limits the accuracy of our earlier statement that rolling the spacecraft eliminates readout errors and gyro suspension torques. Obviously, if the temperatures of key components vary at roll–rate the statement is no longer quite true. In building the electronics, therefore, thermal risks of two distinct kinds must be mitigated. They must have the engineering robustness to work under extreme conditions. They must also have the scientific refinement to maintain temperatures within a very narrow and known range over each roll cycle. Both constraints must be met in a situation where the electronics components are functioning in hard vacuum: any heat has to be removed by radiation or solid conduction, there is no atmospheric convection to aid us. These conditions require a ground verification program

of very considerable sophistication and have a critical influence on spacecraft design. Key electronics boxes are enclosed in a Forward Electronics Enclosure (FEE) at the front end of the spacecraft with passive temperature control; within the individual boxes active temperature control is applied to certain particularly sensitive elements. As part of the on–orbit verification, temperature data is telemetered to ground.

C. The Charged Particle Environment. The presence of the Van Allen belts at or a little above the 640 km Gravity Probe B orbit altitude places requirements on the GP-B electronics different from anything encountered on Earth. Particularly critical are the effects, well known to space engineers but less familiar to most physicists, of the South Atlantic Anomaly (SAA). The old discovery by Gauss in the 1840s that the Earth's magnetic field can be represented by a tilted dipole displaced toward North modifies the lives of space scientists. Over the South Atlantic there is a region where for a period of 5 to 15 minutes the spacecraft e ncounters a much higher density of charged particles than elsewhere in its orbit. Since the Earth is rotating the SAA does not affect GP-B in every orbit; nevertheless there are several each day in which the charged particle environment is hostile. An obvious question is what effect this may have on the gyro and telescope readouts. Tests of the SQUID magnetometers in a special facility at the Paul Scherrer Institute in Villingen, Switzerland designed to provide in a few hours a particle dosage equivalent to a year's passage through the SAA were pleasantly reassuring. The SQUIDS continued to function with only a small and acceptable number of measurable 'flux–jumps'. On the other hand, with the telescope detectors some periods of interruption will certainly take place. Throughout the electronics 'rad–hardened' components must be used in all critical areas, and the design must be made 'robust' so that any temporary electronics failure will not have catastrophic effects. Most critical of all is the digital gyr o suspension system. To guard against upsets of the computers controlling the suspension it is designed with two parallel control computers, a backup analog suspension system, and an arbiter to monitor performance and determine which system shall be used.

Such are the complexities of gyro suspension. In addition, it is necessary to prevent excessive charging of the rotor itself. For this purpose the charge on the rotor is measured by injecting a signal into the suspension system and the information so obtained is used to control an ultraviolet discharge mechanism based on the photoelectric effect.

D. Electromagnetic Interference (EMI) and Compatibility (EMC). The issue of electromagnetic interference is familiar to physicists engaged in ground–based experiments. Nevertheless space in general, and Gravity Probe B in particular, have special problems. The compactness of the spacecraft means that unusual care is needed to insure that signals from the telemetry of the science data to Earth do not interfere with the payload electronics. There is also the bizarre–seeming fact that in low–Earth orbit signals from television stations on the ground ca n be a formidable problem. In Gravity Probe B a special issue is that the gyroscope readout based on the London

> moment depends on accurately measuring very minute changes in magnetic fields and therefore extremes of magnetic shielding. In addition, each gyroscope must be shielded from the next and the readout of each must be shielding from its own gyro suspension signals. Without superconducting shielding and frequency separation between the suspension and readout signals, these requirements could probably never have been met. The fact that they were, as demonstrated in the tests of Probe-C discussed below, is one of the most reassuring results of the GP-B development. It is a success critically dependent on incremental prototyping.

Mitigating risk is a total process requiring judgments of many diverse kinds. Risks need to be identified but they need also to be weighed intellectually and sometimes even literally. If a particular mitigation adds to the weight of the spacecraft, it may in that form be impracticable. Then there is the complicated question of complexity. Does the mitigation achieved by introducing a redundancy so increase the complexity that in the end it increases rather than reduces risk? All such issues have to be addressed in a context of cost, recognizing that in space programs cost penalties are of two kinds. There is the immediate cost of designing, fabricating, and testing some previously unplanned part or system but there is also the secondary, often far greater, problem of the 'marching army'. Any schedule delay increases cost far more than one at first imagines because it entails keeping a large spacecraft team together longer than originally planned.

Mitigation involves assessment and planning; it also involves verification. Reflection on that fact leads to the important concept of a 'verification matrix'. In any actual spacecraft there are many systems. Some have been flown many times. Others are new but very similar to systems that have been flown before. Others again are new but capable of being tested with greater or lesser completeness in the laboratory. Finally, others may not be capable of being tested at all before launch and can only be verifie d through calculation. The idea of a verification matrix is to lay out in a orderly way all the systems, their requirements and the nature of the verifications to be performed: inspection, similarity, analysis, test, and once that is done to establish, and where necessary enhance, the realism of the verification process.

As an example consider Newtonian disturbances of the Gravity Probe B gyroscopes. Almost by definition gyro performance cannot be fully verified until GP-B is on orbit: if it could, one would be seeking to perform the experiment in a ground–based laboratory. That, however, does not preclude a rationally conceived and experimentally well–founded verification process. During the long history of Gravity Probe B an exhaustive error analysis has been conducted identifying no fewer than 120 different known class ical gyro torques, most of them exceedingly small. The verification questions are whether all these have been calculated correctly and whether any significant torques might have been left out.

A broader view reveals that the torques can be classified in a number of ways which allow plausible comparisons of their relative values and aid in setting reasonable limits on unknown terms. One distinction is between support–dependent

torques (which decrease in space) and support–independent ones. Another is between torques due to pressures on the surface of the gyro rotor and those arising from momentum transfer through the volume of the material. A third is between torques varying linearly with time a nd those like Brownian motion, photon bombardment, and the impact of cosmic rays on the rotor which tend to cause 'random–walk' effects evolving as the square root of time. Simple as these distinctions are, they throw a flood of light on the expected performance of the experiment.

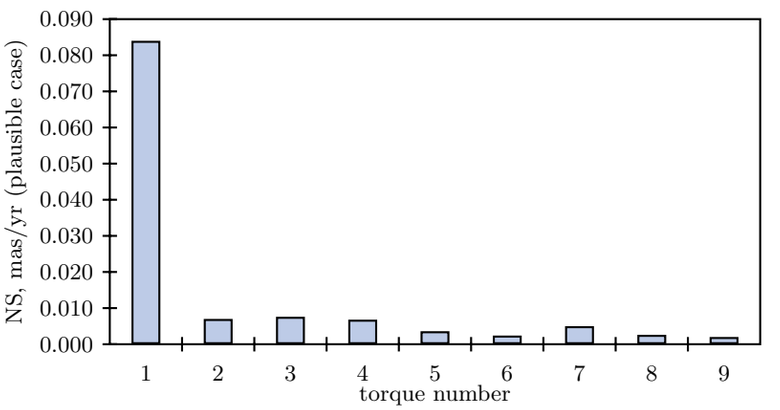
Thus, the two obvious classes of torque involving surface pressure are those from gyro suspension and from residual gas in the housing. The 'electrical pressure' has to be such as to apply to the rotor accelerations of order 10^{-8} g. At the planned operating gas pressure (10^{-12} torr) the corresponding acceleration under the extreme assumption that all the pressure acts on one side of the rotor is 10^{-13} g. Hence, it is inherently plausible – and true – that gas–*pressure* torques are neg ligible and all our attention should be focused on electrical pressures, i.e. on the suspension torques.

In fact, the suspension torques dominate the gyro performance. The analysis by Keiser and Silbergleit demonstrates that these electrical suspension torques are governed by 15 distinct coefficients. A laboratory verification, therefore, should evaluate each one and establish their consistency with other measured parameters, e.g. out–of–roundness of the gyro rotor, rotor centering, spin–axis alignment, etc. Measurements on a flight type gyroscope suspended against 1g, performed at Stanford by Dr. Yoshimi Ohshima, have successfully determined all 15 coefficients.

With the coefficients known, classical electrical formulae based on Coulomb's law determine the resultant gyro drifts to be expected on orbit. This is a powerful verification. Fig.7, based on the analysis of G.M. Keiser and Silbergleit, gives the top eight non–relativistic terms contributing to the precession of the gyroscope in the plane of the frame–dragging effect.

Of these no fewer than seven are support dependent torques and of the seven, four originate in imperfect roll–averaging due to heating and cooling of the electronics as the spacecraft rolls in the presence of the Sun. We are thus led to a further verification. By determining in the laboratory the temperature coefficients of key electronics systems, one can set criteria on their allowed temperature variations in space. This, in turn, leads to an on–orbit verification through measuring and transmitting to E arth a record of the temperatures. It is even possible to use this record as a mode of on–orbit calibration allowing errors to be backed out by calculation after the fact.

Although the argument from pressure eliminates the most obvious gas torque it does not guarantee that the residual gas has no effect. Molecules continually moving back and forth between the rotor and housing cause the gyroscope to slow down. If there is any asymmetry in this process it will result in a gyro drift due to 'differential damping'. At 10^{-12} torr the characteristic spin–down time of the gyroscope is 320,000 years. For GP-B the rolling of the spacecraft around the line of sight to the star h as a powerful symmetrizing effect and the



A. Support–dependent torques from temperature variations of suspension electronics at spacecraft roll–rate (1,5,7,8)

1. Roll–variation of preload voltage difference $(h_a - h_b)$
5. Fixed sums of preloads $(h_a + h_b, h_c)$
7. Roll–variation of sensing bridge voltages
8. Roll–variation of preload sums $(h_a + h_b, h_c)$ plus fixed miscentering of rotor

Note: all above are conservative upper limits, awaiting calibration of flight Electronics. Also, temperature variations of electronics will be measured On orbit.

B. Support–dependent torques from non–temperature–dependent effects (3,4,6)

3. Gyro spin axis misaligned with spacecraft roll axis
4. Constant transverse acceleration on gyro parallel to Earth's gravity gradient. Acting through misalignment between orbit plane and line of sight to star.
6. Sagging of gyroscope at twice orbital period due to Earth's gravity gradient

C. Support–independent torque (2)

2. London moment coupling to local magnetic shield–gyro axis misaligned with line of sight to star

Fig. 7. The Top Eight Gyro Drift Torques.

computed value of this differential damping torque is below 0.0001 mas/year. While this in itself is beyond ground verification, there is an important ground verification related to it, namely, to measure the characteristic spin down time of the gyroscope under laboratory conditions. The measurements confirm the accuracy (at low temperature) of the gas drag formula. They also provide a crucial check on whether any debris are present in the housing.

5 Probe-C Assembly March–August 1999

Assembling the SIA and the Flight Probe has been a lengthy process, lasting more than 6 months. Individual systems and subsystems had to be fully qualified before assembly, all under the discipline of record keeping and Quality Assurance unfamiliar to most physicists. In total Gravity Probe B had 32 distinct mechanical and electrical systems that required verification through the Probe-C tests. It is only through dealing with an apparatus of this sophistication that the physicist comes to realize the benefi ts of the rigorous QA procedures developed through hard–won experience in the aerospace business.

Among the systems tests that were performed before the assembly of the SIA and Probe-C began some of the more important were as follows:

A. Gyro Qualification. To date, more than 110,000 hours of gyro testing have been conducted in a variety of cryogenic and room temperature test facilities. The final qualification of the individual gyroscopes to be inserted in Probe-C was carried out in a special Gyro Acceptance Facility designed to allow the verification that each gyroscope meets five flight requirements: (1) asymptotic spin speed > 100 Hz; (2) spin–down rate after evacuation < -2 mHz/hr (verifying gyro cleanliness); (3) low trapped magnetic field $< 3 \times 10^{-10}$ T; (4) discharge rate by ultraviolet system > 10 fA/μW (in combination with the charge measurement system, this will allow the mean rotor voltage to be kept $< \pm 10$ mV); (5) SQUID–dc coupling $< 1\%$ decay of 200 flux quanta after 15 min (verifying integrity of the readout loop).

B. Qualification of SQUID Assemblies for Gyro Readout. Each SQUID assembly consists of a SQUID microchip sensor mounted on a sapphire circuit board with a superconducting thin–film input transformer, electronic components, and connectors, filter components for rejection of electromagnetic interference, all contained within a superconducting box made of "reactor grade" niobium and sealed with an indium alloy gasket. The SQUID chip was initially selected for low noise by cryogenic testing. It was then mounted into the SQUID assembly, which was subjected to cryogenic qualification testing to determine sensor noise, linearity, transfer function, temperature sensitivity, bias current level, stability, the superconducting persistence of the input circuit, and magnetic and EMI attenuation. Only assemblies that meet all qualification criteria are accepted for use in Probe-C, and only SQUID sensors that have remained stable through multiple thermal cycles to cryogenic temperatures are incorporated in these assemblies.

C. Magnetic Qualification of Probe Components. Control of the magnetic properties of materials and assemblies used in Probe-C is essential for the London moment readout. To meet the necessary stringent requirements, a Magnetic Control Plan was set up dividing the Probe into magnetic zones based on distance from the gyroscope rotors. Within each zone, limits were placed on the residual magnetic moment and magnetic susceptibility of each component, and on the maximum level of stray ac magnetic field. Samples of the actual material stocks, from which probe components had been fabricated, were characterized magnetically at temperatures down to 2 K using a

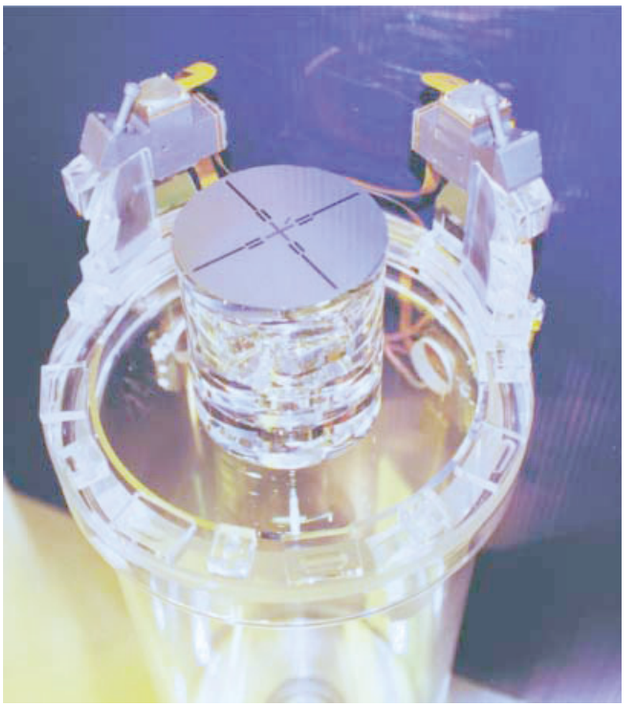


Fig. 8. GP-B Telescope: The cylindrical structure in the center contains the image dividers. The mirror on top of it aids in telescope testings. The two DPA's, each with redundant detectors, are mounted in the two structures set at 90 degrees of each other on the edge of the telescope.

SQUID magnetometer. Once fabricated, the assemblies were further tested for magnetic cleanliness with a SQUID gradiometer to verify that machining and assembly operations did not introduce unacceptable magnetic contamination.

D. Telescope Qualification. Fig. 8 illustrates the front end of the completed Flight Telescope, fabricated entirely of fused quartz and held together by a silicate bonding technique invented by Dr. Jason Gwo. It is a folded Cassegrainian system, of 3.75 m focal length and 0.15 m aperture, and a physical length of 0.32 m. Two star images are formed, one for each readout axis; the images are divided at two orthogonal roof prisms and the intensities of the divided beams are compared in redundant D etector Package Assemblies (DPA), operating at a temperature of about 70 K, mounted on the front end of the Telescope.

The Flight Telescope was the third in a series whose evolution is another illustration of incremental prototyping. In fact, incremental prototyping appears equally in testing the telescope testing. No fewer than three artificial

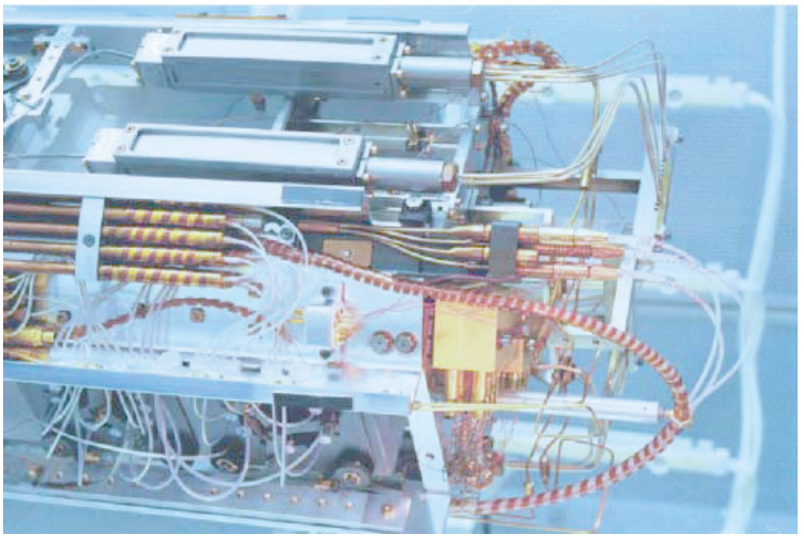


Fig. 9. SIA in Advanced Stage of Assembly: Note the suspension cables and two of the four SQUID boxes mounted on the upper side of the 'bird–cage'.

stars have been constructed to evaluate its performance. The one used for qualification was the second, on which a vibration–isolated dewar vessel was mounted to allow testing both at room temperature and low temperatures. Tests of the DPAs were performed separately to determine no ise performance prior to their installation on the telescope. Linearity measurements with the artificial star demonstrated that the telescope would remain linear to < 3 mas over a range of ± 20 mas, meeting the requirements of spacecraft pointing control.

With the individual systems for the Probe and SIA all qualified, it became possible to proceed to final assembly conducted in a class 10 clean room at Stanford. The Quartz Block/Telescope structure was inserted into an inner framework (the 'bird–cage') in the Probe, followed by the exacting process of inserting and aligning the gyroscopes into the Quartz Block with their centers within 50 μm of a common line. (Hence, when the spacecraft has been mass–balanced on orbit, the gyroscopes will be w ithin that 50 μm radial distance from the roll axis, minimizing the centrifugal acceleration to which they are subjected). Fig.9 illustrates the SIA in an advanced stage of assembly showing the suspension cables, SQUID assemblies, gyro caging line, and fiber optic cable for uv charge control. The final tests performed before inserting the Probe in the dewar were a room temperature spin test for all four gyroscopes, measuring the spin–down rate after evacuation and thus verifying once more that the gyr os were free from particle contamination, and a telescope field–of–view test.

Fig. 10. Probe-C in Airlock Ready for Placement on Science Mission Dewar.

6 Probe-C Testing in Science Mission Dewar August–December 1999

Because the ultra low magnetic field shield is maintained permanently in the inner well of the SMD, the process of inserting the Probe is, as remarked earlier, an unusual one. Whereas in almost all cryogenic systems the probe is inserted before the dewar has cooled down, here a method had to be developed for inserting a warm probe into a cold dewar. In doing so, it is essential to avoid condensing any solid air into the dewar well. The method (devised by B. Cabrera) is to make use of a cylindrical air lock within which the probe is fitted with a tight piston seal. By mounting the probe and air lock on top of the dewar it is then possible to pump out any air and backfill the air lock with helium gas prior to insertion. Insertion has to be performed sufficiently slowly not to generate excessive heat and warm up the superconducting lead shield. To aid in this process, the dewar well is filled temporarily with liquid helium prior to insertion. Fig.10 shows Probe-C mounted in the airlock ready for placement on the dewar in its vertical position.

Probe-C was inserted into the Science Mission Dewar on August 24, 1999 and testing was continued until December 15. Fig.11 shows the assembled probe and dewar under test at Stanford in a horizontal position.

At an early stage it became clear that there was a significant, quite unexpected anomaly in the thermal performance of Probe-C, which would require a recycling to room temperature. However, rather than immediately warming up we decided to 'retire risk' by completing the entire range of planned qualification tests on every one of the 32 major subassemblies contained in Probe C and the SMD. 28 out of 32 met or surpassed the stated requirements; 3 more

Fig. 11. Probe-C under test in the SMD in a Horizontal Position: The dewar can be tilted through a range of angles to perform different tests.

discrepancies (1 significant, 2 minor) were uncovered in a ddition to the thermal anomaly named above. These discrepancies were of a character that would not have prevented launching a successful mission but the necessity of warming up has given the opportunity to address them all.

A. SIA. Tests of the SIA covered 7 major subassemblies and 1 systems test as follows:

- *Telescope:* testing with artificial star # 3 (designed to mount on the SMD) verified performance under final operating conditions. The optical transmissivity and the strehl ratio (which measures the quality of the final image) surpassed the requirements. All 8 detectors in the two DPAs functioned correctly. The noise performance for the selected guide star HR 8703 meets the 10 mas/$\sqrt{\text{Hz}}$ pointing requirement.
- *SQUIDs:* all 4 SQUIDs function correctly, with noise performance meeting, and in 2 of the 4 surpassing, the requirement by a factor of 4. In separate tests of electromagnetic compatibility (EMC) with the Gyro Suspension System the shielding was shown to fully met stated requirements.
- *Gyro levitation and spin–down:* all 4 gyroscopes met the levitation requirements and the stated spin–down requirements, showing that at low

temperature as well as in the earlier room–temperature tests the gyros were free from particle contamination[3].

- *Gyro caging:* tests of the caging mechanism met requirements both at room temperature and low temperature.
- *Gyro uv charge control system:* all 8 systems (2 per gyro for redundancy) functioned correctly yielding a photocurrent $> 100\text{fA}/\mu\text{W}$, surpassing the requirement by more than a factor of 3.
- *Gyro readout coupling:* for 3 of the 4 gyros (## 1, 2, 3) the coupling of the SQUIDs to the readout loop met the superconducting requirement; for # 4 it showed resistance $> 1,000\ \Omega$ which is unacceptable.
- *SIA sensors and wiring:* all of the thermometric and other sensors on the SIA functioned correctly, and all of the wiring met specifications; there was one electrical short on the gyro # 4 suspension shields, which, however, had no practical impact.
- *Trapped flux in gyro rotors:* As noted in C below, the ambient field in the lead bag was $< 2 \times 10^{-10}$ T. Trapped fields in 3 of the 4 gyroscopes were below the 3×10^{-10} T requirement but in gyro # 1 (the one nearest the top of the bag) field levels were anomalously high (10.6×10^{-10} T).

B. Probe-C. The 3 principal system entities: (1) plumbing; (2) sensors and wiring; (3) radiation windows, met their requirements, with the exception of minor leaks in the caging lines and one valve. The practical impact of these minor leaks would have been negligible but since the probe has been warmed up they are being repaired. Transmission through the windows exceeded 80%. The vacuum integrity of the probe fully met its requirement. The failure in thermal performance is discussed below.

C. Dewar. The 4 principal system entities: (1) plumbing; (2) sensors and wiring; (3) low–temperature valves; (4) ultralow magnetic field shield met all their requirements. The magnetic field level in the lead shield was $< 2 \times 10^{-10}$ T. The vacuum integrity and thermal performance also met all requirements.

D. Interfaces. Interface requirements had to be, and were, met for 3 systems: (1) SIA–to–Probe; (2) Probe–to–Dewar; (3) the Belleville–washer preload system to maintain thermal contact between Probe-C and the SMD.

The foregoing tests were highly successful in retiring risk. At the conclusion of this phase of payload testing 90% of the required systems verifications had been completed.

7 The Four Discrepancies and Their Resolution August 1999–June 2000

Considerable work was done to characterize the four discrepancies while Probe-C was still in the SMD. After its removal on December 15, 1999 new diagnostic tests

[3] At a late stage of testing, some contamination from frozen gas occurred in two of the gyros for reasons that are well understood. The gas disappeared on warming to room temperature, leaving the gyroscopes once more free of contamination. Pr ocedures are in place for preventing this contamination happening again.

became possible which further clarified the problems and various experiments and rehearsal tests were performed to validate the redesign. Rework commenced on March 27 and was completed on June 2, 2000. Reinsertion of Probe-C into the SMD will take place before October 18, 2000.

A. Probe-C Thermal Anomaly. To understand the thermal anomaly one needs to grasp the principles governing the design of long hold–time helium dewars.

Liquid helium, because of its quantum mechanical zero point energy, has very low latent heat. Far more cooling–power is available from the warming of the boil–off gas to room temperature than from the latent heat. In designing helium dewars one makes use of this fact by intercepting much of the incoming heat in 'vapor–cooled shields'. With 3 to 4 suitably spaced shields embedded in the multilayer insulation, it is possible to recover about 35% of the available gaseous refrigeration and increase the dewa r hold–time by a factor of about 25.

In Gravity Probe B this vapor–refrigeration must be applied not only to the dewar but also to intercept heat radiated and conducted down the neck–tube of the probe. Accordingly, as Fig.12 shows, radiation windows at 3 locations in the neck are connected to copper rings epoxied to its inner surface. Matching copper rings on the outside of the neck connect these thermally to corresponding cooled rings on the dewar neck. On orbit, where the dewar skin temperature is 220 K, the nominal temperatures of the 3 w indows are 33 K, 77 K, and 132 K; a fourth ring (HEX # 4) without a window operates at 157 K. On Earth with a skin temperature of 293 K, the window temperatures are usually slightly higher; however, there is also a special operating condition used mainly on the launch pad, in which the lowest vapor–cooled shield is artificially cooled by means of a 'guard tank' full of 'normal' (i.e. non–superfluid helium) and therefore to operates at or near 4.2 K. In the experiments to be discussed now, this was in fac t the condition and window # 1 should have been at around 6 K.

This sophisticated design was prototyped in Probes A and B, and successfully demonstrated in both the EDD and the Science Mission Dewar. When Probe-C was cooled down, however, very serious temperature anomalies were observed. Instead of being at 6 K, as it should have been with the guard–tank cold, window # 1 ran at 86.5 K! The temperatures of windows # 2 and # 3 and of Hex # 4 were also far too high. These drastic discrepancies strongly suggested that the windows were receiving effectively no cooling from the boil–off gas (in other words that there was a thermal disconnect somewhere between the window and the vapor–cooled shields. A variety of tests pointed to a thermal disconnect internal to Probe-C, probably between the copper rings and the fiberglass–epoxy neck tube. Conclusive evidence came from tests in which helium exchange gas was introduced into the probe at various pressures with the system vertical to prevent convection. At 0.31 torr the temperatures are reduced dramatically: the exchange g as provides an excellent conductive path between the inner ring and the neck tube.

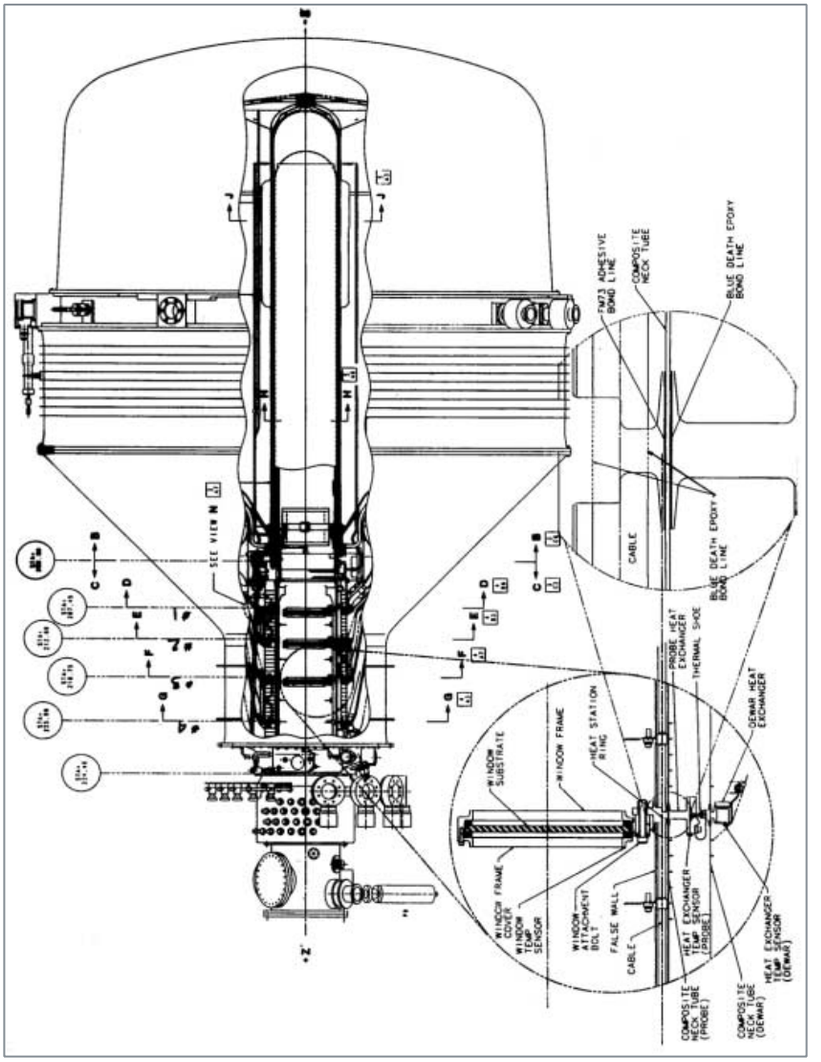


Fig. 12. Schematic of Probe-C Repair: The three radiation windows can be seen within the neck in the main drawing. The left inset shows details of the window frame; the right inset is an enlarged section of the neck–wall and rings.

Location	Expected (with guard–tank cold)	Observed (high vacuum)	Observed (0.31 torr)
Window # 1	6 K	86.5 K	26 K
Window # 2	22.5 K	122.2 K	31.5 K
Window # 3	72.3 K	171.3 K	77.3 K
Hex # 4	151.4 K	196.3 K	166.4 K

The conclusion has to be that the epoxying that had worked so well for Probes A and B went wrong in Probe C. The very detailed written and photographic QA record of the construction showed no discernible cause of failure; nevertheless, failure there was, with three very serious implications. First, the loss of thermal conductivity reduces the dewar hold–time on orbit from 19 months to possibly as low as 10 months. Second, just as serious, it reduces the ground hold–time in such a way as to make operations on the launch pad almost impossibly difficult. Third and gravest of all, the break between heat stations and neck tube makes the probe vulnerable to mechanical failure under the vibration of launch.

A failure of this kind puts a space program into a new situation. To disassemble Probe-C and remake the epoxy bonds would take over a year and still not guarantee success. Instead, the method GP-B has adopted is to short–circuit the problem by a redesign which establishes direct thermal links between the outer and inner copper rings. At each heat station four holes are drilled through the outer ring, the neck tube, and the inner ring ending as a 'blind hole' that stops just short of penetrating through to the far side of the inner ring. Thermal contact and mechanical integrity are established by means of copper pins inserted into each hole with a tight 'press–fit'. For technical reasons two of the pins are inserted horizontally and two inclined at 51 degrees. Fig.13 illustrates a cross–section of the pin and 'blind hole' for one of the 51–degree pins. Vacuum integrity, which has of course been lost by drilling through the neck tube, is restored by two stages of epoxy seal. A low viscosity epoxy is injecte d into the hole around the pin and over this, following a standard cryogenic practice, a 'doubler' in the form either of a flat copper sheet or copper plug is sealed over each pin with an extremely robust thermally matched epoxy rejoicing in the name "Blue Death".

To assure success, rehearsals were performed on two prototype assemblies with two alternative pinning procedures. Critical to the pinning solution is the complex drilling process which has to be undertaken to penetrate first a copper ring, then the composite neck tube, then the second ring, all with perfect alignment and without penetrating beyond the blind end of the hole. Since the pin has to be a 'press–fit', alignment and finish of the holes are critical: it has been found necessary to drill each hole in six stages, with three different kinds of drill. For the horizontal holes two of the stages required carbide drills which are notoriously brittle; great care had to be applied to avoid breaking off a drill partway through a hole. After many trials the procedure was perfected and demonstrated to meet the three essential re-

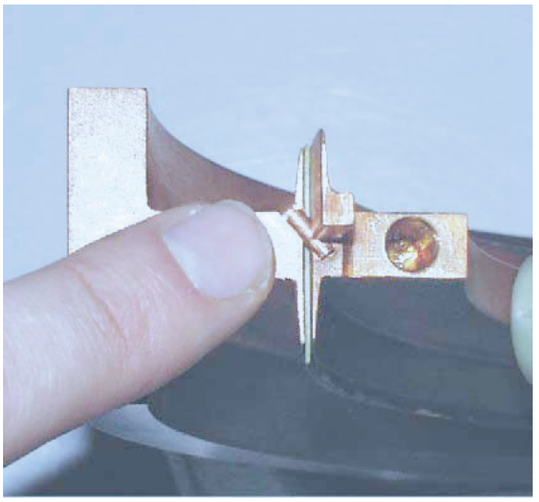


Fig. 13. Cross Section of Model of Probe Necktube with 510 Copper Pin. The section shows the pin in position with the blind hole in the inner copper ring. Not shown is the 'doubler' which completes the vacuum seal on the outside of the outer ring.

quirements, thermal, mechanical, and vacuum. Comparative experiments on Probes B and C provided an unequivocal room–temperature diagnostic to distinguish acceptable and unacceptable the rmal performance at low temperature. Mechanical and vacuum integrity were validated as follows: Each prototype assembly was vacuum tested; subjected to a vibration test at flight qualification levels; vacuum tested again; twice cycled to low temperatures; vacuum tested for a third time. Both pinning techniques met all requirements. Finally, on March 23, 2000 Probe C Redesign Readiness Review was held and approval given to proceed with the rework using the particular method discussed above.

In applying to Probe-C the experience gained during the rehearsals, it was essential to take into consideration the issue of cleanliness. Debris or turnings from the drilling process must, at all costs, be prevented from contaminating the gyroscopes or soiling the surfaces of the telescope and radiation windows. The drilling was performed in an ultra high quality (class 10) clean room at Stanford, with the probe in an inverted position and at a slight overpressure to drive any contamination outwards. It was accompanied by a continuous vacuum cleaning around the hole. Drilling proceeded from the outside in, removing one radiation window after another to gain access for measurement of the position and alignment of the inner rings.

In fact, the entire operation went remarkably smoothly. Pinning was completed on May 15, 2000. After cleaning and reinstallation of the windows the following tests were performed by June 2: (1) a static load test on window frame # 3 at flight qualification levels; (2) the room temperature thermal

performance test mentioned above; (3) a vacuum integrity test of the entire probe. Prior to the installation of the pins the thermal conductance between inner and outer rings was 0.028 W/K, after pinning it was 0. 42 W/K. This factor of 15 in improvement was in excellent agreement with the data obtained with the earlier validation units. The computed helium hold–time on orbit (including a 30% allowance for conservatism) is 17 months. Vacuum performance was excellent: there was no detectable leak even on the most sensitive 3×10^{-10} standard cm^3/sec scale of the leak detector, a margin of at least 4 orders of magnitude on the GP-B requirement.

As of mid–June 2000 Probe C is being prepared for room–temperature spin tests of the gyroscopes to check whether cleanliness has been properly maintained.

B. Gyro # 4 Readout Ring Discrepancy. Beside offering a good opportunity for scientific detective work, the readout failure for Gyro # 4 nicely illustrates the range of issues that need to be weighed in executing a space program.

In the initial planning of Gravity Probe B the decision to have four gyroscopes was in some degree arbitrary (a tradeoff between redundancy and complexity). Obviously, there ought to be more than one, but since each gyroscope will measure both relativity effects a case can be made for flying with only three. If there had been no other reason for removing the probe from the dewar that would almost certainly have been the decision. However, given the fact that the probe did have to be removed the option of re pair becomes worth careful consideration.

The tests demonstrated that the entire SQUID readout unit was functioning perfectly; the fault lay in the failure of a superconducting bond on the housing itself. The location of the high–resistance connection in the gyro–to–SQUID input circuit was determined by measuring the frequency dependence of the SQUID response to high frequency inductively coupled ac signals. Frequencies above 100 kHz are outside the SQUID feedback circuit bandwidth, so these measurements were made open loop (i.e., without feedba ck to the SQUID). To obtain consistent measurements, the SQUID was dc biased to the most linear portion of its characteristic curve. Input signals were limited to no more than one tenth of a SQUID flux quantum in order to obtain a reasonably linear output. The resulting data of SQUID output amplitude versus frequency were compared to the predictions obtained from SPICE circuit simulation results based on a circuit model with a resistive contact at the gyroscope and a second model with the resistive contact in the SQUID assembly. The experimental data fit the former model excellently and the latter model not at all.

It is worth emphasizing that the failure of the superconducting bond on gyro # 4 was not entirely unexpected. During earlier qualifications of gyroscopes it had been found that the original design was somewhat unreliable and occasionally failed on initial cool–down. Accordingly, a new more reliable design of bond was developed, which has shown no failures, and this was used for gyros # # 1, 2, and 3. For gyro # 4 a compromise was made. A

gyroscope was available which was so excellent in other respects that we decided to use it despite having a bond of the original design, the argument being that in the past, once such bonds were qualified, they always withstood repeated temperature cycling.

Repairing the bond simply means replacing the gyroscope. Although the SIA is designed to allow such a substitution, the process has never actually been tried and does entail risk. The cables and connections to the old gyroscope have to be severed and unstaked without damaging or contaminating the assembly; the gyro has to be removed and replaced; and numerous tests have to be performed prior to resealing the probe. Before finalizing the decision, a rework team has been set up to rehearse the process in a ba ckup Quartz Block Assembly containing two gyroscopes left over from earlier prototyping activities. The results are encouraging; a decision will be taken in June 2000.

C. High Trapped Flux in Gyro # 1. The 10.6×10^{-10} T trapped field in gyro # 1 was rather a surprise because in Probe-B, where less attention had been paid to magnetic cleanliness, there had been little difficulty in obtaining 3×10^{-10} T trapped fields in all four gyros. Three possible explanations suggest themselves: (1) a change in magnetic field level in the lead bag; (2) magnetic contamination in the probe; (3) a thermal effect connected with the Probe-C anomaly.

The first conjecture was that there had been a field change in the lead bag. The reason for suspecting this was that during cryogenic operation a 'thermal spike' had occurred which could have heated the top end of the bag above its superconducting transition temperature. However, after removing Probe-C from the dewar and carefully remeasuring the field at all levels we found the field completely unchanged. Magnetic contamination remains a possibility, though in view of the extreme care taken in material screening it seems unlikely. The most probable explanation is a temperature effect.

The hypothesized source is the Thomson effect. As is well–known, Thomson in the 1850s showed from thermodynamical reasoning that when any metal is subjected to a temperature gradient, a voltage difference will be established across it. What is less well–known is Thomson's second discovery (1895) that if the metal is anisotropic circulating currents can be set up generating magnetic fields. At low temperatures, this effect becomes much larger as was discovered by a GP-B graduate student, P.M. Selzer, in 19 74. The argument, therefore, is that the thermal anomaly in Probe-C produces just such a temperature gradient in the metal around the SIA, at its upper end, and that the most likely cause of the anomalous field in Gyro # 1 is this Thomson–Selzer effect. Final verification will only be possible after insertion of the reworked probe into the SMD.

Two possibilities therefore confront us. First that on cool–down the anomaly will have disappeared. Second that for magnetic contamination or some other reason the trapped field in Gyro # 1 will still be anomalously high, in which case the question is, how problematical that will be for the experiment. Detailed analysis by J.M. Lockhart demonstrates that while the 3×10^{-10} T nominal requirement remains a desirable goal, the readout is actually more

'forgiving' than had originally been thought, and there will be little or no loss in final accuracy of the measurement.

D. Leaks in Caging Lines, etc. During the course of the reworking of Probe-C in the clean room at Stanford repairs were effected to the caging lines and these were further secured by the addition of 'doublers' similar in principle to the doublers on the pinning system.

8 Spacecraft, Electronics Systems, Integration & Test

Reflection about Gravity Probe B rightly concentrates first on the Science Instrument with its cryogenic and mechanical refinements but no account of a space mission can end there. The following is a brief summary of the main other aspects of the mission.

8.1 Spacecraft

Gravity Probe B is unusual in that its main structural element is the Science Mission Dewar. To it is attached the Spacecraft, a welded aluminum framework that fits around the lower end of the SMD and carries the solar panels, harnesses, spacecraft electronics, and other support equipment Fig. 14.

Mounted on the front end of the SMD in close proximity to Probe-C is a Forward Electronics Enclosure (FEE), which provides a highly stable thermal environment with extensive shielding against electromagnetic interference and coupling for the key payload electronics boxes. Also mounted on the SMD is the telescope sunshield. The combined weight of spacecraft and payload is 3241 kg. The power requirements are 293 W for the spacecraft and 313 W for the payload, (606 W total) during normal operation; during initial setup, however, the requirement rises at certain intervals to 713 W. For an inertially fixed rolling spacecraft such as GP-B the power output from the sola r panels at any moment depends on their geometry and the orientation of the spacecraft with respect to the Sun. Allowance has to be made for degradation over the course of the mission. The system, as designed, has beginning–of–life and end–of–life capabilities at worst–case seasonal minimum of 731 W and 703 W respectively. At the currently scheduled May 1, 2002 launch–date the actual capability is 873 W at launch declining to 737 W in early June, a sufficient margin over the 713 W initial requirement.

As of June 2000 the spacecraft is 85% complete. Most of the subsystems, mechanical and electrical, are flight–proven. In total there are 21 electronics boxes representing 12 distinct electrical systems all (except for Attitude/Translational Control) flight–proven or close derivatives of flight–proven systems. All have been installed on the Spacecraft. All electrical harnesses are complete. Two and a half of the four solar panels are complete. The spacecraft flight software coding is complete and in veri fication test.

Straightforward as most aspects of the GP-B spacecraft are, two of the mechanical systems essential to the success of the relativity measurement are decidedly unusual. One is the use of helium proportional thrusters for attitude and

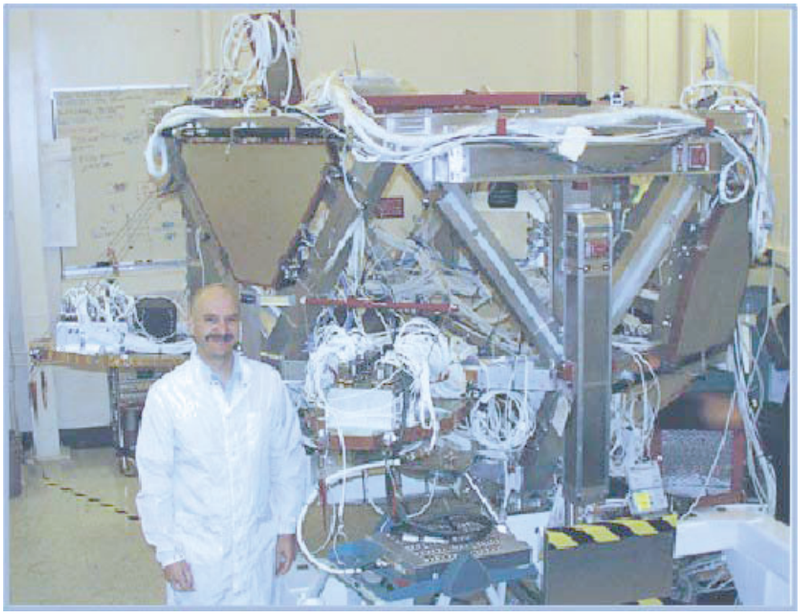


Fig. 14. The GP-B Spacecraft: The spacecraft fits around the lower end of the SMD. Among the items visible in the photo are two long rectangular boxes containing two of the seven mass–trim mechanisms.

translational control; the other is having a set of 7 mass–trim mechanisms to balance and align the principal axes of the spacecraft. Each mass trim mechanism consists of a 20 kg weight mounted in a closed rectangular box and adjusted in position by a lead screw driven by a ste pper motor. These mechanisms have been subject to a very extensive qualification program and are now installed on the spacecraft; 4 are mounted transversely to move the axis of rotation laterally until it coincides with the line through the center of the gyroscopes to within 0.8 mm. The other 3 are parallel to the spacecraft axis and adjust the direction of the principal axes of inertia. Trimming is performed intermittently as the helium is depleted, probably a few times during the course of a year.

In most spacecraft with gas–jet attitude control systems constraints on gas consumption set by weight dictate the use of on–off valves fired only on demand. To apply this method to GP-B with its very fine pointing requirement ($\sim \pm 20$ mas) would take a space–qualified valve capable of reseating perfectly hundreds of millions of times, a severe reliability problem. Fortunately, there is a way. Already on board is a supply of gas that must be vented – the helium boil–off from the dewar. By directing this continuously through pairs of opposed nozzles operated as 'proportional thrusters' one obtains a control system that is at once smoother and mechanically more reliable than the conventional kind. A striking feature is the low Reynolds' number (~ 30) in the throat of the valve. Work by

a succession of GP-B graduate students, J.S. Bull, J.–H. Chen, P. Wiktor, and Y. Jafry, has led to a design subsequently further developed and space–qualified by Lockheed Martin that meets all of GP-B's r equirements, provided the orbit altitude exceeds about 600 km. Fig.15 illustrates the flight thrusters.

Control is maintained in all three modes, pointing, drag–free, and roll, from signals derived respectively from: (1) the science telescope, (2) either of two science gyroscopes operated drag–free, (3) conventional rate–gyroscopes mounted on the spacecraft, updated by signals from a 'star–blipper' picking up a band of stars spread over the heavens at an angle to the roll axis. The pointing accuracy good to about 10 mas, the residual cross–track average acceleration to better than 10^{-12} g, the roll- -rate to about 1 part in 10^5. The translational controller is designed to force the center of rotation of the Spacecraft into coincidence with the line through the gyro centers. The mass–trim mechanism brings it within 0.8 mm; the translational controller brings it to 50 μm, that is, to within the limit to which the gyros are aligned.

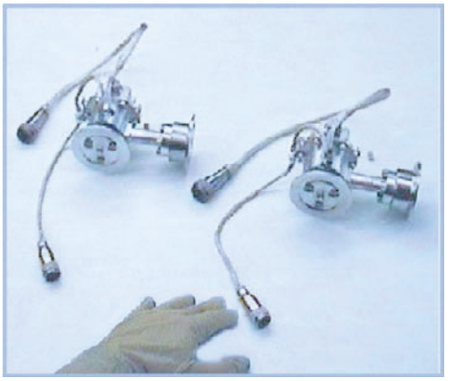


Fig. 15. Proportional Helium Thrusters (Flight Unit).

GP-B has 16 thrusters mounted on fixed struts extending out from the Spacecraft, with geometrical and internal redundancies such that any four systems can be allowed to fail with no loss of control performance.

8.2 Payload Electronics

In contrast to the spacecraft electronics, based on flight–proven hardware, almost all of the payload electronics are new. Altogether there are 5 systems, containing a total of 15 electronics boxes distributed between the FEE and a corresponding Aft Electronics Enclosure (AEE) on the Spacecraft (the reason for this division between forward and aft enclosures is that if all the boxes were mounted forward it would unbalance the mass distribution of the space vehicle):

- Telescope Readout Electronics (TRE), 2 boxes, both forward;
- SQUID Readout Electronics (SRE), 4 boxes, 2 forward, 2 aft;
- Gyro Suspension System (GSS), 8 boxes, 4 forward, 4 aft;
- Experiment Control Unit (ECU), 2 boxes, 1 forward, 1 aft;
- Ultraviolet Discharge (UVD), 2 systems, 1 in each aft ECU box.

Only the UVD has been flown before.

An important lesson for the physicist in the world of space electronics is that even when an engineering model of a flight unit has met all requirements

there is still a long way to go before the actual flight unit is ready for installation. Procuring the right space–hardened components, assembling the flight unit under full Quality Assurance, and completing all the tests that are necessary, including vibration tests, is a major task. Among several common surprises is the extraordinarily long procurement t imes for 'mil–standard' components. For GP-B, engineering models of each unit have been fully tested in the laboratory but as of June 2000 the only flight unit available is the TRE. For the other three all components are in–house; the state of assembly is as follows. (1) SRE: all 16 boards complete. Aft–box passed all functional tests except vibration test; forward–box in functional testing. (2) GSS: 10 of 14 boards assembled. For aft–box, all 5 boards tested, conformally coded, and ready for board int egration. For forward–box, 4 of 9 assembled, 2 completed testing, 1 also conformally coded. (3) ECU: all 19 boards assembled, tested, and conformally coded. Box testing ready to begin.

The most challenging to design of the Payload Electronics systems is the GSS. It has to support the gyroscope with very low voltage (about 0.1 V) in order not to generate suspension torques yet be able to switch instantly to a higher level in emergencies. During gyro spin–up it has to exert an acceleration of approximately 0.3 g on the rotor to balance the pressure from the spin–up gas. Also during spin–up the center of the rotor has to be displaced in a controlled manner toward the spin–up channel to r educe gas leakage over the wall of the channel. All these operations have to be performed automatically without ever allowing the spinning ball to touch the housing. The final design is a digital system with two separate computers and an analog back–up, with an arbiter to decide which of the three parallel systems should take command. Testing it on Earth has been an interesting task. In addition to high voltage operation with live gyroscopes, a gyro test bed has been constructed in which variable flat plat e capacitors driven by piazoelectric actuators mimic the actual gyroscope under controlled conditions.

The current schedule has the GSS completed by December 2000. The GSS and SRE will be used in Probe-C evaluation on the SMD at Stanford early in 2001 prior to shipment of the completed Payload to Lockheed Martin for integration with the Spacecraft in April 2001.

8.3 Integration & Test, Ground Station, and Launch

The process of testing a completed payload and spacecraft is in general terms well–established, though logistically it involves complications that a physicist might hardly expect. For Gravity Probe B, in addition to elaborate acoustic, modal, and end–to–end thermal vacuum tests, very careful spin balancing of the spacecraft is necessary to bring it within the range of the mass–trim mechanisms.

The most unconventional aspect of the GP-B test program is the use of an Integrated Test Facility (ITF) to check software and command/telemetry signals at an early stage prior to transmitting new untested signals to the Spacecraft. From the software point of view GP-B is an unusually complex spacecraft with

7 separate on–board computers, as well as an unusually large number of electronics boxes. To mitigate the difficulty the ITF was commenced very early in the program and brought close to completion in 19 98, though even now not all the hardware is installed for verifying performance of the 1553 computer and command and telemetry signals. A constant difficulty in any test program is to separate problems in the electronics from problems in the Test Facility itself. A method of alleviating this has been devised. It consists of mounting in the ITF the already tested engineering models of the different electronics systems, so that it is possible to switch back and forth between them and the flight units.

No less challenging has been the development at Stanford of the Mission Operation Center (MOC) and ground control software. The MOC communicates through NASA Wallops Flight Facility's Ground Network to a series of ground stations located around the Earth. The primary stations in the Ground Network are the Alaska Ground Station and the Svalbard, Norway Ground Station. During set–up and, if necessary, emergency the Ground Network can be supplemented by the NASA Tracking and Relay Satellite System (TRSS) Spac e Network. And, finally, there is all the planning for the launch. The elaborate preparations required, including safety, preparation of ground support equipment, training of personnel, and the transportation of the spacecraft to the launch site, are a whole story to itself.

Acknowledgement

This work was supported by NASA contract NAS8–39225 through George C. Marshall Space Flight Center, Huntsville, Alabama. We are indebted to the MSFC Gravity Probe B team led by Mr. Rex Geveden. The Lockheed Martin Program Manager is Dr. Hugh Dougherty.

Gravitomagnetism and the Clock Effect

Bahram Mashhoon[1], Frank Gronwald[2],, and Herbert I.M. Lichtenegger[3]

[1] Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211, USA
[2] Institut GET, Universität Magdeburg, D-39106 Magdeburg, Germany
[3] Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, A-8010 Graz, Austria

Abstract. The main theoretical aspects of gravitomagnetism are reviewed. It is shown that the gravitomagnetic precession of a gyroscope is intimately connected with the special temporal structure around a rotating mass that is revealed by the gravitomagnetic clock effect. This remarkable effect, which involves the difference in the proper periods of a standard clock in prograde and retrograde circular geodesic orbits around a rotating mass, is discussed in detail. The implications of this effect for the notion of "inertial dragging" in the general theory of relativity are presented. The theory of the clock effect is developed within the PPN framework and the possibility of measuring it via spaceborne clocks is examined.

1 Introduction

The close formal similarity between Coulomb's law of electricity and Newton's law of gravitation has led to a description of Newtonian gravitation in terms of a gravitoelectric field. The classical tests of general relativity can all be described via post-Newtonian gravitoelectric corrections brought about by relativity theory. Moreover, any theory that combines Newtonian gravitation and Lorentz invariance in a consistent framework must involve a gravitomagnetic field in close analogy with electrodynamics. The gravitomagnetic field is generated by the motion of matter. For instance, the mass current in the rotating Earth generates a dipolar gravitomagnetic field that has not yet been directly observed; in fact, the main objective of the GP-B is to measure this field in a polar Earth orbit via the gravitomagnetic precession of superconducting gyroscopes on board a drag–free satellite.

Gravitomagnetism had its beginning in the second half of the last century. Developments in electrodynamics led Holzmüller [1] and Tisserand [2] to postulate the existence of a solar gravitomagnetic field [3]. In fact, attempts were made to account for the excess perihelion precession of Mercury since the planetary orbits would be affected by the gravitomagnetic field of the Sun. However, the excess perihelion precession of Mercury was successfully explained by Einstein's general relativity theory in terms of a small relativistic correction to the Newtonian gravitoelectric potential of the Sun. It was later shown by Thirring and Lense [4,5] that general relativity also predicts a certain gravitomagnetic field for a rotating mass, but the magnitude of this field in the solar system is generally small and would lead to a retrograde precession of the planetary orbits.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 83–108, 2001.

This Lense-Thirring precession of planetary orbits is too small to be detectable at present.

For the purposes of confronting the theory with observation, gravitomagnetic phenomena are usually described in the framework of the post-Newtonian approximation; however, it is possible to provide a fully covariant treatment of certain aspects of gravitoelectromagnetism [6,7]. In fact, extensions of the Jacobi equation (i.e. the relativistic tidal equation) may be employed to identify the gravitoelectric and gravitomagnetic components of the curvature tensor in close analogy with the Lorentz force law. This analogy is incomplete, however, since the purely spatial components of the curvature tensor do not in general have an analog in the electromagnetic case; in fact, this is expected since linear gravity is a spin-2 field in contrast to the spin-1 character of the electromagnetic field.

Some of the main theoretical aspects of gravitomagnetism are discussed in Section 2. We then turn our attention to how gravitomagnetism affects the spacetime structure in general relativity. Of primary importance in this connection is the gravitomagnetic clock effect, which in its simplest form may be formulated in terms of the difference in the proper periods of two clocks moving on the same circular orbit but in opposite directions about a rotating mass. Let $\tau_+(\tau_-)$ be the period for prograde (retrograde) motion, then for $r \gg 2GM/c^2, \tau_+ - \tau_- \approx 4\pi J/(Mc^2)$. To lowest order, this remarkable result is independent of Newton's constant of gravitation G and the radius of the orbit r. The effect and its consequences are discussed in Section 3 for circular equatorial orbits in the Kerr geometry and the intimate connection between the clock effect and the gravitomagnetic gyroscope precession is demonstrated. The PPN approximation for this effect is developed in Section 4 and a brief discussion of its observability is given in Section 5. The sign of the clock effect is quite intriguing, as it implies that prograde equatorial clocks are slower than retrograde equatorial clocks. This is completely opposite to what would be expected on the basis of "inertial dragging". In fact, gravitomagnetism is historically connected with the question of the origin of inertia as this was Thirring's motivation in his original paper on gravitomagnetism [4]. The present status of the problem of inertia is the subject of Section 6. Finally, Section 7 contains a brief discussion.

Unless specified otherwise, we use units such that $G = c = 1$ for the sake of convenience.

2 Gravitoelectromagnetism

This section is devoted to a brief discussion of certain essential theoretical aspects of gravitoelectromagnetism. The Larmor theorem has played an important role in the field of magnetism; therefore, we begin by an account of the gravitational analog of Larmor's theorem.

2.1 Gravitational Larmor Theorem

A century ago, Larmor established a theorem regarding the local equivalence of magnetism and rotation [8]. That is, the basic electromagnetic force on a slowly moving particle of charge q and mass m can be locally replaced in the linear approximation by the inertial forces that arise if the motion is referred instead to an accelerated system in the absence of the electromagnetic field. The translational acceleration of the system is related to the electric field, $\mathbf{a}_L = -(q/m)\mathbf{E}$, and the rotational (Larmor) frequency is related to the magnetic field via $\boldsymbol{\omega}_L = q\mathbf{B}/(2mc)$. The charge-to-mass ratio is not the same for all particles; otherwise, a geometric theory of electrodynamics could be developed along the same lines as general relativity. It turns out that in general relativity one can provide an interpretation of Einstein's heuristic principle of equivalence via the gravitational Larmor theorem [9]. This is due to the experimentally well-established circumstance that the gravitational charge-to-mass ratio is the same for all particles. Einstein's heuristic principle of equivalence is usually stated in terms of the gravitoelectric field, i.e. the translational acceleration of the "Einstein elevator" in Minkowski spacetime. The gravitational Larmor theorem would also involve the gravitomagnetic field, i.e. a rotation of the elevator as well.

It follows from the theoretical study of the motion of test particles as well as ideal test gyroscopes in a gravitational field that in general relativity the gravitoelectric charge is $q_E = -m$, while the gravitomagnetic charge is $q_B = -2m$; in fact, $q_B/q_E = 2$ since general relativity involves the tensor potential $g_{\mu\nu}$, i.e. (linear) gravitation is a spin-2 field. Thus $\mathbf{a}_L = \mathbf{E}$ and $\boldsymbol{\omega}_L = -\mathbf{B}/c$ in this case. Indeed $\mathbf{B}/c = \boldsymbol{\Omega}_P$ is the gravitomagnetic precession frequency of an ideal test gyroscope at rest in a gravitomagnetic field, i.e. far from a rotating source $d\mathbf{S}/dt = \boldsymbol{\Omega}_P \times \mathbf{S}$, where

$$\boldsymbol{\Omega}_P = \frac{GJ}{c^2 r^3}[3(\hat{\mathbf{r}} \cdot \hat{\mathbf{J}})\hat{\mathbf{r}} - \hat{\mathbf{J}}], \tag{1}$$

and J is the total angular momentum of the source. Let us note that a gyro spin is in effect a gravitomagnetic dipole moment that precesses in a gravitomagnetic field. Locally, the same rotation would be observed in the absence of the gravitomagnetic field but in a frame rotating with frequency $\boldsymbol{\omega}_L = -\boldsymbol{\Omega}_P$ in agreement with the gravitational Larmor theorem.

It is the goal of the GP-B to measure the gravitomagnetic gyroscope precession in a polar orbit about the Earth and thereby provide *direct* observational proof of the existence of the gravitomagnetic field [10].

2.2 Gravitoelectromagnetic Field

Let us consider the gravitational field of a "nonrelativistic" rotating astronomical source in the linear approximation of general relativity. The spacetime metric may be expressed as $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, where $\eta_{\mu\nu}$ is the Minkowski metric. We

define $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$, where $h = \mathrm{tr}(h_{\mu\nu})$; then, the gravitational field equations are given by

$$\Box\bar{h}_{\mu\nu} = -16\frac{\pi G}{c^4}T_{\mu\nu}, \tag{2}$$

where the Lorentz gauge condition $\bar{h}^{\mu\nu}{}_{,\nu} = 0$ has been imposed. We focus attention on the particular retarded solution of the field equations given by

$$\bar{h}_{\mu\nu} = \frac{4G}{c^4}\int \frac{T_{\mu\nu}(ct - |\mathbf{x}-\mathbf{x}'|,\ \mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\, d^3x', \tag{3}$$

where the nature and distribution of the "nonrelativistic" source must be taken into account.

We are interested in sources such that $\bar{h}_{00} = 4\Phi/c^2$, $\bar{h}_{0i} = -2A_i/c^2$ and $\bar{h}_{ij} = O(c^{-4})$, where $\Phi(t,\mathbf{x})$ is the gravitoelectric potential, $\mathbf{A}(t,\mathbf{x})$ is the gravitomagnetic vector potential and we neglect all terms of order c^{-4} and lower including the tensor potential $\bar{h}_{ij}(t,\mathbf{x})$. It follows that $T^{00}/c^2 = \rho$ is the effective gravitational charge density and $T^{0i}/c = j^i$ is the corresponding current. Thus, far from the source

$$\Phi \sim \frac{GM}{r}, \qquad \mathbf{A} \sim \frac{G}{c}\frac{\mathbf{J}\times\mathbf{r}}{r^3}, \tag{4}$$

where M and $\mathbf{J}$ are the total mass and angular momentum of the source, respectively. It follows from the Lorentz gauge condition that

$$\frac{1}{c}\frac{\partial\Phi}{\partial t} + \boldsymbol{\nabla}\cdot\left(\frac{1}{2}\mathbf{A}\right) = 0, \tag{5}$$

since the other three equations ($\bar{h}^{i\mu}{}_{,\mu} = 0$) all involve terms that are of $O(c^{-4})$ and therefore neglected. The spacetime metric involving the gravitoelectromagnetic ("GEM") potentials is then given by

$$-ds^2 = -c^2\left(1 - \frac{2}{c^2}\Phi\right)dt^2 - \frac{4}{c}(\mathbf{A}\cdot d\mathbf{x})dt + \left(1 + \frac{2}{c^2}\Phi\right)\delta_{ij}dx^i dx^j. \tag{6}$$

The GEM fields are defined by

$$\mathbf{E} = -\boldsymbol{\nabla}\Phi - \frac{1}{c}\frac{\partial}{\partial t}\left(\frac{1}{2}\mathbf{A}\right), \qquad \mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A}, \tag{7}$$

in close analogy with electrodynamics. It follows from the field equations (2) and the gauge condition (5) that

$$\boldsymbol{\nabla}\cdot\mathbf{E} = 4\pi G\rho, \tag{8}$$

$$\boldsymbol{\nabla}\cdot\left(\frac{1}{2}\mathbf{B}\right) = 0, \tag{9}$$

$$\boldsymbol{\nabla}\times\mathbf{E} = -\frac{1}{c}\frac{\partial}{\partial t}\left(\frac{1}{2}\mathbf{B}\right), \tag{10}$$

$$\boldsymbol{\nabla}\times\left(\frac{1}{2}\mathbf{B}\right) = \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E} + \frac{4\pi}{c}G\,\mathbf{j}\,, \tag{11}$$

which are the Maxwell equations for the GEM field. Using classical electrodynamics as a guide, one can investigate the various implications of these equations [11]. A thorough approach to the determination of the gravitomagnetic field of a rotating mass (such as the Earth) is contained in the papers of Teyssandier [12].

The fact that the magnetic parts of equations (8) - (11) always appear with a factor of 1/2 as compared to standard electrodynamics is due to the circumstance that the effective gravitomagnetic charge is twice the gravitoelectric charge. That is, $Q_E = M$ and $Q_B = 2M$ are the effective gravitoelectric and gravitomagnetic charges of the source.

The linear approximation of general relativity involves a spin-2 field. This field, once its spatial components are neglected, can be interpreted in terms of a gravitoelectromagnetic vector potential. To sustain the electromagnetic analogy, however, we need to require that the gravitomagnetic charge be twice the gravitoelectric charge. This factor of 2 is a remnant of the spin-2 character of the original field, while for a pure spin-1 field (i.e. the electromagnetic field) the ratio of the magnetic charge to the electric charge is unity.

The equation of motion of a test particle of mass m in this linear gravitational field can be obtained from the variational principle $\delta \int \mathcal{L} dt = 0$, where $\mathcal{L} = -mcds/dt$ is given by

$$\mathcal{L} = -mc^2 \left[1 - \frac{v^2}{c^2} - \frac{2}{c^2} \left(1 + \frac{v^2}{c^2} \right) \Phi + \frac{4}{c^3} \mathbf{v} \cdot \mathbf{A} \right]^{1/2}, \tag{12}$$

using equation (6). To linear order in Φ and $\mathbf{A}$, one can write (12) as

$$\mathcal{L} = -mc^2 \left(1 - \frac{v^2}{c^2} \right)^{1/2} + m\gamma \left(1 + \frac{v^2}{c^2} \right) \Phi - \frac{2m}{c} \gamma \mathbf{v} \cdot \mathbf{A}. \tag{13}$$

Let us note that the deviation of equation (13) from a free-particle Lagrangian is given to lowest order in $\mathbf{v}/c$ by $m\Phi - 2m\mathbf{A} \cdot \mathbf{v}/c$. This deviation would be of the form $j_\mu A^\mu$ in electrodynamics; therefore, the slow motion of the test particle is very similar to that of a charged particle in electrodynamics except that here $q_E = -m$ and $q_B = -2m$ as expected. It thus follows from the geodesic motion of a test particle of mass m far from the source in this gravitational background that the canonical momentum of the particle is given approximately by $\mathbf{p} + (-2m/c)\mathbf{A}$, where $\mathbf{p}$ is the kinetic momentum. In this electrodynamic analogy, the attractive nature of gravity is reflected in our convention of positive gravitational charges for the source and negative gravitational charges for the test particle. The gravitomagnetic charge is always twice the gravitoelectric charge as a consequence of the tensorial character of the gravitational potentials in general relativity.

The gauge transformations

$$\Phi \to \Phi - \frac{1}{c} \frac{\partial \psi}{\partial t}, \qquad \mathbf{A} \to \mathbf{A} + 2\boldsymbol{\nabla} \psi, \tag{14}$$

leave the GEM fields (7) and hence the GEM equations (8)-(11) invariant. The Lorentz gauge condition (5) is also satisfied provided $\Box \psi = 0$. However, the

quantity $-q_E\Phi + q_B\mathbf{A}\cdot\mathbf{v}/c$ in the Lagrangian is not invariant under the gauge transformation (14). The gauge invariance of this Lagrangian is restored, however, if the gauge function ψ is independent of time, $\partial\psi/\partial t = 0$. In this case, we can start from a coordinate transformation $t \to t - 4\psi(\mathbf{x})/c^3$ in the metric (6) resulting in the gauge transformations (14) with Φ left invariant.

The gravitational field corresponding to the metric (6) is given by the Riemann curvature tensor

$$R_{\mu\nu\rho\sigma} = \frac{1}{2}(h_{\mu\sigma,\,\nu\rho} + h_{\nu\rho,\,\mu\sigma} - h_{\nu\sigma,\,\mu\rho} - h_{\mu\rho,\nu\sigma}), \tag{15}$$

where $h_{00} = 2\Phi/c^2$ and $h_{ij} = (2\Phi/c^2)\delta_{ij}$ are gravitoelectric and of $O(c^{-2})$, while $h_{0i} = -2A_i/c^2$ is gravitomagnetic and of $O(c^{-3})$. The components of the curvature tensor as measured by the *standard* geodesic observers are given by $R_{\mu\nu\rho\sigma}\lambda^{\mu}_{(\alpha)}\lambda^{\nu}_{(\beta)}\lambda^{\rho}_{(\gamma)}\lambda^{\sigma}_{(\delta)}$, where $\lambda^{\mu}_{(\alpha)}$ is the tetrad frame of the test observer. In the linear approximation under consideration here, $\lambda^{\mu}_{(\alpha)}$ is in effect equal to δ^{μ}_{α} in the calculation of the measured curvature. The components of this tensor may be expressed in the form of a symmetric 6×6 matrix $\mathcal{R} = (\mathcal{R}_{AB})$, where A and B range over (01, 02, 03, 23, 31, 12); hence,

$$\mathcal{R} = \begin{pmatrix} \mathcal{E} & \mathcal{B} \\ \mathcal{B}^T & \mathcal{S} \end{pmatrix}, \tag{16}$$

where $\mathcal{E}$ and $\mathcal{S}$ are symmetric 3×3 matrices and $\mathcal{B}$ is traceless. We find that the electric and magnetic components of the curvature are given by

$$\mathcal{E}_{ij} = \frac{1}{c^2}E_{j,i} + O(c^{-4}), \tag{17}$$

$$\mathcal{B}_{ij} = -\frac{1}{c^2}B_{j,i} + \frac{1}{c^3}\epsilon_{ijk}\frac{\partial E_k}{\partial t} + O(c^{-4}), \tag{18}$$

and the spatial components are given by

$$\mathcal{S}_{ij} = -\frac{1}{c^2}E_{j,i} + \frac{1}{c^2}(\boldsymbol{\nabla}\cdot\mathbf{E})\delta_{ij} + O(c^{-4}). \tag{19}$$

That $\mathcal{B}$ is traceless is consistent with equation (9) and the fact that $\mathcal{E}$ and $\mathcal{S}$ are symmetric is consistent with equation (10) at $O(c^{-4})$. It is therefore clear that gravitoelectromagnetism permeates every aspect of general relativity: the gravitational potentials (GEM potentials), the connection (GEM field) and the curvature. In the exterior of the rotating source, the spacetime is Ricci-flat and hence $\mathcal{S} = -\mathcal{E}$, $\mathcal{E}$ is traceless and $\mathcal{B}$ is symmetric. These restrictions on the curvature are consistent with the GEM field equations (8)-(11) in the source-free region.

The general treatment of gravitoelectromagnetism presented here has been based on a certain approximate form of the linear gravitation theory and can be used in the theoretical description of many interesting gravitational phenomena. In particular, we use this formalism below to investigate the microphysical implications of the gravitomagnetic precession of spin.

2.3 Free Fall Is Not Universal

The assumption that all free test particles fall in the same way in a gravitational field is reflected in general relativity via the geodesic hypothesis. That is, the worldline of a free test particle is an intrinsic property of the spacetime manifold and is independent of the intrinsic aspects of the particle. In this way, general relativity is a geometric theory of gravitation. This circumstance originates from the well-tested equality of inertial and gravitational masses.

An important consequence of Einstein's geometric theory of gravitation is the fact that an ideal test gyroscope would precess in the gravitomagnetic field of a rotating source. Here we pose the question of whether all spins should "precess" like a gyroscope; evidently, the treatment of intrinsic spin would go beyond classical general relativity. It follows from the consideration of spin-rotation-gravity coupling that the intrinsic spin of a particle (e.g. a nucleus) would couple to the gravitomagnetic field of a rotating source (such as the Earth) via the interaction Hamiltonian

$$\mathcal{H} = \boldsymbol{\sigma} \cdot \boldsymbol{\Omega}_P \tag{20}$$

such that the Heisenberg equations of motion for the spin would be formally the same as that of an ideal test gyro [13]. Intuitively, this interaction is due to the coupling of the gravitomagnetic dipole moment of the particle with the gravitomagnetic field just as would be expected from the electromagnetic analogy. It follows from equation (20) that the particle is subject to a gravitational Stern-Gerlach force given by

$$\mathbf{F} = -\boldsymbol{\nabla}(\boldsymbol{\sigma} \cdot \boldsymbol{\Omega}_P) \tag{21}$$

that is purely dependent upon its spin and not its mass and therefore violates the universality of free fall.

The point is that a particle is in general endowed with mass and spin in addition to other intrinsic properties; indeed, the irreducible unitary representations of the inhomogeneous Lorentz group are characterized by mass and spin. In its interaction with a gravitational field, the mass interacts primarily with the gravitoelectric field while the spin interacts primarily with the gravitomagnetic field. Whereas the former dominant interaction is consistent with the universality of free fall, the latter is not. For instance, the bending of light by the gravitational field of a rotating source depends on the state of polarization of the radiation. The differential deflection of polarized radiation by the Sun is too small to be measurable at present. The predicted violation is also extremely small for a nucleus in a laboratory on the Earth: the weight of the particle is $w = mg_{\oplus}(1 \pm \epsilon)$, depending on whether the spin is polarized vertically up or down and $\epsilon \sim 10^{-29}$. Thus the predicted violation of the universality of free fall is extremely small.

It may still be possible to measure this relativistic quantum gravitational effect by detecting the change in the energy of a particle in the laboratory when its spin is flipped. This would require, for instance, significant refinements in

modern variations of NMR and optical pumping techniques, since

$$\hbar\Omega_P \sim \frac{\hbar GJ}{c^2R^3} = \frac{cJ}{R}\left(\frac{L_P}{R}\right)^2 \sim 10^{-28}\text{eV} \tag{22}$$

is a factor of 10^4 below the sensitivity of recent experiments [14]. Here L_P is the Planck length, $L_P^2 = \hbar G/c^3$, J is the angular momentum of the Earth and R is its average radius. The smallest detected energy shift is about 10^{-24}eV corresponding to a frequency shift of 2 nHz [14]. However, it appears that significantly lower energy shifts may soon be detectable [15].

To clarify the nature of the force (21), let us consider the motion of a classical spinning test body in a *stationary* gravitational field. Such a system is necessarily extended and thus couples to spacetime curvature resulting in a Mathisson-Papapetrou force

$$F_\alpha = \frac{c}{2}R_{\alpha\beta\mu\nu}u^\beta S^{\mu\nu} = cR_{\alpha\mu\beta\nu}u^\beta S^{\mu\nu}, \tag{23}$$

where $S^{\mu\nu}$ is the spin tensor of the system, u^μ is the velocity vector such that $S^{\mu\nu}u_\nu = 0$ and the spin vector is given by

$$S_\mu = \frac{1}{2}(-g)^{1/2}\epsilon_{\mu\nu\rho\sigma}u^\nu S^{\rho\sigma}. \tag{24}$$

For the calculation of F_α, it suffices to set, in the linear approximation, $u^\alpha \approx (1,\ 0,\ 0,\ 0)$, $S^{0i} \approx 0$ and $S^{ij} \approx -\epsilon^{ijk}S_k$. Then, $F_0 \approx 0$ and

$$F_i \approx c\mathcal{B}_{ij}S^j = -\frac{1}{c}B_{j,i}S^j = -(\boldsymbol{\Omega}_P)_{j,i}S^j, \tag{25}$$

in agreement with equation (21). Thus the existence of the force (21) may be ascribed to the intrinsic nonlocality of a particle in the quantum theory and hence the coupling of spin to the magnetic part of the spacetime curvature in a stationary field.

It is important to remark here that our considerations are distinct from proposals to measure the classical spin-spin force as discussed in [11]. Our results ultimately follow from detailed considerations of Dirac-type wave equations in the gravitational field of a rotating mass (see the references cited in [13]); however, one can arrive at equations (20)-(21) on the basis of certain general arguments such as the local isotropy of space, the extended hypothesis of locality and the gravitational Larmor theorem [13].

Assuming the approximate validity of equations (20)-(21), it would be difficult to imagine a basic gravitational theory founded purely on the universality of free fall and Riemannian geometry. However, such a theoretical structure would clearly be an excellent effective theory in the macrophysical domain.

2.4 GEM Stress-Energy Tensor

Let us imagine a congruence of geodesic test particles in a gravitational field. Taking one of the test particles as the reference observer, how does the motion of

the other neighboring test particles appear to the fiducial observer? The result is best expressed in a Fermi coordinate system that is set up along the reference worldline. Let $X^\mu = (\tau, \mathbf{X})$ be the Fermi coordinates of the test particles, while the reference observer is at the origin of spatial Fermi coordinates. Then the equation of motion of the test particles is given by the generalized Jacobi equation

$$\frac{d^2X^i}{d\tau^2} + {}^F\!R_{0i0j}X^j + 2\,{}^F\!R_{ikj0}V^kX^j + (2\,{}^F\!R_{0kj0}V^iV^k + \frac{2}{3}\,{}^F\!R_{ikjl}V^kV^l + \frac{2}{3}\,{}^F\!R_{0kjl}V^iV^kV^l)X^j = 0, \tag{26}$$

which is valid to first order in the relative separation $\mathbf{X}$ and to all orders in the relative velocity $\mathbf{V} = d\mathbf{X}/d\tau$. Here ${}^F\!R_{\alpha\beta\gamma\delta}(\tau)$ are components of the curvature tensor as measured by the fiducial observer, i.e. they are the projections of the Riemann tensor onto the nonrotating tetrad frame of the reference observer. Neglecting the second and third order terms in the relative rate of separation, equation (26) can be written as the GEM analog of the Lorentz force law

$$m\frac{d^2\mathbf{X}}{d\tau^2} = q_E\mathbf{E} + q_B\mathbf{V}\times\mathbf{B}, \tag{27}$$

where $q_E = -m$, $q_B = -2m$ and

$$E_i = {}^F\!R_{0i0j}(\tau)X^j, \quad B_i = -\frac{1}{2}\epsilon_{ijk}\,{}^F\!R_{jk0l}(\tau)X^l. \tag{28}$$

It is important to notice that the spacetime interval in the neighborhood of the reference worldline can be expressed in Fermi coordinates as $-ds^2 = {}^F\!g_{\mu\nu}\,dX^\mu dX^\nu$, where

$${}^F\!g_{00} = -1 - {}^F\!R_{0i0j}(\tau)X^iX^j + \cdots, \tag{29}$$

$${}^F\!g_{0i} = -\frac{2}{3}\,{}^F\!R_{0jik}(\tau)X^jX^k + \cdots, \tag{30}$$

$${}^F\!g_{ij} = \delta_{ij} - \frac{1}{3}\,{}^F\!R_{ikjl}(\tau)X^kX^l + \cdots. \tag{31}$$

Letting ${}^F\!g_{00} = -1 + 2\Phi$ and ${}^F\!g_{0i} = -2A_i$, we find that

$$\Phi = -\frac{1}{2}\,{}^F\!R_{0i0j}X^iX^j, \quad A_i = \frac{1}{3}\,{}^F\!R_{0jik}X^jX^k, \tag{32}$$

so that the corresponding GEM fields using equation (7) agree with the results in equation (28) to linear order in the separation $\mathbf{X}$. One can verify directly that $\boldsymbol{\nabla}\cdot\mathbf{B} = 0$ and $\boldsymbol{\nabla}\times\mathbf{E} = 0$, so that the source-free pair of Maxwell's equations are satisfied *along the reference worldline*. Moreover, it is possible to combine the GEM fields together to form a GEM Faraday tensor $F_{\alpha\beta}$,

$$F_{\alpha\beta} = -\,{}^F\!R_{\alpha\beta0l}X^l, \tag{33}$$

such that $F_{0i} = -E_i$ and $F_{ij} = \epsilon_{ijk}B_k$ as in standard electrodynamics. Then the other pair of Maxwell's equations is given by $F^{\alpha\beta}{}_{,\beta} = 4\pi J^\alpha$, where $J^\alpha(\tau, \mathbf{X})$ is easily obtained to linear order in $\mathbf{X}$ using equation (33). J^α is a conserved current such that $J_\alpha(\tau, \mathbf{0}) = -\,{}^F R_{\alpha 0}/4\pi$ along the fiducial trajectory. This treatment should be compared and contrasted with the linear approximation developed above, in particular, the GEM current is different here.

It is now possible to develop the classical field theory of the GEM field in the Fermi frame; in particular, one can define the Maxwell stress-energy tensor and the corresponding angular momentum for the GEM field. Thus

$$\mathcal{T}^{\alpha\beta}(\tau, \mathbf{X}) = \frac{1}{4\pi}\left(F^\alpha{}_\gamma F^{\beta\gamma} - \frac{1}{4}\eta^{\alpha\beta}F_{\gamma\delta}F^{\gamma\delta}\right) \tag{34}$$

is the Maxwell stress-energy tensor for the GEM field that is quadratic in the spatial separation and vanishes at the location of the fiducial observer. Physical measurements do not occur at a point, as already emphasized by Bohr and Rosenfeld [16]; moreover, the fiducial observer is arbitrary here. Therefore, a physically more meaningful quantity is obtained by averaging equation (34) over a sphere of radius ϵL in the Fermi system. Here L is a constant invariant length-scale associated with the gravitational field. We find that

$$< \mathcal{T}_{\alpha\beta}(\tau, \mathbf{X}) > = \epsilon^2 C_0 L^2 \tilde{T}_{\mu\nu\rho\sigma}\lambda^\mu_{(\alpha)}\lambda^\nu_{(\beta)}\lambda^\rho_{(0)}\lambda^\sigma_{(0)}, \tag{35}$$

where C_0 is a constant numerical factor and

$$\tilde{T}_{\mu\nu\rho\sigma} = \frac{1}{2}\left(R_{\mu\xi\rho\zeta}R_\nu{}^{\xi}{}_\sigma{}^{\zeta} + R_{\mu\xi\sigma\zeta}R_\nu{}^{\xi}{}_\rho{}^{\zeta}\right) - \frac{1}{4}g_{\mu\nu}R_{\alpha\beta\rho\gamma}R^{\alpha\beta}{}_\sigma{}^{\gamma}. \tag{36}$$

For a Ricci-flat spacetime, $\tilde{T}_{\mu\nu\rho\sigma}$ reduces to the Bel-Robinson tensor $T_{\mu\nu\rho\sigma}$; in this case, $R_{\alpha\beta\gamma\delta}$ reduces to the Weyl tensor $C_{\alpha\beta\gamma\delta}$ and in equation (36) $C_{\alpha\beta\rho\gamma}C^{\alpha\beta}{}_\sigma{}^{\gamma} = (K/4)g_{\rho\sigma}$ with $K = C_{\alpha\beta\gamma\delta}C^{\alpha\beta\gamma\delta}$.

The magnitude of C_0 depends on whether we average over the surface or the volume of the sphere; in any case, one can always absorb C_0 into the definition of L. Thus the pseudo-local GEM stress-energy tensor may be defined for any observer with the tetrad frame $\lambda^\mu_{(\alpha)}$ as

$$T_{(\alpha)(\beta)} = L^2\tilde{T}_{\mu\nu\rho\sigma}\lambda^\mu_{(\alpha)}\lambda^\nu_{(\beta)}\lambda^\rho_{(0)}\lambda^\sigma_{(0)}. \tag{37}$$

In this way, the pseudo-local GEM energy density, Poynting flux and stresses are defined up to a common multiplicative factor.

It is important to note that the spatial components of the curvature have been ignored in our construction of the GEM tensor $T_{(\alpha)(\beta)}$. For a Ricci-flat spacetime, however, the spatial components of the curvature are simply related to its electric components; therefore, the pseudo-local tensor defined via equation (37) using the Bel-Robinson tensor contains the full (Weyl) curvature tensor and is thus the *gravitational stress-energy tensor*.

It follows from a simple application of these results to the field of a rotating mass that there exists a steady Poynting flux of gravitational energy in the exterior field of a rotating mass.

2.5 Oscillations of a Charged Rotating Black Hole

Imagine a black hole of mass M, charge Q and angular momentum J that is perturbed by external radiation. The black hole is stationary and axisymmetric; therefore, the perturbation is expressible in terms of eigenmodes $\mathcal{P}(r,\theta)\exp(-i\omega t + im_j\phi)$, where $\mathcal{P}$ depends upon the frequency of the radiation, the total angular momentum parameters of the eigenmode (j, m_j) and the spin of the external field. It turns out that for a Fourier sum of such eigenmodes, the response of the black hole far away and at late times is dominated by a superposition of certain damped oscillations of the form $A\exp(-i\omega t)$, where $\omega = \omega_{\rm BH} - i\Gamma_{\rm BH}$ with $\Gamma_{\rm BH} \geq 0$. For these quasinormal modes, the amplitude A depends, among other things, on the strength of the perturbation while ω depends only on the black hole parameters (M, Q, J). Moreover, these black hole oscillations are in general denumerably infinite and are numbered as $n = 0, 1, 2, \cdots$, such that $n = 0$ is least damped and $\Gamma_{\rm BH}$ increases with n. The intrinsic ringing of a black hole is due to the fact that once perturbed, the black hole undergoes characteristic damped oscillations in order to return to a stationary state.

The fundamental modes of oscillations of black holes were originally found by numerical experiments and initial attempts to explain the numerical results via the properties of black hole effective potentials were unsuccessful [17]. The solution of the problem was first given around 1980 [18]. This work provided the stimulus for many subsequent investigations by a number of authors [19]. A detailed discussion of black hole oscillations is contained in [20].

For the modes of oscillation of a charged rotating black hole, the only reliable results are for $j \geq |m_j| \gg 1$. Expressions for $(\omega_{\rm BH}, \Gamma_{\rm BH})$ have been obtained in the case of $j = |m_j| \gg 1$ for a general Kerr-Newman black hole; however, the results have been generalized to the case of $j > |m_j| \gg 1$ only for a slowly rotating charged black hole [21]. To express $(\omega_{\rm BH}, \Gamma_{\rm BH})$ in terms of (M, J, Q) in the latter case, let $\omega_{\rm K}(r) = (Mr^{-3} - Q^2r^{-4})^{1/2}$ be the "Keplerian" frequency for the motion of a neutral particle in a circular geodesic orbit of radius r about a Reissner-Nordström black hole of mass M and charge Q. Here we use Boyer-Lindquist type of coordinates for the Kerr-Newman geometry. Timelike circular geodesic orbits exist down to a null orbit of radius r_N such that $2r_N = 3M + (9M^2 - 8Q^2)^{1/2}$. Let $\omega_N = \omega_{\rm K}(r_N)$, then it can be shown that for a slowly rotating black hole

$$\omega_{\rm BH} \approx \pm j\omega_N + m_j\Omega_N, \tag{38}$$

where

$$\Omega_N = \frac{J}{r_N^3}\left(1 - \frac{Q^2}{Mr_N}\right)\frac{r_N + M}{r_N - M}, \tag{39}$$

is an effective black hole rotation frequency. The $(2j+1)$-fold degeneracy in the spectrum of oscillations of the spherical black hole is removed by its rotation. We note that Ω_N is proportional to the gravitomagnetic precession frequency at r_N. Moreover,

$$\Gamma_{\rm BH} \approx \left(n + \frac{1}{2}\right)\left(2 - 3\frac{M}{r_N}\right)^{1/2}\left(\omega_N \mp \frac{m_j}{j}Q_*\Omega_N\right), \tag{40}$$

where $n = 0, 1, 2, \cdots$ is the mode number and $Q_* = 6MQ^2/[r_N(9M^2 - 8Q^2)]$. It is interesting to note that if the black hole is charged, the rotation removes the degeneracy of the damping factor as well. Moreover, in the formulas (38) and (40) if $(\omega_{\rm BH}, \Gamma_{\rm BH})$ is a ringing mode, then so is $(-\omega_{\rm BH}, \Gamma_{\rm BH})$. These results are independent of the spin of the perturbing field, since they are valid for states of high total angular momentum $j \geq |m_j| \gg 1$.

3 Structure of Time and Relativistic Precession

Let us now return to the gravitomagnetic temporal structure around a rotating source. The gravitomagnetic clock effect involves a coupling between the *orbital* motion of clocks and the rotation of the source. On the other hand, the gravitomagnetic gyroscope precession occurs even for a gyroscope at rest in the exterior field of a rotating source. Nevertheless, there is a general physical connection between relativistic precession and temporal structure. This is not surprising since the operational definition of time ultimately involves counting a definite period and simple precession is uniform periodic motion. It is the purpose of this section to explain this relationship. We do this in several steps in the context of Kerr geometry with

$$-ds^2 = -dt^2 + \frac{\Sigma}{\Delta}\, dr^2 + \Sigma\, d\theta^2 + (r^2 + a^2)\sin^2\theta\, d\phi^2 + \frac{R_g r}{\Sigma}(c\, dt - a\sin^2\theta\, d\phi)^2, \qquad (41)$$

where $\Sigma = r^2 + a^2\cos^2\theta$ and $\Delta = r^2 - R_g r + a^2$. Here $R_g = 2GM/c^2$ is the gravitational radius of the source and the Kerr parameter $a = J/Mc$ is a lengthscale characteristic of the rotation of the source. For $M = 0$ and $a \neq 0$, the spacetime given by (41) is flat as expected. For $a = 0$ and $M \neq 0$, the metric (41) represents the Schwarzschild geometry. Finally, for $a = 0$ and $M = 0$ we have the metric of an inertial frame expressed in spherical coordinates (r, θ, ϕ).

Let us first imagine an accelerated observer in an inertial frame. Suppose that this observer carries along its worldline an ideal pointlike test gyroscope so that there is no net torque on the gyroscope and its spin axis is therefore nonrotating. To simplify matters, let us first assume that the path is a circle of radius r in the (x, y)-plane with its center at the origin of coordinates. According to the standard static observers in the inertial frame, the accelerated observer moves with uniform frequency $\omega_*\hat{\mathbf{z}}$. The transformation between the inertial frame and the rest frame of the rotating observer involves a simple rotation of frequency ω_*; therefore, from the viewpoint of the standard (i.e. static) observers in the rotating frame a natural operational way to keep the direction of the gyroscope spin axis nonrotating is to imagine fixing this axis at some initial time with respect to the axes of the *rotating frame*, but then continuously rotating it backward with frequency ω_*. In this way, the spin direction would remain fixed in the inertial frame if the rotation of the observer were virtual. In reality, however, the observer's proper time τ is related to t by $d\tau = dt\ (1 - v^2/c^2)^{1/2}$, where $v = r\omega_*$;

hence, the backward rotation of the spin occurs with respect to the rotating observer's proper time, i.e. with frequency $\omega_*(dt/d\tau)$. From the standpoint of the standard inertial observers, the time dilation causes the spin direction to overcompensate and hence the spin direction is not fixed but precesses with the Thomas precession frequency $\omega_{\rm T} = -\omega_*(dt/d\tau) + \omega_*$. This amounts to a precession of frequency $\omega_*(\gamma - 1)$ in a sense that is opposite to that of the rotation of the comoving observer; moreover, the generalization to arbitrary acceleration can be simply carried out by means of the Frenet procedure. That is, a Frenet frame can be set up along the path of the observer in space; then, $\omega_* = v/R(t)$, where $R(t)$ is the radius of the curvature at each instant of time t.

The intimate connection between time dilation and Thomas precession in Minkowski spacetime can be extended to a gravitational field. Therefore, let us imagine next that the motion described above is the geodesic motion of a free test particle carrying an ideal test gyroscope around a spherical mass M. Let $\omega_{\rm K} = d\phi/dt$ be the Keplerian frequency as perceived by static inertial observers at infinity; then, $\omega_{\rm K}^2 = GM/r^3$, where r is the Schwarzschild radius of the circular orbit. The proper frequency in this case is $\omega = \Gamma\omega_{\rm K}$, where $\Gamma = dt/d\tau = (1 - 3GM/c^2r)^{-1/2}$. The gravitational time dilation involves the static "gravitational redshift" effect of $-g_{00} = 1 - 2GM/c^2r$ in the Schwarzschild geometry as well as the azimuthal motion $r^2(d\phi/dt)^2 = GM/r$ resulting in the factor of 3 in Γ. This situation is reminiscent of the spin-orbit coupling in the motion of the electron around the nucleus in the hydrogen atom; however, there are subtle differences between the electromagnetic and gravitational cases. In this case, the spin precession frequency is given by the Fokker frequency $\omega_{\rm F} = \omega - \omega_{\rm K}$, and the sense of precession is in the *same* sense as the orbital motion. This gravitational analog of the Thomas precession has a simple and transparent explanation in terms of Einstein's principle of equivalence. According to this heuristic principle, an observer $\mathcal{O}$ in a gravitational field is locally equivalent to an observer $\mathcal{O}'$ in Minkowski spacetime with an acceleration that is equal in magnitude but opposite in direction to the Newtonian gravitational "acceleration" of $\mathcal{O}$. The gravitational (Fokker) precession is thus locally equivalent to Thomas precession with the direction of acceleration reversed. It follows that the Fokker precession is in the same sense as the orbital motion. For an arbitrary accelerated observer in a gravitational field with velocity $u^\mu = dx^\mu/d\tau$ and acceleration $a^\mu = Du^\mu/d\tau$, the nonrotating equation of motion for the torque-free pointlike spin vector ($u_\mu S^\mu = 0$) is

$$\frac{dS^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta}\, u^\alpha S^\beta = u^\mu a_\nu S^\nu, \tag{42}$$

so that both Fokker and Thomas precessions would be present for accelerated motion in Schwarzschild geometry.

The gravitomagnetic precession of a gyroscope is in a similar way related to the temporal structure brought about by the rotation of the source. However, the situation here is more complicated than the gravitoelectric Fokker precession since the temporal structure is affected by the coupling of orbital motion with the angular momentum of the source.

Specifically, let us imagine stable circular geodesic orbits in the equatorial plane of the Kerr source (41). It can be shown that

$$\frac{dt}{d\phi} = \frac{a}{c} \pm \frac{1}{2\pi} T_{\rm K} \quad , \quad \frac{d\tau}{d\phi} = \pm \frac{1}{2\pi} T_{\rm K} \, (1 \pm 2\alpha - 3R_g/2r)^{1/2} \, , \tag{43}$$

where $T_{\rm K} = 2\pi/\omega_{\rm K}$ is the Keplerian period of the orbit and $\alpha = a \, \omega_{\rm K}/c$. Here the upper (lower) sign refers to a prograde (retrograde) orbit. It follows that the orbital period is given by

$$t_\pm = T_{\rm K} \, (1 \pm \alpha) \quad , \quad \tau_\pm = T_{\rm K} \, (1 \pm 2\alpha - 3R_g/2r)^{1/2} \, . \tag{44}$$

Let us first note that $t_+ - t_- = 4\pi a/c$ and $\tau_+^2 - \tau_-^2 = 4\alpha T_{\rm K}^2$. Since $\tau_+ + \tau_- = 2T_{\rm K} + O(c^{-2})$, we find that $\tau_+ - \tau_- \approx 4\pi a/c$. In fact, $\tau_+ - \tau_-$ monotonically decreases as a function of r and approaches $4\pi a/c$ as $r \to \infty$. Thus a prograde clock moves more slowly than a retrograde clock according to comoving observers as well as the standard asymptotically inertial observers at infinity; moreover, $\tau_+ - \tau_- \approx 4\pi a/c$ for $r \gg R_g$. This remarkable gravitomagnetic clock effect has been discussed in some detail in recent publications [22]-[27]. This classical effect is in some sense the gravitomagnetic analog of the topological Aharonov-Bohm effect; in fact, let us note that far from a finite rotating source $\tau_+ - \tau_-$ is nearly a constant independent of the "distance" r and the gravitational coupling constant G. These aspects of the clock effect have been discussed in detail elsewhere [26].

Let us now imagine two clocks moving in opposite directions on a stable circular geodesic orbit of radius r in the equatorial plane of the Kerr source. Suppose that at $t_0 = 0$, they are both at $\phi_0 = 0$; let us denote the event at which the clocks next meet again by (t_1, ϕ_1). It follows from equation (43) that $2\pi t_1 = \phi_1 T_{\rm K}(1 + \alpha)$ for the prograde clock and $2\pi t_1 = (2\pi - \phi_1) T_{\rm K}(1 - \alpha)$ for the retrograde clock. Thus $\phi_1 = \pi(1 - \alpha)$ and $t_1 = T_{\rm K}(1 - \alpha^2)/2$. Moreover, $\tau_+^2(\phi_1) - \tau_-^2(\phi_1) = \alpha T_{\rm K}^2(\alpha^2 + 3R_g/2r)$, which is negligibly small for clocks in orbit about astronomical sources in the solar system; in fact, $\tau_+(\phi_1) - \tau_-(\phi_1) \sim O(c^{-4})$. The next time the clocks meet is at (t_2, ϕ_2), which bears the same relationship to (t_1, ϕ_1) as (t_1, ϕ_1) to (t_0, ϕ_0); therefore, $\phi_2 = 2\pi(1 - \alpha)$ and $t_2 = T_{\rm K}(1 - \alpha^2)$. In general, the nth time the clocks meet is at (t_n, ϕ_n) with $\phi_n = n\pi(1 - \alpha)$ modulo 2π and $t_n = nT_{\rm K}(1 - \alpha^2)/2$.

Consider now the behavior of the diametrical line joining these events to the origin of the spatial coordinates. For $a = 0$, i.e. in the Schwarzschild case, this line is fixed as the clocks repeatedly meet at two diametrically opposite points. However, for $a \neq 0$ the line precesses in the opposite sense as the rotation of the source with the precession frequency given approximately by (cf. Fig. 1)

$$\frac{\pi - \phi_1}{\tau_+(\phi_1)} = \frac{GJ}{c^2 r^3} + O(c^{-4}), \tag{45}$$

which at this order is in agreement with the precession frequency of an ideal torque-free gyroscope that is fixed in the equatorial plane of the Kerr source [28].

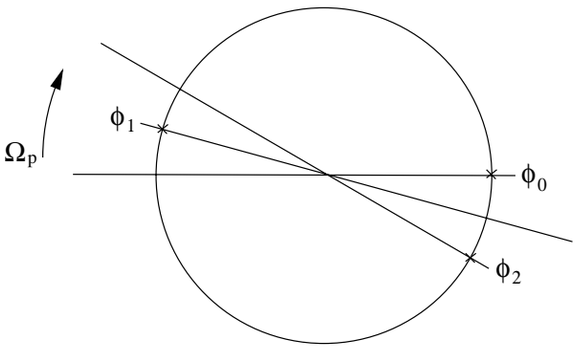


Fig. 1. Gravitomagnetic precession of the diametrical lines indicating the points $(\phi_0, \phi_1, \phi_2, \dots)$ at which the clocks would meet.

Our treatment of the clock effect has been limited thus far to circular orbits in the equatorial plane of the source. Off this plane, even circular orbits are not generally closed and the discussion of the clock effect as well as its intimate connection with the gravitomagnetic gyroscope precession becomes more complicated. In fact, the clock effect can be extended to such orbits using the notion of *azimuthal closure* [26].

Finally, it should be mentioned that the general motion of an ideal pointlike torque-free gyroscope in the Kerr field would, in accordance with equation (42), involve Thomas and Fokker precessions as well as a complicated gravitomagnetic motion that consists of both precession and nutation. Indeed, the notion of *relativistic nutation* has been introduced in the post-Schwarzschild approximation scheme in order to describe the nutational part of the gravitomagnetic spin motion [29]. The complex gravitomagnetic spin motion reduces to the simple (Schiff) precession in the lowest post-Newtonian order.

4 Clock Effect in the PPN Approximation

In view of the possibility of detecting the gravitomagnetic clock effect via spaceborne clocks, it is interesting to develop the theory of the clock effect within the parametrized post-Newtonian (PPN) framework. The PPN formalism contains a set of parameters that characterize different metric theories of gravitation in the post-Newtonian approximation. The general form of the PPN metric is described in [30,31]; it includes theories and effects that are not of primary interest for our treatment of the clock effect. Therefore we will start from a simplified PPN metric of the form

$$g_{00} = -1 + 2U - 2\beta U^2 \,, \tag{46}$$

$$g_{0i} = -\frac{1}{4}(4\gamma + 4 + \alpha_1) H_i^* \,, \tag{47}$$

$$g_{ij} = (1 + 2\gamma U)\delta_{ij} \,, \tag{48}$$

which describes a rotating body in an underlying Cartesian coordinate system $x^\mu = (ct, \boldsymbol{\varrho})$ with $\boldsymbol{\varrho} = (x, y, z)$. In the following, we will use spherical coordinates (ϱ, θ, ϕ); the isotropic radial coordinate ϱ should not be confused with the Schwarzschild radial coordinate r. In general relativity, the PPN parameters α_1, β and γ are given by $\alpha_1 = 0$ and $\beta = \gamma = 1$.

The PPN metric (46)-(48) is restricted to theories that exhibit conservation laws for total momentum and ignores the Whitehead and preferred-frame effects [32]. We assume that the gravitational source is a uniformly rotating and nearly spherical body that is symmetric about the axis of rotation (i.e. the z-axis). We are interested in the exterior gravitational field of the source when its center of mass is at the origin of spatial coordinates. The gravitoelectric potential $U(\varrho, \theta)$ is given in this case by [12]

$$U(\varrho,\theta) = \frac{GM}{c^2\varrho}\Big[1 + \sum_{n=2}^{\infty} J_n \Big(\frac{\varrho_e}{\varrho}\Big)^n P_n(\cos\theta)\Big]\,, \tag{49}$$

where ϱ_e is the equatorial radius of the source, M is in effect the asymptotically measured mass of the source, P_n is the Legendre polynomial of degree n and

$$J_n := \frac{1}{M\varrho_e^n}\int \mu(\varrho',\theta')\varrho'^n P_n(\cos\theta')\, d^3\varrho'\,. \tag{50}$$

Here μ denotes the effective mass-energy density of the source. In a similar way the multipole expansion of the gravitomagnetic vector potential H_i^* can be expressed as [12]

$$H_i^*(\varrho,\theta) = \frac{G(\boldsymbol{J}\times\boldsymbol{\varrho})_i}{c^3\varrho^3}\Big[1 + \sum_{n=1}^{\infty} K_n \Big(\frac{\varrho_e}{\varrho}\Big)^n P'_{n+1}(\cos\theta)\Big]\,, \tag{51}$$

where $\boldsymbol{J} = J\hat{\mathbf{z}}$ is in effect the asymptotically measured angular momentum of the source and $P'_n(x) = dP_n(x)/dx$. Here

$$K_n := \frac{2}{2n+3}\frac{M\varrho_e^2}{J}(L_n + J_{n+2}) \tag{52}$$

and

$$L_n := \frac{1}{M\varrho_e^{n+2}}\int \mu(\varrho',\theta')\varrho'^{n+2} P_n(\cos\theta') d^3\varrho'\,. \tag{53}$$

The derivation of the clock effect involves the computation of the proper time τ over a complete azimuthal cycle along geodesic orbits about the source. For simplicity, we limit our discussion to circular geodesic orbits in the equatorial plane, i.e. $\varrho =$ constant and $\theta = \pi/2$.

The radial geodesic equation, corresponding to a circular orbit in the equatorial plane, is given by

$$\Gamma^{\varrho}_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0\,, \tag{54}$$

which can be written as

$$\Big(\frac{c\,dt}{d\phi}\Big)^2+2\Big(\frac{c\,dt}{d\phi}\Big)\frac{\Gamma^{\varrho}_{0\phi}}{\Gamma^{\varrho}_{00}}+\frac{\Gamma^{\varrho}_{\phi\phi}}{\Gamma^{\varrho}_{00}}=0\,. \tag{55}$$

It is straightforward to show that $\Gamma^{\varrho}_{0\phi}/\Gamma^{\varrho}_{00}=g_{0\phi,\varrho}/g_{00,\varrho}$ and $\Gamma^{\varrho}_{\phi\phi}/\Gamma^{\varrho}_{00}=g_{\phi\phi,\varrho}/g_{00,\varrho}$. Using equations (46)-(48), we find that

$$g_{0\phi}=-\frac{1}{4}\Big(4\gamma+4+\alpha_1\Big)H(\varrho,\theta)\sin^2\theta\,, \tag{56}$$

where

$$H(\varrho,\theta)=\frac{GJ}{c^3\varrho}\Big[1+\sum_{n=1}^{\infty}K_n\Big(\frac{\varrho_e}{\varrho}\Big)^n P'_{n+1}(\cos\theta)\Big]\,, \tag{57}$$

and $g_{\phi\phi}=\varrho^2(1+2\gamma U)\sin^2\theta$. The solution of equation (55) can then be written as

$$\begin{aligned}\frac{dt}{d\phi}=&\pm\Big|\frac{c^2}{\varrho}U_{,\varrho}\Big|^{-\frac{1}{2}}\Big[1+(\beta+\gamma)U+\frac{1}{2}\gamma\varrho U_{,\varrho}\Big]\\&+\frac{1}{8c}\Big(4\gamma+4+\alpha_1\Big)\frac{H_{,\varrho}}{U_{,\varrho}}+O(c^{-3})\,.\end{aligned} \tag{58}$$

It follows from the PPN metric $-c^2d\tau^2=c^2g_{00}dt^2+2cg_{0\phi}dtd\phi+g_{\phi\phi}d\phi^2$ that

$$\Big(\frac{d\tau}{d\phi}\Big)^2=(1-2U)\Big(\frac{dt}{d\phi}\Big)^2-\frac{1}{c^2}\varrho^2+O(c^{-3})\,. \tag{59}$$

Using equation (58), we find after some algebra that

$$\begin{aligned}\frac{d\tau}{d\phi}=&\pm\Big|\frac{c^2}{\varrho}U_{,\varrho}\Big|^{-\frac{1}{2}}\Big[1+(\beta+\gamma-1)U+\frac{1}{2}\varrho(\gamma U_{,\varrho}-|U_{,\varrho}|)\Big]\\&+\frac{1}{8c}\Big(4\gamma+4+\alpha_1\Big)\frac{H_{,\varrho}}{U_{,\varrho}}+O(c^{-3})\,.\end{aligned} \tag{60}$$

Integration of this equation immediately yields $\tau_{\pm}$; hence,

$$\tau_+-\tau_-=\frac{\pi}{2c}\Big(4\gamma+4+\alpha_1\Big)\frac{H_{,\varrho}}{U_{,\varrho}}+O(c^{-3}) \tag{61}$$

gives the gravitomagnetic clock effect within the restricted PPN framework adopted here. The explicit dependence of the gravitomagnetic clock effect on the PPN parameters is through the proportionality factor of $(4\gamma+4+\alpha_1)$; in fact, the clock effect has this feature in common with other main gravitomagnetic effects [32].

It is interesting to note that in general relativity the gravitomagnetic clock effect in the post-Newtonian approximation is given by

$$\tau_+-\tau_-\approx 4\pi\frac{J}{Mc^2}\Big[1+\Big(\frac{3}{2}J_2-\frac{9}{2}K_2\Big)\frac{\varrho_e^2}{\varrho^2}\Big]\,, \tag{62}$$

when the source is assumed to be symmetric about the equatorial plane and all moments higher than the quadrupole are neglected. Using data given in [12], we find that for the Earth $J_2 \approx -10^{-3}$ and $K_2 \approx -10^{-3}$, so that $(3/2)J_2 - (9/2)K_2 \approx 3 \times 10^{-3}$ gives the relative contribution of the oblateness of the Earth to the clock effect for a near-Earth equatorial orbit.

5 Detection of the Gravitomagnetic Temporal Structure

According to Eq. (43) and the discussion following it, the orbital motion of free clocks around a rotating mass gives rise to the gravitomagnetic clock effect which shows up in the difference between the proper orbital periods of co- and counter-orbiting clocks. This is given by $4\pi a/c$ for equatorial trajectories. Inserting the specific angular momentum of the Earth ($a \sim 3$ m) yields an amazingly "large" value of $\tau_+ - \tau_- \sim 10^{-7}$s.

Despite this seemingly large effect, the actual measurement of this time difference encounters severe practical difficulties. Since the two clocks are assumed to move along opposite but identical orbits, their Kepler periods exactly cancel upon forming the difference $\tau_+ - \tau_-$, thereby revealing the gravitomagnetic clock effect. In reality, however, clocks cannot be injected into identical trajectories and the resulting difference in the Kepler periods will readily exceed the time difference induced by the rotation of the Earth. Since for near-Earth orbits a radial separation of 0.1 mm of the clocks involves a time difference in the Kepler periods of the same order of magnitude as the gravitomagnetic clock effect, the position of the clocks has to be known at the submillimeter level in order to filter the effect which is caused by the rotation of the Earth out of the data. Similary, as the satellite moves just under 1 mm within 10^{-7}s along its track (or $\sim 10^{-2}$ milliarcseconds in angular distance), the azimuthal position has to be known at the same accuracy as the radial one. On the other hand, the gravitomagnetic time difference accumulates with the number of revolutions and after hundreds or thousands of periods a knowledge of the position of the clocks at the centimeter level will be sufficient to overcome the difference in the Kepler periods.

Another difficulty arises from the determination of all the forces that act on the satellites carrying the clocks. Since the period of revolution for orbits of $\sim 10^3$ km altitude is of the order of $\sim 10^4$ s, accelerations as weak as 10^{-12} m/s^2 will already cover the gravitomagnetic clock effect. Among these forces, gravitational perturbations due to the nonsphericity of the Earth, solid and ocean Earth tides as well as the interaction with the Sun, Moon and planets will cause the most significant deviations from an ideal orbit. Depending on the altitude of the satellites, the atmospheric drag effect can also considerably change the shape of the orbit. Moreover, this latter effect is quite difficult to model because it strongly depends on the atmospheric density which is not only correlated to the orbital height but also subject to temporal variations. Other non-gravitational perturbations like solar and terrestrial radiation pressure, thermal thrust, charged particle

drag etc. must also be taken into account despite their less distinct influence, since they likely induce accelerations in excess of 10^{-12} m/s^2.

In practice, the effect of all these perturbations will be modeled by determining a precise orbit based on the actual spacecraft observations. This will be accomplished by generating an orbital trajectory following Newton's equations of motion and by including all perturbing forces acting on the satellite, using the most accurate models available. In the next step, this predicted orbit has to be best fitted to the one observed, where some force parameters may be solved for during the orbit adjustment process in order to obtain an improved or tailored force model for the specific mission. From the resulting precise orbit the effects of the individual perturbations are removed step by step thus yielding a quasi-Keplerian orbit, but still carrying the signatures of the relativistic effects. Finally, a comparison with the corresponding clock predictions for an appropriate synthetic orbit is performed which is expected to confirm the clock effect being investigated.

Therefore, in order to meet the very stringent conditions for the observation of the clock effect, many tiny perturbing sources have to be considered and investigated that are usually absent in most of the present orbit determination systems.

6 Quantum Origin of Inertia

The sign of the gravitomagnetic clock effect has a remarkable consequence that will be elucidated in this section. It follows from $t_+ > t_-$ and $\tau_+ > \tau_-$ that the uniform motion of the prograde clock is slower than that of the retrograde clock. Thus in comparison with motion around a nonrotating mass, a rotating mass would "drag" free test particles along such that it would take longer (shorter) to go once around it on an equatorial circular orbit in the prograde (retrograde) direction. We may call this circumstance virtual "inertial antidragging," since it is the exact opposite of what would be expected on the basis of the so-called "inertial dragging" [33]. In fact, as Fig. 2 clearly demonstrates, for a given r (i.e. fixed orbital radius), the faster the Kerr source spins, the slower the prograde motion and the faster the retrograde motion.

Rotational or translational inertial dragging refers to the circumstance that an accelerating mass would somehow induce acceleration in the same sense in nearby masses as a consequence of the so-called "Mach's principle". The clock effect indicates that precisely the opposite situation is predicted for rotational motion in the equatorial plane by the general theory of relativity. Translational inertial dragging has been discussed by a number of authors [34]; again, such notions are foreign to the standard geometric interpretation of general relativity [35]. In general relativity, accelerated motion is absolute in the sense that it is nonrelative. Thus the term "absolute" as employed here only signifies the opposite of the term "relative" and is devoid of any metaphysical connotations. The gravitomagnetic clock effect and the gyroscope precession indicate the absolute rotation of the source. That is, a direct gravitomagnetic verification of Einstein's

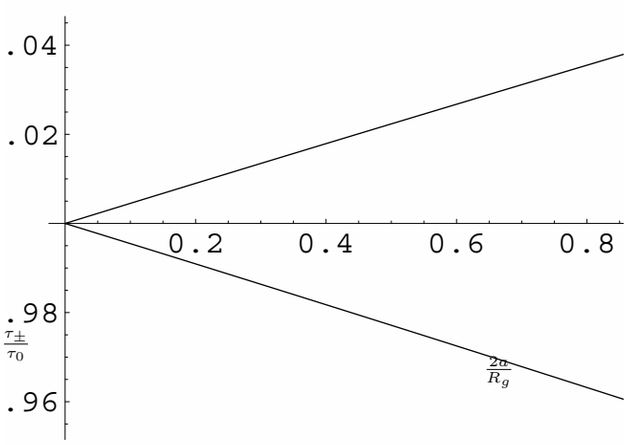


Fig. 2. Plot of the clock rate versus the rotation of the source. Imagine an ensemble of Kerr fields with the same mass but different angular momenta. For a fixed stable circular geodesic orbit with "radius" r, $\tau_+(\tau_-)$ monotonically increases (decreases) over the ensemble with increasing angular momentum of the source. Stable orbits of this type occur for $r \geq r_\pm$, where $r_\pm - 3R_g \pm 4a\ (2R_g/r_\pm)^{1/2} = 3a^2/r_\pm$. Let us note that $r_- \geq r_+$ and the equality occurs for $a = 0$ and $r_\pm = 3R_g$; moreover, $r_\pm/a \to (3)^{1/2}$ as $2a/R_g \to \infty$. The graph illustrates the behavior of $\tau_\pm$ for an ensemble of Kerr black holes with $2a \leq R_g$; for $2a = R_g$, $r_+ = a$ and $r_- = 9a$. In the graph, $\tau_0 = \tau_\pm(a = 0)$ and the radius r is chosen to be $5R_g$.

theory of gravitation—e.g. via NASA's GP-B—would constitute observational proof that the rotation of the Earth is absolute and not merely relative to the distant matter of the universe [36].

Mach's profound analysis of the foundations of Newtonian mechanics occasioned a thorough re-examination of the basic classical notions of space, time and motion that had been prevalent since Newton provided a rational basis for the Copernican revolution and Kepler's laws of planetary motion. This re-evaluation culminated in Einstein's theory of *general relativity*. It is therefore of great importance to recognize that general relativity—which agrees with all experimental data to date—does not contain the idea of relativity of arbitrary motion. That is, this concept — which was so crucial in the historical development of Einstein's theory — is absent in the standard geometric interpretation of general relativity in the sense that it is neither a part of the foundations of the theory nor follows from it. The re-emergence of absolute motion may be taken to mean that general relativity is still not completely devoid of certain "metaphysical" elements. Should general relativity therefore be modified or abandoned in favor of a theory based on the relativity of arbitrary motion? To do so would be unwise. One

should recognize instead that physics has progressed far beyond the early days of relativity theory and the observational successes of general relativity must now be integrated within a quantum framework that involves the vacuum state of microphysics as well as the rest frame of the cosmic microwave background radiation.

Mach noted that in Newtonian mechanics, the intrinsic state of a classical particle characterized by its mass m has no direct connection with the extrinsic state of the particle characterized by its position and velocity $(\mathbf{x}, \mathbf{v})$ in absolute space and time. Hence the same extrinsic state can be occupied by other masses comoving with the particle. Thus an observer can change its perspective by comoving with each particle in turn. In Newtonian mechanics, the particles are thus "placed" on the absolute space and time continuum and remain external to it. On the other hand, classical particles are "connected" to each other via interactions such as gravity and electromagnetism. Mach therefore concluded that only the motion of a particle relative to other particles should have ultimate physical significance. Mach's basic analysis has been restated in modern form in [37].

In classical physics, motion takes place via classical particles as well as electromagnetic waves. It appears that Mach did not extend his analysis of classical particle motion to electromagnetic wave propagation; in this connection, the issues that arise in the examination of the historical record are briefly mentioned in the Appendix. Let us therefore apply Mach's argument to the motion of electromagnetic waves. The intrinsic aspects of the wave are its amplitude, period, wavelength and polarization, which therefore characterize its intrinsic state. The extrinsic state of the wave is given by its wave function $\Psi(t, \mathbf{x})$ in absolute time and space, and we note that the wave's intrinsic state is directly related to its extrinsic state, i.e. electric and magnetic field components, since the former cannot be defined independently of the latter. The conclusion is that the motion of classical electromagnetic waves is absolute, i.e. nonrelative.

Classical motion can be either relative or absolute. In Einstein's discussion of the so-called "Mach's principle," only "ponderable" masses are considered [33], whereas classical motion occurs via classical particles as well as electromagnetic waves. It is natural to think of the motion of classical particles (i.e. "ponderable" masses) as relative, since one can change one's perspective by moving with each mass in turn. In the same sense, the motion of electromagnetic waves must be considered absolute due to its observer-independent status. The development of these simple notions taking due account of wave-particle duality leads to the principle of complementarity of absolute and relative motion [38]. In this connection, let us note that Mach's analysis of classical particle motion may be restated in terms of the complete kinematic independence of the absolute position $\mathbf{x}$ of a particle from its momentum $\mathbf{p} = m\mathbf{v}$. In contrast, quantum kinematics can be consistently formulated only by imposing the fundamental quantum condition on the operators characterizing the position and momentum of a particle in absolute space and time, i.e. $[\hat{x}_j, \hat{p}_k] = i\hbar\delta_{jk}$. For instance, in the nonrelativistic motion of a free particle in the Heisenberg picture $\hat{\mathbf{p}} = m\hat{\mathbf{v}}$ and

$[\hat{x}_j, \hat{v}_k] = i\hbar m^{-1}\delta_{jk}$. Thus in contrast to the situation in classical mechanics, the mass of a particle is related to its position and velocity in quantum mechanics due to the fact that the particle has wave characteristics as well. This idea naturally extends to the specific orbital angular momentum of the particle, $\hat{l}_i = \epsilon_{ijk}\hat{x}_j\hat{v}_k$, so that $[\hat{l}_j, \hat{l}_k] = i\hbar m^{-1}\epsilon_{jkn}\hat{l}_n$. In the limit of an infinitely massive particle, the connection disappears and the position and velocity commute; that is, one recovers classical mechanics when the system is so massive that the perturbation due to an act of observation on the system is negligible and the system therefore behaves classically.

Mach's argument involves classical quantities (c-numbers), whereas the quantum condition involves operators (q-numbers); nevertheless, the quantum condition implies that the intrinsic and extrinsic aspects of the particle are directly related through Planck's constant. For instance, in the Schrödinger picture the extrinsic state of the particle is given by the wave function $\Psi(t, \mathbf{x})$ and the Schrödinger equation for Ψ involves m, which characterizes the intrinsic state of the particle in Mach's analysis. The relationship under discussion here is not merely formal but can be verified observationally. In fact, this kinematic connection is particularly well illustrated by the example of a free particle passing through a slit. The resulting diffraction angle is inversely proportional to the mass of the particle, so that the diffraction is absent in the limit of large mass and the particle behaves classically. To the extent that classical mechanics can be thought of as a limiting form of quantum mechanics, the epistemological problem of Newtonian mechanics — so clearly brought out by Mach — disappears. That is, the problem of the origin of inertia is resolved through the wave nature of matter.

Thus far the inertial mass of the particle has provided the quantum connection to the inertial reference frames of Newtonian mechanics. The invariance group of Minkowski spacetime is the Poincaré group whose irreducible unitary representations can be described in terms of mass and spin. Thus in the relativistic theory the inertial properties of a particle are characterized by mass and spin. The inertial properties of intrinsic spin have been discussed in [39].

Inertia has its origin in the fact that matter is intrinsically extended in space and time and through this nonlocality inertial reference frames can be "recognized"; then, a physical system tends to preserve its state with respect to such frames. This is beautifully illustrated by experiments involving macroscopic quantum systems that have phase coherence, such as the recent demonstration of Earth's absolute rotation via superfluid He^4 [40]. The quantum aspects of the origin of inertia are further developed in [41].

7 Discussion

In this paper we have examined some of the main theoretical aspects of gravitomagnetism in general relativity. The influence of the proper rotation of a source on the spacetime structure can be studied in various ways. Attention has been focused here on certain features of the gravitomagnetic clock effect and its relation

with the gravitomagnetic gyro precession. However, other approaches exist and should be mentioned. The detection of the gravitomagnetic field of the Earth via the Lense-Thirring precession of satellite orbits has been investigated by Ciufolini *et al.* [42]. Moreover, gravitomagnetic effects in the spacetime curvature can be measured in principle using gravity gradiometry [43]. In this connection, it is interesting to note that gravity gradiometers of high sensitivity that are based on atomic interferometry are being developed for space applications [44,45].

Appendix: Mach and the Absolute Motion of Light

Newton's introduction of the concepts of absolute space and time was truly revolutionary in his day and allowed him to formulate the basic classical laws of particle motion. Leibniz [46] and Berkeley [47] criticized the notions of absolute space and absolute time and emphasized instead the idea of relativity of all motion. Later, however, Maxwell [48] relied on absolute space and time for his fundamental extension of the Newtonian ideas of motion to electromagnetic field propagation. On the other hand, Mach revived the principle of relativity of all motion on the basis of a profound analysis of the foundations of classical mechanics [49]. Mach's work played a significant role in Einstein's development of the theory of relativity [50].

Mach's deep physical treatment of the relativity of classical particle motion was motivated by his epistemological stance on the relativity of all measurement. According to Mach, the result of a measurement is the establishment of a *relation* and not of "absolute" notions, since in Mach's view the latter refer to processes or objects that are not empirically verifiable in principle [51]. Mach's analysis of the relativity of particle motion in his great work on classical mechanics [49] was not extended to electromagnetic wave motion in his later work on physical optics [52]. In this book, Mach discussed the wave theory of light as well as the speed of light; however, he apparently made no attempt to put these in the context of his epistemological stance on the relativity of all motion. There is no evidence that Mach ever wavered in his opposition to absolute motion [53]. However, a number of Mach's contemporaries pointed out the absolute character of the constancy of the speed of light and were troubled by the fact that this aspect of the relativity theory was in conflict with the relativity of all motion. Among the physicists and philosophers who raised such doubts about the epistemological stance of *the theory of relativity* one can mention Friedrich Adler, Hugo Dingler, Philipp Frank, Anton Lampa and Joseph Petzoldt. Although it is claimed in the book of Blackmore [53] that Mach rejected the principle of the constancy of the speed of light because it was in contradiction to his phenomenalistic epistemology due to its constant validity independent of all sensations and conscious data, there is actually no evidence that Mach ever directly or indirectly commented on the constancy of the speed of light [54]. A historical analysis of the situation and the influence of these criticisms on Mach can be found in the monographs of Blackmore [53] and Wolters [54].

An exposition of the reasons for the supposed opposition to the theory of relativity based on epistemological considerations and experimental facts was promised to appear in a sequel to Mach's book on optics [52] in collaboration with his son Ludwig, but this was never published. Although the preface to the "Optics" is generally regarded as the most obvious evidence of Mach's reluctance to accept the relativity theory, there is every reason to believe that it was written by Ludwig only after the death of his father and expresses Ludwig's opinion on the theory of relativity, despite the fact that Ernst Mach is stated to be the author of this preface. More on this conjecture can be found in [54].

Finally, the position of Mach vis-à-vis the theory of relativity is also discussed in the paper of Thiele [55].

References

1. G. Holzmüller: *Z. Math. Phys.* **15**, 69 (1870).
2. F. Tisserand: *Compt. Rend.* **75**, 760 (1872); **110**, 313 (1890).
3. O. Heaviside: *Electromagnetic Theory* (The Electrician Printing and Publishing Co., London, 1894); J.D. North: *The Measure of the Universe* (Clarendon Press, Oxford, 1965), ch. 3; R. Anderson, H.R. Bilger and G.E. Stedman: *Am. J. Phys.* **62**, 975 (1994).
4. H. Thirring: *Phys. Z.* **19**, 33 (1918); **22**, 29 (1921); J. Lense and H. Thirring: *Phys. Z.* **19**, 156 (1918).
5. B. Mashhoon, F.W. Hehl and D.S. Theiss: *Gen. Rel. Grav.* **16**, 711 (1984).
6. A. Matte: *Canadian J. Math.* **5**, 1 (1953); L. Bel, *C. R. Acad. Sci., Paris* **247** 1094, (1958); R. Debever: *Bull. Soc. Math. Belg.* **10**, 12 (1958); R.T. Jantzen, P. Carini and D. Bini: *Ann. Phys. (NY)* **215**, 1 (1992); W.B. Bonnor: *Class. Quantum Grav.* **12**, 1483 (1995); R. Maartens and B.A. Bassett: *Class. Quantum Grav.* **15**, 705 (1998).
7. B. Mashhoon, J.C. McClune and H. Quevedo: *Phys. Lett. A* **231**, 47 (1997); *Class. Quantum Grav.* **16**, 1137 (1999).
8. J. Larmor: *Phil. Mag.* **44**, 503 (1897).
9. B. Mashhoon: *Phys. Lett. A* **173**, 347 (1993).
10. C.W.F. Everitt *et al.*: in *Near Zero: Festschrift for William M. Fairbank*, ed. C.W.F. Everitt (Freeman, San Francisco, 1986).
11. V. B. Braginsky, C.M. Caves and K.S. Thorne: *Phys. Rev. D* **15**, 2047 (1977).
12. P. Teyssandier: *Phys. Rev. D* **16**, 946 (1977); *ibid.* **18**, 1037 (1978).
13. B. Mashhoon: *Gen. Rel. Grav.* **31**, 681 (1999).
14. J.P. Jacobs, W.M. Klipstein, S.K. Lamoreaux, B.R. Heckel and E.N. Fortson: *Phys. Rev.* **A 52**, 3521 (1995).
15. R.L. Mallet: "Optically induced gravitational frame-dragging," preprint (1999).
16. N. Bohr and L. Rosenfeld: *Det. Kgl. dansk. Vid. Selakab.* **12**, No. 8 (1933).
17. S. Chandrasekhar and S. Detweiler: *Proc. R. Soc. London A* **344**, 441 (1975).
18. B. Mashhoon: in Hu Ning (ed.): *Proc. Third Marcel Grossmann Meeting on General Relativity* (Shanghai, 1982), (Science Press and North-Holland, Amsterdam 1983), pp. 599.
19. Hongya Liu and B. Mashhoon: *Class. Quantum Grav.* **13**, 233 (1996).
20. V.P. Frolov and I.D. Novikov: *Black Hole Physics* (Kluwer Academic Publishers, Dordrecht, 1998).

21. B. Mashhoon: *Phys. Rev. D* **31**, 290 (1985).
22. J.M. Cohen and B. Mashhoon: *Phys. Lett. A* **181**, 353 (1993).
23. B. Mashhoon: in L. Maleki (ed.): *Proceedings of the Workshop on the Scientific Applications of Clocks in Space* (JPL Publication 97-15, NASA, 1997), p.41.
24. F. Gronwald, E. Gruber, H. Lichtenegger and R.A. Puntigam: in *Proc. Alpbach School on Fundamental Physics in Space* (ESA, SP-420, 1997), p.29.
25. H.I.M. Lichtenegger, F. Gronwald and B. Mashhoon: *Adv. Space Res.*, in press (1999); gr-qc/9808017.
26. B. Mashhoon, F. Gronwald and D.S. Theiss: *Ann. Physik* **8**, 135 (1999); gr-qc/9804008.
27. W.B. Bonnor and B.R. Steadman: *Class. Quantum Grav.* **16** (1999) 1853; O. Semerák, *ibid.* **16**, 3769 (1999).
28. N.V. Mitskevich and I. Pulido Garcia: *Sov. Phys. Dokl.* **15**, 591 (1970); V. Ferrari and B. Mashhoon: *Phys. Rev. D* **30**, 295 (1984), see Ref. 21.
29. B. Mashhoon and D.S. Theiss: *Nuovo Cimento B* **106**, 545 (1991).
30. C.W. Misner, K.S. Thorne and J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973).
31. C.M. Will, *Theory and Experiment in Gravitational Physics* (Cambridge University Press, Cambridge 1993).
32. B. Mashhoon, H.J. Paik and C.M. Will: *Phys. Rev. D* **39**, 2825 (1989).
33. A. Einstein: *The Meaning of Relativity* (Princeton University Press, Princeton, 1950), pp. 100-103; W. Davidson: *Mon. Not. Roy. Astron. Soc.* **117**, 212 (1957); J.D. Nightingale: *Am. J. Phys.* **45**, 376 (1977).
34. Ø. Grøn and E. Eriksen: *Gen. Rel. Grav.* **21**, 105 (1989); D. Lynden-Bell, J. Bičák and J. Katz: *Ann. Phys. (NY)* **271**, 1 (1999).
35. B. Mashhoon: *Nuovo Cimento B* **109**, 187 (1994).
36. B. Mashhoon: in B.L. Hu and T.A. Jacobson (eds.): *Directions in General Relativity: Papers in Honor of Dieter Brill* (Cambridge University Press, Cambridge 1993), p. 182.
37. Hongya Liu and B. Mashhoon: *Ann. Physik* **4**, 565 (1995).
38. B. Mashhoon: *Phys. Lett.* **A 126**, 393 (1988).
39. B. Mashhoon: *Phys. Lett.* **A 198**, 9 (1995).
40. K. Schwab, N. Bruckner and R.E. Packard: *Nature* **386**, 585 (1997).
41. B. Mashhoon: *Found. Phys. Lett.* **6**, 545 (1993).
42. I. Ciufolini, E. Pavlis, F. Chieppa, E. Fernandesvieira and J. Perezmercader: *Science* **279**, 2100 (1998).
43. B. Mashhoon and D.S. Theiss: *Phys. Rev. Lett.* **49**, 1542 (1982); B. Mashhoon, H.J. Paik and C.M. Will: *Phys. Rev.* **D 39**, 2825 (1989).
44. M.J. Snadden, J.M. McGuirk, P. Bouyer, K.G. Haritos and M.A. Kasevich: *Phys. Rev. Lett.* **81**, 971 (1998).
45. Ch. J. Bordé: this volume, p. 403; *Phys. Lett.* **A 140**, 10 (1989).
46. G. Leibniz: *Leibniz: Selections*, edited by P.P. Wiener (Charles Scribner's Sons, New York, 1951), p. 222.
47. G. Berkeley: *Berkeley's Philosophical Writings*, edited by D.M. Armstrong (Collier-MacMillan, New York, 1965), p. 250.
48. J.C. Maxwell: *Matter and Motion* (Dover, New York, 1952).
49. E. Mach: *The Science of Mechanics* (Open Court, La Salle, 1960).
50. A. Einstein: *The Meaning of Relativity* (Princeton University Press, Princeton, 1950).
51. E. Mach: *Space and Geometry* (Open Court, La Salle, 1960).

52. E. Mach: *The Principles of Physical Optics* (Dover, New York, 1953).
53. J.T. Blackmore: *Ernst Mach* (University of California Press, Berkeley, 1972).
54. G. Wolters: *Mach I, Mach II, Einstein und die Relativitätstheorie* (Walter de Gruyter, Berlin, 1987).
55. J. Thiele: "Bemerkungen zu einer Äusserung im Vorwort der '*Optik*' von Ernst Mach," *Schriftenreihe für Geschichte der Naturwissenschaften, Technik und Medizin* **2**, 10 (1965).

Spinning Relativistic Particles in External Fields

Iosip B. Khriplovich

Budker Institute of Nuclear Physics, 630090 Novosibirsk, Russia,
and Novosibirsk University

Abstract. The motion of spinning relativistic particles in external electromagnetic and gravitational fields is considered. The noncovariant spin formalism is crucial for the correct description of the influence of the spin on the particle trajectory. It is shown that the true coordinate of a relativistic spinning particle is the naive, common coordinate $\boldsymbol{r}$. A simple derivation is presented for the gravitational interaction of first order in spin, for a relativistic particle. The equations of motion obtained for a relativistic spinning particle in an external gravitational field differ essentially from the Papapetrou equations. Effects of higher order in spin are discussed, including the gravimagnetic moment, a special spin effect in general relativity. We consider also the contributions of the spin interactions of first and second order to the gravitational radiation of compact binary stars.

1 Introduction

The problem of the motion of a particle with internal angular momentum (spin) in an external field consists of two parts: the description of the spin precession and accounting for the spin influence on the trajectory of motion. To lowest nonvanishing order in c^{-2} the complete solution for the case of an external electromagnetic field was given more than 70 years ago [1]. The gyroscope precession in a centrally symmetric gravitational field had been considered to the same approximation even earlier [2]. Then, much later the spin precession was investigated in the case of the gravitational spin–spin interaction [3]. The fully relativistic problem of the spin precession in an external electromagnetic field was also solved more than 70 years ago [4] and then, in a more convenient formalism, using the covariant vector of spin, in [5].

The situation with the second part of the problem, which refers to how the spin influences the trajectory, is different. Covariant equations of motion for a relativistic spinning particle in an electromagnetic field were written in the same paper [4] and for the case of a gravitational field in [6]. These equations have been discussed repeatedly from various points of view in numerous papers (see, e.g., [7-18]). The problem of the influence of the spin on the trajectory of a particle in external fields is not only of purely theoretical interest. In particular, it attracts attention being related to the description of the motion of relativistic particles in accelerators [19] (see also recent review [20]).

In fact, it is far from being obvious whether one can observe in practice the discussed spin corrections to the equations of motion of elementary particles as, for instance, an electron or proton. According to the well–known argument by

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 109–128, 2001.

Bohr (see [21]), an additional Lorentz force due to the finite size of the wave packet of a charged particle and to the uncertainty relation exceeds the corresponding component of the Stern–Gerlach force. However, this argument by itself does not exclude in principle the possibility to observe a common Stern–Gerlach effect, even a small one, in the presence of a larger background due to the uncertainty relation. In particular, in a recent paper [22] this possibility is claimed to be supported by numerical calculations. Moreover, spin–dependent correlations certainly exist in differential cross sections of scattering processes. So, it was proposed long ago to separate charged particles of different polarizations through the spin interaction with external fields in a storage ring [23]. Though this proposal is being discussed rather actively (see review [20]), it is not clear up to now whether it is feasible technically.

There are however macroscopic objects for which internal rotation certainly influences their trajectories. We mean the motion of Kerr black holes in external gravitational fields. This problem is of importance in particular for the calculation of the gravitational radiation of binary stars. In this connection it was considered in [24-27]. However, when turning to these calculations, we found [28] that the equations of motion taking account of spin to the lowest nonvanishing order in c^{-2}, used in these papers, lead to results which differ from the well–known gravitational spin–orbit interaction even in the simpler case of an external field. The problem is essentially related to the correct definition of the center–of–mass coordinate. Moreover, it turned out that the widely used Papapetrou equations [6] also fail to reproduce in the same c^{-2} approximation the result for the gravitational spin–orbit interaction found in the classical work [2]. This discrepancy was pointed out long ago in [29]; however the explanation suggested in [29] does not appear satisfactory (see [28]).

The present talk is essentially based on recent work [28,30,31] where the equations of motion of a relativistic particle were derived with a noncovariant description of spin. These equations agree with well–known limiting cases. Though for an external electromagnetic field such equations (in the linear in spin approximation) have been obtained previously [19] (see also [20]), we would like to start with comments related to this approximation in electrodynamics.

2 Covariant and Noncovariant Equations of Motion of a Spinning Particle in an Electromagnetic Field

2.1 The Problem with the Covariant Equations of Motion

The interaction of spin with external electromagnetic field is described, up to terms on the order of c^{-2} included, by the well–known Hamiltonian (see, e.g., [32])

$$H = -\frac{eg}{2m}\boldsymbol{s}\cdot\boldsymbol{B} + \frac{e(g-1)}{2m^2}\boldsymbol{s}\cdot[\boldsymbol{p}\times\boldsymbol{E}]\,. \tag{1}$$

Here $\boldsymbol{B}$ and $\boldsymbol{E}$ are external magnetic and electric fields; e, m, $\boldsymbol{s}$, and $\boldsymbol{p}$ are the particle charge, mass, spin, and momentum, respectively; g is its gyromagnetic

ratio. Let us emphasize that the structure of the second (Thomas) term in this expression not only has been firmly established theoretically, but has also been confirmed with high accuracy experimentally, at any rate in atomic physics. To avoid misunderstandings, let us note that, generally speaking, the last term in formula (1) should be rewritten in a Hermitian form (see, e.g., [33]):

$$[\boldsymbol{p}\times\boldsymbol{E}]\to\frac{1}{2}\left([\boldsymbol{p}\times\boldsymbol{E}]-[\boldsymbol{E}\times\boldsymbol{p}]\right)=[\boldsymbol{p}\times\boldsymbol{E}]+\frac{i}{2}\,\boldsymbol{\nabla}\times\boldsymbol{E}\,.$$

We will be interested, however, in the semiclassical approximation mainly, when in the interaction linear in spin, field derivatives are neglected. (Besides, the correction with $\boldsymbol{\nabla}\times\boldsymbol{E}$ vanishes in the case of potential electric field considered in [32].)

Let us try to construct a covariant equation of motion accounting for spin, which would reproduce in the same approximation the force

$$f_m=\ \frac{eg}{2m}\boldsymbol{s}\boldsymbol{B}_{,m}+\frac{e(g-1)}{2m}\ \left(\frac{d}{dt}[\boldsymbol{E}\times\boldsymbol{s}\,]_m-\ \boldsymbol{s}[\boldsymbol{v}\times\boldsymbol{E}_{,m}\,]\right), \tag{2}$$

corresponding to the Hamiltonian (1) (here and below a comma with a subscript denotes a partial derivative). A covariant correction f^μ to the Lorentz force $eF^{\mu\nu}u_\nu$ should be linear in the tensor of spin $S_{\mu\nu}$ and in the gradient of the tensor of electromagnetic field $F_{\mu\nu,\lambda}$, it may depend also on the 4–velocity u^μ. Since $u^\mu u_\mu=1$, this correction must satisfy the condition $u_\mu f^\mu=0$. From the mentioned tensors one can construct only two independent structures meeting the last condition. The first,

$$\eta^{\mu\kappa}F_{\nu\lambda,\kappa}S^{\nu\lambda}\ -\ F_{\lambda\nu,\kappa}u^\kappa S^{\lambda\nu}u^\mu, \tag{3}$$

reduces in the c^{-2} approximation to

$$2\boldsymbol{s}(\boldsymbol{B}_{,m}-[\boldsymbol{v}\times\boldsymbol{E}_{,m}\,]), \tag{4}$$

and the second, v

$$u^\lambda F_{\lambda\nu,\kappa}u^\kappa S^{\nu\mu}, \tag{5}$$

reduces to

$$\frac{d}{dt}[\boldsymbol{s}\times\boldsymbol{E}]_{\,m}\,. \tag{6}$$

Let us note that possible structures with the contraction $F_{\nu\kappa,\lambda}S^{\kappa\lambda}$ reduce to (3) and (5), due to the Maxwell equations and the antisymmetry of $S_{\kappa\lambda}$.

Obviously, no linear combination of (4) and (6) can reproduce the correct expression (2) for the spin–dependent force. In a somewhat less general way this was shown in [28].

But why is it that the correct (in the c^{-2} approximation) formula (2) cannot be obtained from a covariant expression for the force? Obviously, one can easily reproduce by a linear combination of (4) and (6) those terms in (2) which are proportional to g. In other words, there is no problem to present in a covariant form

the terms which describe, so to say, direct interaction of a magnetic moment with external fields. It is the terms in (2) independent of g and corresponding to the Thomas precession which cannot be written covariantly. Certainly, the Thomas precession can be described beyond the c^{-2} approximation, for arbitrary velocities. But there are no reasons why this essentially noncovariant phenomenon should have a covariant description. This is the point.

2.2 What Is the Correct Definition of the Coordinate of a Spinning Particle?

It was noted in [28] that the covariant formalism can be reconciled with the correct results if the coordinate $\boldsymbol{x}$ entering the covariant equation is related to the usual one $\boldsymbol{r}$ in the c^{-2} approximation as follows:

$$\boldsymbol{x} \,=\, \boldsymbol{r} + \frac{1}{2m}\boldsymbol{s}\times\boldsymbol{v}. \tag{7}$$

The generalization of this substitution to the case of arbitrary velocities

$$\boldsymbol{x} \,=\, \boldsymbol{r} + \frac{\gamma}{m(\gamma+1)}\boldsymbol{s}\times\boldsymbol{v}, \quad \gamma = \frac{1}{\sqrt{1-v^2}}. \tag{8}$$

was pointed out in [20].

But why can the spin precession itself (as distinct from the spin influence on the trajectory) be described covariantly [4,5] without any concern for the coordinate definition? First of all, the covariant equations of spin precession

$$\frac{dS_\mu}{d\tau} \,=\, \frac{e}{2m}\left[gF_{\mu\nu}S^\nu - (g-2)u_\mu F_{\lambda\nu}u^\lambda S^\nu\right] \tag{9}$$

(here S_μ is the covariant 4–vector of spin) are written in the semiclassical approximation, i.e., the coordinate dependence of external fields is completely neglected. Second, equations (9) are linear and homogeneous in spin. So, even if one went here beyond the semiclassical approach, but stayed within the approximation linear in spin, the use of the usual coordinate $\boldsymbol{r}$, which differs from $\boldsymbol{x}$ in terms proportional to $\boldsymbol{s}$ only, would be completely legitimate.

Of course, the choice of the variable, $\boldsymbol{r}$ or $\boldsymbol{x}$, is by itself a matter of convention. But still, which of them is the true center–of–mass coordinate of a relativistic spinning body?

We note first of all that relation (7) is valid for a free particle as well. So, to answer the question, it is sufficient to consider a simple example of the free Dirac particle with the Hamiltonian

$$H \,=\, \boldsymbol{\alpha}\cdot\boldsymbol{p} \,+\, \beta\, m\,.$$

Here, the operator whose expectation value equals to $\boldsymbol{r}$, is not $\boldsymbol{r}$ itself, but [34]

$$\boldsymbol{x} = \boldsymbol{r} + \frac{i\beta\boldsymbol{\alpha}}{2\varepsilon} - \frac{i\beta(\boldsymbol{\alpha}\cdot\boldsymbol{p})\boldsymbol{p} + [\boldsymbol{\Sigma}\times\boldsymbol{p}]\,\varepsilon}{2\varepsilon^2(\varepsilon+m)}; \;\; \varepsilon = \sqrt{p^2+m^2}; \;\; \boldsymbol{\Sigma} = \frac{1}{2i}[\boldsymbol{\alpha}\times\boldsymbol{\alpha}]. \tag{10}$$

To lowest nonvanishing order in c^{-2} expression (10) reduces to

$$\boldsymbol{x} = \boldsymbol{r} + \frac{1}{2m}\,\boldsymbol{s}\times\boldsymbol{v}\,, \quad \boldsymbol{s} = \frac{1}{2}\boldsymbol{\sigma}\,, \tag{11}$$

which might prompt indeed substitution (7). However, under the Foldy–Wouthuysen (FW) transformation, which separates positive–energy states from negative–energy ones, the relativistic operator $\boldsymbol{x}$ goes over into mere $\boldsymbol{r}$. And the transition from the exact Dirac equation in an external field to its approximate form containing only the first–order correction in c^{-2}, is performed just by means of the FW transformation. Thus, in the resulting c^{-2} Hamiltonian the coordinate of a spinning electron is the same $\boldsymbol{r}$ as in the completely nonrelativistic case. Nobody makes substitution (7) in the Coulomb potential when treating the spin–orbit interaction in the hydrogen atom.

As to a classical particle, it is in fact a well–localized wave packet constructed from positive–energy states, i.e., it is properly described in the FW representation. Therefore, it is $\boldsymbol{r}$ which is the true coordinate of a classical relativistic spinning particle. The same conclusion is made in [35], starting from quite different arguments.

2.3 The Noncovariant Formalism

The correct equations of motion in an electromagnetic field including spin to first order are known for a fairly long time [19]. Though being fully relativistic, they are noncovariant and based on the initial physical definition of spin. According to this definition, spin is the 3–dimensional vector $\boldsymbol{s}$ (or 3–dimensional antisymmetric tensor s_{mn}) of the internal angular momentum defined in the rest frame of the particle. The covariant vector of spin S_μ (or the covariant antisymmetric tensor $S_{\mu\nu}$) are obtained from $\boldsymbol{s}$ (or s_{mn}) by the Lorentz transformation. By the way, an advantage of this approach is that the constraints $u^\mu S_\mu = 0$ and $u^\mu S_{\mu\nu} = 0$ hold identically. The precession frequency for spin $\boldsymbol{s}$ at an arbitrary velocity is well–known (see, for instance, [32]):

$$\boldsymbol{\Omega} = \frac{e}{2m}\,\left\{(g-2)\,\left[\boldsymbol{B} - \frac{\gamma}{\gamma+1}\,\boldsymbol{v}(\boldsymbol{v}\boldsymbol{B}) - \boldsymbol{v}\times\boldsymbol{E}\right]\right.$$

$$\left. +2\,\left[\frac{1}{\gamma}\,\boldsymbol{B} - \frac{1}{\gamma+1}\,\boldsymbol{v}\times\boldsymbol{E}\right]\right\}. \tag{12}$$

Naturally, the corresponding interaction Lagrangian (here the Lagrangian formulation is somewhat more convenient than the Hamiltonian one) equals

$$L_{e1} = \boldsymbol{\Omega}\cdot\boldsymbol{s} = \frac{e}{2m}\,\boldsymbol{s}\cdot\left\{(g-2)\,\left[\boldsymbol{B} - \frac{\gamma}{\gamma+1}\,\boldsymbol{v}(\boldsymbol{v}\boldsymbol{B}) - \boldsymbol{v}\times\boldsymbol{E}\right]\right.$$

$$\left. +2\,\left[\frac{1}{\gamma}\,\boldsymbol{B} - \frac{1}{\gamma+1}\,\boldsymbol{v}\times\boldsymbol{E}\right]\right\}. \tag{13}$$

The equation of motion for the position has the usual form:

$$(\boldsymbol{\nabla} - \frac{d}{dt}\boldsymbol{\nabla}_{\boldsymbol{v}})L_{\rm tot} = 0, \tag{14}$$

where $L_{\rm tot}$ is the total Lagrangian of the system. The equation of motion for the spin in general form is

$$\dot{\boldsymbol{s}} = -\{L_{\rm tot}, \boldsymbol{s}\}, \tag{15}$$

where $\{... , ...\}$ is the Poisson bracket, or

$$\dot{\boldsymbol{s}} = -i[L_{\rm tot}, \boldsymbol{s}] \tag{16}$$

in the quantum problem. This applicability of a common canonical formalism is one more advantage of the noncovariant approach. Meanwhile, in the covariant approach we have to deal in particular with higher time derivatives, which is obvious already from relationship (7).

3 Spin Precession in a Gravitational Field

In this section we present a simple and general derivation of the equations of the spin precession in a gravitational field (restricting to first order in spin), based on a remarkable analogy between gravitational and electromagnetic fields. Due to this correspondence, the formulae of the previous section are naturally adapted for the case of an external gravitational field. In this way we easily reproduce and generalize known results for gravitational spin effects.

3.1 General Relationships

It follows from the angular momentum conservation in flat space–time taken together with the equivalence principle that the 4–vector of spin S^μ is parallel transported along the particle world–line. The parallel transport of a vector along a geodesic $x^\mu(\tau)$ means that its covariant derivative vanishes:

$$\frac{DS^\mu}{D\tau} = 0\,. \tag{17}$$

We will use the tetrad formalism natural for the description of spin. In view of relation (17), the equation for the tetrad components of spin $S^a = S^\mu e^a_\mu$ is

$$\frac{DS^a}{D\tau} = \frac{dS^a}{d\tau} = S^\mu e^a_{\mu;\nu} u^\nu = \eta^{ab}\gamma_{bcd}u^d S^c\,. \tag{18}$$

Here

$$\gamma_{abc} = e_{a\mu;\nu}e^\mu_b e^\nu_c = -\gamma_{bac} \tag{19}$$

are the Ricci rotation coefficients [36], §98. Certainly, the equation for the tetrad 4–velocity components is exactly the same:

$$\frac{du^a}{d\tau} = \eta^{ab}\gamma_{bcd}u^d u^c\,. \tag{20}$$

The meaning of Eqs. (18), (20) is clear: the tetrad components of both vectors vary in the same way, due only to the rotation of the local Lorentz vierbein.

In exactly the same way, the 4–dimensional spin and velocity of a charged particle with the gyromagnetic ratio $g = 2$ precess with the same angular velocity in an external electromagnetic field, by virtue of equation (9) at $g = 2$ and the Lorentz equation:

$$\frac{dS_a}{d\tau} = \frac{e}{m} F_{ab} S^b; \qquad \frac{du_a}{d\tau} = \frac{e}{m} F_{ab} u^b.$$

Thus, the correspondence:

$$\frac{e}{m} F_{ab} \longleftrightarrow \gamma_{abc} u^c. \tag{21}$$

becomes obvious. This correspondence makes it possible to obtain the precession frequency $\boldsymbol{\omega}$ of the 3–dimensional vector of spin $\boldsymbol{s}$ in external gravitational field from expression (12) via the simple substitution

$$\frac{e}{m} B_i \longrightarrow -\frac{1}{2} \varepsilon_{ikl} \gamma_{klc} u^c; \quad \frac{e}{m} E_i \longrightarrow \gamma_{0ic} u^c. \tag{22}$$

This frequency is

$$\omega_i = -\varepsilon_{ikl} \left(\frac{1}{2} \gamma_{klc} + \frac{u^k}{u^0 + 1} \gamma_{0lc} \right) \frac{u^c}{u^0_w}. \tag{23}$$

The factor $1/u^0_w$ in this expression is related to the transition in the left–hand side of Eq. (18) to the differentiation with respect to the world time t:

$$\frac{d}{d\tau} = \frac{dt}{d\tau} \frac{d}{dt} = u^0_w \frac{d}{dt}.$$

A subscript w is attached to the quantity u^0_w to emphasize that u^0_w is a world, but not a tetrad, component of 4–velocity. All other indices in expression (23) are tetrad ones, $c = 0, 1, 2, 3$; $i, k, l = 1, 2, 3$. The corresponding spin–dependent correction to the Lagrangian is

$$L_{sg} = \boldsymbol{\omega} \cdot \boldsymbol{s}\,. \tag{24}$$

However, in some respect, the first–order spin interaction with a gravitational field differs essentiallly from that with an electromagnetic field. In the electromagnetic case, the interaction depends, generally speaking, on a free phenomenological parameter, g–factor. Moreover, if one allows for the violation of P and T invariances, one more parameter arises here, the value of the electric dipole moment of the particle. The point is that both magnetic and electric dipole moments interact with the electromagnetic field strength, thus this interaction is gauge–invariant for any value of these moments. Only the spin–independent interaction with the electromagnetic vector potential is fixed by the charge conservation and gauge invariance. Meanwhile, the Ricci rotation coefficients γ_{abc}

entering the gravitational first–order spin interaction (24), as distinct from the Riemann tensor, are not covariant. This interaction is fixed in a unique way by the angular momentum conservation in flat space–time, taken together with the equivalence principle, it has no free parameters [37,38].

The tetrads $e_{a\mu}$ are related to the metric as follows:

$$e_{a\mu}e_{b\nu}\eta^{ab} = g_{\mu\nu}.$$

In the weak–field approximation there is no difference between the tetrad and world indices in $e_{a\mu}$. The ambiguity in the choice of tetrads will be fixed by choosing the symmetric gauge $e_{\mu\nu} = e_{\nu\mu}$. Then (with $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$)

$$e_{\mu\nu} = \eta_{\mu\nu} + \frac{1}{2}h_{\mu\nu}\,.$$

Using expression (19) for the Ricci coefficients, we find in the weak–field approximation

$$\gamma_{abc} = \frac{1}{2}(h_{bc,a} - h_{ac,b})\,. \tag{25}$$

This approach is applied below to the problems of spin–orbit and spin–spin interactions, as well as to the spin precession in a plane gravitational wave[1]. We restrict mostly to the weak–field approximation. However, as distinct from the standard approaches, all three problems can be easily solved now at arbitrary particle velocities. The combination of a high velocity for a spinning particle with a weak gravitational field obviously refers to a scattering problem. Another possible application is to a spinning particle bound by other forces, for instance, by electromagnetic ones, when we are looking for the correction to the precession frequency due to the gravitational interaction.

3.2 Spin–Orbit Interaction

In the centrally symmetric field created by a mass M, the metric is

$$h_{00} = -\frac{r_g}{r} = \frac{2kM}{r}; \quad h_{mn} = -\frac{r_g}{r}\,\delta_{mn} = -\frac{2kM}{r}\,\delta_{mn}. \tag{26}$$

Here the nonvanishing Ricci coefficients are

$$\gamma_{ijk} = \frac{kM}{r^3}(\delta_{jk}r_i - \delta_{ik}r_j)\,, \qquad \gamma_{0i0} = -\frac{kM}{r^3}r_i\,. \tag{27}$$

Plugging these expressions into formula (23) yields the following result for the precession frequency:

$$\boldsymbol{\omega}_{ls} = \frac{2\gamma+1}{\gamma+1}\,\frac{kM}{r^3}\,\boldsymbol{v}\times\boldsymbol{r}\,. \tag{28}$$

In the limit of low velocities, $\gamma \to 1$, the answer goes over into the classical result of [2].

[1] I am grateful to T. Vargas for attracting my attention to the last problem.

Let us note here that in the case of an external gravitational field there is no covariant expression for the force linear in the particle spin. In other words, the deviation from geodesics of the trajectory of a spinning particle is not described by the Riemann tensor. If this was the case, there would be a unique possible covariant structure, to within a factor (in [6] it equals $-1/2m$): $R_{\mu\nu ab}u^{\nu}S^{ab}$. As mentioned already in Section 1, the covariant description contradicts in particular the classical result [2] for the gravitational spin–orbit interaction.

3.3 Spin–Orbit Interaction in the Schwarzschild Field

In the present subsection we treat the spin precession in the Schwarzschild field beyond the weak–field approximation (though neglecting the spin influence on the trajectory). The 3–dimensional components of the Schwarzschild metric can be conveniently written as

$$g_{mn} = -\left(\delta_{mn} - \frac{r_m r_n}{r^2}\right) - \frac{r_m r_n}{r^2}\,\frac{1}{1-r_g/r} = -\delta^{\perp}_{mn} - n_m n_n\,\frac{1}{1-r_g/r}\,. \quad (29)$$

Nonvanishing tetrads are chosen as follows:

$$e^{(0)}_0 = \sqrt{1-r_g/r}; \quad e^{(k)}_m = \delta^{\perp}_{km} + n_k n_m\,\frac{1}{\sqrt{1-r_g/r}}\,; \quad (30)$$

in this subsection the tetrad indices are singled out by brackets. Now the nonvanishing Ricci coefficients (here their last indices are world ones) are

$$\gamma_{(0)(i)0} = -\frac{kM}{r^3}r_i\,; \quad \gamma_{(i)(j)k} = \frac{1-\sqrt{1-r_g/r}}{r^2}\,(\delta_{jk}r_i - \delta_{ik}r_j)\,. \quad (31)$$

At last, the precession frequency in this case is

$$\boldsymbol{\omega} = -\boldsymbol{L}\,\frac{r_g}{2mr^3}\left\{\frac{2}{u^0 + u^0\sqrt{1-r_g/r}} + \frac{1}{1+u^0\sqrt{1-r_g/r}}\right\}. \quad (32)$$

Here m and $\boldsymbol{L}$ are the particle mass and orbital angular momentum, respectively;

$$u^0 = \frac{dt}{d\tau} = \left\{1 - r_g/r - (\boldsymbol{n}\cdot\boldsymbol{v})^2(1-r_g/r)^{-1} - (\boldsymbol{v}^{\perp})^2\right\}^{-1/2}.$$

The rather cumbersome general expression (32) simplifies for a circular orbit. Here

$$u^0 = \left(1 - \frac{3kM}{r}\right)^{-1/2}; \quad L = mr\left(\frac{kM}{r}\right)^{1/2}\left(1 - \frac{3kM}{r}\right)^{-1/2},$$

so that

$$\omega = \frac{(kM)^{1/2}}{r^{3/2}}\left[1 - \left(1 - \frac{3kM}{r}\right)^{1/2}\right]. \quad (33)$$

The general case of spin precession in the Schwarzschild field was considered previously in [39]. Our expression (33) agrees with the corresponding result of [39] (the precession is considered there with respect to the proper time τ, but not with respect to t).

3.4 Spin–Spin Interaction

Let the spin of the central body be $\boldsymbol{s}_0$. The components of the metric linear in $\boldsymbol{s}_0$, which are responsible for the spin–spin interaction, are

$$h_{0i} = 2k\,\frac{[\boldsymbol{s}_0 \times \boldsymbol{r}]_i}{r^3}\,.$$

Here the nonvanishing Ricci coefficients are

$$\gamma_{ij0} = k\left(\boldsymbol{\nabla}_i \frac{[\boldsymbol{s}_0 \times \boldsymbol{r}]_j}{r^3} - \boldsymbol{\nabla}_j \frac{[\boldsymbol{s}_0 \times \boldsymbol{r}]_i}{r^3}\right), \qquad \gamma_{0ij} = -k\boldsymbol{\nabla}_i \frac{[\boldsymbol{s}_0 \times \boldsymbol{r}]_j}{r^3}\,. \tag{34}$$

The frequency of the spin–spin precession is

$$\begin{aligned}\boldsymbol{\omega}_{ss} = &-k\left(2 - \frac{1}{\gamma}\right)(\boldsymbol{s}_0\boldsymbol{\nabla})\boldsymbol{\nabla}\frac{1}{r} \\ &+k\,\frac{\gamma}{\gamma+1}\left[\boldsymbol{v}(\boldsymbol{s}_0\cdot\boldsymbol{\nabla}) - \boldsymbol{s}_0(\boldsymbol{v}\cdot\boldsymbol{\nabla}) + (\boldsymbol{v}\cdot\boldsymbol{s}_0)\boldsymbol{\nabla}\right](\boldsymbol{v}\cdot\boldsymbol{\nabla})\,\frac{1}{r}\,.\end{aligned} \tag{35}$$

In the low–velocity limit this formula also goes over into the corresponding classical result [3].

3.5 Spin Precession in a Plane Gravitational Wave

Let a weak gravitational wave propagate along the axis 3. It is well known (see, for instance, [36], §107) that here coordinate conditions can be choosen in such a way that the only nonvanishing components of $h_{\mu\nu}$ are

$$h_{11} = -\,h_{22} = f_1(t-z), \quad h_{12} = h_{21} = f_2(t-z).$$

Straightforward (though rather tedious) calculation with formulae (23), (25), results in the following expressions for the components of the angular velocity:

$$\begin{aligned}\omega_{w1} &= \left(1 - \frac{\gamma}{\gamma+1}v_3\right)\frac{1}{2}\,(\dot{f}_1 v_2 - \dot{f}_2 v_1); \\ \omega_{w2} &= \left(1 - \frac{\gamma}{\gamma+1}v_3\right)\frac{1}{2}\,(\dot{f}_1 v_1 + \dot{f}_2 v_2); \\ \omega_{w3} &= \frac{\gamma}{\gamma+1}\left[\dot{f}_1 v_1 v_2 - \frac{1}{2}\dot{f}_2\,(v_1^2 - v_2^2)\right].\end{aligned} \tag{36}$$

The equations of motion in a plane gravitational wave, following from the corresponding Lagrangian $L = \boldsymbol{\omega}_w \boldsymbol{s}$, differ essentially from those obtained in [40,41] within the covariant approach.

4 Effects of Higher Order in Spin

4.1 Outline of the General Formalism

The above description of the effects linear in spin was obtained by rather simple means. As to the interaction of second order in spin, it may manifest itself at least in the motion of rotating black holes (and possibly in some subtle spin effects for polarized nuclei of high spin in storage rings). Anyway, going beyond the linear approximation in spin is of a certain theoretical interest. To study this general problem, we have to resort to a more sophisticated approach [30,31]. It is based on the following physically obvious argument. As long as we do not consider excitations of internal degrees of freedom of a body moving in an external field, this body (even if it is a macroscopic one!) can be treated as an elementary particle with spin. Thus, the Lagrangian of the spin interaction with an external field can be derived from the elastic scattering amplitude of a particle with spin s by external field. In this way we can describe the interaction of a relativistic particle to first order in the external field, but to arbitrary order in the spin. Explicit closed formulae were obtained in [30,31] for the interaction of second order in spin. According to the arguments presented in Section 1, the discussion of the effects nonlinear in spin may be physically meaningful first of all in the classical limit $s \gg 1$. This limit is certainly adequate for rotating black holes. However, having in mind the mentioned problem of polarized nuclei, as well as some theoretical questions, the results were derived in [31] for arbitrary spins.

The accurate derivation of the second-order spin Lagrangians in electrodynamics and gravity, based on rather sophisticated technique of high spins, can be found in [30,31]. Here we confine ourselves mainly to a qualitative discussion of some curious features of the second-order spin interactions.

4.2 Second-Order Spin Effects in an Electromagnetic Field

Even the final formula for the discussed interaction is lengthy:

$$\begin{aligned} L_{e2} = \frac{Q}{2s(2s-1)} & \left[(\boldsymbol{s}\cdot\boldsymbol{\nabla}) - \frac{\gamma}{\gamma+1}(\boldsymbol{v}\cdot\boldsymbol{s})(\boldsymbol{v}\cdot\boldsymbol{\nabla}) \right] \times \\ & \times \left[(\boldsymbol{s}\cdot\boldsymbol{E}) - \frac{\gamma}{\gamma+1}(\boldsymbol{s}\cdot\boldsymbol{v})(\boldsymbol{v}\cdot\boldsymbol{E}) + (\boldsymbol{s}\cdot[\boldsymbol{v}\times\boldsymbol{B}]) \right] \\ & + \frac{e}{2m^2}\frac{\gamma}{\gamma+1}(\boldsymbol{s}\cdot[\boldsymbol{v}\times\boldsymbol{\nabla}]) \left[\left(g-1+\frac{1}{\gamma} \right)(\boldsymbol{s}\cdot\boldsymbol{B}) \right. \\ & \left. -(g-1)\frac{\gamma}{\gamma+1}(\boldsymbol{s}\cdot\boldsymbol{v})(\boldsymbol{v}\cdot\boldsymbol{B}) - \left(g-\frac{\gamma}{\gamma+1} \right)(\boldsymbol{s}\cdot[\boldsymbol{v}\times\boldsymbol{E}]) \right]. \end{aligned} \tag{37}$$

Here the particle quadrupole moment Q is defined as usual: $Q = Q_{zz}|_{s_z=s}$.

It is well-known that the electromagnetic interaction of the convection current and magnetic moment also contributes in the nonrelativistic limit to the

quadrupole interaction. The value of this, induced contribution to the quadrupole moment (already included into Q in formula (37)) is [42]

$$\Delta Q = -e\,(g-1)\left(\frac{\hbar}{mc}\right)^2 \begin{cases} s, & \text{integer spin,} \\ s-1/2, & \text{half-integer spin.} \end{cases} \tag{38}$$

We have singled out explicitly in this formula the Planck constant $\hbar$ to demonstrate that the induced quadrupole moment ΔQ vanishes in the classical limit $\hbar \to 0,\ s \to \infty,\ \hbar s \to \text{const}$. Therefore, the contribution proportional to ΔQ does not influence in fact equations of motion of a classical particle (though it plays a role in atomic spectroscopy [42]).

On the other hand, the electromagnetic interactions of the convection current and spin current also induce an interaction of second order in spin which has a classical limit and is described by the last two lines of formula (37). This Q-independent part of the interaction (37) tends to zero in the nonrelativistic limit. Besides, it is reducible in spin; in other words, the structure $s_i s_j$ in it cannot be rewritten as an irreducible tensor $s_i s_j - (1/3)\delta_{ij}\boldsymbol{s}^2$. The Q-independent part of the interaction (37) does not have a quadrupole structure at all.

Of great interest is the asymptotic behaviour of the interaction (37) at $\gamma \to \infty$. Surprisingly, though both Q-dependent and Q-independent parts of the interaction (37) by themselves grow up with energy, there is a singled out value of the quadrupole moment for which this interaction as a whole falls down at $\gamma \to \infty$.

The situation resembles that which takes place for the interaction linear in spin. It is well-known (see, e.g., [11,43,44]) that there is a special value of the g-factor, $g = 2$, at which the electromagnetic interaction linear in spin, decreases with increasing energy. This follows immediately from formula (12) for $\gamma \to \infty$. Thus, the choice $g = 2$ for the bare magnetic moment is a necessary (but insufficient!) condition of unitarity and renormalizability in quantum electrodynamics. It holds not only for the electron, but also for the charged vector boson in the renormalizable electroweak theory. Other arguments in favour of $g = 2$ are given in [45-49].

The same situation takes place with the second-order spin interaction in electrodynamics. There is a special value of the quadrupole moment Q at which this interaction as well decreases with increasing energy. If we also assume $g = 2$, this value is

$$Q = -s(2s-1)\,\frac{e}{m^2}. \tag{39}$$

The same preferred value of the quadrupole moment was derived also otherwise, basing on the supersymmetric sum rules [46,48,49]. Again, (39) is a necessary condition of unitarity and renormalizability. And indeed, this is the value of the quadrupole moment of the charged vector boson in the renormalizable electroweak theory. For it

$$g = 2\,, \quad s = 1\,, \quad Q = -e/m^2\,.$$

4.3 Second-Order Spin Effects in a Gravitational Field

For a binary star effects of second-order in spin are of the same order of magnitude as the spin-spin interaction when the spins of the components of the system are comparable [28]. The influence of the latter on the characteristics of the gravitational radiation becomes noticeable for a system of two extreme black holes [25]. Correspondingly, second–order spin effects in the equations of motion become substantial if at least one component of a binary is close to an extreme black hole [28]. Therefore, the investigation of these effects is not of a purely theoretical interest only. In principle they can be observed with the gravitational wave detectors under construction.

Though the second–order gravitational spin interaction can be obtained within our general approach as well, we will resort here to an instructive short–cut which allows to derive easily the so–called gravimagnetic interaction [11], a gravitational analogue of the Q–dependent terms in formula (37). It was mentioned already that the analogy between first–order spin interactions in electrodynamics and gravity is incomplete. The electromagnetic interaction depends on the field strength, which is gauge–invariant. However, the gravitational one depends not on the Riemann tensor, which is generally covariant, but on the Ricci rotation coefficients, which are not. In this respect, the second–order spin interaction discussed below, the gravimagnetic one, which depends on the Riemann tensor, is the gravitational analogue of the first–order spin interactions in electrodynamics.

The starting point of the derivation is the observation that the canonical momentum p_μ enters a relativistic wave equation for a particle in external electromagnetic and gravitational fields through the combination

$$\Pi_\mu = p_\mu - eA_\mu - \frac{1}{2}\,\Sigma^{ab}\gamma_{ab\mu}.$$

Here Σ^{ab} are the generators of the Lorentz group; $\gamma_{ab\mu} = e^c_\mu \gamma_{abc}$. The commutation relation

$$[\Pi_\mu, \Pi_\nu] = -ieF_{\mu\nu} + \frac{i}{2}\,\Sigma^{ab}R_{ab\mu\nu} \tag{40}$$

demonstrates the remarkable correspondence

$$eF_{\mu\nu} \leftrightarrow -\frac{1}{2}\,\Sigma^{ab}R_{ab\mu\nu}. \tag{41}$$

The squared form of the Dirac equation in an external electromagnetic field

$$(-g^{\mu\nu}\Pi_\mu\Pi_\nu + m^2 + e\,\Sigma^{ab}F_{ab})\,\psi = 0$$

prompts the correct conclusion that for an arbitrary spin s the Lagrangian $-e/(2m)\,\Sigma^{ab}F_{ab}$ describes the magnetic moment interaction for $g = 2$; the factor $1/(2m)$ in this Lagrangian, additional to the above wave equation, becomes obvious from the comparison with the nonrelativistic limit. Clearly, for an arbitrary g–factor this covariant magnetic moment interaction is

$$\mathcal{L}_{e1} = -\frac{eg}{4m}\,F_{ab}\Sigma^{ab}. \tag{42}$$

This is in fact a covariant form of g–dependent terms in the Lagrangian (13). As to the g–independent, Thomas terms in (13), it was pointed out already that they cannot be presented in a covariant form with the usual, physical definition of the coordinate $\boldsymbol{r}$. It is natural to define, in analogy with the magnetic moment

$$\frac{eg}{2m}\Sigma^{ab}\,,$$

the gravimagnetic moment

$$-\frac{\kappa}{2m}\Sigma^{ab}\Sigma^{cd}\,.$$

Now, the correspondence (41) prompts the following gravitational analogue of the Lagrangian (42):

$$\mathcal{L}_{gm} = \frac{\kappa}{8m}\Sigma^{ab}\Sigma^{cd}R_{abcd}. \tag{43}$$

This is what we call the gravimagnetic interaction. Let us note that in the classical limit $\Sigma^{ab} \to S^{ab} = \varepsilon^{abcd}S_c u_d$.

The gravimagnetic ratio κ, like the gyromagnetic ratio g in electrodynamics, may have in general any value. Still, it is natural that in gravity the value $\kappa = 1$ is as singled out as $g = 2$ in electrodynamics. Indeed, the analysis of the complete Lagrangian for the gravitational interaction of second order in spin, including of course κ–independent terms which correspond to the Q–independent terms in (37), demonstrate that just for $\kappa = 1$ this total interaction asymptotically tends to zero with increasing energy [11,30,31]. The same conclusion is made in [50-53]. Unfortunately, the gravitational interaction for any spin is not renormalizable even at $\kappa = 1$.

In any case, at $g = 2$ and $\kappa = 1$ the equations of motion have the simplest form. Moreover, it has been shown in [11] that just this value of the gravimagnetic ratio, $\kappa = 1$, follows from the wave equations in the Feynman gauge both for the photon and graviton in an external gravitational field, as well as from the Rarita–Schwinger equation for $s = 3/2$ in a gravitational field.

The situation for spin 1/2 is rather tricky. Obviously, no second–order spin interaction is possible here. Indeed, for spin 1/2 the properties of the spin matrices $\Sigma^{ab} = i/4(\gamma^a\gamma^b - \gamma^b\gamma^a)$ are such that $\Sigma^{ab}\Sigma^{cd}R_{abcd}$ degenerates into the scalar curvature R (times 1/2) without any spin dependence at all. So, our arguments in favour of $\kappa = 1$ do not apply for spin 1/2. And indeed the squared Dirac equation contains $1/4\,R$, but with $\kappa = 1$ one obtains here $1/8\,R$. Nevertheless, we cannot see any real physical meaning in the recent proposal [54] to ascribe to the electron (which in fact has no gravimagnetic interaction at all) the gravimagnetic ratio $\kappa = 2$.

Wave equations for particles of arbitrary spins in an external gravitational field were previously considered in [55]. The equation for integer spins proposed in [55] corresponds also to the gravimagnetic ratio $\kappa = 1$. However, the value of κ prescribed in [55] for half–integer spins is different. Even in the classical limit $s \to \infty$ it does not tend to unity. This obviously does not comply with the correspondence principle according to which at least in this classical limit there should be no difference between integer and half–integer spins.

5 Multipoles of Black Holes

Let us come back from elementary particles to macroscopic bodies. For a classical object the values of both parameters g and κ depend in general on the various properties of the body. However, for black holes the situation is different. It has been shown in [56] from an analysis of the Kerr–Newman solution that the gyromagnetic ratio of a charged rotating black hole is universal (and equal to that of the electron!): $g = 2$.

We will show that for the Kerr black hole the gravimagnetic ratio is $\kappa = 1$. This value follows in fact from the analysis of the motion of the spin of a black hole in an external field in [24] (though this statement was not explicitly formulated there). We present here an independent and, in our opinion, simpler derivation of this important result.

At large distances from a Kerr hole, the hole can be considered as a point source of a weak gravitational field. To linear approximation in the field of a hole at rest, the Lagrangian density corresponding to the interaction (43) can be written as

$$\tilde{L} = \frac{\kappa}{4m} (\boldsymbol{s} \cdot \boldsymbol{\nabla})^2 h_{00} \, \delta(\boldsymbol{r}) \,. \tag{44}$$

This interaction unduces a correction to the energy–momentum tensor in a single component:

$$\delta T_{00} = -\frac{\kappa}{2m} (\boldsymbol{s} \cdot \boldsymbol{\nabla})^2 \, \delta(\boldsymbol{r}) \,. \tag{45}$$

In the gauge

$$\bar{h}^{\mu\nu}{}_{,\nu} = 0, \;\; \bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2} \eta_{\mu\nu} h^{\alpha}_{\alpha} \,, \tag{46}$$

the static Einstein equation for the corresponding correction h_{00} to the 00–component of the metric is

$$\Delta h_{00} = 8\pi k T_{00} \,.$$

The correction itself is

$$h_{00} = \kappa \frac{k}{m} (\boldsymbol{s} \cdot \boldsymbol{\nabla})^2 \frac{1}{r} \,. \tag{47}$$

Let us compare h_{00} with the corresponding contribution to the Kerr metric. In the Boyer–Lindquist coordinates this metric is

$$\begin{aligned} ds^2 = (1 - \frac{r_g r}{\Sigma}) \, dt^2 - \frac{\Sigma}{\Delta} \, dr^2 - \Sigma \, d\theta^2 \\ - (r^2 + a^2 + \frac{r_g r a^2}{\Sigma} \sin^2\theta) \, r^2 \sin^2\theta + \frac{2 r_g r a}{\Sigma} \sin^2\theta \, d\phi \, dt \,, \end{aligned} \tag{48}$$

where $\Delta = r^2 - r_g r + a^2 \,, \;\; \Sigma = r^2 + a^2 \cos^2\theta \,, \;\; \boldsymbol{a} = \boldsymbol{s}/m$. At $r_g = 0$ the metric (48) describes a flat space in spheroidal coordinates [36]. Meanwhile, it is

Cartesian coordinates which correspond in the flat space to the gauge (46). The transition from the spheroidal coordinates to Cartesian ones is carried out with the required accuracy by the substitution

$$\boldsymbol{r} \to \boldsymbol{r} + \frac{\boldsymbol{a}(\boldsymbol{a}\cdot\boldsymbol{r}) - \boldsymbol{r}a^2}{2r^2}\,.$$

In the Cartesian coordinates the spin–dependent part of the 00–component of the metric

$$g_{00} = 1 - \frac{r_g}{r} + \frac{r_g a^2}{2r^3}\,(3\cos^2\theta - 1)$$

obviously coincides with h_{00} from formula (47) at $\kappa = 1$. Somewhat more tedious consideration of the space components of the Kerr metric leads to the same result, $\kappa = 1$.

Let us note that the motion of the Kerr black hole in an external gravitational field is not described by the Papapetrou equation even if one leaves aside the problem of the spin–orbit interaction linear in spin. The point is that this equation refers to the case $\kappa = 0$ [14].

It is proven in the same way that for a charged Kerr hole the gravimagnetic ratio κ is also unity. Moreover, the electric quadrupole moment of a charged Kerr hole also equals

$$Q = -2\,\frac{es^2}{m^2}, \tag{49}$$

the value, at which the interaction quadratic in spin decreases with energy (this is the obvious limit of the general formula (39) at $s \to \infty$). It can be demonstrated [57] that other, higher multipoles of a charged Kerr hole, both electromagnetic and gravitational, as well possess just those values which guarantee that the interaction of any order in spin (but of course, linear in an external field) asymptotically decreases with increasing energy.

6 Gravitational Interaction of Spinning Bodies, and Radiation of Compact Binary Stars

It is expected that in a few years the gravitational radiation from coalescing binary stars will be observed by laser interferometer systems. Its successful detection depends crucially on the accurate theoretical prediction of the exact form of the signal. In this way the observed effect becomes sensitive to the relativistic corrections of the orders c^{-2}, c^{-3} and c^{-4} to the motion of a binary system and to the radiation intensity. In particular, the spin–orbit interaction becomes essential. Moreover, effects of second order in spin may be observed in the gravitational radiation, in the case of two extreme Kerr black holes [25].

6.1 Spin Interactions in a Two–Body Problem

The spin interactions in a two–body problem can be easily obtained from the well–known results for the limiting case when one of the bodies (say, 2) is very

heavy. In this limit we have the usual spin–orbit interaction with the frequency $\boldsymbol{\omega}_{ls}$ given in fact by formula (28) (the limit $\gamma \to 1$ is sufficient here):

$$V^1_{1ls} = -\boldsymbol{\omega}_{ls}\boldsymbol{s}_1 = \frac{3}{2}\frac{k}{r^3}\frac{m_2}{m_1}\boldsymbol{l}\boldsymbol{s}_1. \tag{50}$$

Then, there is the well–known Lense–Thirring interaction of the orbital angular momentum $\boldsymbol{l}$ with the spin $\boldsymbol{s}_2$ of the central body [58]

$$V^1_{2ls} = 2\frac{k}{r^3}\boldsymbol{l}\boldsymbol{s}_2. \tag{51}$$

Simple symmetry arguments dictate now the form of the total spin–orbit interaction for the two–body problem:

$$V_{ls} = \frac{k}{r^3}\boldsymbol{l}\cdot\left[\frac{3}{2}\left(\frac{m_2}{m_1}\boldsymbol{s}_1 + \frac{m_1}{m_2}\boldsymbol{s}_2\right) + 2\left(\boldsymbol{s}_1 + \boldsymbol{s}_2\right)\right]. \tag{52}$$

As to the spin–spin interaction, it is of the usual form, with $\boldsymbol{\omega}_{ss}$ given by formula (35) (again the lowest nonvanishing order in $1/c$ is implied):

$$V_{ss} = \frac{k}{r^3}\left[3(\boldsymbol{s}_1\boldsymbol{n})(\boldsymbol{s}_2\boldsymbol{n}) - \boldsymbol{s}_1\boldsymbol{s}_2\right]. \tag{53}$$

Of course, both expressions (52) and (53) can be derived directly, following, for instance, the approach of [36] (§106, Problem 4).

Let us go over now to the gravimagnetic interaction. This interaction (43) for a particle 1 with the field created by a heavy mass m_2 reduces in lowest, first order in c^{-2} to the quadrupole form:

$$V^1_s = \frac{3km_2}{2r^3}\,Q^s_{1mn}n_m n_n \tag{54}$$

where the effective gravitational quadrupole moment of the particle 1 is

$$Q^s_{1mn} = \frac{\kappa_1}{m_1}\left(\,s_{1m}s_{1n} - \frac{1}{3}\,\delta_{mn}\boldsymbol{s}_1^2\right).$$

For the two–body problem under discussion, expression (54) generalizes to the following self–interaction of spin:

$$V_s = \frac{3k}{2r^3}\left(\kappa_1\frac{m_2}{m_1}\,s_{1m}s_{1n} + \kappa_2\frac{m_1}{m_2}\,s_{2m}s_{2n}\right)\left(n_m n_n - \frac{1}{3}\,\delta_{mn}\right), \tag{55}$$

resembling the usual spin–spin interaction (53).

At $\kappa_{1,2} \sim 1$ the effective quadrupole interaction (55) is of the same order of magnitude as the spin–spin one (53). Even in the most favourable case when they can become important, that of two extreme Kerr black holes, both interactions are of the c^{-4} order. The star rotation velocity is here $\sim c$, but the star radius is close to the gravitational one $r_g \sim c^{-2}$, so that each spin $s \sim c^{-1}$ [25]. The same argument demonstrates that the spin–orbit interaction is of the c^{-3} order [25].

As to the common quadrupole interaction, due to quadrupole deformations of the stars, it is suppressed by the small value of these deformations and, according to [59], can also manifest itself only in the case of two extreme Kerr black holes.

6.2 Contribution of Spin Interactions to Gravitational Radiation

The spin interactions contribute in various ways to the gravitational radiation: through spin–dependent corrections to the orbit radius r and to the equations of motion used to evaluate the time derivatives, which enter the usual expression for the gravitational quadrupole radiation; through the corrections to the 00 component of the energy–momentum tensor of the particles; through the gravitational analogue of the magnetic quadrupole radiation in electrodynamics; through retardation effects.

In all our discussions of gravitational radiation we restrict ourselves to the case of circular orbits which is the most interesting one from the physical point of view [25]. Besides, the assumption of circular orbits simplifies essentially the calculations. Still the calculations remain tedious, so only the final results are presented here.

The relative correction to the radiation intensity generated by the spin–orbit interaction (52) is [28]

$$\frac{I_{ls}}{I_q} = -\frac{\boldsymbol{l}\,(\,73\,\boldsymbol{s}\,+\,45\,\boldsymbol{\xi}\,)}{12\,m_1 m_2 r^2}. \tag{56}$$

Here

$$I_q = \frac{32k^4 m_1^2 m_2^2 (m_1+m_2)}{5r^5}$$

is the unperturbed quadrupole intensity and

$$\boldsymbol{s} = \boldsymbol{s}_1 + \boldsymbol{s}_2; \;\; \boldsymbol{\xi} = \frac{m_2}{m_1}\,\boldsymbol{s}_1 + \frac{m_1}{m_2}\,\boldsymbol{s}_2.$$

It can be easily checked that the corresponding result of [25,26] would be reconciled with this one under the proper definition of the center–of–mass coordinate.

The correction due to the spin–spin interaction (53) is [25,26]

$$\frac{I_{ss}}{I_q} = \frac{1}{48\,m_1 m_2 r^2}\,(\,649\,s_{1t}s_{2t} - \,223\,\boldsymbol{s}_1\boldsymbol{s}_2). \tag{57}$$

The expressions for I_{ss}, as well as that for I_s below, have been averaged over the period of rotation. That is why both of them contain the spin components s_t orthogonal to the orbit plane.

And at last, the spin–self–interaction correction, generated by the gravimagnetic interaction (55) and by the above–mentioned gravitational analogue of the magnetic quadrupole radiation in electrodynamics, is [28]

$$\frac{I_s}{I_q} = \frac{1}{4m_1 m_2 r^2}\left[\left(27\,\kappa_1 - \frac{1}{24}\right)\frac{m_2}{m_1}\,s_{1t}^2 + \left(27\,\kappa_2 - \frac{1}{24}\right)\frac{m_1}{m_2}\,s_{2t}^2\right.$$

$$\left. - \left(9\,\kappa_1 - \frac{7}{24}\right)\frac{m_2}{m_1}\,\boldsymbol{s}_1^2 - \left(9\,\kappa_2 - \frac{7}{24}\right)\frac{m_1}{m_2}\,\boldsymbol{s}_2^2\right]. \tag{58}$$

This correction is discussed also in [60].

Acknowledgements

I am grateful to A.A. Pomeransky and R.A. Sen'kov for useful discussions and to R. Golub for the efforts to improve the English of the manuscript. The work was supported by the Russian Foundation for Basic Research through Grant No. 98-02-17797, through Grant No. 96-15-96317 for Leading Scientific Schools, by the Ministry of Education Grant No. 3N-224-98, and by the Federal Program Integration-1998 through Project No. 274.

References

1. L.H. Thomas: *Nature* **117**, 514 (1926); Phil. Mag. **3**, 1 (1926).
2. A.D. Fokker: *Kon. Akad. Weten. Amsterdam, Proc.* **23**, 729 (1921).
3. L. Schiff: *Phys. Rev. Lett.* **4**, 435 (1959).
4. J. Frenkel: *Z. Phys.* **37**, 243 (1926).
5. V. Bargmann, L. Michel, V. Telegdi: *Phys. Rev. Lett.* **2**, 435 (1959).
6. A. Papapetrou: *Proc. Roy. Soc. London* **A 209**, 248 (1951).
7. A. Barducci, R. Casalbuoni, L. Lusanna: *Nuovo Cimento* **A 35**, 389 (1976).
8. F. Ravndal: *Phys. Rev.* **D 21**, 2823 (1980).
9. P.L. Nash: *J. Math. Phys.* **25**, 2104 (1984).
10. U. Heinz: *Phys. Lett.* **B 144**, 228 (1984); *Ann. Phys.* (N.Y.) **161**, 48 (1985).
11. I.B. Khriplovich: *Zh. Eksp. Teor. Fiz.* **96**, 385 (1989) [*Sov. Phys. JETP* **69**, 217 (1989)].
12. J.W. van Holten: *Nucl. Phys.* **B 356**, 3 (1991).
13. R.H. Rietdijk, J.W. van Holten: *Class. Quantum Grav.* **9**, 575 (1992).
14. K. Yee, M. Bander: *Phys. Rev.* **D 48**, 2797 (1993).
15. J.P. Costella, B.H.J. McKellar: *Int. J. Mod. Phys.* **A 9**, 461 (1994).
16. M. Chaichian, R. Gonzales Felipe, D. Louis Martinez: *Phys. Lett.* **A 236**, 188 (1997); E-print archive hep-th/9601119.
17. Ya.I. Azimov, R.M. Ryndin: In *Proceedings of the XXXI PNPI Winter School* (St. Petersburg 1997); E-print archive hep-ph/9710433, hep-ph/9707468.
18. D.Bini, G.Gemelli, R.Ruffini: to be published.
19. Ya.S. Derbenev, A.M. Kondratenko: *Zh. Eksp. Teor. Fiz.* **64**, 1918 (1973) [*Sov. Phys. JETP* **37**, 968 (1973)].
20. K. Heinemann: DESY report DESY 96-229; E-print archive physics/9611001.
21. W. Pauli: *Collected Scientific Papers*, ed. by R. Kronig, V.F. Weisskopf (John Wiley and Sons, 1964), v. 2, p. 544.
22. H. Batelaan, T.J. Gay, J.J. Schwendiman: *Phys. Rev. Lett.* **79**, 4517 (1997).
23. T.O. Niinikoski, R. Rosmanith: *Nucl. Instr. Meth.* **A 225**, 460 (1987).
24. K.S. Thorne, J.B. Hartle: *Phys. Rev.* **D 31**, 1815 (1985).
25. L.E. Kidder, C.M. Will, A.G. Wiseman: *Phys. Rev.* **D 47**, R4183 (1993).
26. L. Blanchet, T. Damour, B.R. Iyer, C.M. Will, A.G. Wiseman: *Phys. Rev. Lett.* **74**, 3515 (1995).
27. H.T. Cho: *Class. Quantum Grav.* **15**, 2465, (1998); E-print archive gr-qc/9703071.
28. I.B. Khriplovich, A.A. Pomeransky: *Phys. Lett.* **A 216**, 7 (1996); E-print archive gr-qc/9602004.
29. B.M. Barker, R.F. O'Connell: *Gen. Rel. Grav.* **5**, 539 (1974).
30. A.A. Pomeransky, I.B. Khriplovich: *Zh. Eksp. Teor. Fiz.* **113**, 1537 (1998). [*Sov. Phys. JETP* **86**, 839 (1998)], E-print archive gr-qc/9710098.

31. A.A. Pomeransky, R.A. Sen'kov: *Phys. Lett.* **B** in press; E-print archive gr-qc/9909090.
32. V.B. Berestetskii, E.M. Lifshitz, L.P. Pitaevskii, *Quantum Electrodynamics* (Pergamon Press, 1994).
33. J.D. Bjorken, S.D. Drell: *Relativistic Quantum Mechanics* (McGraw-Hill, New York 1964).
34. L.L. Foldy, S.A. Wouthuysen: *Phys. Rev.* **78**, 248 (1951).
35. A.J. Bracken, G.F. Melloy: *J. Phys.* **A 32**, 6127 (1999); E-print archive quant-ph/9903087.
36. L.D. Landau, E.M. Lifshitz: *The Classical Theory of Fields* (Butterworth-Heinemann, 1975).
37. I.Yu. Kobzarev, L.B. Okun': *Zh. Eksp. Teor. Fiz.* **43**, 1904 (1962). [*Sov. Phys. JETP* **16**, 1343 (1962)].
38. F.W. Hehl, A. Macias, E.W. Mielke, Yu.N. Obukhov: In A. Harvey (ed.): *On Einstein's Path, Festschrift for E. Schucking on the Occasion of His 70th Birthday* (Springer, 1998); E-print archive gr-qc/9706009.
39. T.A. Apostolatos: *Class. Quantum. Grav.* **13**, 799 (1996).
40. J.A. Nieto, M.P. Ryan: *Nuovo Cimento* **63** A, 71 (1981).
41. D. Bini, G. Gemelli: *Nuovo Cimento* **112** B, 165 (1997).
42. I.B. Khriplovich, A.I. Milstein, R.A. Sen'kov: *Phys. Lett.* **A 221**, 370 (1996); *Zh. Eksp. Teor. Fiz.* **111**, 1935 (1997) [*Sov. Phys. JETP* **84**, 1054 (1997)].
43. S. Weinberg: In S. Deser, M. Grisaru, H. Pendleton (eds.): *Lectures on Elementary Particles and Quantum Field Theory* (MIT Press, 1970).
44. S. Ferrara, M. Porrati, V.L. Telegdi: *Phys. Rev.* **D 46**, 3529 (1992).
45. S. Ferrara, E. Remiddi: *Phys. Lett.* **B 53**, 347 (1974).
46. S. Ferrara, M. Porrati: *Phys. Lett.* **B 288**, 85 (1995).
47. R. Jackiw: *Phys. Rev.* **D 57**, 2635 (1998); E-print archive hep-th/9708097.
48. I. Giannakis, J.T. Liu: *Phys. Rev.* **D 58**, 025009 (1998); E-print archive hep-th/9711173.
49. I. Giannakis, J.T. Liu, M. Porrati: *Phys. Rev.* **D 58**, 045016 (1998); E-print archive hep-th/9803073.
50. M. Porrati: *Phys. Lett.* **B 304**, 77 (1993); E-print archive gr-qc/9301012.
51. A. Cucchieri, M. Porrati, S. Deser: *Phys. Rev.* **D 51**, 4543 (1995); E-print archive hep-th/9408073.
52. I. Giannakis, J.T. Liu, M. Porrati: *Phys. Rev.* **D 59**, 104013 (1999); E-print archive hep-th/9809142.
53. I. Giannakis, J.T. Liu, M. Porrati: E-print archive hep-th/9909012.
54. R. Aldrovandi, V.C. de Andrade, J.G. Pereira: E-print archive hep-th/9804117.
55. M. Christensen, M.J. Duff: *Nucl. Phys.* **B 154**, 301 (1979).
56. B. Carter: *Phys. Rev.* **174**, 1559 (1968).
57. A.A. Pomeransky: Talk at the 3rd William Fairbank Meeting, Rome, Italy, 29 June – 4 July 1998.
58. H. Thirring, J. Lense: *Phys. Z.* **19**, 156 (1918).
59. L. Bildsten, C. Cutler: *Astrophys. J.* **400**, 175 (1992).
60. E. Poisson: *Phys. Rev.* **D 57**, 5287 (1998).

The GEO 600 Gravitational Wave Detector Status, Research, Development

Albrecht Rüdiger and Karsten Danzmann for the GEO team

Max-Planck–Institut für Quantenoptik, Garching and Hannover, Germany

Abstract. The last few years have brought a great break–through in the quest for earth–bound detection of gravitational waves: at five sites, laser–interferometric detectors, of armlengths from 0.3 to 4 km, are being built. These projects have in common that one prominent noise source, the shot noise, is reduced by the use of *power recycling*. By using advanced optical technologies early on, the German–British project GEO 600, although only intermediate in size (600 m), has good chances for a competitive sensitivity, at least with the first versions of the larger detectors. Particularly the use of the so–called *signal recycling* technique will allow to search for faint sources of only slowly varying frequency (pulsars, close binaries). The talk will describe the particular topology of the GEO 600 interferometer, characterized by the use of a four–pass delay line and *signal recycling*. The major noise sources, and the experimental effort aiming at their reduction, will be discussed. The current status of the construction of GEO 600 will be outlined (civil engineering, vacuum, optics). The research and development activities at the experimental sites (Garching, Glasgow, Hannover) will be given broad emphasis. First science runs of GEO 600, well in time with those of other ground–based interferometers, are expected in the year 2001.

1 Introduction

The construction of laser interferometers for the detection of gravitational waves is being pursued by various groups all over the world and, due to their particular preferences, their past history, and their funding, differing approaches have been developed. This diversity can, at the present stage, be seen as a great asset, as it will allow the community to develop and investigate a wide variety of promising schemes and evaluate their respective merits.

It is particularly in this field of *research and development* that a fruitful collaboration within the whole gravitational–wave community can and should be intensified.

This paper will give an account of the current status of the German–British project GEO 600 and elaborate on the various fields of research and development at the main centers of GEO activity: Hannover, Glasgow, and Garching. In many of these fields, the GEO groups have taken a leading role.

2 The GEO 600 Concept

At five sites, large–scale gravitational wave detectors using laser interferometry are being built and are nearing completion: at two sites in the US, interferometers of 4 km are being built (LIGO), the French–Italian collaboration VIRGO is

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 131–140, 2001.

building a 3 km detector near Pisa, and on a somewhat smaller scale, with 600 m and 300 m armlength: the German–British antenna GEO 600 near Hannover, and the Japanese antenna TAMA near Tokyo.

In the case of GEO 600, the smaller armlength was not a matter of choice, but one of necessity. The site (on grounds belonging to the University of Hannover) cannot accommodate a larger antenna, and the funds did not allow buying or leasing ground elsewhere. This paper will try to outline how this shortcoming in length is to be compensated, at least partially, by the application of advanced interferometric techniques.

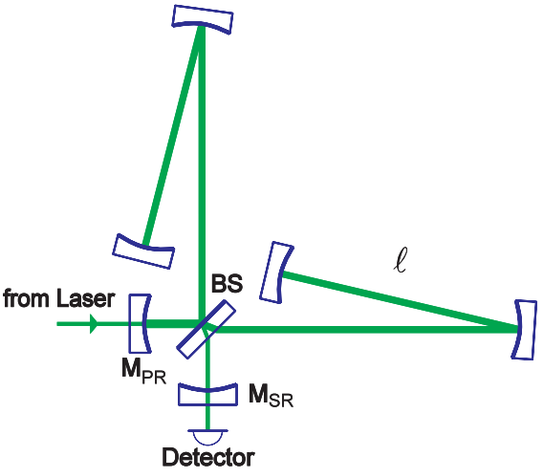


Fig. 1. DL 4 configuration of the GEO 600 interferometer. Two extra mirrors M_{PR} and M_{SR} for power and signal recycling.

In its arms, GEO 600 will employ an optical delay line with only four light transits (DL 4), as shown in Fig.1. After being bounced off the distant mirror the beam is reflected at the near mirror and retraces its path back to the beam–splitter BS. This DL4 scheme differs from the conventional Herriot delay line, which would pose problems in size, in scattered light effects, and in separating incoming and outgoing beam.

The DL4 configuration is distinct in so far as it has the highest number of beams (4) that one can support in a delay line made only of 'small' mirrors (the GEO 600 mirrors will nevertheless have a diameter of 18 cm).

An outstanding characteristic of the GEO 600 concept is the use of 'dual recycling' [1,2]: In addition to the scheme of power recycling, which is now standard in all the large detectors, a further mirror (M_{SR}) is introduced in the output port, to allow a resonant enhancement of the sidebands that the gravitational wave produces from the carrier beam.

3 Noise Contributions

To obtain the projected sensitivity, extensive theoretical and experimental work is required to reduce the contributions of noise that would limit the sensitivity of GEO 600. These efforts (some are very similar to work going on at other

institutions) are being carried out at Hannover, Glasgow, and Garching, with assistance also from the institutions at Cardiff and Potsdam.

Some of the more prominent noise sources (but by no means an exhaustive list) will be discussed below. Although these noise contributions are very serious, current technologies can reduce them to a level at which an interesting sensitivity of GEO 600 can be attained.

3.1 Seismic Noise

The influence from *external* mechanical vibrations, i.e. from *seismic noise*, is strongly reduced by a multiple pendulum suspension, indicated in Fig.2. With triple pendulum suspensions for the most critical optical components (mirrors, beam splitter), and double pendulums for the less critical components, GEO 600 is making a very ambitious effort, topped only by the even more extreme design of the VIRGO project with its seven pendulum stages. Installation at GEO has begun.

GEO's multiple pendulums are hung from a set of cantilever springs for better *vertical* isolation. These springs are supported by a top frame that itself is isolated via an encapsulated stack of 'rubber' and metal layers.

Underneath this stack, an active seismic control (feedback plus perhaps feed-forward) will provide further isolation down to very low frequencies.

The forces to the test mass pendulum masses are applied from reaction masses that are similarly suspended, to avoid direct coupling from ground noise. The forces to the intermediate masses are applied via coil–and–magnet actuators. For the application of only very minute forces to the actual test masses (the lowest stage), an electrostatic scheme is being considered.

Work in this field was led by the Glasgow group, with contributions also from JILA (active isolation) and in close collaboration with the other GEO laboratories. The GEO pendulum design is a candidate for the suspension of the masses in the Advanced LIGO.

3.2 Thermal Noise

The thermal ('Brownian') noise of the suspended optical components can be split into two separate regimes: the *internal vibrations* of the solid suspended bodies (mirrors, beamsplitters), and the *pendulation modes* due to the pendulum suspension. Both must be kept as low as possible, to the very limit of current technology (or beyond).

The *internal* mechanical vibrations of the mirrors (the 'test masses') can be kept low by choosing materials of very high mechanical Q, such as pure fused silica, sapphire, or even silicon. (Note that the mirrors used in the DL4 scheme do not need to be transparent.) Elaborate methods of attaching the suspending 'wires' are required so as not to compromise the intrinsic Q. Glasgow and Hannover have worked out viable solutions, and Q values of $5{\times}10^6$ and better have been achieved with fused silica mirrors. This internal noise is responsible for the rather shallow portion of the curve 'thermal noise' in Fig.3, above 50 Hz.

Also for the *pendulum motion*, extremely low–loss wires or fibers, and very elaborate techniques for attaching these 'wires' to the solid masses were developed, and very promising results were obtained, again by research carried out both at Glasgow and Hannover [3].

Low–loss silica fibers (or thin ribbons) were welded on to small silica prisms, and these were, in turn, attached to the mirrors using a bonding technique developed at Stanford, by 'hydroxy–catalysis bonding of silicates' [4]. This method turned out to be very successful. The measurements on these pendulums were done in close collaboration with the VIRGO group at Perugia. The pendulum thermal noise is the cause of the steep rise of the thermal noise curve as one goes to lower frequencies (in GEO 600 below, say, 30 Hz).

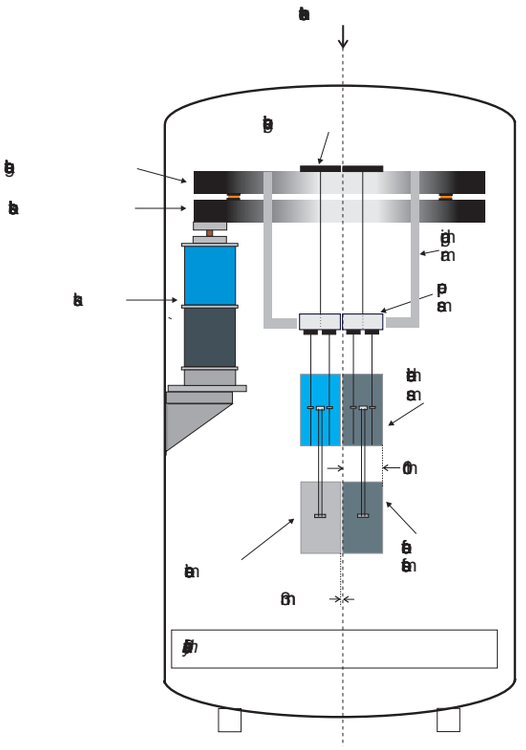


Fig. 2. Schematic of the GEO suspension. The actual test mass is on the right–hand side, forces are applied from equally quiet reaction masses, to avoid introducing ground noise.

3.3 Laser Noise

The development of the laser system for GEO 600 is mainly in the hands of Laser Zentrum Hannover (LZH). The master/slave system, with its highly stabilized Nd:YAG MISER as 'master' and a powerful slave, are now completed, 14 W are realiably available [5]. The master is already out at the site. Improved stabilization of the laser with respect to frequency and power is the goal of work going on at Hannover.

The geometrical noise of the beam, i.e. fluctuations in position, orientation, and shape of the beam, are greatly reduced by the use of 'mode–cleaners' first introduced by the GEO groups [6,7]. GEO 600 will use two in series, and intensive work was done at Glasgow. These mode–cleaners are the first units being optically tested, in the late months of 1999. Mirrors and seismic isolation are installed, and locking of the first stage has been achieved.

4 Shot Noise

A very fundamental noise source is the shot noise produced by the 'graininess' of the detected light. Its effect decreases with the square root of the available light power. To reach sensitivities that make the detection of events from as far out as the Virgo cluster possible, light powers in the order of 10 kW are needed.

As these light powers cannot be supplied by today's lasers, the scheme of 'power recycling' will be applied in GEO 600, as in all other detector projects. The cavity composed of the Michelson interferometer and the *power recycling mirror* M_{PR} is designed to resonantly enhance the light power by three orders of magnitude. Recycling gains in the order of 300 have been obtained [8] in the Garching 30–meter prototype.

5 The GEO 600 Sensitivity

Fig.3 shows the predicted sensitivity (i.e. the spectral density of the apparent strain noise) of GEO 600, resulting from the expected contributions of the various noise sources.

Having the signal sidebands resonate in the cavity formed by the Michelson interferometer as the one (albeit very complex) mirror, and the signal–recycling mirror M_{SR} as the other, will make the signal response a rather complicated function of the GW frequency, which can be adapted to a variety of requirements.

On the left–hand side, the broadband operation of GEO 600 is shown, exhibiting a broad noise minimum around a frequency of 200 Hz. The limiting effects are thermal noise between 40 and 200 Hz and shot noise from 300 Hz on upward.

By appropriate choice of the transmission of the 'signal recycling mirror', a narrowband operation is possible, and the antenna can (by microscopic positioning of the mirror M_{SR}) be tuned to a given frequency, e.g. to a known or expected GW source. This is indicated in the righthand diagram, showing a dip in the shot–noise contribution at 600 Hz.

In this way, the limitation due to shot–noise can be reduced, if one is willing to "sacrifice" measuring bandwidth. There can be very good reasons to do that: for a nearly continuous–wave signal, as in the early stages of a binary inspiral, the signal frequency changes only very slowly, and narrow–banding at the 'response' level reduces noise much more effectively than narrow–banding at the data–analysis level. But, of course, the thermal noise (intrinsic noise of the mirrors) does then limit the sensitivity.

6 Interferometry

A major subject of research at GEO 600 was the design, the analysis, and a critical assessment of various advanced interferometric techniques: dual recycling [1,2] and resonant sideband extraction [9] appear to have great potential

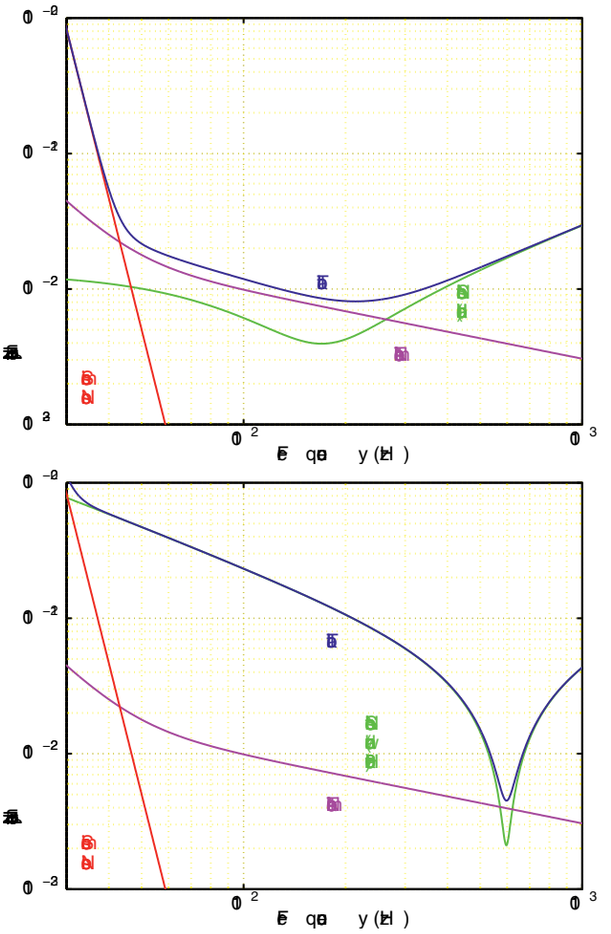


Fig. 3. Noise spectral density of GEO 600, shown for two operating modes: for broadband (top) and narrowband operation (bottom).

also for future large detectors. But also the feasibility of other interferometric schemes (e.g. Sagnac [10]) was explored.

Of the schemes investigated, it will be particularly the scheme of 'dual recycling' that will find application in GEO 600. Elaborate simulation programs were written (e.g. [11]), compared at a Workshop in Garching, and shared with colleagues from other groups.

At that workshop, also a variety of other software tools were presented to assist in the optical design [12] and development of controls [13] for advanced interferometry. These have been (or will be) made available to everybody in the community.

7 Civil Engineering

GEO 600 was, from the beginning, a project that had to take into account a very low budget (the lowest of the world–wide community). Many of the features in GEO's implementation are thus dictated by budget considerations. This was, at times, a great burden, but it also gave an incentive for designing and developing very cost–efficient techniques that can become the basis also for future large–scale detectors.

7.1 Construction

The construction of the *central house* and the *end houses* had been completed in 1997. The size of these buildings is at the lower limit, the accommodation of electronics and other equipment is scant, but sufficient. Some mechanical equipment is available at the site, but the main workshops are at the university site in Hannover.

The *clean–room* environment is a low–cost design which, however, turned out to be working very satisfactorily; a clean–room class 100 was readily achieved, and further improvement during critical work can be achieved by using laminar–flow tents above the vacuum tanks.

The connecting *trenches* that house the vacuum tubes were finished even earlier. The trenches are open constructions, the sides consisting of steel panels driven into the ground. For weather protection, they are loosely covered with a corrugated steel roofing.

7.2 Electronics and Data Management

Most of the electronics is designed and built by the labs in Glasgow, Hannover, and Garching. Communication between the buildings will be exclusively via glass fiber, to avoid crosstalk. The data can be sent via direct radio communication to the University of Hannover for on–line monitoring and possible remote interaction, and from there via fast data links to Potsdam and Cardiff for analysis.

8 The Vacuum System

8.1 Vacuum Pumps

The good vacuum required in the whole apparatus is made possible by powerful turbo–molecular pumps that are situated in the central house and in the end houses. They will run continuously, i.e. also during measurements; the magnetic bearings allow a very quiet operation.

8.2 Vacuum Tubes

GEO 600 has tried a very cost–effective way of producing the vacuum tubing. The tube is made of pieces of 60 cm diameter stainless steel tubes of less than 1 mm thickness. The necessary stiffness against air pressure stems from a bellows–type convolution with a few cm period (see Fig.4). The tube is suspended from crossbars that traverse the trench.

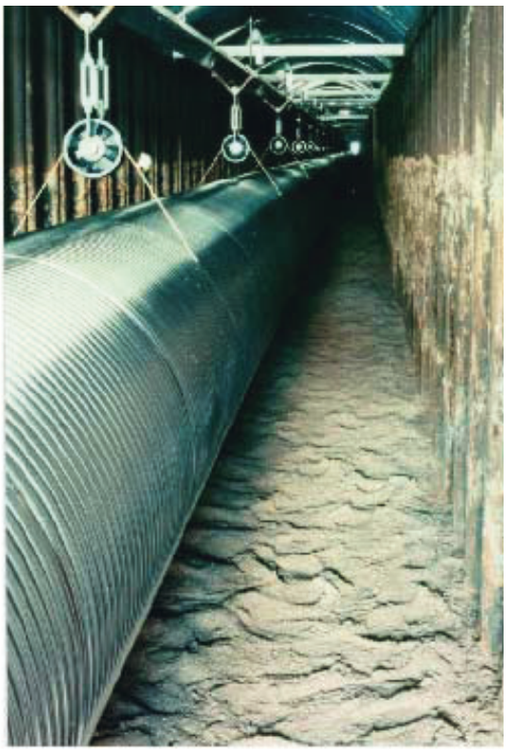


Fig. 4. The convoluted tubes during installation; with each new 4.5 m piece welded on, the tube was slid 4.5 m further into the 600 m long trench, using trolleys along the aluminum rail at the top.

8.3 The Vacuum Achieved

The tubes, wrapped in a 20 cm layer of rock wool, were first air–baked at 200 °C for one week, and then baked under vacuum for several more days. The vacuum achieved, in the range of 10^{-9} mbar, is by far sufficient. The tubes are supposed never to be let up to air again: they are shut off with gate valves when opening the vacuum tanks becomes necessary, as in the present phase of installation of suspensions.

8.4 Vacuum Tanks

The vacuum tanks (9 in the central house, one each in the end houses) have been leak–tested, baked out, and installed. The metal–and–rubber stacks, three per tank, are encapsulated in steel bellows to protect the vacuum. They are finished, most already installed. They will carry the top suspension structures, from which double or triple pendulums will carry the main optical components.

9 The 30 Meter Prototype

Many of the advanced interferometric concepts require verification and optimization by laboratory experiments. Such experimental development went on at the Glasgow 10 m prototype, as well as at the Garching 30 m prototype. It is the

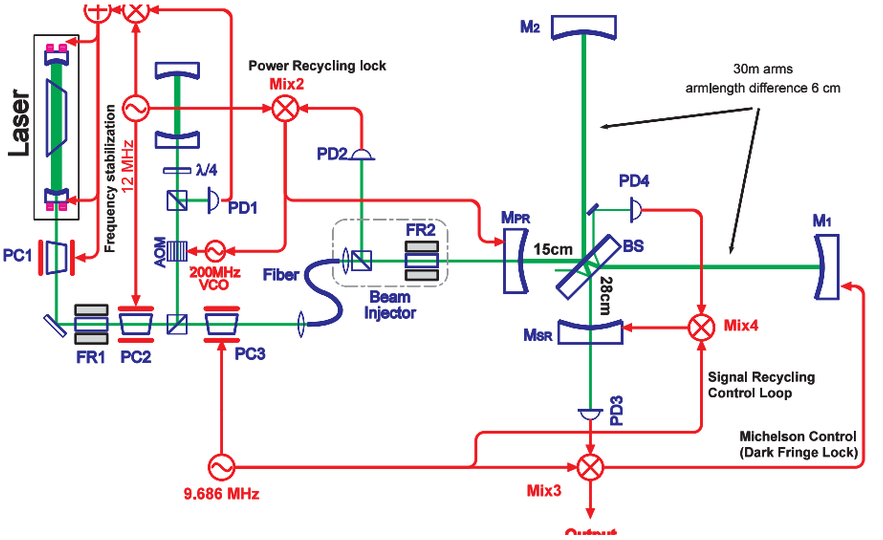


Fig. 5. The Garching 30 m prototype.

latter that will be treated here in some more detail. For a full discussion, see [16].

The Garching 30 m prototype (Fig.5) was used for verifying the concepts of power and signal recycling, and for experimentally investigating the demands on the overall control system.

The laser (here still an Argon ion laser) is first stabilised against a reference cavity, and then against the power recycling cavity consisting of the interferometer and the mirror M_{PR}. A fully automatic alignment system for controlling 10 mechanical degrees of freedom provides the stability of the settings of the optical components [14]. The frontal ('Schnupp') modulation simplifies the read–out with which the Michelson interferometer is kept in lock, as compared with the previous concept of 'external modulation'.

Only after this set–up was sufficiently robust, the next step, the introduction of the signal recycling mirror M_{SR}, was started. Error signals for controlling its position are obtained with the same Schnupp modulation, using an additional photodetector, PD4. One big problem in such complicated interferometers is the fact that many of the control loops can work properly only after all other loops are locked. The operation of prototypes is essential in finding feasible schemes for this lock acquisition problem. The appropriate arm lengths between beam splitter and the four mirrors were chosen with the help of an elaborate simulation code, developed in close collaboration between Glasgow and Garching.

With all of this cautious preparatory work done, acquiring lock of the signal recycling loop was then surprisingly easy, and gave the expected enhancement of the signal response [15]. It furthermore gave evidence of the important effect of 'mode healing' when signal recycling is introduced: Only those transversal

modes that are supported by the signal recycling cavity will be dominating at the output photo diode, which results in a contrast enhancement.

At the beginning [15], the output was still dominated by an excessive frequency noise contribution from the laser. Solving this problem gave a considerable improvement in sensitivity, but studies for further noise reduction are going on.

10 Data Management

An integral part of GEO 600 are the groups at Cardiff and Potsdam. In working on the theory of gravitational wave sources, they help to choose the right settings for the detector parameters. The choice of the data acquisition hardware and the development of the software is made in close collaboration with these groups. Strategies for sensitive and efficient signal detection are being developed both at Potsdam and Cardiff.

11 Outlook

The aim of GEO 600 is to have first science runs of the interferometer in the year 2001. This is well within the time when also other large detectors will begin taking data. It will be an exciting time to see gravitational wave astronomy to come about.

An interesting mix of theoretical, experimental, and technical problems is yet ahead of us, and these tasks could very well, and with benefit for all, be tackled in an international collaboration, by an exchange of ideas as well as of personnel. The GEO groups explicitly endorse any such collaboration.

References

1. B.J. Meers: *Phys. Rev.* **D 38**, 2317 (1988).
2. K.A. Strain, B.J. Meers: *Phys. Rev. Lett.* **66**, 1391 (1991).
3. S. Rowan *et al.*: *Phys. Lett.* **A 233**, 303 (1997).
4. S. Rowan *et al.*: *Phys. Lett.* **A 246**, 471 (1998).
5. O.S. Brozek *et al.*: Proc. 18th Moriond Workshop, in press (1999).
6. A. Rüdiger *et al.*: *Optica Acta* **28**, 641 (1981).
7. K. Skeldon *et al.*: *Rev. Sci. Instrum.* **67**, 2443 (1996).
8. D. Schnier *et al.*: *Phys. Lett.* **A 225**, 210 (1997).
9. J. Mizuno *et al.*: *Phys. Lett.* **A 175**, 273 (1993).
10. J. Mizuno *et al.*: *Opt. Comm.* **138**, 383 (1997).
11. J. Mizuno et al.: paper in preparation (1999).
12. R. Schilling: Optocad, paper in preparation (1999).
13. G. Heinzel: LISO, available from `ftp.rzg.mpg.de/pub/grav/ghh/liso` (1999).
14. G. Heinzel et al.: Automatic beam alignment in the Garching 30-m prototype of a laser-interferometric gravitational wave detector, *Opt. Comm.* **160**, 321 (1999).
15. G. Heinzel et al.: Experimental Demonstration of a Suspended Dual Recycling Interferometer, *Phys. Rev. Lett.* **81**, 5493 (1998).
16. G. Heinzel: *Advanced optical techniques for laser-interferometric gravitational-wave detectors*, Thesis, Univ. Hannover (1999); available as MPQ 243.

Gravitational Radiation Theory and Light Propagation

Luc Blanchet[1], Sergei Kopeikin[2], and Gerhard Schäfer[3]

[1] Département d'Astrophysique Relativiste et de Cosmologie (CNRS), Observatoire de Paris, 92195 Meudon Cedex, France
[2] Department of Physics & Astronomy, University of Missouri-Columbia, Physics Building 223, Columbia, MO 65211, USA
[3] Theoretisch–Physikalisches Institut, Friedrich–Schiller–Universität, Max–Wien–Platz 1, 07743 Jena, Germany

Abstract. The paper gives an introduction to the gravitational radiation theory of isolated sources and to the propagation properties of light rays in radiative gravitational fields. It presents a theoretical study of the generation, propagation, back–reaction, and detection of gravitational waves from astrophysical sources. After reviewing the various quadrupole–moment laws for gravitational radiation in the Newtonian approximation, we show how to incorporate post–Newtonian corrections into the source multipole moments, the radiative multipole moments at infinity, and the back-reaction potentials. We further treat the light propagation in the linearized gravitational field outside a gravitational wave emitting source. The effects of time delay, bending of light, and moving source frequency shift are presented in terms of the gravitational lens potential. Time delay results are applied in the description of the procedure of the detection of gravitational waves.
Pacs Numbers : 04.25.-g, 04.25.Nx

1 Introduction

It was only in the late fifties of the twentieth century that by the work of Hermann Bondi and Joseph Weber gravitational radiation entered the domain of physics. Before that time gravitational radiation was not considered to be of observational relevance and the gravitational radiation theory was not developed very deeply.

The supposed detection of gravitational radiation by J. Weber in the late sixties triggered strong and still on-going efforts both in the building of gravitational wave detectors and in the elaboration of the gravitational radiation theory, including investigations of the most reliable sources of detectable gravitational waves, calculations of wave forms, and analysis of data from detectors (cf. [1]). It turned out that coalescing neutron stars and/or stellar-mass black holes, together with gravitationally collapsing objects (type II supernovae), are the most relevant sources for detectable gravitational waves on Earth because they are strong and fit well to the frequency band of the Earth-based detectors which ranges from 10 Hz to 10 kHz. The strength of these sources is such high that several detection events per year might be expected in future fully developed detectors. The most sensitive Earth-based detectors will go into operation

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 141–166, 2001.

in the first few years of the new millenium. These are the laser-interferometric detectors in Germany, GEO600, built by a German/British consortium, in Italy, VIRGO, built by a Italian/French consortium, the two LIGO detectors in the United States, and the TAMA300 detector in Japan. There are several bar detectors already operating on Earth (ALLEGRO in the United States, AURIGA and NAUTILUS in Italy, EXPLORER at CERN, NIOBE in Australia). These detectors are being permanently upgraded and will be supplementing the measurements of the interferometric detectors later. For the measurement of gravitational waves in the frequency range between 0.1 Hz and 0.1 mHz the space-borne laser-interferometric detector LISA is devised which is expected to be flown around 2010 by NASA/ESA. The astrophysical sources of the gravitational waves to be detected by this detector are a variety of orbiting stars (interacting white dwarf binaries, compact binaries), orbiting massive black holes, as well as the formation and coalescence of supermassive black holes. Stochastic gravitational waves from the early universe are expected to exist in the whole measurable frequency range from 10^4 Hz down to 10^{-18} Hz. The tools to possibly measure the primordial waves are the mentioned Earth-based and space-borne detectors, Doppler tracking, pulsar timing, very long baseline interferometry, as well as the cosmic microwave background.

On the theoretical side there are essentially two approaches which permit to investigate the properties of and to make predictions about gravitational waves from various sources. The first approach, that we can qualify as "exact", stays within the exact theory, solving or establishing theorems about the complete non-linear Einstein field equations. Within this approach one can distinguish the work dealing with exact solutions of the field equations in the form of plane gravitational waves, and especially colliding plane waves. Since the waves are never planar in nature, this work is not very relevant to real astrophysics, but its academic interest is important in that it permits notably the study of the appearance of singularities triggered by collisions of waves. Also within the exact approach, but more important for applications in astrophysics, is all the work concerned with the study of the asymptotic structure of the gravitational field of isolated radiating systems. The work on asymptotics started with the papers of Bondi *et al.* [2] and Penrose [3]. The second approach is much more general, in the sense that it is not restricted to any particular symmetry of the system, nor it is applicable only in the far region of the system. However, the drawback of this approach is that it is only *approximate* and essentially looks for the solutions of the Einstein field equations in the form of formal expansions when $c \to \infty$ (post-Newtonian approximation). This approximate post-Newtonian method can be applied to the study of all theoretical aspects of gravitational radiation: the equations of motion of the source including the gravitational radiation reaction (works of Einstein, Infeld, and Hoffmann [4], Chandrasekhar and Esposito [5], Burke and Thorne [6–8], Ehlers [9], Papapetrou and Linet [10], Damour and Deruelle [11,12], Schäfer [13], Kopejkin [14]); the structure of the radiation field (work of Bonnor [15], Thorne [16], Blanchet and Damour [17]), and, more recently, accurate post-Newtonian wave generation formalisms [18–22].

To the lowest, Newtonian order, the wave-generation formalism is called the quadrupole formalism, because as a consequence of the equality of the inertial and gravitational mass of all bodies the dominant radiating moment of any system is the (mass-type) quadrupole, which simply is at this approximation the standard Newtonian quadrupole moment. We are very much confident in using the theoretical framework of the post-Newtonian approximation because, marvellously enough, the framework of the Newtonian, quadrupole formalism has been checked by astronomical observations.

In fact, there are two observational tests of the validity of the quadrupole formalism. The first test concerns the famous Hulse–Taylor binary pulsar whose decrease of the orbital period P_b by gravitational radiation is predicted from the quadrupole formula to be [23–26]

$$\dot{P}_b = -\frac{192\pi}{5c^5}\left(\frac{2\pi G}{P_b}\right)^{5/3}\frac{M_p M_c}{(M_p+M_c)^{1/3}}\frac{1+\frac{73}{24}e^2+\frac{37}{96}e^4}{(1-e^2)^{7/2}}\ , \tag{1}$$

where M_p and M_c are the pulsar and companion masses, e is the orbit eccentricity, and G, c are the universal gravitational constant and the speed of light. Numerically, one finds $\dot{P}_b = -2.4\times10^{-12}$ sec/sec, in excellent agreement (0.35% precision) with the observations by Taylor *et al.* [27,28]. The second test concerns the so-called cataclysmic variables. There we have binary systems in which a star filling its Roche lobe (the "secondary" with mass M_2) transfers mass onto a more massive white dwarf (the "primary" with mass $M_1 > M_2$). From the formula for the angular momentum in Newtonian theory $J = GM_1M_2(a/GM)^{1/2}$ (where $M = M_1+M_2$), we deduce the secular evolution of the orbital semi-major radius a (whatever may be the mechanism for the variation of J),

$$\frac{\dot{a}}{a} = \frac{2\dot{J}}{J} - \frac{2\dot{M}_2}{M_2}\left(1-\frac{M_2}{M_1}\right)\ , \tag{2}$$

where $\dot{M}_2$ is the rate at which the secondary transfers mass to the primary ($\dot{M}_2 < 0$). Since $M_1 > M_2$, the mass transfer tends to increase the radius a of the orbit, hence to increase the radius of the secondary's Roche lobe, and, thus, to stop the mass transfer. Therefore a long lived mass transfer is possible only if the system looses angular momentum to compensate for the increase of a. For cataclismic binaries with periods longer than about two hours, the loss of angular momentum is explained by standard astrophysical theory (interaction between the magnetic field and the stellar wind of the secondary). But for short-period binaries, with period less than about two hours, the only way to explain the loss of angular momentum is to invoke gravitational radiation. Now, from the quadrupole formula, we have

$$\left(\frac{\dot{J}}{J}\right)_{\rm GW} = -\frac{32G^3}{5c^5}\frac{M_1M_2M}{a^4}\ . \tag{3}$$

Inserting this into (2) one can then predict what should be $\dot{M}_2$ in order that $\dot{a}/a \sim 0$, and the result is in good agreement with the mass transfer measured from the X-rays observations of cataclysmic binaries.

Another important aspect of the theory of gravitational radiation, with obvious implications in astronomy, is the interaction of the gravitational wave field with photons. In this article we present the results of a thorough investigation of light propagation in the gravitational wave field generated by some isolated system. Our motivation is that electromagnetic waves are still the main carrier of astrophysically important information from very remote domains of our universe. Also, the operation of interferometric gravitational wave detectors and other techniques used for making experiments in gravitational physics (lunar laser ranging, very long baseline interferometry, pulsar timing, Doppler tracking, etc.) are fully based on the degree of our understanding of how light propagates in variable, time-dependent gravitational fields generated by various celestial bodies. Although quite a lot of work has been done on this subject (see, for example, [29]-[36]) a real progress and much deeper insight into the nature of the problem has been achieved only recently [37]-[39]. The main advantage of the integration technique which has been developed for finding the light-ray trajectory perturbed by the gravitational field is its account for the important physical property of gravitational radiation, namely, its retardation character. Previous authors, apart from Damour and Esposito-Farèse [40], accounted for the retardation of the gravitational field only in form of plane gravitational waves. Hence, effects produced in the near and induction zones of isolated astronomical sources emitting waves could not be treated in full detail. As a particular example of importance of such effects we note the problem of detection of gravitational waves created by g-modes of the Sun. The space interferometer LISA will be able to detect those waves. However, the problem is that LISA will fly in the induction zone of the emission process of these gravitational waves and, hence, a much more complete theoretical analysis of the working of the detector is needed. The approximation of a plane gravitational wave for the description of the detection procedure is definitely not sufficient. The other example could be effects caused by the time-dependent gravitational field of the ensemble of binary stars in our galaxy. Timing of high-stable millisecond pulsars might be a tool for the detection of stochastic effects produced by that field [41].

2 Wave generation from Isolated Sources

2.1 Einstein Field Equations

The gravitational field is described in general relativity solely by the metric tensor $g_{\mu\nu}$ (and its inverse $g^{\mu\nu}$). It is generated by the stress-energy tensor of the matter fields $T^{\mu\nu}$ via the second-order differential equations

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = \frac{8\pi G}{c^4}\, T^{\mu\nu}\ , \tag{4}$$

where $R^{\mu\nu}$ and $R = g_{\rho\sigma} R^{\rho\sigma}$ denote, respectively, the Ricci tensor and the Ricci scalar. We assume that the matter tensor $T^{\mu\nu}$ corresponds to an isolated source, i.e. $T^{\mu\nu}$ has a spatially compact support with maximal radius a, and that the

internal gravity of the source is weak in the sense that its mass M satisfies $GM \ll ac^2$. Within these conditions it is appropriate to write the metric $g^{\mu\nu}$ in the form of a small deformation of the flat metric $\eta^{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. We pose $h^{\mu\nu} = \sqrt{-g}g^{\mu\nu} - \eta^{\mu\nu}$ (g = determinant of $g_{\mu\nu}$) and assume that each component of $h^{\mu\nu}$ is numerically small: $|h^{\mu\nu}| \ll 1$. We lower and raise all indices of our metric perturbation $h^{\mu\nu}$ with the flat metric; for instance, $h_{\mu\nu} = \eta_{\mu\rho}\eta_{\nu\sigma}h^{\rho\sigma}$ and $h = \eta^{\rho\sigma}h_{\rho\sigma}$. Then the field equations (4) can be re-written in terms of the metric perturbation $h^{\mu\nu}$ by separating out a second-order linear operator acting on $h^{\mu\nu}$, and the remaining part of the equations, which is at least quadratic in $h^{\mu\nu}$ and its first and second derivatives, we conventionally set to the right side of the equations together with the matter tensor. This yields

$$\Box\, h^{\mu\nu} - \partial^\mu H^\nu - \partial^\nu H^\mu + \eta^{\mu\nu}\partial_\rho H^\rho = \frac{16\pi G}{c^4}\,\tau^{\mu\nu}\,, \tag{5}$$

where $\Box = \Box_\eta$ denotes the flat d'Alembertian operator and where $H^\mu \equiv \partial_\nu h^{\mu\nu}$; on the right side of the equation we have put

$$\tau^{\mu\nu} = (-g)\,T^{\mu\nu} + \frac{c^4}{16\pi G}\,\Lambda^{\mu\nu}\,, \tag{6}$$

which represents the total stress-energy distribution of both the matter fields – first term in (6) – and the gravitational field itself – second term involving the non-linear gravitational source $\Lambda^{\mu\nu} = O(h^2)$ (note that $\tau^{\mu\nu}$ transforms as a Minkowskian tensor under Lorentz transformations). The divergence of the left side of (6) is identically zero by virtue of the Bianchi identities, therefore the pseudo-tensor $\tau^{\mu\nu}$ is conserved in the ordinary sense,

$$\partial_\nu \tau^{\mu\nu} = 0\,, \tag{7}$$

which is equivalent to the covariant conservation of the matter tensor, $\nabla_\nu T^{\mu\nu} = 0$. A gauge transformation $h^{\mu\nu} \to h^{\mu\nu} + \partial^\mu\xi^\nu + \partial^\nu\xi^\mu - \eta^{\mu\nu}\partial_\lambda\xi^\lambda$ does not affect the left side of (5), and consequently by solving for a vector ξ^μ the wave equation $\Box\xi^\mu = -H^\mu$ one can arrange that $h^{\mu\nu}$ satisfies the *harmonic-gauge* condition $\partial_\nu h^{\mu\nu} = 0$. In this gauge the field equations (5) simply become

$$\Box\, h^{\mu\nu} = \frac{16\pi G}{c^4}\,\tau^{\mu\nu}\,. \tag{8}$$

We want now to formulate the condition that the source is really isolated, i.e. it does not receive any radiation from other sources located far away, at infinity. Recall that we can express any homogeneous regular solution of the wave equation $\Box h_{\mathrm{hom}} = 0$ at a given field point in terms of the values of h_{hom} at some source points forming a surrounding surface at retarded times. This is the Kirchhoff formula (see e.g. [42]), which reads in the case where the surrounding surface is a sphere,

$$h_{\mathrm{hom}}(\mathbf{x}', t') = \int\int \frac{d\Omega}{4\pi}\left[\frac{\partial}{\partial\rho}(\rho h_{\mathrm{hom}}) + \frac{\partial}{c\partial t}(\rho h_{\mathrm{hom}})\right](\mathbf{x}, t), \tag{9}$$

where $\rho = |\mathbf{x} - \mathbf{x}'|$ and $t = t' - \rho/c$. To formulate the no-incoming radiation condition we say that there should be no such homogeneous regular solutions $h_{\rm hom}$ (since they correspond physically to waves propagating from sources at infinity). Taking the limit $r \to +\infty$ with $t + r/c = $ const. in Kirchhoff's formula, we then arrive at the physical conditions that

$$\lim_{\substack{r\to+\infty \\ t+\frac{r}{c}=\text{const.}}} \left[\frac{\partial}{\partial r}(rh^{\mu\nu}) + \frac{\partial}{c\partial t}(rh^{\mu\nu})\right](\mathbf{x}, t) = 0 \,, \tag{10}$$

and that $r\partial_\lambda h^{\mu\nu}$ should be bounded in this limit. The no-incoming radiation condition (10) is thus imposed at (Minkowskian) past null infinity $\mathcal{J}^-$ in a conformally rescaled space-time diagram.

2.2 Multipole Expansion in Linearized Gravity

For the rest of this Section (and also in Section 5) we shall restrict ourselves to the case of linearized gravity, defined in particular by the neglect of the non-linear gravitational source term $\Lambda^{\mu\nu}$, for which the field equations in harmonic gauge $\partial_\nu h^{\mu\nu} = 0$ read

$$\Box\, h^{\mu\nu} = \frac{16\pi G}{c^4}\, T^{\mu\nu} \,. \tag{11}$$

Within the linearized approximation the matter stress-energy tensor is divergenceless: $\partial_\nu T^{\mu\nu} = 0$. Therefore the linearized approximation is inconsistent as regards the motion of the matter source, which in this approximation stays unaffected by the gravitational field. However this approximation is quite adequate for describing the generation of waves by a given source (for instance acted on by non-gravitational forces). From the no-incoming radiation condition (10), we find that the unique solution of (11) is the retarded one:

$$h^{\mu\nu}(\mathbf{x}, t) = -\frac{4G}{c^4}\int \frac{d^3\mathbf{x}'}{|\mathbf{x} - \mathbf{x}'|}\, T^{\mu\nu}(\mathbf{x}', t - \frac{1}{c}|\mathbf{x} - \mathbf{x}'|) \,. \tag{12}$$

Since we are being interested in the wave-generation problem, we choose the field point outside the source, that is $r = |\mathbf{x}| > a$ (with the origin of the spatial coordinates at the center of the ball with radius a, so $a > |\mathbf{x}'|$), and we decompose in that region $h^{\mu\nu}$ into "multipole moments". The straightforward way to do this is to employ the standard Taylor formula for the formal limit $\mathbf{x}' \to 0$,

$$\frac{T(\mathbf{x}', t - |\mathbf{x} - \mathbf{x}'|/c)}{|\mathbf{x} - \mathbf{x}'|} = \sum_{l=0}^{+\infty} \frac{(-)^l}{l!} x'_L \partial_L \left[\frac{T(\mathbf{x}', t - r/c)}{r}\right] \,. \tag{13}$$

Notice the short-hand notation $L = i_1 i_2 \cdots i_l$ for a multi-index with l indices, as well as $x'_L = x'^L = x'^{i_1} x'^{i_2} \cdots x'^{i_l}$, $\partial_L = \partial_{i_1}\partial_{i_2}\cdots\partial_{i_l}$ where $\partial_i = \partial/\partial x^i$. From this Taylor expansion we immediately arrive at the following expression for the multipole decomposition of the metric perturbation,

$$\mathcal{M}(h^{\mu\nu})(\mathbf{x}, t) = -\frac{4G}{c^4}\sum_{l=0}^{+\infty} \frac{(-)^l}{l!}\partial_L \left[\frac{1}{r}\mathcal{H}^{\mu\nu}_L(t - \frac{r}{c})\right] \,, \tag{14}$$

where the "multipole moments" depend on the retarded time $u \equiv t - r/c$ and are given by

$$\mathcal{H}_L^{\mu\nu}(u) = \int d^3\mathbf{x}' x'_L T^{\mu\nu}(\mathbf{x}', u) \,, \tag{15}$$

In (14) we employ the notation $\mathcal{M}$ to distinguish the multipole expansion $\mathcal{M}(h)$ from h itself. Of course, we have numerically $\mathcal{M}(h) = h$ outside the source, however inside the source $\mathcal{M}(h)$ and h will differ from each other; indeed h is a smooth solution of the equations (12) while $\mathcal{M}(h)$ satisfies $\Box\mathcal{M}(h) = 0$ and becomes infinite when $r \to 0$ [as it is clear from (14)]. In Section 4, dealing with the general case of the non-linear theory, we shall prefer to use the multipole expansion in terms of symmetric and trace-free (STF) multipole moments. In the present case it is simpler to use the non-STF moments $\mathcal{H}_L^{\mu\nu}$. A systematic investigation of the STF multipole expansion in linearized gravity can be found in [43].

Applying the (linearized) conservation law $\partial_\nu T^{\mu\nu} = 0$ we easily find certain physical evolution equations (and conservation laws) to be satisfied by the multipole moments (15). Making use of the Gauss theorem to discard spatial divergences of compact-support terms we successively obtain (with (n) referring to n time-derivatives)

$$\frac{1}{c} \overset{(1)}{\mathcal{H}}{}_L^{\mu 0} = l\mathcal{H}_{L-1)}^{\mu(i_l} \,, \tag{16}$$

$$\frac{1}{c} \overset{(2)}{\mathcal{H}}{}_L^{00} = l(l-1)\mathcal{H}_{L-2)}^{(i_l i_{l-1}} \,, \tag{17}$$

where the round brackets around spatial indices denote the symmetrization (and where $L - 1 = i_1 \cdots i_{l-1}$; $L - 2 = i_1 \cdots i_{l-2}$). As a consequence of (16) we see that the anti-symmetric part of $\mathcal{H}_j^{i0}$ in the indices ij is constant. A more general consequence is

$$\frac{1}{c}\epsilon_{ijk} \overset{(1)}{\mathcal{H}}{}_{kL-1}^{j0} = (l-1)\epsilon_{ijk}\mathcal{H}_{L-2)k}^{j(i_{l-1}} \,. \tag{18}$$

In the case of the lowest-order ($l = 0$ and $l = 1$) multipole moments the right sides vanish, and therefore these equations represent the conservation laws for the corresponding moments. Over all we find ten conservation laws, one for the mass-type monopole or total mass M, three for the mass-type dipole or center of mass position X_i (times M), three for the time derivative of the mass dipole or linear momentum P_i, and three for the current-type dipole or total angular momentum S_i. Specifically, we define

$$M \equiv \quad \frac{1}{c^2}\mathcal{H}^{00} = \int d^3\mathbf{x} \, \frac{T^{00}}{c^2} \,, \tag{19}$$

$$P_i \equiv \quad \frac{1}{c}\mathcal{H}^{0i} = \int d^3\mathbf{x} \, \frac{T^{0i}}{c} \,, \tag{20}$$

$$S_i \equiv \frac{1}{c} \, \epsilon_{ijk}\mathcal{H}_j^{0k} = \epsilon_{ijk} \int d^3\mathbf{x} \, x_j \frac{T^{0k}}{c} \,, \tag{21}$$

$$MX_i \equiv \frac{1}{c^2}\mathcal{H}_i^{00} = \int d^3\mathbf{x}\, x_i \frac{T^{00}}{c^2}\ . \tag{22}$$

Then, from (16)-(18), we have

$$\dot{M} = 0, \qquad \dot{P}_i = 0, \qquad \dot{S}_i = 0, \qquad \text{and} \qquad \dot{X}_i = \frac{P_i}{M}\ . \tag{23}$$

3 The Quadrupole Moment Formalism

3.1 Multipole Expansion in the Far Region

We analyze the gravitational field in the far-zone of the source, in which we perform the expansion of the multipolar expansion $\mathcal{M}(h^{\mu\nu})$ when $r \to +\infty$ with $t - r/c =$const (Minkowskian future null infinity $\mathcal{J}^+$). To leading order $1/r$ the formula (14) yields

$$\mathcal{M}(h^{\mu\nu}) = -\frac{4G}{c^4 r}\sum_{l=0}^{+\infty}\frac{n_L}{c^l l!}\overset{(l)}{\mathcal{H}}{}_L^{\mu\nu}(u) + O\left(\frac{1}{r^2}\right)\ , \tag{24}$$

[where (l) represents the lth time-derivative and where $n_L \equiv n^L = n^{i_1}n^{i_2}\cdots n^{i_l}$ with $n^i = x^i/r$]. Because of the powers of $1/c$ in front of each multipolar piece, it is clear that the far-zone expansion of $\mathcal{M}(h)$ is especially useful when the numerical value of each term of the formula (24) really scales with the factor $1/c^l$ in front of it. This will be the case when the typical velocities of the particles composing the system are small with respect to the speed of light, $v/c \equiv \varepsilon \ll 1$, or, equivalently, when the maximal radius a of the system is much smaller than the wavelength λ of the emitted gravitational radiation ($\lambda = cP$ where P is the typical period of the internal motion). In particular, this "slow motion" assumption is always realized in the case of a self-gravitating system with weak internal gravity, for which we have

$$\varepsilon \equiv \frac{v}{c} \sim \frac{a}{\lambda} \sim \sqrt{\frac{GM}{c^2 a}} \ll 1\ . \tag{25}$$

Thus, for slowly-moving systems we can retain only the first few terms in the multipolar-post-Newtonian expansion (24). Let us restrict ourselves to the terms

$$h^{00} = -\frac{4G}{c^4 r}\left\{\mathcal{H}^{00} + \frac{n_a}{c}\overset{(1)}{\mathcal{H}}{}_a^{00} + \frac{n_{ab}}{2c^2}\overset{(2)}{\mathcal{H}}{}_{ab}^{00}\right\}, \tag{26}$$

$$h^{0i} = -\frac{4G}{c^4 r}\left\{\mathcal{H}^{0i} + \frac{n_a}{c}\overset{(1)}{\mathcal{H}}{}_a^{0i}\right\}, \tag{27}$$

$$h^{ij} = -\frac{4G}{c^4 r}\left\{\mathcal{H}^{ij}\right\}\ . \tag{28}$$

For easier notation we do not indicate the multipole expansion $\mathcal{M}$, nor the neglected $O(1/r^2)$ in the distance to the source. Using the conservation laws

(16)-(17) we can easily re-express the latter expressions in terms of the total mass M, total linear-momentum P_i, and the Newtonian quadrupole moment

$$Q_{ij} \equiv \frac{1}{c^2}\mathcal{H}^{00}_{ij} = \int d^3\mathbf{x}\frac{T^{00}}{c^2}x_i x_j \ . \tag{29}$$

Since M and P_i are conserved they do not participate to the radiation field which is therefore dominantly quadrupolar. Restoring the neglected post-Newtonian error terms we obtain

$$h^{00} = -\frac{4G}{c^2 r}\left\{M + \frac{n_a}{c}P_a + \frac{n_{ab}}{2c^2}\overset{(2)}{Q}_{ab}(u) + O\left(\varepsilon^3\right)\right\}, \tag{30}$$

$$h^{0i} = -\frac{4G}{c^3 r}\left\{P_i + \frac{n_a}{2c}\overset{(2)}{Q}_{ai}(u) + O\left(\varepsilon^2\right)\right\}, \tag{31}$$

$$h^{ij} = -\frac{4G}{c^4 r}\left\{\frac{1}{2}\overset{(2)}{Q}_{ij}(u) + O\left(\varepsilon\right)\right\}. \tag{32}$$

Note that the contribution of the angular momentum S_i appears only at the sub-dominant order $O(1/r^2)$ [see for instance (78) below]. When acting on terms of order $1/r$ in the distance like in (30)-(32) the derivative ∂_ν is proportional to the (Minkowskian) null vector $k_\nu = (-1, \mathbf{n})$; namely $\partial_\nu = -k_\nu\partial_0 + O(1/r^2)$. Using this, the harmonic gauge condition $\partial_\nu h^{\mu\nu} = 0$ reads

$$k_\nu \overset{(1)}{h}{}^{\mu\nu} = O\left(\frac{1}{r^2}\right) \tag{33}$$

which is checked directly to be satisfied by the expressions (30)-(32).

3.2 The Far-Field Quadrupole Formula

Using the gauge freedom $h^{\mu\nu} \to h^{\mu\nu} + \partial^\mu\xi^\nu + \partial^\nu\xi^\mu - \eta^{\mu\nu}\partial_\lambda\xi^\lambda$, we apply a gauge transformation to what is called the transverse-traceless (TT) gauge. Namely we pose

$$\xi^0 = \frac{G}{2rc^3}\left[-n_{ab}\overset{(2)}{Q}_{ab} - \overset{(2)}{Q}_{aa}\right], \tag{34}$$

$$\xi^i = \frac{G}{2rc^3}\left[n_{iab}\overset{(2)}{Q}_{ab} + n_i\overset{(2)}{Q}_{aa} - 4n_a\overset{(2)}{Q}_{ia}\right]. \tag{35}$$

The new metric in TT coordinates, say $h^{\rm TT}_{\mu\nu}$ (where we lower indices with the flat metric), is straightforwardly seen to involve in its 00 and $0i$ components only the static contributions of the mass monopole and dipoles, namely

$$h^{\rm TT}_{00} = -\frac{4GM}{c^2 r}\left(1 + \frac{n_a}{c}\dot{X}_a\right), \qquad h^{\rm TT}_{0i} = -\frac{4GP_i}{c^3 r}\ . \tag{36}$$

In the TT gauge the only radiating components of the field are the spatial ones, ij, and we obtain

$$h_{ij}^{\rm TT} = -\frac{2G}{c^4 r} P_{ijab}(\mathbf{n}) \left\{ \overset{(2)}{I}_{ab}(u) + O(\varepsilon) \right\} + O\left(\frac{1}{r^2}\right) . \tag{37}$$

The latter equation is known as the far-field quadrupole equation. The TT projection operator is defined by $P_{ijab} = P_{ia}P_{jb} - \frac{1}{2}P_{ij}P_{ab}$, where $P_{ij} = \delta_{ij} - n_i n_j$. This is a projector: namely $P_{ijkl}P_{klab} = P_{ijab}$, onto the plane orthogonal to $\mathbf{n}$: thus for instance, $n_i P_{ijab} = 0$; and it is trace free: $P_{ijab}\delta_{ab} = 0$, so we have substituted in (37) the trace free part of the quadrupole moment Q_{ab}, i.e.

$$I_{ij} = Q_{ij} - \frac{\delta_{ij}}{3} Q_{aa} + O\left(\varepsilon^2\right) . \tag{38}$$

We have added a remainder $O(\varepsilon^2)$ to indicate the post-Newtonian corrections in the source moment I_L computed in Section 4 [note that the remainder in (37) is only $O(\varepsilon)$].

As we see all the physical properties of the gravitational wave are contained into the TT projection (37). As a consequence, the effects of the wave on matter fields are transverse: the motion of matter induced by the wave takes place only in the plane orthogonal to the propagation of the wave. Furthermore, from the trace-less character of the wave we see there can be only two independent components or polarization states. We introduce two polarization vectors, $\mathbf{p}$ and $\mathbf{q}$, in the plane orthogonal to the direction of propagation $\mathbf{n}$, forming an orthonormal right-handed triad. In terms of these polarization vectors the projector onto the transverse plane reads $P_{ij} = p_i p_j + q_i q_j$. The two polarization states (customarily referred to as the "plus" and "cross" polarizations) are defined by

$$h_+ = \frac{p_i p_j - q_i q_j}{2} h_{ij}^{\rm TT} , \qquad h_\times = \frac{p_i q_j + p_j q_i}{2} h_{ij}^{\rm TT} . \tag{39}$$

Until very recently all expectations was that any astrophysical (slowly-moving) source would emit gravitational radiation according to (at least dominantly) the quadrupole formula (37), involving the *mass-type* quadrupole moment I_{ij}. For instance the waves from the binary pulsar obey this formula. However, it has been realized by Andersson [44] and Friedman and Morsink [45] that in the case of the secular instability of the r-modes (rotation, or Rossby modes) of isolated newly-born neutron stars, the gravitational radiation is dominated by the variation of the *current-type* quadrupole moment J_{ij}. Here we give, without proof, the formula analogous to (37) but for the current quadrupole:

$$h_{ij|\rm current}^{\rm TT} = \frac{8G}{3c^5 r} P_{ijab}(\mathbf{n}) \, \epsilon_{acd} \, n_c \overset{(2)}{J}_{bd}(u) , \tag{40}$$

where the (trace-free) current quadrupole moment is given by

$$J_{ij} = \epsilon_{ab(i} \int d^3\mathbf{x} \, x_{j)} x_a \frac{T^{0b}}{c} + O\left(\varepsilon^2\right) . \tag{41}$$

It could be, rather ironically, that the first direct detection of gravitational waves would follow the formula (40) rather than the classic formula (37) appearing in all text-books such as [8].

3.3 Energy Balance Equation and Radiation Reaction

The stress-energy pseudo-tensor of all matter and gravitational fields (in harmonic coordinates) $\tau^{\mu\nu}$ is defined by (6). Now, for gravitational waves propagating in vacuum at large distances from their sources (in regions where the waves are almost planar), it is appropriate to define the stress-energy tensor of the waves as the gravitational source term [involving $\Lambda^{\mu\nu}(h)$] in the definition of $\tau^{\mu\nu}$, in which $h_{\mu\nu}$ is replaced by the far-field metric (37). Since the expression of the metric is valid up to fractional terms $O(1/r^2)$ in the distance, and since $\Lambda^{\mu\nu}$ is at least quadratic in h, the stress-energy tensor of gravitational waves will be valid up to $O(1/r^3)$. Thus, we define, in the far-zone,

$$T^{\mu\nu}_{\rm GW} = \frac{c^4}{16\pi G}\,\Lambda^{\mu\nu} + O\left(\frac{1}{r^3}\right)\,. \tag{42}$$

Now to quadratic order $\Lambda^{\mu\nu}$ is a complicated sum of terms like $h\partial\partial h + \partial h\partial h$. But when using (33) together with the fact that $k^2 = 0$, this sum simplifies drastically and we end up with [still neglecting $O(1/r^3)$]

$$T^{\mu\nu}_{\rm GW} = \frac{c^2}{32\pi G}\,k^\mu k^\nu\,\overset{(1)}{h}{}^{\rm TT}_{ij}\overset{(1)}{h}{}^{\rm TT}_{ij} = \frac{c^2}{16\pi G}\,k^\mu k^\nu\left[(\overset{(1)}{h}{}_+)^2 + (\overset{(1)}{h}{}_\times)^2\right]. \tag{43}$$

The second form is obtained from the inverse of (39): $h^{\rm TT}_{ij} = (p_ip_j - q_iq_j)h_+ + (p_iq_j + p_jq_i)h_\times$. The expression (43) takes the classic form $\sigma\,k^\mu k^\nu$ of the stress-energy tensor for a swarm of massless particles (gravitons) moving with the speed of light. Notice from (43) that the energy density of waves is positive definite. In the general case where we do not neglect the terms $O(1/r^3)$ the previous expressions of $T^{\mu\nu}_{\rm GW}$ are still valid, but provided that one performs a suitable average over several gravitational wavelengths (see [8]). For quadrupole waves, substituting the quadrupole formula (37), we get

$$T^{\mu\nu}_{\rm GW} = \frac{G}{8\pi r^2c^4}\,k^\mu k^\nu P_{ijkl}(\mathbf{n})\,\overset{(3)}{I}{}_{ij}\overset{(3)}{I}{}_{kl}\,. \tag{44}$$

We can integrate the conservation law $\partial_\nu\tau^{\mu\nu} = 0$ over the usual three-dimensional space (volume element d^3x), and use the Gauss theorem to obtain a flux of $T^{\mu\nu}_{\rm GW}$ through a surface at infinity (exterior surface element dS_i), so that

$$\frac{d}{dt}\int d^3\mathbf{x}\;\tau^{\mu 0} = -c\int dS_i\,T^{\mu i}_{\rm GW}\,. \tag{45}$$

We consider the $\mu = 0$ component of this law, substitute for $T^{\mu\nu}_{\rm GW}$ the expression (44) at the quadrupole approximation, perform the angular integration assuming

for simplicity a coordinate sphere at infinity (i.e. $dS_i = r^2 n_i d\Omega$), and obtain the famous Einstein (mass-type) quadrupole formula

$$\frac{dE}{dt} = -\frac{G}{5c^5}\left\{ \overset{(3)}{I}_{ij} \overset{(3)}{I}_{ij} + O\left(\varepsilon^2\right)\right\}, \tag{46}$$

where $E = \int d^3x\, \tau^{00}$ represents the energy (matter + gravitation) of the source, and where we re-installed the correct post-Newtonian remainder $O(\varepsilon^2)$. Without proof we give also the formula corresponding to the current-type quadrupole moment,

$$\left(\frac{dE}{dt}\right)_{|\text{current}} = -\frac{16G}{45c^7} \overset{(3)}{J}_{ij} \overset{(3)}{J}_{ij}\,, \tag{47}$$

where the current quadrupole moment is defined by (41).

Interestingly, we can treat the decrease of the energy as the result of the back-action of a radiative force (cf. [46]). We operate by parts the time-derivatives in (46) so as to obtain

$$\frac{d}{dt}\left(E + \frac{\delta E_5}{c^5}\right) = -\frac{G}{5c^5} \overset{(1)}{I}_{ij} \overset{(5)}{I}_{ij} \tag{48}$$

where we put on the left side a term in the form of a total time-derivative, representing a correction of order $1/c^5$ to the energy E, given by

$$\frac{\delta E_5}{c^5} = \frac{G}{5c^5}\left[\overset{(3)}{I}_{ij} \overset{(2)}{I}_{ij} - \overset{(4)}{I}_{ij} \overset{(1)}{I}_{ij}\right]. \tag{49}$$

Now, after a time much longer than the characteristic period of the source (for definiteness one can consider a quasi-periodic source or perform a suitable average; see e.g. [47]), the contribution due to the correcting term (49) will become negligible as compared to the right hand side of (48). Therefore, *in the long term*, we can ignore this term and finally, the equation (48) can be re-written equivalently in a form where the energy loss in the source is the result of the work of a radiation reaction force $\mathbf{F}_{\text{reac}}$, namely

$$\frac{dE}{dt} = -\int d^3x\, \mathbf{F}_{\text{reac}} \cdot \mathbf{v} \tag{50}$$

where

$$F^i_{\text{reac}}(t, \mathbf{x}) = \frac{2G}{5c^5}\, \rho\, x_j \overset{(5)}{I}_{ij}(t)\,, \qquad \rho \equiv T^{00}/c^2\,. \tag{51}$$

The equation (50)-(51) is called the radiation-reaction quadrupole formula; the specific expression (51) of the radiation reaction force is called after Burke and Thorne [6–8]. This force is to be interpreted as a small Newtonian-like force superposed to the usual Newtonian force at the 2.5PN order (or ε^5). Actually, the Burke-Thorne radiation reaction force is valid only in a special gauge. That is, only in a special gauge, differing for instance from the harmonic or ADM gauges, does the source equation of motion involve at 2.5PN order the correcting

force (50)-(51). (See [48] for a discussion of various expressions of the radiation-reaction force in different gauges.) Notice that the reaction force (51) contains time derivatives up to the fifth order inclusively. In practice, for implementation in numerical codes, high time-derivatives have the tendency of decreasing the precision of a numerical computation, and therefore it is advantageous to choose other expressions of the reaction force for implementation in numerical codes [49,50]. On the other hand, the order of second, third, and higher time-derivatives can be reduced by making use of the Newtonian equations of motion of the matter source. Subsequent implication of such a form of the radiation reaction in binary systems leads, for example, to the theoretical prediction of the rate of orbital decay shown in (1).

4 Post-Newtonian Gravitational Radiation

4.1 The Multipole Moments in the Post-Newtonian Approximation

In Section 2 we presented the formula for the multipole expansion of the field outside the source in linearized gravity. In the present section let us present, without proof, the corresponding formula in the full *non-linear* theory, i.e. when the Einstein field equations (8) are solved taking into account the gravitational source term $\Lambda^{\mu\nu}$. The formula will be valid whenever the post-Newtonian expansion is valid, i.e. when (25) holds. Under this assumption the field in the near-zone of a slowly-moving source can be expanded in non-analytic (involving logarithms) series of $1/c$ [17]. The general structure of the expansion is

$$\overline{h}^{\mu\nu}(t,\mathbf{x},c) = \sum_{p,q} \frac{(\ln c)^q}{c^p}\, h^{\mu\nu}_{pq}(t,\mathbf{x}) \;, \tag{52}$$

where $h^{\mu\nu}_{pq}$ are the functional coefficients of the expansion ($p, q =$ integers, including the zero). The general multipole expansion of the metric field $\mathcal{M}(h)$ is found by requiring that when re-developed in the near-zone in the limit of the parameter $r/c \to 0$ (which is equivalent with the formal re-expansion in the limit $c \to \infty$) it *matches* with the previous post-Newtonian expansion (52) in the sense of the mathematical techniques of matched asymptotic expansions, i.e.

$$\overline{\mathcal{M}(h)} = \mathcal{M}(\overline{h}) \;. \tag{53}$$

It is worthwhile noting that the equality (53) should be true in the sense of formal series, i.e. term by term in each coefficient after both sides of the equation are re-arranged with respect to the same expansion parameter.

We find [21,51] that the multipole expansion generalizing (14) to the full theory is composed of two terms,

$$\mathcal{M}(h^{\mu\nu}) = \text{finite part}\, \Box_R^{-1}[\mathcal{M}(\Lambda^{\mu\nu})] - \frac{4G}{c^4}\sum_{l=0}^{+\infty}\frac{(-)^l}{l!}\partial_L\left\{\frac{1}{r}\mathcal{H}^{\mu\nu}_L(t-r/c)\right\} \;, \tag{54}$$

where $\Box_R^{-1}$ is the inverse flat space-time retarded operator. Herein, the first term is a particular solution of the Einstein field equations outside the matter compact support, i.e. it satisfies $\Box h^{\mu\nu}_{\rm part} = \Lambda^{\mu\nu}$, and the second term is a retarded solution of the source-free (homogeneous) wave equation, i.e. $\Box h^{\mu\nu}_{\rm hom} = 0$. The "multipole moments" parametrizing this homogeneous solution are given explicitly by an expression similar to (15),

$$\mathcal{H}^{\mu\nu}_L(u) = \text{finite part} \int d^3\mathbf{x}\; x_L\, \overline{\tau}^{\mu\nu}(\mathbf{x}, u)\;, \tag{55}$$

but involving in place of the matter stress-energy tensor $T^{\mu\nu}$ the *post-Newtonian* expansion $\overline{\tau}^{\mu\nu}$, in the sense of (52), of the *total* (matter+gravitation) pseudo-tensor $\tau^{\mu\nu}$ defined by (6). Both terms in (54) involve an operation of taking the finite part. This finite part can be defined precisely by means of an analytic continuation (see [51] for details), but in fact it is basically equivalent to taking the finite part of a divergent integral in the sense of Hadamard [52]. Notice in particular that the finite part in the expression of the multipole moments (55) deals with the behaviour of the integral *at infinity*: $r \to \infty$ (without the finite part the integral would be divergent because of the factor $x_L \sim r^l$ in the integrand and the fact that the pseudo-tensor $\overline{\tau}^{\mu\nu}$ is not of compact support). One can show that the multipole expansion (54)-(55) is equivalent with a different one proposed recently by Will and Wiseman [22].

Generally, it is more useful (for applications) to express the multipole expansion not in terms of the moments (55), but in terms of symmetric trace-free (STF) moments. We denote the STF projection with a hat, $\hat{x}_L \equiv \text{STF}(x^L)$, so that, for instance, $\hat{x}_{ij} = x_i x_j - \frac{1}{3}\delta_{ij}\mathbf{x}^2$. Then it can be shown that the STF multipole expansion equivalent to (54)-(55) reads,

$$\mathcal{M}(h^{\mu\nu}) = \text{finite part}\, \Box_R^{-1}[\mathcal{M}(\Lambda^{\mu\nu})] - \frac{4G}{c^4}\sum_{l=0}^{+\infty}\frac{(-)^l}{l!}\partial_L\left\{\frac{1}{r}\mathcal{F}^{\mu\nu}_L(t-r/c)\right\}\;, \tag{56}$$

where the parametrizing multipole moments are a bit more complicated,

$$\mathcal{F}^{\mu\nu}_L(u) = \text{finite part} \int d^3\mathbf{x}\; \hat{x}_L \int_{-1}^{1} dz\; \delta_l(z)\overline{\tau}^{\mu\nu}(\mathbf{x}, u + z|\mathbf{x}|/c)\;. \tag{57}$$

With respect to the non-tracefree expression (55) this involves an extra integration over the variable z, with weighting function

$$\delta_l(z) = \frac{(2l+1)!!}{2^{l+1}l!}(1-z^2)^l\;, \qquad \int_{-1}^{1} dz\delta_l(z) = 1\;, \qquad \lim_{l\to+\infty}\delta_l(z) = \delta(z)\;. \tag{58}$$

The results (56)-(58) permit us to define a very convenient notion of the *source multipole moments*. Quite naturally, these are constructed from the ten components of $\mathcal{F}^{\mu\nu}_L(u)$. First of all, we reduce the number of independent components to only six by using the four relations given by the harmonic gauge

condition $\partial_\nu h^{\mu\nu} = 0$. Next we apply standard STF techniques (see [18,43,51] for details), and, in this way, we are able to define *six* STF-irreducible multipole moments, denoted $I_L, J_L, W_L, X_L, Y_L, Z_L$, which are given by *explicit* integrals extending over the post-Newtonian-expanded pseudo-tensor $\overline{\tau}^{\mu\nu}$ like in (57). All of the moments $I_L, J_L, \cdots, Z_L$ are referred to as the moments of the source; however notice that, among them, only the moments I_L (mass-type moment) and J_L (current-type) play a physical role at the *linearized* level. The other four moments W_L, X_L, Y_L, Z_L simply parametrize a linear gauge transformation and can often be omitted from the consideration. Only at the order 2.5PN or ε^5 do they start playing a physical role. The complete formulas for the moments I_L, J_L are [51]

$$\begin{aligned} I_L(u) &= \text{finite part} \int d^3\mathbf{x} \int_{-1}^{1} dz \left\{ \delta_l \hat{x}_L \Sigma - \frac{4(2l+1)}{c^2(l+1)(2l+3)} \delta_{l+1} \hat{x}_{iL} \partial_t \Sigma_i \right. \\ &\left. + \frac{2(2l+1)}{c^4(l+1)(l+2)(2l+5)} \delta_{l+2} \hat{x}_{ijL} \partial_t^2 \Sigma_{ij} \right\} (\mathbf{x}, u + z|\mathbf{x}|/c) \ , \quad (59) \\ J_L(u) &= \text{finite part} \int d^3\mathbf{x} \int_{-1}^{1} dz \ \varepsilon_{ab<i_l} \left\{ \delta_l \hat{x}_{L-1>a} \Sigma_b \right. \\ &\left. - \frac{2l+1}{c^2(l+2)(2l+3)} \delta_{l+1} \hat{x}_{L-1>ac} \partial_t \Sigma_{bc} \right\} (\mathbf{x}, u + z|\mathbf{x}|/c) \ . \quad (60) \end{aligned}$$

In these expressions, $<>$ refers to the STF projection, and we have posed

$$\Sigma \equiv \frac{\overline{\tau}^{00} + \overline{\tau}^{ii}}{c^2} \quad (\text{where } \overline{\tau}^{ii} \equiv \delta_{ij} \overline{\tau}^{ij}) \ , \qquad \Sigma_i \equiv \frac{\overline{\tau}^{0i}}{c} \ , \qquad \Sigma_{ij} \equiv \overline{\tau}^{ij} \ . \quad (61)$$

The moments I_L, J_L given by these formulas are valid formally up to any post-Newtonian order. They constitute the generalization in the non-linear theory of the Newtonian moments introduced earlier in (38) and (41).

In order to apply usefully these moments to a given problem, one must find the explicit expressions of the moments at a given post-Newtonian order by inserting into them the components of the pseudo-tensor $\overline{\tau}^{\mu\nu}$ obtained from an explicit post-Newtonian algorithm. Without entering into details, we find for instance that at the 1PN order the mass-type source moment I_L is given (rather remarquably) by a simple compact-support formula [17,21], on which we can, therefore, remove the finite part prescription:

$$I_L = \int d^3\mathbf{x} \left\{ \hat{x}_L \sigma + \frac{|\mathbf{x}|^2 \hat{x}_L}{2c^2(2l+3)} \partial_t^2 \sigma - \frac{4(2l+1)\hat{x}_{iL}}{c^2(l+1)(2l+3)} \partial_t \sigma_i \right\} + O\left(\varepsilon^4\right) \ . \quad (62)$$

We denote the compact-support parts of the source scalar and vector densities in (61) by

$$\sigma \equiv \frac{T^{00} + T^{ii}}{c^2} \ , \qquad \sigma_i \equiv \frac{T^{0i}}{c} \ . \quad (63)$$

See Blanchet and Schäfer [53] for application of the formula (62) to the computation of the relativistic correction in the $\dot{P}_b$ of a binary pulsar [given to lowest

order by (1)]. On the other hand, Damour, Soffel, and Xu [54] (extending previous work of Brumberg and Kopeikin [55], [56]) used the formula in their study of the Solar-system dynamics at 1PN order. The property of being of compact support is a special feature of the 1PN mass-moment I_L. To higher-order (2PN and higher) the mass moment I_L is intrinsically of non-compact support (see its expression in [21]); hence the finite part prescription in the definition of the moment I_L plays a crucial role at 2PN. Similarly, starting already at 1PN order, the current-moment J_L is intrinsically of non-compact support [19,21].

In a linear theory, the source multipole moments coincide evidently with the radiative multipole moments, defined as the coefficients of the multipole expansion of the $1/r$ term in the distance to the source at retarded times $t - r/c$ =const. (this is evident from Section 2). However, in a non-linear theory like general relativity, the source multipole moments interact with each other in the exterior field through the non-linearities. This is clear from the presence of the first term in (56), containing the gravitational source $\Lambda^{\mu\nu}$, and which does contribute to the $1/r$ part of the metric at infinity. Therefore the source multipole moments must be related to the radiative ones, the latter constituting in this approach the actual observables of the field at infinity.

In the TT projection of the metric field one can define two sets of radiative moments U_L (mass-type) and V_L (current-type). The *definition* of these moments is that they parametrize the $1/r$-term of the ij components of the metric in TT gauge. Thus, extending the formula (37), the radiative moments are given by the decomposition

$$h_{ij}^{\rm TT} = -\frac{4G}{c^2 r} P_{ijab} \sum_{l\geq 2} \frac{1}{c^l l!} \left\{ n_{L-2} U_{abL-2} - \frac{2l}{c(l+1)} n_{cL-2} \varepsilon_{cd(a} V_{b)dL-2} \right\} + O\left(\frac{1}{r^2}\right). \tag{64}$$

The radiative moments U_L, V_L are related to the l-th time-derivatives of the corresponding source moments. Let us give, without proof, the result of the connection of the radiative moments to the source moments (59)-(60) to the order ε^3 (or 1.5PN) inclusively. To this order some non-linear "monopole-radiative l-pole" interactions appear, which correspond physically to the scattering of the l-pole wave on the static curvature induced by the total mass of the source (i.e. the mass monopole $M \equiv I$) – an effect well known under the name of tail of gravitational waves. We find [20,21]

$$U_L(u) = \overset{(2)}{I}_L(u) + \frac{2GM}{c^3} \int_0^{+\infty} d\tau \overset{(4)}{I}_L(u-\tau) \left[\ln\left(\frac{\tau}{2b}\right) + \kappa_l \right] + O\left(\varepsilon^4\right), \tag{65}$$

$$V_L(u) = \overset{(2)}{J}_L(u) + \frac{2GM}{c^3} \int_0^{+\infty} d\tau \overset{(4)}{J}_L(u-\tau) \left[\ln\left(\frac{\tau}{2b}\right) + \pi_l \right] + O\left(\varepsilon^4\right). \tag{66}$$

The same expressions come out from the Will and Wiseman formalism [22]. Here, b is a normalization constant (essentially irrelevant since it corresponds to

a choice of the origin of time in the far zone), and κ_l, π_l are given by

$$\kappa_l = \frac{2l^2+5l+4}{l(l+1)(l+2)} + \sum_{k=1}^{l-2} \frac{1}{k} \,, \qquad \pi_l = \frac{l-1}{l(l+1)} + \sum_{k=1}^{l-1} \frac{1}{k} \,. \tag{67}$$

The first term in (65) relates essentially to the original quadrupole formula (see (37)). From (65)-(66) one sees that the first non-linearity in the propagation of the waves is at 1.5PN order with respect to the quadrupole formula. Thus, with enough precision, one can replace in (65) the mass-type moment I_L by its compact-support expression that we computed in (62) [since the post-Newtonian remainder in (62) is $O(\varepsilon^4)$].

4.2 Post-Newtonian Radiation Reaction

Emission of gravitational radiation affects the equations of motion of an isolated system dominantly at the 2.5PN order beyond the Newtonian acceleration. In a suitable gauge the radiation-reaction force density at the 2.5PN order is given by the quadrupole formula (51). In this Section we extend this quadrupole formula to include the relativistic corrections up to the relative 1.5PN order, which means the absolute 4PN order with respect to the Newtonian force. The method is to compute the radiation reaction by means of the matching [in the sense of (53)] of the post-Newtonian field to the exterior multipolar field. Indeed, recall that the post-Newtonian field is valid only in the near zone, and, thus, only via a matching can it incorporate information from the correct boundary condition, *viz* the no-incoming radiation condition imposed at infinity by equation (10), which specifies the braking character of gravitational radiation reaction.

To the relative 1.5PN order, and in a suitable gauge, it can be shown that the reaction force derives from some "electromagnetic-like" scalar and vector reaction potentials $V^{\rm reac}$ and $V_i^{\rm reac}$. Explicitly we have [57]

$$\begin{aligned} V^{\rm reac}(\mathbf{x},t) = &-\frac{G}{5c^5} x_{ij} \left\{ \overset{(5)}{I}_{ij}(t) + \frac{4GM}{c^3} \int_0^{+\infty} d\tau \ln\left(\frac{\tau}{2b}\right) \overset{(7)}{I}_{ij}(t-\tau) \right\} \\ &+ \frac{G}{c^7} \left[\frac{1}{189} x_{ijk} \overset{(7)}{I}_{ijk}(t) - \frac{1}{70} \mathbf{x}^2 x_{ij} \overset{(7)}{I}_{ij}(t) \right] + O\left(\varepsilon^9\right) \,, \end{aligned} \tag{68}$$

$$V_i^{\rm reac}(\mathbf{x},t) = \frac{G}{c^5} \left[\frac{1}{21} \hat{x}_{ijk} \overset{(6)}{I}_{jk}(t) - \frac{4}{45} \epsilon_{ijk} x_{jm} \overset{(5)}{J}_{km}(t) \right] + O\left(\varepsilon^7\right) \,. \tag{69}$$

The dominant term in the formula (68)-(69) is the standard Burke-Thorne reactive *scalar* potential at 2.5PN order [compare (68) with (51)]. In this term, consistently with the approximation, one must insert the 1PN expression of the moment as given by (62). The Burke-Thorne term is of "odd"-parity-type as it corresponds to an odd power of $1/c$, and thus changes sign upon a time reversal (or more precisely when we replace the retarded potentials by advanced ones). Similarly is the next term in (68)-(69) is 3.5PN (i.e. ε^7), involving both the mass-quadrupole, mass-octupole and current-quadrupole moments (the term ε^5

in the vector potential $V_i^{\rm reac}$ corresponds really to ε^7 in the equations of motion). However, notice that the next term in the reaction scalar potential $V^{\rm reac}$, at 4PN or ε^8 order, belongs to the "even"-parity-type. Nevertheless this term is really part of the radiation reaction for it is not invariant under a time reversal, as is involves an integration over the "past" history of the source, so that when changing the retarded potentials to advanced ones the integration range would change to the whole "future"; hence, this term does not stay invariant. It represents the contribution of tails in the radiation-reaction force, and is nicely consistent, in the sense of energy conservation, with the tails in the far zone [equation (65)]. For explicit computations of the back-reaction to 3.5PN order in the case of point-mass binary systems see Iyer and Will [58,59], and Jaranowski and Schäfer [60].

Using the matching (53) one finds that the near-zone post-Newtonian metric (to 1.5PN relative order in both the "damping" and "conservative" effects) is parametrized in this gauge by some generalized potentials

$$\mathcal{V}_\mu = \Box_{\rm sym}^{-1}\left[-4\pi G\sigma_\mu\right] + V_\mu^{\rm reac}\ . \tag{70}$$

The first term represents, to this post-Newtonian order, the conservative part of the metric; it is of the normal "even"-parity-type and is given by the usual symmetric integral (half-retarded plus half-advanced) of the source densities $\sigma_\mu = (\sigma, \sigma_i)$ given by (63). The second term $V_\mu^{\rm reac}$ denotes the radiation-reaction potentials (68)-(69). By inserting the metric parametrized by (70) into the equations of motion of the source (i.e. $\partial_\nu \tau^{\mu\nu} = 0 \Leftrightarrow \nabla_\nu T^{\mu\nu} = 0$), and considering the integral of energy, we obtain the balance equation

$$\frac{dE}{dt} = \int d^3\mathbf{x}\left\{-\sigma\partial_t V^{\rm reac} + \frac{4}{c^2}\sigma_j\partial_t V_j^{\rm reac}\right\} + O\left(\varepsilon^9\right)\ . \tag{71}$$

Here E denotes the energy of the source at the 1PN (or even 1.5PN) order. Actually what we obtain is not E but some $E + \delta E_5/c^5 + \delta E_7/c^7$ like in (48). Arguing as before we neglect these δE_5 and δE_7. Substituting now the expressions (68)-(69) for the reactive potentials (and neglecting other δE's) we get

$$\begin{aligned}\frac{dE}{dt} = &-\frac{G}{5c^5}\left\{\overset{(3)}{I}_{ij} + \frac{2GM}{c^3}\int_0^{+\infty} d\tau \ln\left(\frac{\tau}{2b}\right)\overset{(5)}{I}_{ij}(t-\tau)\right\}^2 \\ &-\frac{G}{c^7}\left[\frac{1}{189}\overset{(4)}{I}_{ijk}\overset{(4)}{I}_{ijk} + \frac{16}{45}\overset{(3)}{J}_{ij}\overset{(3)}{J}_{ij}\right] + O\left(\varepsilon^9\right)\ ,\end{aligned} \tag{72}$$

The right-side is exactly in agreement with the computation of the total flux energy emitted in gravitational waves at infinity, which is computed making use of the stress-energy tensor of gravitational waves (43). In particular we recover in the brackets of the first term of (72) the third time-derivative of the radiative moment U_{ij} including its tail contribution. The difference with the standard derivation of the flux is that instead of computing a surface integral at infinity we have performed the computation completely within the source, using the local source equations of motion.

5 Light Propagation in Gravitational Fields of Isolated Sources

5.1 General Solution of the Light Propagation Equation

We are going now to calculate in linearized approximation the propagation of a light ray in the gravitational field of an isolated source at rest showing a mass-monopole, a spin-dipole and a time-dependent mass-quadrupole moment. In linearized approximation, the propagation equation for a particle (massless or with mass) with space coordinates $x^i(t)$ reads

$$\ddot{x}^i(t) = \frac{1}{2}g_{00,i} - g_{0i,t} - \frac{1}{2}g_{00,t}\dot{x}^i - g_{ik,t}\dot{x}^k - (g_{0i,k} - g_{0k,i})\,\dot{x}^k - g_{00,k}\dot{x}^k\dot{x}^i - \left(g_{ik,j} - \frac{1}{2}g_{kj,i}\right)\dot{x}^k\dot{x}^j + \left(\frac{1}{2}g_{kj,t} - g_{0k,j}\right)\dot{x}^k\dot{x}^j\dot{x}^i \,, \tag{73}$$

where the metric coefficients $g_{\mu\nu} = \eta_{\mu\nu} + f_{\mu\nu}$, in linear approximation, are related with $h^{\mu\nu}$ from Sections 2–4 through $f_{\mu\nu} = -h_{\mu\nu} + \frac{1}{2}\eta_{\mu\nu}h$. The dots denote differentiation with respect to time t and $c = 1$ has been put for simplicity. In the linear approximation scheme, the velocity $\dot{x}^i$ appearing on the right-hand-side of (73) can be treated as a constant vector. For massless particles (photons) it has unit length, i.e. $\dot{x}^i\dot{x}^i = 1$. In the following we use $\dot{x}^i = k^i$ in the right hand side of (73).

The unperturbed motion of photons reads

$$x^i(t) = x_0^i + k^i(t - t_0) \,, \tag{74}$$

where x_0^i denotes the position of the photon at time of emission t_0. For solving the light propagation equation (73) it is very convenient to introduce the new time parameter τ defined by $\tau = t - t^*$, where t^* denotes the time of the closest approach of the photon to the source of the gravitational field. Then it holds

$$x^i(\tau) = \xi^i + k^i\tau \,, \tag{75}$$

where ξ^i is the vector pointing from the position of the source to the position of the photon at the closest approach. Its length is the impact parameter $|\xi| = d$. Obviously, ξ^i and k^i are orthogonal to each other in the euclidean sense, i.e. $\xi^i k^i = 0$. Therefore, the length of x^i, $r = |x|$, takes the simple form $r = \sqrt{\tau^2 + d^2}$. Introducing the derivatives $\hat{\partial}_i = \hat{P}_{ij}\partial/\partial\xi^j$ and $\hat{\partial}_\tau = \partial/\partial\tau$, where $\hat{P}_{ij} = \delta_{ij} - k_ik_j$ is the projection operator onto the plane orthogonal to k^i, allows the light-propagation equation to be written as

$$\ddot{x}^i(\tau) = \frac{1}{2}k^\alpha k^\beta\hat{\partial}_i f_{\alpha\beta} - \hat{\partial}_\tau\left(k^\alpha f_{i\alpha} + \frac{1}{2}k^i f_{00} - \frac{1}{2}k^i k^j k^p f_{jp}\right) \,, \tag{76}$$

where the four-dimensional vector k^α reads $k^\alpha = (1, k^i)$.

To get a complete overview of the influence of a gravitational wave, emitted from an isolated source, on the propagation of light rays, we use the representation of the metric coefficients (14) which is valid all-over in the space

outside a domain which includes the matter source. Splitting $f_{\mu\nu}$ into a canonical part $f^{\rm can}_{\mu\nu}$ which contains trace-free tensors only, and a gauge part, i.e. $f_{\mu\nu} = f^{\rm can}_{\mu\nu} + \partial_\mu w_\nu + \partial_\nu w_\mu$, we obtain in case of mass-monopole, spin-dipole, and mass-quadrupole source moments (remember the source being at rest)

$$f^{\rm can}_{00} = \frac{2M}{r} + \partial_{pq}\left[\frac{I_{pq}(t-r)}{r}\right] , \tag{77}$$

$$f^{\rm can}_{0i} = -\frac{2\varepsilon_{ipq}S_p n_q}{r^2} + 2\partial_j\left[\frac{\dot I_{ij}(t-r)}{r}\right] , \tag{78}$$

$$f^{\rm can}_{ij} = \delta_{ij} f^{\rm can}_{00} + \frac{2}{r}\ddot I_{ij}(t-r) . \tag{79}$$

Herein, for simplicity, we have put $G = 1$, and $\partial_i = \partial/\partial x^i$. The mass M, spin S^i, and the quadrupole moment I_{ij} of the source of gravitational waves are given (in the Newtonian approximation) by the expressions (19), (21), and (38). The explicit expressions for the gauge functions w^μ relating $f^{\rm can}_{\mu\nu}$ with $f_{\mu\nu}$ are important for general discussion of light propagation in the field of gravitational waves emitted by the isolated source. However, for the sake of simplicity they will be omitted. Their precise form can be found in [37].

The insertion of these expressions (77)-(79) into the equation (76) results in the equation

$$\begin{aligned}\ddot x^i(\tau) = {}& \left[2M\left(\hat\partial_i - k_i\hat\partial_\tau\right) - 2S^p\left(\varepsilon_{ipq}\hat\partial_{q\tau} - k_j\varepsilon_{jpq}\hat\partial_{iq}\right)\right]\left\{\frac{1}{r}\right\} \\ & + \left(\hat\partial_{ipq} - k_i\hat\partial_{pq\tau} + 2k_p\hat\partial_{iq\tau}\right)\left\{\frac{I_{pq}(t-r)}{r}\right\} - 2\hat P_{ij}\hat\partial_{q\tau}\left\{\frac{\dot I_{jq}(t-r)}{r}\right\} \\ & - \hat\partial_{\tau\tau}\left[w^i + \varphi^i - k^i\ \left(w^0 + \varphi^0\right)\right] ,\end{aligned} \tag{80}$$

where the vector φ^μ denotes terms which are of gauge type. The precise form of φ^μ is not important here and can be found in [37].

The solution of equation (80), using the boundary conditions $\dot x^i(-\infty) = k^i$ and $x^i(\tau_0) = x^i_0$ – emission point in space of the light ray at time τ_0, reads

$$\dot x^i(\tau) = k^i + \dot\Xi^i(\tau) , \tag{81}$$

$$x^i(\tau) = x^i_N(\tau) + \Xi^i(\tau) - \Xi^i(\tau_0) , \tag{82}$$

where $x^i_N(\tau)$ denotes the unperturbed trajectory (75). The relativistic perturbation of the trajectory is given by

$$\begin{aligned}\dot\Xi^i(\tau) = {}& \left(2M\hat\partial_i + 2S^p k_j\varepsilon_{jpq}\hat\partial_{iq}\right)A(\tau,\xi) + \hat\partial_{ipq}B_{pq}(\tau,\xi) - \\ & \left(2Mk_i + 2S^p\varepsilon_{ipq}\hat\partial_q\right)\left\{\frac{1}{r}\right\} - \left(k_i\hat\partial_{pq} - 2k_p\hat\partial_{iq}\right)\left\{\frac{I_{pq}(t-r)}{r}\right\} \\ & - 2\hat P_{ij}\hat\partial_q\left\{\frac{\dot I_{jq}(t-r)}{r}\right\} - \hat\partial_\tau\left[w^i + \varphi^i - k^i\ \left(w^0 + \varphi^0\right)\right] ,\end{aligned} \tag{83}$$

$$\Xi^i(\tau) = \left(2M\hat{\partial}_i + 2S^p k_j \varepsilon_{jpq}\hat{\partial}_{iq}\right) B(\tau,\xi) - \left(2Mk_i + 2S^p\varepsilon_{ipq}\hat{\partial}_q\right) A(\tau,\xi) \quad (84)$$
$$+ \hat{\partial}_{ipq} D_{pq}(\tau,\xi) - \left(k_i\hat{\partial}_{pq} - 2k_p\hat{\partial}_{iq}\right) B_{pq}(\tau,\xi) - 2\hat{P}_{ij}\hat{\partial}_q C_{jq}(\tau,\xi) -$$
$$- w^i(\tau,\xi) - \varphi^i(\tau,\xi) + k^i \ \left[w^0(\tau,\xi) + \varphi^0(\tau,\xi)\right] \ ,$$

whereby the scalar functions A and B and derivatives of the tensors B_{ij}, C_{ij}, D_{ij} are known fully explicitly. They are given by

$$A(\tau,\xi) \equiv \int \frac{d\tau}{r} = \int \frac{d\tau}{\sqrt{d^2+\tau^2}} = -\ln\left(\sqrt{d^2+\tau^2} - \tau\right) \ , \quad (85)$$

$$B(\tau,\xi) \equiv \int A(\tau,\xi)d\tau = -\tau\ln\left(\sqrt{d^2+\tau^2} - \tau\right) - \sqrt{d^2+\tau^2} \ , \quad (86)$$

$$\hat{\partial}_k B_{ij}(\tau,\xi) = (yr)^{-1} I_{ij}(t-r)\xi^k \ , \quad (87)$$

$$\hat{\partial}_k C_{ij}(\tau,\xi) = (yr)^{-1} \dot{I}_{ij}(t-r)\xi^k \ , \quad (88)$$

$$\hat{\partial}_{ijk} D_{pq}(\tau,\xi) = \frac{1}{y}\left[\left(\hat{P}^{ij} + \frac{\xi^i\xi^j}{yr}\right)\hat{\partial}_k B_{pq}(\tau,\xi) + \hat{P}^{jk}\hat{\partial}_i B_{pq}(\tau,\xi) + \xi^j\hat{\partial}_{ik}B_{pq}(\tau,\xi)\right] , \quad (89)$$

where the variable $y = \tau - \sqrt{\tau^2+d^2}$ is the retarded time argument for the photon which passes through the point of closest approach to the source of the gravitational radiation at time $t^* = 0$. More details concerning the method of calculation of light ray trajectory in time dependent gravitational fields can be found in [37,38].

5.2 Time Delay and Bending of Light

The time delay results in the form

$$t - t_0 = |\mathbf{x} - \mathbf{x}_0| - \mathbf{k}\cdot\Xi(\tau) + \mathbf{k}\cdot\Xi(\tau_0) \ , \quad (90)$$

or

$$t - t_0 = |\mathbf{x} - \mathbf{x}_0| + \Delta_M(t,t_0) + \Delta_S(t,t_0) + \Delta_Q(t,t_0) \ , \quad (91)$$

where $|\mathbf{x} - \mathbf{x}_0|$ is the usual Euclidean distance between the points of emission, $\mathbf{x}_0$, and reception, $\mathbf{x}$, of the photon, Δ_M is the classical Shapiro delay produced by the (constant) spherically symmetric part of the gravitational field of the deflector (see, e.g. [8]), Δ_S is the Lense-Thirring or Kerr delay due to the (constant) spin of the localized source of gravitational waves [39], and Δ_Q describes an additional delay caused by the time dependent quadrupole moment of the

source [37]. Specifically we obtain

$$\Delta_M = 2M \ln\left[\frac{r+\tau}{r_0+\tau_0}\right] , \tag{92}$$

$$\Delta_S = -2\varepsilon_{ijk} k^j S^k \hat{\partial}_i \ln\left[\frac{r+\tau}{r_0+\tau_0}\right] , \tag{93}$$

$$\Delta_Q = \hat{\partial}_{ij}\left[B_{ij}(\tau,\xi) - B_{ij}(\tau_0,\xi)\right] + \delta_Q(\tau,\xi) - \delta_Q(\tau_0,\xi) , \tag{94}$$

where

$$\delta_Q(\tau,\xi) = k^i\left(w^i + \varphi^i\right) - w^0 - \varphi^0 . \tag{95}$$

Let us now denote by α^i the dimensionless vector describing the total angle of deflection of the light ray measured at the point of observation and calculated with respect to vector k^i given at past null infinity. It is defined according to the relationship

$$\alpha^i(\tau,\xi) = k^i[\mathbf{k}\cdot\dot{\Xi}(\tau,\xi)] - \dot{\Xi}^i(\tau,\xi) = -\,\hat{P}_{ij}\,\dot{\Xi}^j(\tau,\xi) . \tag{96}$$

For observers being far away from the source of the gravitational wave the projection of the mass-quadrupole tensor of the source of gravitational radiation onto the plane orthogonal to the propagation direction of the gravitational wave is the crucial object which enters into the observable effects. It reads

$$I_{ij}^{\rm TT} = P_{ijpq} I_{pq} = I_{ij} + \frac{1}{2}\left(\delta_{ij} + n_i n_j\right) n_p n_q\; I_{pq} - \left(\delta_{ip} n_j n_q + \delta_{jp} n_i n_q\right)\; I_{pq} , \tag{97}$$

where again we denote $n_i = x^i/r$.

In the case of small impact parameter d ($d/r_0 \ll 1, d/r \ll 1$) we respectively obtain for the time delay and the angle of deflection

$$t - t_0 - |\mathbf{x} - \mathbf{x}_0| = -4\psi + 2M\ln(4rr_0) , \tag{98}$$

$$\alpha_i = 4\hat{\partial}_i\psi , \tag{99}$$

where ψ is the gravitational lens potential having the form

$$\psi = \left[M + \varepsilon_{jpq} k^p S^q \hat{\partial}_j + \frac{1}{2}\, I_{pq}^{\rm TT}(t^*)\,\hat{\partial}_{pq}\right]\ln d . \tag{100}$$

(Notice that in this gravitational lens approximation $n^i = k^i$ holds.) Remarquably, the gravitational lens potential does depend on the gravitational source mass-quadrupole tensor only through its value at the time of closest approach. Furthermore, the gravitational lens potential decays like $1/d^2$, i.e. it is not being influenced by the wave part of the gravitational field [40,37].

A direct consequence of the time delay formula is the frequency shift formula for a moving gravitational source

$$\frac{\delta\nu}{\nu} = 4\frac{\partial\psi}{\partial t^*} + v^i\alpha^i , \tag{101}$$

where v^i is the velocity of the observer. It is worthwhile noting that the expression (101) holds for the source of electromagnetic waves being at past null infinity. An exhaustive treatment of the gravitational frequency shift for arbitrary locations of observer, source of light, and the source of gravitational waves is rather complicated and has been done only recently in [38].

6 Detection of Gravitational Waves

In the asymptotic regime of a gravitational wave field, the time delay reads

$$\Delta_Q(\mathbf{k}; t, t_0) = \frac{k^i k^j}{1-\cos\theta}\left[\frac{\dot{I}_{ij}^{\mathrm{TT}}(t-r)}{r} - \frac{\dot{I}_{ij}^{\mathrm{TT}}(t_0-r_0)}{r_0}\right] , \tag{102}$$

where θ is the angle between the receiver - (light) emitter and receiver - (gravitational wave) source direction ($\cos\theta = -N^i k^i$) and where the assumption $|\mathbf{x}-\mathbf{x}_0| \ll r$ has been made. Let us now apply the above formula to the time delay in a Michelson interferometer. Obviously, in this case, $r = r_0$ holds. For simplicity we assume that the interferometer device is oriented orthogonal to the propagation direction of the gravitational wave, i.e. $N^i k_1^i = N^i k_2^i$, where k_1^i and k_2^i denote the directions of the two interferometer arms which are taken to be orthogonal ($k_1^i k_2^i = 0$) and of equal length L. We also assume that the light, emitted from the beam-splitter, is reflected once at the end mirrors. Furthermore, the interferometer arms are to be oriented such that they coincide with the main axes of the plus-polarization. Then the relative time delay of the reflected light beams reads

$$\Delta_Q(\mathbf{k}_1; t, t-2L) = \frac{k_1^i k_1^j}{r_0}[\dot{I}_{ij}^{\mathrm{TT}}(t-r_0) - \dot{I}_{ij}^{\mathrm{TT}}(t-r_0-2L)] . \tag{103}$$

The multiplication of the relative time delay by the angular frequency of the laser light, ω, which is treated as constant in the approximation under consideration, results in the measurable phase shift at time t of $\Delta\Phi(t) = 2\omega\Delta_Q(\mathbf{k}_1; t, t-2L)$. If we assume for the plus-component of the gravitational wave the expression $h_+(t-r_0) = A_+\cos(\omega_g t)$, where ω_g is the constant frequency of the wave and A_+ its constant amplitude, we obtain for the phase shift (cf. [61])

$$\Delta\Phi(t) = 2A_+\frac{\omega}{\omega_g}\sin(\omega_g L)\cos[\omega_g(t-L)] . \tag{104}$$

The maximal amplitude is achieved if the condition $\omega_g L = \pi/2$ holds. This yields

$$\Delta_{\max}\Phi(t) = 2A_+\frac{\omega}{\omega_g}\sin(\omega_g t) . \tag{105}$$

At the photo-diode the following photo-current results [62]

$$I_{\mathrm{ph}}(t) = I_{\min} + \frac{I_{\max}-I_{\min}}{2}\left[1-\cos\phi(t)\right] , \tag{106}$$

where the phase $\phi(t)$ is composed out of the signal $\Delta\Phi$ plus a modulation term from Pockels cells, $\phi_m \sin(\omega_m t)$, i.e.

$$\phi(t) = \Delta\Phi(t) + \phi_m \sin(\omega_m t) \; . \tag{107}$$

The approximate decomposition of the expression (106) into a *dc* ("direct current" or non-alternate) part and a signal part reads $I_{\rm ph}(t) = I_{\rm dc} + I_{\omega_m}$, where

$$I_{\rm dc} = I_{\rm min} + \frac{I_{\rm eff}}{2}\left[1 - J_0(\phi_m)\cos\Delta\Phi(t)\right] \; , \tag{108}$$

$$I_{\omega_m} = I_{\rm eff} J_1(\phi_m)\sin\Delta\Phi(t)\sin(\omega_m t) \; , \tag{109}$$

with Bessel functions J_0 and J_1; $I_{\rm eff} = I_{\rm max} - I_{\rm min}$. Because of the smallness of the gravitational phase shift we get

$$I_{\rm dc} = I_{\rm min} + \frac{I_{\rm eff}}{2}\left[1 - J_0(\phi_m)\right] \; , \tag{110}$$

$$I_{\omega_m} = I_{\rm eff} J_1(\phi_m)\Delta\Phi(t)\sin(\omega_m t) \; . \tag{111}$$

Using the equation (105), the latter equation can be written

$$I_{\omega_m} = I_{\rm eff} J_1(\phi_m) A_+ \frac{\omega}{\omega_g}\left\{\cos\left[(\omega_m - \omega_g)t\right] - \cos\left[(\omega_m + \omega_g)t\right]\right\} \; . \tag{112}$$

In this side-band form, the signal from the gravitational wave is being detected. For more details we refer to [61].

References

1. D.G. Blair (ed.): *The Detection of Gravitational Waves* (Cambridge University Press, Cambridge 1991).
2. H. Bondi, M.G.J. van der Burg, and A.W.K. Metzner, *Proc. R. Soc. London* **A 269**, 21 (1962).
3. R. Penrose: *Proc. R. Soc.* (London) **A 284**, 159 (1965).
4. A. Einstein, L. Infeld, and B. Hoffmann: *Ann. Math.* **39**, 65 (1938).
5. S. Chandrasekhar and F.P. Esposito: *Astrophys. J.* **160**, 153 (1970).
6. W.L. Burke: *J. Math. Phys.* **12**, 401 (1971).
7. W.L. Burke and K.S. Thorne: in *Relativity*, M. Carmeli et al. (eds.), Plenum Press: New York (1970) p. 208.
8. C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation*, Freeman: San Francisco (1973).
9. J. Ehlers: *Ann. N.Y. Acad. Sci.* **336**, 279 (1980).
10. A. Papapetrou and B. Linet: *Gen. Relat. Grav.* **13**, 335 (1981).
11. T. Damour and N. Deruelle: *Phys. Lett.* **87A**, 81 (1981).
12. T. Damour, in *Gravitational Radiation*, N. Deruelle and T. Piran (eds.), North-Holland: Amsderdam (1983), p. 59.

13. G. Schäfer: *Ann. Phys.* (N.Y.) **161**, 81 (1985).
14. S.M. Kopejkin (Kopeikin): *Astron. Zh.* **62**, 889 (1985), (In Russian).
15. W.B. Bonnor: *Philos. Trans. R. Soc. London* **A 251**, 233 (1959).
16. K.S. Thorne: *Rev. Mod. Phys.* **52**, 299 (1980).
17. L. Blanchet and T. Damour: *Philos. Trans. R. Soc. London* **A 320**, 379 (1986).
18. L. Blanchet and T. Damour: *Ann. Inst. H. Poincaré* (Phys. Théorique) **50**, 377 (1989).
19. T. Damour and B.R. Iyer: *Ann. Inst. H. Poincaré* (Phys. Théorique) **54**, 115 (1991).
20. L. Blanchet and T. Damour: *Phys. Rev.* **D 46**, 4304 (1992).
21. L. Blanchet: *Phys. Rev.* **D 51**, 2559 (1995).
22. C.M. Will and A.G. Wiseman: *Phys. Rev.* **D 54**, 4813 (1996).
23. P.C. Peters and J. Mathews: *Phys. Rev.* **131**, 435 (1963).
24. L.W. Esposito and E.R. Harrison: *Astrophys. J.* **196**, L1 (1975).
25. R.V. Wagoner: *Astrophys. J.* **196**, L63 (1975).
26. T. Damour: *Phys. Rev. Lett.* **51**, 1019 (1983).
27. J.H. Taylor, L.A. Fowler, and P.M. Mc Culloch: *Nature* **277**, 437 (1979).
28. J.H. Taylor: *Class. Quantum Grav.* **10**, S167 (1993).
29. V.A. Brumberg, *Relativistic Celestial Mechanics*, Nauka: Moscow (1972). In Russian.
30. C.H. Brans: *Astrophys. J.* **197**, 1 (1975).
31. B. Mashhoon and L.P. Grishchuk: *Astrophys. J.* **236**, 990 (1980).
32. A.G. Polnarev: *Sov. Astronomy* **29**, 607 (1985).
33. R.W. Hellings: *Astron. J.* **91**, 650; Erratum **92**, 1446 (1986).
34. V.A. Brumberg, *Essential Relativistic Celestial Mechanics*, Adam Hilger: Bristol (1991).
35. S.A. Klioner: *Astron. Zh.* **68**, 1046 (1991). In Russian.
36. S.A. Klioner and S.M. Kopeikin: *Astron. J.* **104**, 897 (1992).
37. S.M. Kopeikin, G. Schäfer, C.R. Gwinn, and T.M. Eubanks: *Phys. Rev.* **D 59**, 084023 (1999).
38. S.M. Kopeikin and G. Schäfer: *Phys. Rev.* **D 60**, 124002 (1999).
39. S.M. Kopeikin: *J. Math. Phys.* **38**, 2587 (1997).
40. T. Damour and G. Esposito–Farèse: *Phys. Rev.* **D 58**, 042001 (1998).
41. S.M. Kopeikin, *Timing Effects of Gravitational Waves from Localized Sources*, a talk given at the XXXIVth Rencontres de Moriond on "Gravitational Waves and Experimental Gravity", Les Arcs, 23-30 January 1999; e-print gr-qc/9903070.
42. V.A. Fock: *Theory of Space, Time and Gravitation* (Pergamon, London 1959).
43. T. Damour and B.R. Iyer: *Phys. Rev.* **D 43**, 3259 (1991).
44. N. Andersson: *Astrophys. J.* **502**, 708 (1998).
45. J.F. Friedman and S.M. Morsink: *Astrophys. J.* **502**, 714 (1998).
46. L.D. Landau and E.M. Lifshitz, *The Classical Theory of Fields* (Pergamon, New York 1975).
47. R. Breuer and E. Rudolph: *Gen. Relat. Grav.* **13**, 777 (1981).
48. G. Schäfer: *Lett. Nuovo Cimento* **36**, 105 (1983).
49. L. Blanchet, T. Damour, and G. Schäfer: *Mon. Not. R. Astr. Soc.* **242**, 289 (1990).
50. L. Rezzolla, M. Shibata, H. Asada, T.W. Baumgarte, and S.L. Shapiro: *Astrophys. J.* **525**, 935 (1999).
51. L. Blanchet: *Class. Quantum Grav.* **15**, 1971 (1998).
52. J. Hadamard: *Le problème de Cauchy et les équations aux dérivées partielles hyperboliques* (Hermann, Paris 1932).

53. L. Blanchet and G. Schäfer: *Mon. Not. R. Astr. Soc.* **239**, 845 (1989); Erratum **242**, 704 (1990).
54. T. Damour, M. Soffel, and C. Xu: *Phys. Rev.* **D 43**, 3273 (1991).
55. S.M. Kopeikin: *Cel. Mechanics* **44**, 87 (1988).
56. V.A. Brumberg and S.M. Kopeikin: *Nuovo Cimento* **B 103**, 63 (1989).
57. L. Blanchet: *Phys. Rev.* **D 55**, 714 (1997).
58. B.R. Iyer and C.M. Will: *Phys. Rev. Lett.* **70**, 113 (1993).
59. B.R. Iyer and C.M. Will: *Phys. Rev.* **D 52**, 6882 (1995).
60. P. Jaranowski and G. Schäfer: *Phys. Rev.* **D 55**, 4712 (1997).
61. P.R. Saulson: *Fundamentals of Interferometric Gravitational Wave Detection* (World Scientific, Singapore 1994).
62. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rüdiger: *Phys. Rev.* **D 38**, 423 (1988).

Relic Gravitational Waves and Their Detection

Leonid P. Grishchuk

Department of Physics and Astronomy,
Cardiff University, Cardiff CF2 3YB, United Kingdom
and
Sternberg Astronomical Institute, Moscow University, Moscow 119899, Russia

Abstract. The range of expected amplitudes and spectral slopes of relic (squeezed) gravitational waves, predicted by theory and partially supported by observations, is within the reach of sensitive gravity–wave detectors. In the most favorable case, the detection of relic gravitational waves can be achieved by the cross–correlation of outputs of the initial laser interferometers in LIGO, VIRGO, GEO600. In the more realistic case, the sensitivity of advanced ground–based and space–based laser interferometers will be needed. The specific statistical signature of relic gravitational waves, associated with the phenomenon of squeezing, is a potential reserve for further improvement of the signal to noise ratio.

1 Introduction

It is appropriate and timely to discuss the detection of relic gravitatational waves at the experimental meeting like this one. We are in the situation when the advanced laser interferometers, currently under construction or in a design phase, can make the dream of detecting relic gravitons a reality. The detection of relic gravitational waves is the only way to learn about the evolution of the very early Universe, up to the limits of Planck era and Big Bang.

The existence of relic gravitational waves is a consequence of quite general assumptions. Essentially, we rely only on the validity of general relativity and of basic principles of quantum field theory. The strong variable gravitational field of the early Universe amplifies the inevitable zero–point quantum oscillations of the gravitational waves and produces a stochastic background of relic gravitational waves measurable today [1]. It is important to appreciate the fundamental and unavoidable nature of this mechanism. Other physical processes can also generate stochastic backgrounds of gravitational waves. But those processes either involve many additional hypotheses, which may turn out to be not true, or produce a gravitational wave background (like the one from binary stars in the Galaxy) which should be treated as an unwanted noise rather than a useful and interesting signal. The scientific importance of detecting relic gravitational waves has been stressed on several occasions (see, for example, [2]–[4]).

The central notion in the theory of relic gravitons is the phenomenon of superadiabatic (parametric) amplification. The roots of this phenomenon are known in classical physics, and we will remind ourselves of its basic features. As every wave–like process, gravitational waves use the concept of a harmonic

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 167–192, 2001.

oscillator. The fundamental equation for a free harmonic oscillator is

$$\ddot{q} + \omega^2 q = 0, \tag{1}$$

where q can be a displacement of a mechanical pendulum or a time–dependent amplitude of a mode of the physical field. The energy of the oscillator can be changed by an acting force or, alternatively, by a parametric influence, that is, when a parameter of the oscillator, for instance the length of a pendulum, is being changed. In the first case, the fundamental equation takes the form

$$\ddot{q} + \omega^2 q = f(t), \tag{2}$$

whereas in the second case we have

$$\ddot{q} + \omega^2(t) q = 0. \tag{3}$$

Eqs.(2) and (3) are profoundly different, both, mathematically and physically.

Let us concentrate on the parametric influence. We consider a pendulum of length L oscillating in a constant gravitational field g. The unperturbed pendulum oscillates with the constant frequency $\omega = \sqrt{g/L}$. Fig.1a illustrates the variation of the length of the pendulum $L(t)$ by an external agent, shown by alternating arrows. Since $L(t)$ varies, the frequency of the oscillator does also vary: $\omega(t) = \sqrt{g/L(t)}$. The variation $L(t)$ does not need to be periodic, but cannot be too slow (adiabatic) if the result of the process is going to be significant. Otherwise, in the adiabatic regime of slow variations, the energy of the oscillator E and its frequency ω do change slowly, but E/ω remains constant, so one can say that the "number of quanta" $E/\hbar\omega$ in the oscillator remains fixed. In other words, for the creation of new "particles – excitations", the characteristic time of the variation should be comparable with the period of the oscillator and the adiabatic behaviour should be violated. After some duration of the appropriate parametric influence, the pendulum will oscillate at the original frequency, but will have a significantly larger amplitude and energy than before. This is shown in Fig.1b. Obviously, the energy of the oscillator has been increased at the expense of the external agent (pump field). For simplicity, we have considered a familiar case, when the length of the pendulum varies, while the gravitational acceleration g remains constant. Variation of g would represent a gravitational parametric influence and would even be in closer analogy to what we will study below.

A classical oscillator must have a non–zero initial amplitude for the amplification mechanism to work. Otherwise, if the initial amplitude is zero, the final amplitude will also be zero. Indeed, imagine the pendulum strictly at rest, hanging stright down. Whatever the variation of its length is, it will not make the pendulum to oscillate and gain energy. In contrast, a quantum oscillator does not need to be excited from the very beginning. The oscillator can be initially in its quantum–mechanical vacuum state. The inevitable zero–point quantum oscillations are associated with the vacuum state energy $\frac{1}{2}\hbar\omega$. One can imagine a pendulum hanging straight down, but fluctuating with a tiny amplitude determined by the "half of the quantum in the mode". In the classical picture, it is

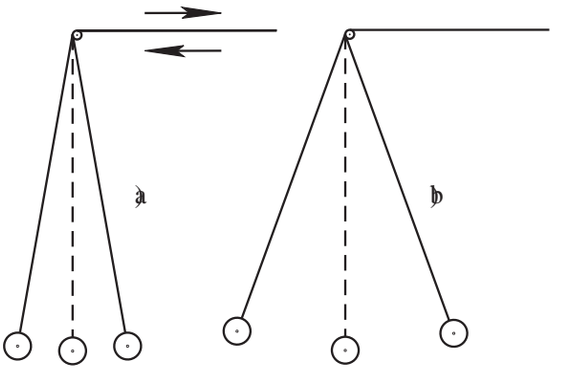


Fig. 1. Parametric amplification. a) variation of the length of the pendulum, b) increased amplitude of oscillations.

this tiny amplitude of quantum–mechanical origin that is being parametrically amplified.

The Schrödinger evolution of a quantum oscillator depends crucially on whether the oscillator is being excited parametrically or by a force. Consider the phase diagram (q, p), where q is the displacement and p is the conjugate momentum. The vacuum state is described by the circle in the center (see Fig.2). The mean values of q and p are zeros, but their variances (zero–point quantum fluctuations) are not zeros and are equal to each other. Their numerical values are represented by the circle in the center. Under the action of a force, the vacuum state evolves into a coherent state. The mean values of p and q increased, but the variances are still equal and are described by the circle of the same size as for the vacuum state. On the other hand, under a parametric influence, the vacuum state evolves into a squeezed vacuum state. [For a recent review of squeezed states see [5], for example, and references therein.] Its variances for the conjugate variables q and p are significantly unequal and are described by an ellipse. As a function of time, the ellipse rotates with respect to the origin of the (q, p) diagram, and the numerical values of the variances oscillate, too. The mean numbers of quanta in the two states, one of which is coherent and another is squeezed vacuum, can be equal (similar to the coherent and squeezed states shown in Fig. 2) but the statistical properties of these states are significantly different. Among other things, the variance of the phase of the oscillator in a squeezed vacuum state is very small (squeezed). Graphically, this is reflected in the fact that the ellipse is very thin, so that that the uncertainty in the angle between the horizontal axis and the orientation of the ellipse is very small. This highly elongated ellipse can be regarded as a portrait of the gravitational wave quantum state that is being inevitably generated by parametric amplification, and which we will be dealing with below.

A wave–field is not a single oscillator, it depends on spatial coordinates and on time and may have several independent components (polarization states).

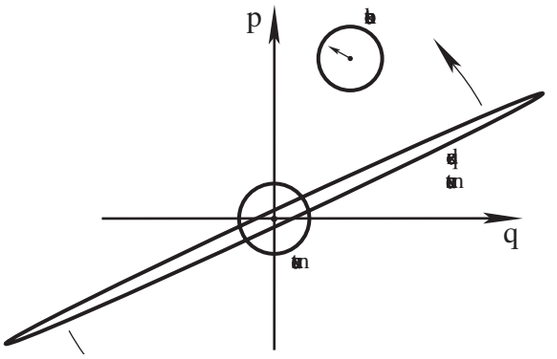


Fig. 2. Some quantum states of a harmonic oscillator.

However, the field can be decomposed into a set of spatial Fourier harmonics. In this way we represent the gravitational wave field as a collection of many modes, many oscillators. Because of the nonlinear character of the Einstein equations, each of these oscillators is coupled to the variable gravitational field of the surrounding Universe. For sufficiently short gravitational waves of experimental interest, this coupling was especially effective in the early Universe, when the condition of the adiabatic behaviour of the oscillator was violated. It is this homogeneous and isotropic gravitational field of all the matter in the early Universe that played the role of the external agent – pump field. The variable pump field acts parametrically on the gravity–wave oscillators and drives them into multiparticle states. Concretely, the initial vacuum state of each pair of waves with oppositely directed momenta evolves into a highly correlated state known as the two–mode squeezed vacuum state [6,7]. The strength and duration of the effective coupling depends on the oscillator's frequency. They all start in the vacuum state but get excited to various amounts. As a result, a broad spectrum of relic gravitational waves is being formed. This spectrum is accessible to our observations today.

Let us formulate the problem in more detail.

2 Cosmological Gravitational Waves

In the framework of general relativity, a homogeneous isotropic gravitational field is decribed by the line element

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a^2(t)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j = a^2(\eta)[\mathrm{d}\eta^2 - \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j]. \tag{4}$$

In cosmology, the function $a(t)$ (or $a(\eta)$) is called scale factor. In our discussion, it will represent the gravitational pump field.

Cosmological gravitational waves are small corrections h_{ij} to the metric tensor. They are defined by the expression

$$\mathrm{d}s^2 = a^2(\eta)[\mathrm{d}\eta^2 - (\delta_{ij} + h_{ij})\mathrm{d}x^i \mathrm{d}x^j]. \tag{5}$$

The functions $h_{ij}(\eta, \mathbf{x})$ can be expanded over spatial Fourier harmonics $e^{i\mathbf{nx}}$ and $e^{-i\mathbf{nx}}$, where $\mathbf{n}$ is a constant wave vector. In this way, we reduce the dynamical problem to the evolution of time–dependent amplitudes for each mode $\mathbf{n}$. Among six functions h_{ij} there are only two independent (polarization) components. This decomposition can be made, both, for real and for quantized field h_{ij}. In the quantum version, the functions h_{ij} are treated as quantum–mechanical operators. We will use the Heisenberg picture in which the time evolution is carried out by the operators while the quantum state is fixed. This picture is fully equivalent to the Schrödinger picture, discussed in the Introduction, in which the vacuum state evolves into a squeezed vacuum state while the operators are time independent.

The Heisenberg operator for the quantized real field h_{ij} can be written as

$$h_{ij}(\eta, \mathbf{x}) = \frac{C}{(2\pi)^{3/2}} \int_{-\infty}^{\infty} d^3\mathbf{n} \sum_{s=1}^{2} \overset{s}{p}_{ij}(\mathbf{n}) \frac{1}{\sqrt{2n}} \left[\overset{s}{h}_n(\eta) e^{i\mathbf{nx}} \overset{s}{c}_{\mathbf{n}} + \overset{s*}{h}_n(\eta) e^{-i\mathbf{nx}} \overset{s\dagger}{c}_{\mathbf{n}} \right], \tag{6}$$

where C is a constant which will be discussed later. The creation and annihilation operators satisfy the conditions $[\overset{s'}{c}_{\mathbf{n}}, \overset{s\dagger}{c}_{\mathbf{m}}] = \delta_{s's}\delta^3(\mathbf{n} - \mathbf{m})$, $\overset{s}{c}_{\mathbf{n}}|0\rangle = 0$, where $|0\rangle$ (for each $\mathbf{n}$ and s) is the fixed initial vacuum state discussed below. The wave number n is related with the wave vector $\mathbf{n}$ by $n = (\delta_{ij} n^i n^j)^{1/2}$. The two polarization tensors $\overset{s}{p}_{ij}(\mathbf{n})$ $(s = 1, 2)$ obey the conditions

$$\overset{s}{p}_{ij} n^j = 0, \quad \overset{s}{p}_{ij} \delta^{ij} = 0, \quad \overset{s'}{p}_{ij} \overset{s}{p}^{ij} = 2\delta_{ss'}, \quad \overset{s}{p}_{ij}(-\mathbf{n}) = \overset{s}{p}_{ij}(\mathbf{n}).$$

The time evolution, one and the same for all $\mathbf{n}$ belonging to a given n, is represented by the complex time–dependent function $\overset{s}{h}_n(\eta)$. This evolution is dictated by the Einstein equations. The nonlinear nature of the Einstein equations leads to the coupling of $\overset{s}{h}_n(\eta)$ with the pump field $a(\eta)$. For every wave number n and each polarization component s, the functions $\overset{s}{h}_n(\eta)$ have the form

$$\overset{s}{h}_n(\eta) = \frac{1}{a(\eta)} [\overset{s}{u}_n(\eta) + \overset{s*}{v}_n(\eta)], \tag{7}$$

where $\overset{s}{u}_n(\eta)$ and $\overset{s}{v}_n(\eta)$ can be expressed in terms of the three real functions (the polarization index s is omitted): r_n - squeeze parameter, ϕ_n - squeeze angle, θ_n - rotation angle,

$$u_n = e^{i\theta_n} \cosh\ r_n, \qquad v_n = e^{-i(\theta_n - 2\phi_n)} \sinh\ r_n. \tag{8}$$

The dynamical equations for $u_n(\eta)$ and $v_n(\eta)$

$$i\frac{\mathrm{d}u_n}{\mathrm{d}\eta} = nu_n + i\frac{a'}{a}v_n^*, \qquad i\frac{\mathrm{d}v_n}{\mathrm{d}\eta} = nv_n + i\frac{a'}{a}u_n^* \tag{9}$$

lead to the dynamical equations governing the functions $r_n(\eta)$, $\phi_n(\eta)$, $\theta_n(\eta)$ [7]:

$$r_n' = \frac{a'}{a}\cos 2\phi_n, \quad \phi_n' = -n - \frac{a'}{a}\sin 2\phi_n \coth\ 2r_n, \quad \theta_n' = -n - \frac{a'}{a}\sin 2\phi_n \tanh\ r_n, \tag{10}$$

where $' = \mathrm{d}/\mathrm{d}\eta$, and the evolution begins from $r_n = 0$. This value of r_n characterizes the initial vacuum state $|0\rangle$ which is defined long before the interaction with the pump field became effective, that is, long before the coupling term a'/a became comparable with n. The constant C should be taken as $C = \sqrt{16\pi}\ l_{\mathrm{Pl}}$ where $l_{\mathrm{Pl}} = (G\hbar/c^3)^{1/2}$ is the Planck length. This particular value of the constant C guarantees the correct quantum normalization of the field: energy $\frac{1}{2}\hbar\omega$ per each mode in the initial vacuum state. The dynamical equations and their solutions are identical for both polarization components s.

Equations (9) can be translated into the more familiar form of the second–order differential equation for the function $\overset{s}{\mu}_n(\eta) \equiv \overset{s}{u}_n(\eta) + \overset{s}{v}{}^*_n(\eta) \equiv a(\eta)\overset{s}{h}_n(\eta)$ [1]:

$$\mu_n'' + \mu_n\left[n^2 - \frac{a''}{a}\right] = 0. \tag{11}$$

Clearly, this is the equation for a parametrically disturbed oscillator (compare with Eq. (3)). In absence of the gravitational parametric influence represented by the term a''/a, the frequency of the oscillator defined in terms of η-time would be a constant: n. Whenever the term a''/a can be neglected, the general solution to Eq. (11) has the usual oscillatory form

$$\mu_n(\eta) = A_n e^{-in\eta} + B_n e^{in\eta}, \tag{12}$$

where the constants A_n, B_n are determined by the initial conditions. On the other hand, whenever the term a''/a is dominant, the general solution to Eq. (11) has the form

$$\mu_n(\eta) = C_n a + D_n a \int^{\eta} \frac{\mathrm{d}\eta}{a^2}. \tag{13}$$

In fact, this approximate solution is valid as long as n is small in comparison with $|a'/a|$. This is more clearly seen from the equivalent form of Eq. (11) written in terms of the function $h_n(\eta)$ [8]:

$$h_n'' + 2\frac{a'}{a}h_n' + n^2 h_n = 0. \tag{14}$$

For growing functions $a(\eta)$, that is, in expanding universes, the second term in Eq.(13) is usually smaller than the first one (see below), so that, as long as $n \ll a'/a$, the dominant solution is the growing function $\mu_n(\eta) = C_n a(\eta)$, and

$$h_n = const. \tag{15}$$

Equation (11) can be also looked at as a kind of the Schrödinger equation for a particle moving in presence of the effective potential $U(\eta) = a''/a$. In the situations that are normally considered, the potential $U(\eta)$ has a bell–like shape and forms a barrier (see Fig. 3). When a given mode n is outside the barrier, its amplitude h_n is adiabatically decreasing with time: $h_n \propto \frac{e^{\pm in\eta}}{a(\eta)}$. This is shown in Fig.3 by oscillating lines with decreasing amplitudes of oscillations. The modes with sufficiently high frequencies do not interact with the barrier, they stay above the barrier. Their amplitudes h_n behave adiabatically all the time. For these high–frequency modes, the initial vacuum state (in the Schrödinger picture) remains the vacuum forever. On the other hand, the modes that interact with the barrier are subject to the superadiabatic amplification. Under the barrier and as long as $n < a'/a$, the function h_n stays constant instead of the adiabatic decrease. For these modes, the initial vacuum state evolves into a squeezed vacuum state.

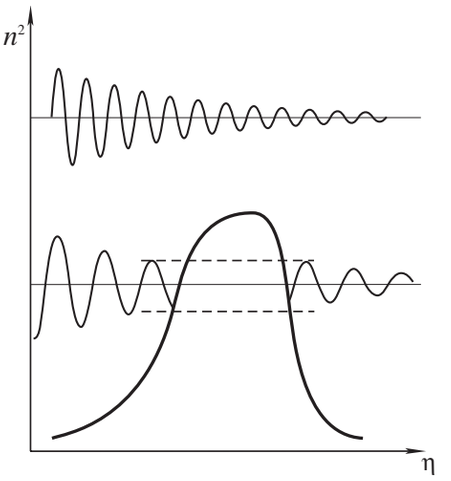


Fig. 3. Effective potential $U(\eta)$.

After having formulated the initial conditions, the present day behaviour of r_n, ϕ_n, θ_n (or, equivalently, the present day behaviour of h_n) is essentially all we need to find. The mean number of particles in a two–mode squeezed state is $2\sinh^2 r_n$ for each s. This number determines the mean square amplitude of the gravitational wave field. The time behaviour of the squeeze angle ϕ_n determines the time dependence of the correlation functions of the field. The amplification (that is, the growth of r_n) governed by Eq. (10) is different for different wave numbers n. Therefore, the present day results depend on the present day frequency ν ($\nu = cn/2\pi a$) measured in Hz.

In cosmology, the function $H \equiv \dot{a}/a \equiv ca'/a^2$ is the time–dependent Hubble parameter. The function $l \equiv c/H$ is the time–dependent Hubble radius. The time–dependent wavelength of the mode n is $\lambda = 2\pi a/n$. The wavelength λ has this universal definition in all regimes. In contrast, the ν defined as $\nu = cn/2\pi a$ has the usual meaning of a frequency of an oscillating process only in the short–wavelength (high–frequency) regime of the mode n, that is, in the regime where $\lambda \ll l$. As we have seen above, the qualitative behaviour of solutions to Eqs. (11), (14) depends crucially on the comparative values of n and a'/a, or, in other words, on the comparative values of $\lambda(\eta)$ and $l(\eta)$. This relationship is also crucial for solutions to Eq. (10) as we shall see now.

In the short–wavelength regime, that is, during intervals of time when the wavelength $\lambda(\eta)$ is shorter than the Hubble radius $l(\eta) = a^2/a'$, the term n in

(10) is dominant. The functions $\phi_n(\eta)$ and $\theta_n(\eta)$ are $\phi_n = -n(\eta + \eta_n)$, $\theta_n = \phi_n$ where η_n is a constant. The factor $\cos 2\phi_n$ is a quickly oscillating function of time, so the squeeze parameter r_n stays practically constant. This is the adiabatic regime for a given mode.

In the opposite, long–wavelength regime, the term n can be neglected. The function ϕ_n is $\tan \phi_n(\eta) \approx const/a^2(\eta)$, and the squeeze angle quickly approaches one of the two values: $\phi_n = 0$ or $\phi_n = \pi$ (analog of "phase bifurcation" [9]). The squeeze parameter $r_n(\eta)$ grows with time according to

$$r_n(\eta) \approx ln\frac{a(\eta)}{a_*} \ , \tag{16}$$

where a_* is the value of $a(\eta)$ at η_*, when the long–wavelength regime, for a given n, begins. The final amount of r_n is

$$r_n \approx ln\frac{a_{**}}{a_*} \ , \tag{17}$$

where a_{**} is the value of $a(\eta)$ at η_{**}, when the long–wavelength regime and amplification come to the end. It is important to emphasize that it is not a "sudden transition" from one cosmological era to another that is responsible for amplification, but the entire interval of the long–wavelength (non–adiabatic) regime.

After the end of amplification, the accumulated (and typically large) squeeze parameter r_n stays approximately constant. The mode is again in the adiabatic regime. In course of the evolution, the complex functions $\overset{s}{u}_n(\eta) + \overset{s^*}{v}_n(\eta)$ become practically real, and one has $\overset{s}{h}_n(\eta) \approx \overset{s^*}{h}_n(\eta) \approx \frac{1}{a}e^{r_n}\cos\phi_n(\eta)$. Every amplified mode n of the field (6) takes the form of a product of a function of time and a (random, operator–valued) function of spatial coordinates; the mode acquires a standing–wave pattern. The periodic dependence $\cos\phi_n(\eta)$ will be further discussed below.

It is clearly seen from the fundamental equations (10), (11), (14) that the final results depend only on $a(\eta)$. Equations do not ask us the names of our favorite cosmological prejudices, they ask us about the pump field $a(\eta)$. Conversely, from the measured relic gravitational waves, we can deduce the behaviour of $a(\eta)$, which is essentially the purpose of detecting the relic gravitons.

3 Cosmological Pump Field

With the chosen initial conditions, the final numerical results for relic gravitational waves depend on the concrete behaviour of the pump field represented by the cosmological scale factor $a(\eta)$. We know a great deal about $a(\eta)$. We know that $a(\eta)$ behaves as $a(\eta) \propto \eta^2$ at the present matter–dominated stage. We know that this stage was preceeded by the radiation–dominated stage $a(\eta) \propto \eta$. At these two stages of evolution the functions $a(\eta)$ are simple power–law functions of η. What we do not know is the function $a(\eta)$ describing the initial stage of

expansion of the very early Universe, that is, before the era of primordial nucleosynthesis. It is convenient to parameterize $a(\eta)$ at this initial stage also by power–law functions of η. First, this is a sufficiently broad class of functions, which, in addition, allows us to find exact solutions to our fundamental equations. Second, it is known [1] that the pump fields $a(\eta)$ which have power–law dependence in terms of η, produce gravitational waves with simple power–law spectra in terms of ν. These spectra are easy to analyze and discuss in the context of detection.

We model cosmological expansion by several successive eras. Concretely, we take $a(\eta)$ at the initial stage of expansion (i–stage) as

$$a(\eta) = l_o|\eta|^{1+\beta}, \tag{18}$$

where η grows from $-\infty$, and $1 + \beta < 0$. We will show later how the available observational data constrain the parameters l_o and β. The i–stage lasts up to a certain $\eta = \eta_1$, $\eta_1 < 0$. To make our analysis more general, we assume that the i–stage was followed by some interval of the z–stage (z from Zeldovich). It is known that an interval of evolution governed by the most "stiff" matter (effective equation of state $p = \epsilon$) advocated by Zeldovich, leads to a relative increase of gravitational wave amplitudes [1]. It is also known that the requirement of conistency of the graviton production with the observational restrictions does not allow the "stiff" matter interval to be too much long [1], [10]. However, we want to investigate any interval of cosmological evolution that can be consistently included. In fact, the z–stage of expansion that we include is quite general. It can be governed by a "stiffer than radiation" [11] matter, as well as by a "softer than radiation" matter. It can also be simply a part of the radiation–dominated era. Concretely, we take $a(\eta)$ at the interval of time from η_1 to some η_s (z–stage) in the form

$$a(\eta) = l_o a_z(\eta - \eta_p)^{1+\beta_s}, \tag{19}$$

where $1 + \beta_s > 0$. For the particular choice $\beta_s = 0$, the z–stage reduces to an interval of expansion governed by the radiation–dominated matter. Starting from η_s and up to η_2 the Universe was governed by the radiation–dominated matter (e–stage). So, at this interval of evolution, we take the scale factor in the form

$$a(\eta) = l_o a_e(\eta - \eta_e). \tag{20}$$

And, finally, from $\eta = \eta_2$ the expansion went over into the matter–dominated era (m–stage):

$$a(\eta) = l_o a_m(\eta - \eta_m)^2. \tag{21}$$

A link between the arbitrary constants participating in Eqs. (18) - (21) is provided by the conditions of continuous joining of the functions $a(\eta)$ and $a'(\eta)$ at points of transitions η_1, η_s, η_2.

We denote the present time by η_R (R from reception). This time is defined by the observationally known value of the present–day Hubble parameter $H(\eta_R)$ and Hubble radius $l_H = c/H(\eta_R)$. For numerical estimates we will be using

$l_H \approx 2 \times 10^{28}$ cm. It is convenient to choose $\eta_R - \eta_m = 1$, so that $a(\eta_R) = 2l_H$. The ratio

$$a(\eta_R)/a(\eta_2) \equiv \zeta_2$$

is believed to be around $\zeta_2 = 10^4$. We also denote

$$a(\eta_2)/a(\eta_s) \equiv \zeta_s \ , \quad a(\eta_s)/a(\eta_1) \equiv \zeta_1 \ .$$

With these definitions, all the constants participating in Eqs. (18) - (21) (except parameters β and β_s which should be chosen from other considerations) are being expressed in terms of l_H, ζ_2, ζ_s, and ζ_1. For example,

$$|\eta_1| = \frac{|1+\beta|}{2\zeta_2^{\frac{1}{2}}\zeta_s\zeta_1^{\frac{1}{1+\beta_s}}} \ .$$

The important constant l_o is expressed as

$$l_o = bl_H\zeta_2^{\frac{\beta-1}{2}}\zeta_s^{\beta}\zeta_1^{\frac{\beta-\beta_s}{1+\beta_s}}, \tag{22}$$

where $b \equiv 2^{2+\beta}/|1+\beta|^{1+\beta}$. Note that $b = 1$ for $\beta = -2$. [This expression for l_o may help to relate formulas written here with the equivalent treatment [12] which was given in slightly different notations.] The sketch of the entire evolution $a(\eta)$ is given in Fig. 4.

We work with the spatially–flat models (4). At every instant of time, the energy density $\epsilon(\eta)$ of matter driving the evolution is related with the Hubble radius $l(\eta)$ by

$$\kappa\epsilon(\eta) = \frac{3}{l^2(\eta)}, \tag{23}$$

where $\kappa = 8\pi G/c^4$. For the case of power–law scale factors $a(\eta) \propto \eta^{1+\beta}$, the effective pressure $p(\eta)$ of the matter is related with the $\epsilon(\eta)$ by the effective equation of state

$$p = \frac{1-\beta}{3(1+\beta)}\epsilon. \tag{24}$$

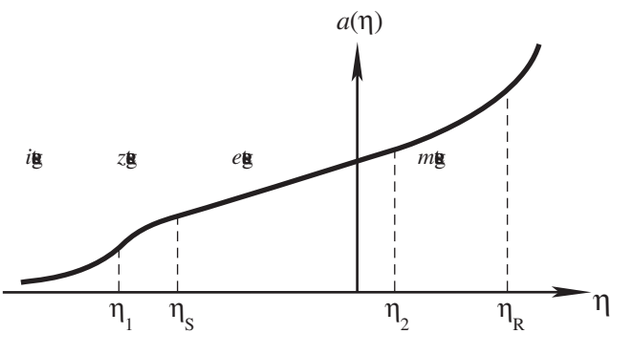


Fig. 4. Scale factor $\alpha(\eta)$.

For instance, $p = 0$ for $\beta = 1$, $p = \frac{1}{3}\epsilon$ for $\beta = 0$, $p = -\epsilon$ for $\beta = -2$, and so on. Each interval of the evolution (18)-(21) is governed by one of these equations of state.

In principle, the function $a(\eta)$ could be even more complicated than the one that we consider. It could even include an interval of the early contraction, instead of expansion, leading to the "bounce" of the scale factor. In case of a decreasing $a(\eta)$ the gravitational–wave equation can still be analyzed and the amplification is still effective [1]. However, the Einstein equations for spatially–flat models do not permit a regular "bounce" of $a(\eta)$ (unless ϵ vanishes at the moment of "bounce"). Possibly, a "bounce" solution can be realized in alternative theories, such, for example, as string–motivated cosmologies [13]. For a recent discussion of spectral slopes of gravitational waves produced in "bounce" cosmologies, see [14].

4 Solving Gravitational Wave Equations

The evolution of the scale factor $a(\eta)$ given by Eqs. (18) - (21) and sketched in Fig. 4 allows us to calculate the function a'/a. This function is sketched in Fig. 5. In all the theoretical generality, the left–hand–side of the barrier in $Fig.5$ could also consist of several pieces, but we do not consider this possibility here. The graph also shows the important wave numbers n_H, n_2, n_s, n_1. The n_H marks the wave whose today's wavelength $\lambda(\eta_R) = 2\pi a(\eta_R)/n_H$ is equal to the today's Hubble radius l_H. With our parametrization $a(\eta_R) = 2l_H$, this wavenumber is $n_H = 4\pi$. The n_2 marks the wave whose wavelength $\lambda(\eta_2) = 2\pi a(\eta_2)/n_2$ at $\eta = \eta_2$ is equal to the Hubble radius $l(\eta_2)$ at $\eta = \eta_2$. Since $\lambda(\eta_R)/\lambda(\eta_2) = (n_2/n_H)[a(\eta_R)/a(\eta_2)]$ and $l(\eta_R)/l(\eta_2) = [a(\eta_R)/a(\eta_2)][a(\eta_R)/a(\eta_2)]^{1/2}$, this gives us $n_2/n_H = [a(\eta_R)/a(\eta_2)]^{1/2} = \zeta_2^{1/2}$. Working out in a similar fashion other ratios, we find

$$\frac{n_2}{n_H} = \zeta_2^{\frac{1}{2}}, \quad \frac{n_s}{n_2} = \zeta_s, \quad \frac{n_1}{n_s} = \zeta_1^{\ \frac{1}{1+\beta_s}}. \tag{25}$$

Solutions to the gravitational wave equations exist for any $a(\eta)$. At intervals of power–law dependence $a(\eta)$, solutions to Eq. (11) have simple form of the Bessel functions. We could have found piece–wise exact solutions to Eq. (11) and join them in the transition points. However, we will use a much simpler treatment which is sufficient for our purposes. We know that the squeeze parameter r_n stays constant in the short–wavelength regimes and grows according to Eq. (16) in the long–wavelength regime. All modes start in the vacuum state, that is, $r_n = 0$ initially. After the end of amplification, the accumulated value (17) stays constant up to today. To find today's value of e^{r_n} we need to calculate the ratio $a_{**}(n)/a_*(n)$. For every given n, the quantity a_* is determined by the condition $\lambda(\eta_*) = l(\eta_*)$, wheras a_{**} is determined by the condition $\lambda(\eta_{**}) = l(\eta_{**})$.

Let us start from the mode $n = n_1$. For this wave number we have $a_* = a_{**} = a(\eta_1)$, and therefore $r_{n_1} = 0$. The higher frequency modes $n > n_1$ (above

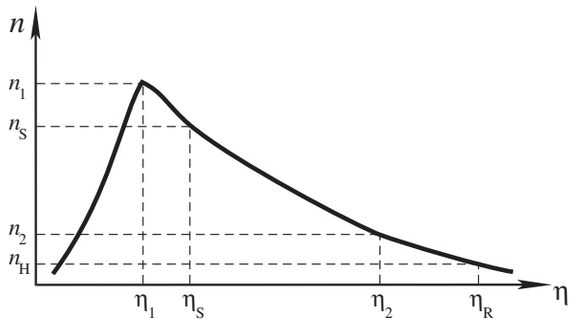


Fig. 5. The function α'/α for the scale factor from Fig. 4.

the barrier in Fig. 5) have never been in the amplifying regime, so we can write

$$e^{r_n} = 1, \quad n \geq n_1. \tag{26}$$

Let us now consider the modes n in the interval $n_1 \geq n \geq n_s$. For a given n we need to know $a_*(n)$ and $a_{**}(n)$. Using Eq. (18) one finds $a_*(n)/a_*(n_1) = (n_1/n)^{1+\beta}$, and using Eq. (19) one finds $a_{**}(n)/a_{**}(n_s) = (n_s/n)^{1+\beta_s}$. Therefore, one finds

$$\frac{a_{**}(n)}{a_*(n)} = \frac{a_{**}(n_s)}{a_*(n_1)} \left(\frac{n_s}{n}\right)^{1+\beta_s} \left(\frac{n}{n_1}\right)^{1+\beta}.$$

Since $a_{**}(n_s) = a(\eta_s)$, $a_*(n_1) = a(\eta_1)$, and $a(\eta_s)/a(\eta_1) = \zeta_1 = (n_1/n_s)^{1+\beta_s}$, we arrive at

$$\frac{a_{**}(n)}{a_*(n)} = \left(\frac{n}{n_1}\right)^{\beta-\beta_s}.$$

Repeating this analysis for other intervals of the decreasing n, we come to the conclusion that

$$\begin{aligned}
e^{r_n} &= \left(\frac{n}{n_1}\right)^{\beta-\beta_s}, \quad n_1 \geq n \geq n_s, \\
e^{r_n} &= \left(\frac{n}{n_s}\right)^{\beta} \left(\frac{n_s}{n_1}\right)^{\beta-\beta_s}, \quad n_s \geq n \geq n_2, \\
e^{r_n} &= \left(\frac{n}{n_2}\right)^{\beta-1} \left(\frac{n_2}{n_1}\right)^{\beta} \left(\frac{n_s}{n_1}\right)^{-\beta_s}, n_2 \geq n \geq n_H.
\end{aligned} \tag{27}$$

The mnemonic rule of constructing e^{r_n} at successive intervals of decreasing n is simple. If the interval begins at n_x, one takes $(n/n_x)^{\beta_*-\beta_{**}}$ and multiples with $e^{r_{n_x}}$, that is, with the previous interval's value e^{r_n} calculated at the end of that interval n_x. For the function a'/a that we are working with, the β_* is always β, whereas the β_{**} takes the values β_s, 0, 1 at the successive intervals.

The modes with $n < n_H$ are still in the long–wavelength regime. For these modes, we should take $a(\eta_R)$ instead of $a_{**}(n)$. Combining with $a_*(n)$, we find

$$e^{r_n} = \left(\frac{n}{n_H}\right)^{\beta+1} \left(\frac{n_H}{n_2}\right)^{\beta-1} \left(\frac{n_2}{n_1}\right)^{\beta} \left(\frac{n_s}{n_1}\right)^{-\beta_s}, \quad n \le n_H. \tag{28}$$

Formulas (26) - (28) give approximate values of r_n for all n. The factor e^{r_n} is $e^{r_n} \ge 1$ for $n \le n_1$, and $e^{r_n} \gg 1$ for $n \ll n_1$. This factor determines the mean square amplitude of the gravitational waves.

The mean value of the field h_{ij} is zero at every moment of time η and in every spatial point $\mathbf{x}$: $\langle 0|h_{ij}(\eta, \mathbf{x})|0\rangle = 0$. The variance

$$\langle 0|h_{ij}(\eta, \mathbf{x})h^{ij}(\eta, \mathbf{x})|0\rangle \equiv \langle h^2\rangle$$

is not zero, and it determines the mean square amplitude of the generated field – the quantity of interest for the experiment. Taking the product of two expressions (6) one can show that

$$\langle h^2\rangle = \frac{C^2}{2\pi^2}\int_0^\infty n \sum_{s=1}^{2} \left|\overset{s}{h}_n(\eta)\right|^2 \mathrm{d}n \equiv \int_0^\infty h^2(n,\eta)\frac{\mathrm{d}n}{n}\ . \tag{29}$$

Using the representation (7), (8) in Eq. (29) one can also write

$$\langle h^2\rangle = \frac{C^2}{\pi^2 a^2}\int_0^\infty n\mathrm{d}n(\cosh 2r_n + \cos 2\phi_n \sinh 2r_n). \tag{30}$$

We can now consider the present era and use the fact that e^{r_n} are large numbers for all n in the interval of our interest $n_1 \ge n \ge n_H$. Then, we can derive

$$h(n,\eta) \approx \frac{C}{\pi}\frac{1}{a(\eta_R)} n e^{r_n}\cos\phi_n(\eta) = 8\sqrt{\pi}\left(\frac{l_{\mathrm{Pl}}}{l_H}\right)\left(\frac{n}{n_H}\right) e^{r_n}\cos\phi_n(\eta)\ . \tag{31}$$

The quantity $h(n,\eta)$ is the dimensionless spectral amplitude of the field whose numerical value is determined by the calculated squeeze parameter r_n. The oscillatory factor $\cos\phi_n(\eta)$ reflects the squeezing (standing wave pattern) acquired by modes with $n_1 > n > n_H$. For modes with $n < n_H$ this factor is approximately 1. For high–frequency modes $n \gg n_H$ one has $\phi_n(\eta) \approx n(\eta - \eta_n) \gg 1$, so that $h(n,\eta)$ makes many oscillations while the scale factor $a(\eta)$ is practically fixed at $a(\eta_R)$.

The integral (30) extends formally from 0 to ∞. Since $r_n \approx 0$ for $n \ge n_1$, the integral diverges at the upper limit. This is a typical ultra–violet divergence. It should be discarded (renormalized to zero) because it comes from the modes which have always been in their vacuum state. At the lower limit, the integral diverges, if $\beta \le -2$. This is an infra–red divergence which comes from the assumption that the amplification process has started from infinitely remote time in the past. One can deal with this divergence either by introducing a lower frequency cut–off (equivalent to the finite duration of the amplification) or by

considering only the parameters $\beta > -2$, in which case the integral is convergent at the lower limit. It appears that the available observational data (see below) favour this second option. The particular case $\beta = -2$ corresponds to the de Sitter evolution $a(\eta) \propto |\eta|^{-1}$. In this case, the $h(n)$ found in Eqs. (31), (28) does not depend on n. This is known as the Harrison–Zeldvich, or scale–invariant, spectrum.

An alternative derivation of the spectral amplitude $h(n)$ uses the approximate solutions (12), (13) to the wave equation (11). This method gives exactly the same, as in Eqs. (31), (26) - (28) numerical values of $h(n)$, but does not reproduce the oscillatory factor $\cos\phi_n(\eta)$.

One begins with the initial spectral amplitude $h_i(n)$ defined by quantum normalization: $h_i(n) = 8\sqrt{\pi}(l_{\rm Pl}/\lambda_i)$. This is the amplitude of the mode n at the moment η_* of entering the long wavelength regime, i.e. when the mode's wavelength λ_i is equal to the Hubble radius $l(\eta_*)$. For λ_i one derives

$$\lambda_i = \frac{1}{b} l_o \left(\frac{n_H}{n}\right)^{2+\beta}. \tag{32}$$

Thus, we have

$$h_i(n) = A\left(\frac{n}{n_H}\right)^{2+\beta}, \tag{33}$$

where A denotes the constant

$$A = b8\sqrt{\pi}\frac{l_P l}{l_o}\ . \tag{34}$$

The numbers $h_i(n)$ are defined at the beginning of the long–wavelength regime. In other words, they are given along the left–hand–side slope of the barrier in Fig. 5. We want to know the final numbers (spectral amplitudes) $h(n)$ which describe the field today, at η_R.

According to the dominant solution $h_n(\eta) = const$ of the long–wavelength regime (see Eq. (15)), the initial amplitude $h_i(n)$ stays practically constant up to the end of the long–wavelength regime at η_{**}, that is, up to the right–hand–side slope of the barrier. [The second term in Eq. (13) could be important only at the z–stage and only for parameters $\beta_s \leq -(1/2)$, which correspond to the effective equations of state $p \geq \epsilon$. In order to keep the analysis simple, we do not consider those cases.] After the completion of the long–wavelength regime, the amplitudes decrease adiabatically in proportion to $1/a(\eta)$, up to the present time. Thus, we have

$$h(n) = A\left(\frac{n}{n_H}\right)^{2+\beta}\frac{a_{**}(n)}{a(\eta_R)}. \tag{35}$$

Let us start from the lower end of the spectrum, $n \leq n_H$, and go upward in n. The modes $n \leq n_H$ have not started yet the adiabatic decrease of the amplitude, so we have

$$h(n) = A\left(\frac{n}{n_H}\right)^{2+\beta}, \quad n \leq n_H. \tag{36}$$

Now consider the interval $n_2 \geq n \geq n_H$. At this interval, the $a_{**}(n)/a(\eta_R)$ scales as $(n_H/n)^2$, so we have

$$h(n) = A\left(\frac{n}{n_H}\right)^{\beta}, \quad n_2 \geq n \geq n_H. \tag{37}$$

At the interval $n_s \geq n \geq n_2$ the ratio

$$\frac{a_{**}(n)}{a(\eta_R)} = \frac{a_{**}(n)}{a(\eta_2)}\ \frac{a(\eta_2)}{a(\eta_R)}$$

scales as $(n_2/n)(n_H/n_2)^2$, so we have

$$h(n) = A\left(\frac{n}{n_H}\right)^{1+\beta}\frac{n_H}{n_2}, \quad n_s \geq n \geq n_2. \tag{38}$$

Repeating the same analysis for the interval $n_1 \geq n \geq n_s$ we find

$$h(n) = A\left(\frac{n}{n_H}\right)^{1+\beta-\beta_s}\left(\frac{n_s}{n_H}\right)^{\beta_s}\frac{n_H}{n_2}, \quad n_1 \geq n \geq n_s. \tag{39}$$

It is seen from Eq. (39) that an interval of the z–stage with $\beta_s < 0$ (the already imposed restrictions require also $(-1/2) < \beta_s$) bends the spectrum $h(n)$ upwards, as compared with Eq. (38), for larger n. If one recalls the relationship (22) between l_o and l_H and uses (27), (28) in Eq. (31) one arrives exactly at Eqs.(36)-(39) up to the oscillating factor $\cos\phi_n(\eta)$.

Different parts of the barrier in Fig.5 are responsible for amplitudes and spectral slopes at different intervals of n. The sketch of the generated spectrum $h(n)$ in conjunction with the form of the barrier is shown in Fig.6.

The present day frequency of the oscillating modes, measured in Hz, is defined as $\nu = cn/2\pi a(\eta_R)$. The lowest frequency (Hubble frequency) is $\nu_H = c/l_H$. For numerical estimates we will be using $\nu_H \approx 10^{-18}$ Hz. The ratios of n are equal to the ratios of ν, so that $n/n_H = \nu/\nu_H$, for example. For high–frequency modes we will now often use the ratios of ν instead of ratios of n.

In addition to the spectral amplitudes $h(n)$ the generated field can be also characterized by the spectral energy density parameter $\Omega_g(n)$. The energy density ϵ_g of the gravitational wave field is

$$\kappa\epsilon_g = \frac{1}{4}h^{ij}_{,0}h_{ij,0} = \frac{1}{4a^2}h^{ij\prime}h_{ij}{}'.$$

The mean value $\langle 0|\epsilon_g(\eta,\mathbf{x})|0\rangle$ is given by

$$\kappa\langle\epsilon_g\rangle = \frac{1}{4a^2}\frac{C^2}{2\pi^2}\int_0^\infty n\sum_{s=1}^{2}\left|\overset{s}{h}{}'_n(\eta)\right|^2 \mathrm{d}n. \tag{40}$$

For high–frequency modes, it is only the factor $e^{\pm in\eta}$ that needs to be differentiated by η. After avaraging out the oscillating factors, one gets $\left|\overset{s}{h}{}'_n\right|^2 = n^2\left|\overset{s}{h}_n\right|^2$, so that

$$\kappa\langle\epsilon_g\rangle = \frac{1}{4a^2}\int_0^\infty n^2h^2(n)\frac{\mathrm{d}n}{n}. \tag{41}$$

In fact, the high–frequency approximation, that has been used, permits integration over lower n only up to n_H. And the upper limit, as was discussed above, is in practice n_1, not infinity. The parameter Ω_g is defined as $\Omega_g = \langle\epsilon_g\rangle/\epsilon$, where ϵ is given by Eq. (23) (critical density). So, we derive

$$\Omega_g = \int_{n_H}^{n_1} \Omega_g(n)\frac{\mathrm{d}n}{n} = \int_{\nu_H}^{\nu_1} \Omega_g(\nu)\frac{\mathrm{d}\nu}{\nu}$$

and

$$\Omega_g(\nu) = \frac{\pi^2}{3}h^2(\nu)\left(\frac{\nu}{\nu_H}\right)^2. \tag{42}$$

The dimensionless quantity $\Omega_g(\nu)$ is useful because it allows us to quickly evaluate the cosmological importance of the generated field in a given frequency interval. However, the primary and more universal concept is $h(\nu)$, not $\Omega_g(\nu)$. It is the field, not its energy density, that is directly measured by the gravity–wave detector. One should also note that some authors use quite a misleading definition $\Omega_g(f) = (1/\rho_c)(d\rho_{gw}/d\ln f)$ which suggests differentiation of the gravity–wave energy density by frequency. This would be incorrect and could cause disagreements in numerical values of Ω_g. Whenever we use $\Omega_g(\nu)$, we mean relationship (42); and for order of magnitude estimates one can use [1]:

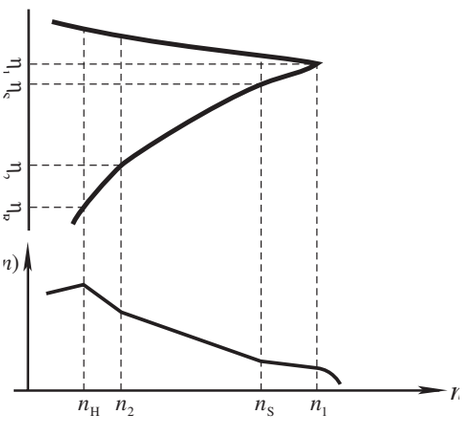


Fig. 6. Amplitudes and spectral slopes of $h(n)$ are determined by different parts of the barrier α'/α.

$$\Omega_g(\nu) \approx h^2(\nu)\left(\frac{\nu}{\nu_H}\right)^2. \tag{43}$$

5 Theoretical and Observational Constraints

The entire theoretical approach is based on the assumption that a weak quantized gravity–wave field interacts with a classical pump field. We should follow the validity of this approximation throughout the analysis. The pump field can be treated as a classical gravitational field as long as the driving energy density ϵ is smaller than the Planck energy density, or, in other words, as long as the Hubble radius $l(\eta)$ is greater than the Planck length l_{Pl}. This is a restriction on the pump field, but it can be used as a restriction on the wavelength λ_i of

the gravity–wave mode n at the time of entry the long–wavelength regime. If $l(\eta_*) > l_{\rm Pl}$, then $\lambda_i > l_{\rm Pl}$. The λ_i is given by Eq. (32). So, we need to ensure that

$$b\,\frac{l_{\rm Pl}}{l_o}\left(\frac{\nu}{\nu_H}\right)^{2+\beta} < 1.$$

At the lowest–frequency end $\nu = \nu_H$ this inequality gives $b(l_{\rm Pl}/l_o) < 1$. In fact, the observational constraints (see below) give a stronger restiction:

$$b\,\frac{l_{\rm Pl}}{l_o} \approx 10^{-6}, \tag{44}$$

which we accept. Then, at the highest–frequency end $\nu = \nu_1$ we need to satisfy

$$\left(\frac{\nu_1}{\nu_H}\right)^{2+\beta} < 10^6. \tag{45}$$

Let us now turn to the generated spectral amplitudes $h(\nu)$. According to Eq. (36) we have $h(\nu_H) \approx b8\sqrt{\pi}(l_{\rm Pl}/l_o)$. The measured microwave beckgound anisotropies, which we discuss below, require this number to be at the level of 10^{-5}, which gives the already mentioned Eq. (44). The quantity $h(\nu_1)$ at the highest frequency ν_1 is given by Eq. (39):

$$h(\nu_1) = b8\sqrt{\pi}\,\frac{l_{\rm Pl}}{l_o}\left(\frac{\nu_1}{\nu_H}\right)^{1+\beta-\beta_s}\left(\frac{\nu_s}{\nu_H}\right)^{\beta_s}\frac{\nu_H}{\nu_2}.$$

Using Eq. (22) this expression for $h(\nu_1)$ can be rewritten as

$$h(\nu_1) = 8\sqrt{\pi}\,\frac{l_{\rm Pl}}{l_H}\frac{\nu_1}{\nu_H} = 8\sqrt{\pi}\frac{l_{\rm Pl}}{\lambda_1}, \tag{46}$$

where $\lambda_1 = c/\nu_1$. This last expression for $h(\nu_1)$ is not surprising: the modes with $\nu \geq \nu_1$ are still in the vacuum state, so the numerical value of $h(\nu_1)$ is determined by quantum normalization.

All the amplified modes have started with small initial amplitudes h_i, at the level of zero–point quantum fluctuations. These amplitudes are also small today, since the h_i could only stay constant or decrease. However, even these relatively small amplitudes should obey observational constraints. We do not want the Ω_g in the high–frequency modes, which might affect the rate of the primordial nucleosynthesis, to exceed the level of 10^{-5}. This means that $\Omega_g(\nu_1)$ cannot exceed the level of 10^{-6} or so. The use of Eq. (42) in combination with $\Omega_g(\nu_1) \approx 10^{-6}$ and $h(\nu_1)$ from Eq. (46), gives us the highest allowed frequency $\nu_1 \approx 3 \times 10^{10}$ Hz. We will use this value of ν_1 in our numerical estimates. Returning with this value of ν_1 to Eq. (45) we find that parameter β can only be $\beta \leq -1.8$. We will be treating $\beta = -1.8$ as the upper limit for the allowed values of β.

We can now check whether the accepted parameters leave room for the postulated z–stage with $\beta_s < 0$. Using Eq. (22) we can rewrite Eq. (44) in the form

$$10^{-6}\frac{l_H}{l_{\rm Pl}} = \left(\frac{\nu_1}{\nu_H}\right)^{-\beta}\left(\frac{\nu_1}{\nu_s}\right)^{\beta_s}\frac{\nu_2}{\nu_H}. \qquad (47)$$

We know that $\nu_2/\nu_H = 10^2$ and ν_1/ν_s is not smaller than 1. Substituting all the numbers in Eq. (47) one can find that this equation cannot be satisfied for the largest possible $\beta = -1.8$. In the case $\beta = -1.9$, Eq. (47) is only marginally satisfied, in the sense that a significant deviation from $\beta_s = 0$ toward negative β_s can only last for a relatively short time. For instance, one can accomodate $\beta_s = -0.4$ and $\nu_s = 10^8$ Hz. On the other hand, if one takes $\beta = -2$, a somewhat longer interval of the z–stage with $\beta_s < 0$ can be included. For instance, Eq. (47) is satisfied if one accepts $\nu_s = 10^{-4}$ Hz and $\beta_s = -0.3$. This allows us to slightly increase $h(\nu)$ in the interval $\nu_s < \nu < \nu_1$, as compared with the values of $h(\nu)$ reached in the more traditional case $\beta = -2$, $\beta_s = 0$. In what follows, we will consider consequences of this assumption for the prospects of detection of the produced gravitational wave signal. Finally, let us see what the available information on the microwave background anisotropies [15,16] allows us to conclude about the parameters β and l_o.

Usually, cosmologists operate with the spectral index n (not to be confused with the wave number n) of primordial cosmological perturbations. Taking into account the way in which the spectral index n is defined, one can relate n with the spectral index $\beta + 2$ that shows up in Eq.(36). The relationship between them is $\mathrm{n} = 2\beta + 5$. This relationship is valid independently of the nature of cosmological perturbations. In particular, it is valid for density perturbations, in which case the $h(n)$ of Eq.(36) is the dimensionless spectral amplitude of metric perturbations associated with density perturbations. If primordial gravitational waves and density perturbations were generated by the mechanism that we discuss here (the assumption that is likely to be true) than the parameter β that participates in the spectral index is the same one that participates in the scale factor of Eq. (18). Primordial gravitational waves and primordial density perturbations with the same spectral index produce approximately the same lower–order multipole distributions of large–scale anisotropies.

The evaluation of the spectral index n of primordial perturbations have resulted in $\mathrm{n} = 1.2 \pm 0.3$ [16] or even in a somewhat higher value. A recent analysis [17] of all available data favors $\mathrm{n} = 1.2$ and the quadrupole contribution of gravitational waves twice as large as that of density perturbations. One can interpret these evaluations as indication that the true value of n lies somewhere near $\mathrm{n} = 1.2$ (hopefully, the planned new observational missions will determine this index more accurately). This gives us the parameter β somewhere near $\beta = -1.9$. We will be using $\beta = -1.9$ in our estimates below, as the observationally preferred value. The parameter β can be somewhat larger than $\beta = -1.9$. However, as we already discussed, the value $\beta = -1.8$ ($\mathrm{n} = 1.4$) is the largest one for which the entire approch is well posed. The Harrison–Zeldovich spectral index $\mathrm{n} = 1$ corresponds to $\beta = -2$.

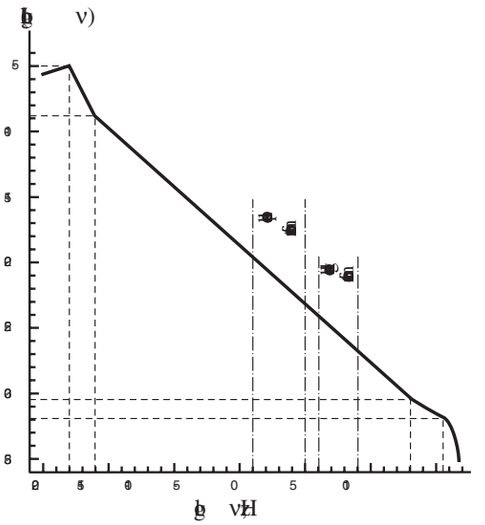


Fig. 7. Expected spectrum $h(\nu)$ for the case $\beta = -1.9$.

The observed quadrupole anisotropy of the microwave background radiation is at the level $\delta T/T \approx 10^{-5}$. The quadrupole anisotropy that would be produced by the spectrum (36) - (39) is mainly accounted for by the wave numbers near n_H. Thus, the numerical value of the quadrupole anisotropy produced by relic gravitational waves is approximately equal to A. According to general physical considerations and detailed calculations [18], the metric amplitudes of long–wavelength gravitational waves and density perturbations generated by the discussed amplification mechanism are of the same order of magnitude. Therefore, they contribute roughly equally to the anisotropy at lower multipoles. This gives us the estimate $A \approx 10^{-5}$, that we have already used in Eq. (44). It is not yet proven observationally that a significant part of the observed anosotropies at lower multipoles is indeed provided by relic gravitational waves, but we can at least assume this with some degree of confidence. It is likely that the future measurements of the microwave background radiation will help us to verify this theoretical conclusion.

Combining all the evaluated parameters together, we show in Fig. 7 the expected spectrum of $h(\nu)$ for the case $\beta = -1.9$. A small allowed interval of the z–stage is also included. The intervals of the spectrum accessible to space–based and ground–based interferometers are indicated by vertical lines.

It is necessary to note [18,19] that the confirmation of any n > 1 ($\beta > -2$) would mean that the very early Universe was not driven by a scalar field – the cornerstone of inflationary considerations. This is because the n > 1 ($\beta > -2$) requires the effective equation of state at the initial stage of expansion to be $\epsilon + p < 0$ (see Eq. (24)), but this cannot be accomodated by any scalar field with

whichever scalar field potential. The available data do not prove yet that $\mathrm{n} > 1$, but this possibility seems likely.

It is also necessary to say that a certain damage to gravitational wave research was inflicted by the so called "standard inflationary result". The "standard inflationary result" predicts infinitely large amplitudes of density perturbations in the interval of spectrum with the Harrison–Zeldovich slope $\mathrm{n} = 1$ ($\beta = -2$): $\delta\rho/\rho \propto 1/\sqrt{1-\mathrm{n}}$. The metric (gravitational field) amplitudes of density perturbations are also predicted to be infinitely large, in the same proportion. Through the so–called "consistency relation" this divergence leads to the vanishingly small amplitudes of relic gravitational waves. Thus, the "standard" inflationary theory predicts zero for relic gravitational waves; the spectrum similar in shape to the one shown in Fig. 7 would have been shifted down by many orders of magnitude. This prediction is hanging on the "standard inflationary result", but the "result" itself is in a severe conflict not only with theory but with observations too: when the observers marginalize their data to $\mathrm{n} = 1$ (enforce this value of n in data analysis) they find finite and small density perturbations instead of infinitely large perturbations predicted by inflationary theorists. (For analytical expressions of the "standard inflationary result" see any inflationary article, including recent reviews. For graphical illustration of the predicted divergent density perturbations and quadrupole anisotropies see [20], for example. For critical analysis and disagreement with the "standard inflationary result" see [18].) General relativity and quantum field theory do not produce the "standard inflationary result", so we shall better return to what they say.

6 Detectability of Relic Gravitational Waves

We switch now from cosmology to prospects of detecting the predicted relic gravitational waves. The ground–based [21]-[23] and space–based [24], [25] laser interferometers (see also [26]-[28]) will be in the focus of our attention. We use laboratory frequencies ν and intervals of laboratory time t ($cdt = a(\eta_R)\mathrm{d}\eta$). Formulas (38) and (39), with $A = 10^{-5}$, $\nu_2/\nu_H = 10^2$, and the oscillating factor restored, can be written as

$$h(\nu, t) \approx 10^{-7} \cos[2\pi\nu(t - t_\nu)] \left(\frac{\nu}{\nu_H}\right)^{\beta+1}, \quad \nu_2 \le \nu \le \nu_s \tag{48}$$

and

$$h(\nu, t) \approx 10^{-7} \cos[2\pi\nu(t - t_\nu)] \left(\frac{\nu}{\nu_H}\right)^{1+\beta-\beta_s} \left(\frac{\nu_s}{\nu_H}\right)^{\beta_s}. \quad \nu_s \le \nu \le \nu_1 \tag{49}$$

where the deterministic (not random) constant t_ν does not vary significantly from one frequency to another at the intervals $\Delta\nu \approx \nu$. The explicit time dependence of the spectral variance $h^2(\nu, t)$ of the field, or, in other words, the explicit time dependence of the (zero–lag) temporal correlation function of the field at every given frequency, demonstrates that we are dealing with a non–stationary process (a consequence of squeezing and severe reduction of the phase uncertainty). We will first ignore the oscillating factor and will compare the predicted

amplitudes with the sensitivity curves of advanced detectors. The potential reserve of improving the signal to noise ratio by expoloiting the squeezing will be discussed later.

Let us start from the Laser Interferometer Space Antenna (LISA) [24]. The instrument will be most sensitive in the interval, roughly, from 10^{-3} Hz to 10^{-1} Hz, and will be reasonably sensitive in a broader range, up to frequencies 10^{-4} Hz and 1 Hz. The sensitivity graph of LISA to a stochastic background is usually plotted under the assumption of a 1–year observation time, that is, the root–mean–square (r.m.s.) instrumental noise is being evaluated in frequency bins $\Delta\nu = 3 \times 10^{-8}$ Hz around each frequency ν. We need to rescale our predicted amplitude $h(\nu)$ to these bins.

The mean square amplitude of the gravitational wave field is given by the integral (29). Thus, the r.m.s. amplitude in the band $\Delta\nu$ centered at a given frequency ν is given by the expression

$$h(\nu, \Delta\nu) = h(\nu)\sqrt{\frac{\Delta\nu}{\nu}}. \tag{50}$$

We use Eqs. (48), (49) and calculate expression (50) assuming $\Delta\nu = 3\times10^{-8}$ Hz. The results are plotted in Fig. 8. Formula (48) has been used throughout the covered frequency interval for the realistic case $\beta = -1.9$ and for the extreme case $\beta = -1.8$. The line marked z–model describes the signal produced in the composite model with $\beta = -2$ up to $\nu_s = 10^{-4}$ Hz (formula (48)) and then followed by formula (49) with $\beta_s = -0.3$. This model gives the signal a factor of 3 higher at $\nu = 10^{-3}$ Hz, than the model $\beta = -2$ extrapolated down to this frequency.

There is no doubt that the signal $\beta = -1.8$ would be easily detectable even with a single instrument. The signal $\beta = -1.9$ is marginally detectable, with the signal to noise ratio around 3 or so, in a quite narrow frequency interval near and above the frequency 3×10^{-3} Hz. However, at lower frequencies one would need to be concerned with the possible gravitational wave noise from unresolved binary stars in our Galaxy. The further improvement of the expected LISA sensitivity by a factor of 3 may prove to be crucial for a confident detection of the predicted signal with $\beta = -1.9$.

Let us now turn to the ground–based interferometers operating in the interval from 10 Hz to 10^4 Hz. The best sensitivity is reached in the band around $\nu = 10^2$ Hz. We take this frequency as the representative frequency for comparison with the predicted signal. We will work directly in terms of the dimensionless quantity $h(\nu)$. If necessary, the r.m.s. amplitude per $\mathrm{Hz}^{1/2}$ at a given ν can be found simply as $h(\nu)/\sqrt{\nu}$. The instrumental noise will also be quoted in terms of the dimensionless quantity $h_{ex}(\nu)$.

The expected sensitivity of the initial instruments at $\nu = 10^2$ Hz is $h_{ex} = 10^{-21}$ or better. The theoretical prediction at this frequency, following from (48), (49) with $\beta_s = 0$, is $h_{th} = 10^{-23}$ for $\beta = -1.8$, and $h_{th} = 10^{-25}$ for $\beta = -1.9$. Therefore, the gap between the signal and noise levels is from 2 to 4 orders of magnitude. The expected sensitivity of the advanced interferometers, such as

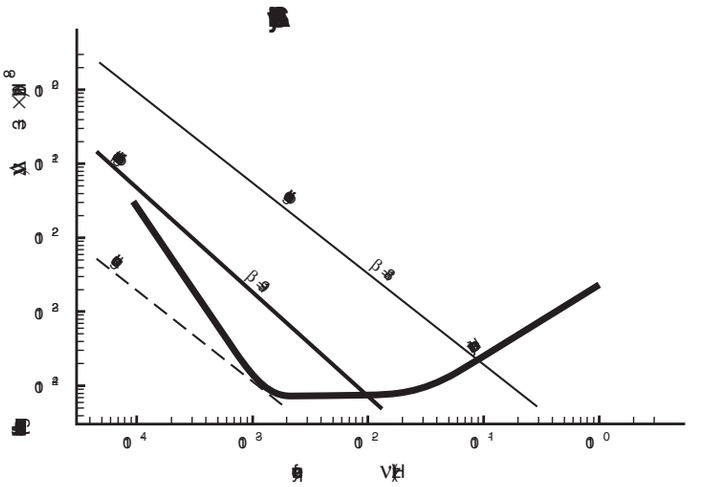


Fig. 8. Expected spectrum $\beta = -1.9$ and other possible spectra in comparison with the LISA sensitivity.

LIGO-II [29], can be as high as $h_{ex} = 10^{-23}$. In this case, the gap vanishes for the $\beta = -1.8$ signal and reduces to 2 orders of magnitude for the $\beta = -1.9$ signal. Fig. 9 illustrates the expected signal in comparison with the LIGO-II sensitivity. Since the signal lines are plotted in terms of $h(\nu)$, the LISA sensitivity curve (shown for periodic sources) should be raised and adjusted in accordance with Fig. 8.

A signal below noise can be detected if the outputs of two or more detectors can be cross correlated. (For the early esimates of detectability of relic gravitational waves see [30].) The cross correlation will be possible for ground–based interferometers, several of which are currently under construction. The gap between the signal and the noise levels should be covered by a sufficiently long observation time τ. The duration τ depends on whether the signal has any temporal signature known in advance, or not. We start from the assumption that no temporal signatures are known in advance. In other words, we first ignore the squeezed nature of the relic background and work under the assumption that the squeezing cannot be exploited to our advantage.

The response of an instrument to the incoming radiation is $s(t) = F_{ij}h^{ij}$ where F_{ij} depends on the position and orientation of the instrument. Since the h^{ij} is a quantum–mechanical operator (see Eq. (6)) we need to calculate the mean value of a quadratic quantity. The mean value of the cross correlation of responses from two instruments $\langle 0|s_1(t)s_2(t)|0\rangle$ will involve the overlap reduction function [31]-[34], which we assume to be not much smaller than 1 [33]. The signal to noise ratio S/N in the measurement of the amplitude of a signal with no specific known features increases as $(\tau\nu)^{1/4}$, where ν is some characteristic

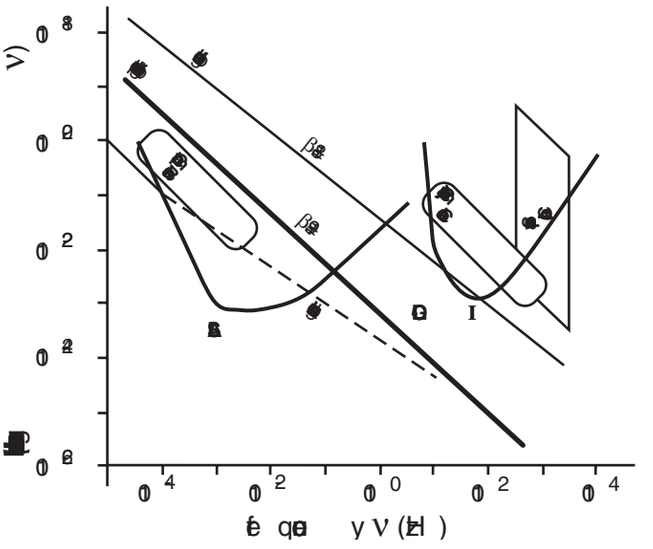


Fig. 9. Full spectrum $h(\nu)$ accessible to laser interferometers.

central frequency. If the signal has features known in advance and exploited by the matched filtering technique, the S/N increases as $(\tau\nu)^{1/2}$.

We apply the guaranteed law $(\tau\nu)^{1/4}$ to initial and advanced instruments at the representative frequency $\nu = 10^2$ Hz. This law requires a reasonably short time $\tau = 10^6$ sec in order to improve the S/N in initial instruments by two orders of magnitude and to reach the level of the signal with extreme spectral index $\beta = -1.8$. The longer integration time or a better sensitivity will make the S/N larger than 1. In the case of a realistic spectral index $\beta = -1.9$ the remaining gap of 4 orders of magnitude can be covered by the combination of a significantly better sensitivity and a longer observation time (not necessarily in one non–interrupted run). The sensitivity of the advanced laser interferometers, such as LIGO II, at the level $h_{ex} = 10^{-23}$ and the same observation time $\tau = 10^6$ sec would be sufficient for reaching the level of the predicted signal with $\beta = -1.9$.

An additional increase of S/N can be achieved if the statistical properties of the signal can be properly exploited. Squeezing is automatically present at all frequencies from ν_H to ν_1. The squeeze parameter r is larger in gravitational waves of cosmological scales, and possibly the periodic structure in Eq. (31) can be better revealed at those scales. However, we are interested here in frequencies accessible to ground based interferometers, say, in the interval 30 Hz − 100 Hz. If our intention were to monitor one given frequency ν from the beginning of its oscillating regime and up till now, then, in order to avoid the destructive interference from neighbouring modes during all that time, the frequency resolution of the instrument should have been increadibly narrow, of the order of

10^{-18} Hz. Certainly, this is not something what we can, or intend to do. Although the amplitudes of the waves have adiabatically decreased and their frequencies redshifted since the beginning of their oscillating regime, the general statistical properties of the discussed signal are essentially the same now as they were 10 years after the Big Bang or will be 1 million years from now.

The periodic structure (48) may survive at some level in the instrumental window of sensitivity from ν_{min} (minimal frequency) to ν_{max} (maximal frequency). The mean square value of the field in this window is

$$\int_{\nu_{\min}}^{\nu_{\max}} h^2(\nu,t)\frac{\mathrm{d}\nu}{\nu} = 10^{-14}\frac{1}{{\nu_H}^{2\beta+2}}\int_{\nu_{\min}}^{\nu_{\max}} \cos^2[2\pi\nu(t-t_\nu)]\nu^{2\beta+1}\mathrm{d}\nu \ . \qquad (51)$$

Because of the strong dependence of the integrand on frequency, $\nu^{-2.6}$ or $\nu^{-2.8}$, the value of the integral (51) is determined by its lower limit. Apparently, the search through the data should be based on the periodic structure that may survive at $\nu = \nu_{\min}$. As an illustration, one can consider such a narrow interval $\Delta\nu = \nu_{\max} - \nu_{\min}$ that the integral (51) can be approximated by the formula

$$\int_{\nu_{\min}}^{\nu_{\max}} h^2(\nu,t)\frac{\mathrm{d}\nu}{\nu} \approx 10^{-14}\left(\frac{\nu_{\min}}{\nu_H}\right)^{2\beta+2}\left(\frac{\Delta\nu}{\nu_{\min}}\right)\cos^2[2\pi\nu_{\min}(t-t_{\min})] \ .$$

Clearly, the correlation function is strictly periodic and its structure is known in advance, in contrast to other possible signals. This is a typical example of using the a priori information. Ideally, the gain in S/N can grow as $(\tau\nu_{\min})^{1/2}$. This would significantly reduce the required observation time τ. For a larger $\Delta\nu$, even an intermediate gain between the guaranteed law $(\tau\nu)^{1/4}$ and the law $(\tau\nu)^{1/2}$, adequate for the matched filtering technique, would help. This could potentially make the signal with $\beta = -1.9$ measurable even by the initial laser interferometers. A straightforward application of (51) for exploiting the squeezing may not be possible, as argued in the recent study [35], but more sophisticated methods are not excluded.

For frequency intervals covered by bar detectors and electromagnetic detectors, the results expected follow from the same formulas (48,49) and have been briefly discussed elsewhere [30,19].

7 Conclusion

It would be strange, if the predicted signal at the level corresponding to $\beta = -1.9$ were not seen by the instruments capable of its detection. There are not so many cosmological assumptions involved in the derivation that could prove wrong, thus invalidating our predictions. On the other hand, it would be even more strange (and even more interesting) if the relic gravitational waves were detected at the level above the $\beta = -1.8$ line. This would mean that there is something fundamentally wrong in our basic cosmological premises. To summarise, it is quite possible that the detection of relic (squeezed) gravitational waves may be awaiting only the first generation of sensitive instruments and an appropriate data processing strategy.

Acknowlegements

I appreciate the help of M. V. Prokhorov in preparation of the figures.

References

1. L.P. Grishchuk: *Zh. Eksp. Teor. Fiz.* **67**, 825 (1974) [JETP **40**, 409 (1975)]; *Ann. NY Acad. Sci.* **302**, 439 (1977); *Uspekhi Fiz. Nauk.* **156**, 297 (1988) [Sov. Phys.-Uspekhi. **31**, 940 (1988)].
2. K.S. Thorne, in S.W. Hawking and W. Israel (eds.): *300 Years of Gravitation* (Cambridge University Press, Cambridge 1987), p. 330.
3. K.S. Thorne, in E. Kolb and R. Peccei (eds.): *Particle and Nuclear Astrophysics and Cosmology in the Next Millenium* (World Scientific, Singapore 1995), p. 160.
4. B.F. Schutz: *Class. Quant. Grav.* **16**, A131 (1999).
5. P.L. Knight: in S. Reynaud, E. Giacobino, and J. Zinn–Justin (eds.): *Quantum Fluctuations* (Elsevier Science 1997), p. 5.
6. L.P. Grishchuk and Yu.V. Sidorov: *Class. Quant. Grav.* **6**, L161 (1989); *Phys. Rev.* **D 42**, 3413 (1990).
7. L.P. Grishchuk, in *Workshop on Squeezed States and Uncertainty Relations*, NASA Conf. Publ. **3135** 329 (1992). *Class. Quant. Grav.* **10**, 2449 (1993); in S. Reynaud, E. Giacobino, and J. Zinn–Justin (eds.): *Quantum Fluctuations* (Elsevier Science 1997), p. 541.
8. L.D. Landau and E.M. Lifshitz: *The Classical Theory of Fields* (New York: Pergamon) 1975.
9. W. Schleich and J.A. Wheeler: *J. Opt. Soc. Am.* **B4**, 1715 (1987); W. Schleich *et. al.*: *Phys. Rev.* **A 40**, 7405 (1989).
10. Ya.B. Zeldovich and I.D. Novikov: *The Structure and Evolution of the Universe* (University of Chicago Press, Chicago 1983).
11. M. Giovannini: *Phys. Rev.* **D 58**, 083504 (1998); Report hep-ph/9912480.
12. L.P. Grishchuk: Report gr-qc/9810055.
13. G. Veneziano: *Phys. Lett.* **B 265**, 287 (1991); M. Gasperini and G. Veneziano: *Astropart. Phys.* **1**, 317 (1993); M. Gasperini and M. Giovannini: *Phys. Rev.* **D 47**, 1519 (1993); M. Gasperini: Report hep-th/9607146.
14. T. Creighton: Report gr-qc/9907045.
15. G.F. Smoot *et. al.*: *Astroph. J.* **396**, L1 (1992).
16. C.L. Bennet *et. al.*: *Astroph. J.* **464**, L1 (1996).
17. A. Melchiori, M.V. Sazhin, V.V. Shulga, and N. Vittorio: *Astroph. J.* **518**, 562 (1999).
18. L.P. Grishchuk: *Phys. Rev.* **D 50**, 7154 (1994); in N. Sanchez and A. Zichichi (eds.): *Current Topics in Astrofundamental Physics: Primordial Cosmology* (Kluwer Academic 1998), p. 539; Report gr-qc/9801011.
19. L.P. Grishchuk: *Class. Quant. Grav.* **14**, 1445 (1997).
20. J. Martin and D.J. Schwarz: Report astro-ph/9911225.
21. A. Abramovici *et. al.*: *Science* **256**, 325 (1992).
22. C. Bradaschia *et. al.*: *Nucl. Instrum. and Methods* **A 289**, 518 (1990).
23. J. Hough and K. Danzmann *et. al.*: *GEO600 Proposal*, 1994.
24. P. Bender *et. al.*: *LISA Pre–Phase A Report*, Second Edition, 1998.
25. S.L. Larson, W. A. Hiscock. and R. W. Hellings: Report gr-qc/9909080.
26. *Gravitational Wave Experiments*, E. Coccia, G. Pizzella, and F. Ronga, eds. (World Scientific, Singapore 1995).

27. *Gravitational Waves: Sources and Detection*, I. Giufolini and F. Fidecaro, eds. (World Scientific, Singapore 1997).
28. W. Folkner (ed.): *Laser Interferometer Space Antenna* (AIP Conference Proceedings **456**, 1998).
29. *LIGO II Conceptual Project Book*, LIGO-M990288-00-M.
30. L.P. Grishchuk: *Pis'ma Zh. Eks. Teor. Fiz.* **23**, 326 (1976) [JETP Lett. **23**, 293 (1976)].
31. P.F. Michelson: *Mon. Not. R. Astr. Soc.* **227**, 933 (1987).
32. N.L. Christensen: *Phys. Rev.* **D 46**, 5250 (1992).
33. E.E. Flanagan: *Phys. Rev.* **D 48**, 2389 (1993).
34. B. Allen, in J–A. Marck and J.–P. Lasota (eds.): *Relativistic Gravitation and Gravitational Radiation* (Cambridge University Press, Cambridghe 1997) p. 373.
35. B. Allen, E.E. Flanagan and M.A. Papa: Report gr-qc/9906054.

Principles of Equivalence: Their Role in Gravitation Physics and Experiments That Test Them

Mark P. Haugan[1] and C. Lämmerzahl[2]

[1] Purdue University, West Lafayette, IN 47907, USA
[2] Department of Physics, University of Konstanz, Fach M674, 78457 Konstanz, Germany

Abstract. Modern formulations of equivalence principles provide the foundation for an efficient approach to understanding and organizing the structural features of gravitation field theories. Since theories' predictions reflect differences in their structures, principles of equivalence also support an efficient experimental strategy for testing gravitation theories and for exploring the range of conceivable gravitation physics. These principles focus attention squarely on empirical consequences of the fundamental structural differences that distinguish one gravitation theory from another. Interestingly, the variety of such consequences makes it possible to design and perform experiments that test equivalence principles stringently but do so in markedly different ways than the most familiar experimental tests.

1 Equivalence Principles and the Structure of Gravitation Theories

1.1 From the Weak to Einstein's Equivalence Principle

Since the time of the Renaissance observations indicating that bodies fall in a gravitational field in a way that is independent of their internal composition and structure have been considered remarkable. Clearly Newton thought so since he deemed it necessary to perform pendulum experiments to verify this property of freefall as precisely as he could before he published laws of motion and universal gravitation that predict it [1]. Einstein also found this property of freefall remarkable. The insight he gained by reflecting on it in his famous elevator *Gedanken* experiment [2] is communicated by what we now call the Einstein equivalence principle (EEP).

Einstein noted that if all bodies fall in the same way in an external gravitational field, an observer in freefall will find that freely falling bodies in his or her neighborhood move with uniform velocities relative to him or her and that the physics of pure particle mechanics in that neighborhood is indistinguishable from mechanics in the absence of gravity. This led him to suggest that a freely falling observer might find all other nongravitational physics in his or her neighborhood to be indistinguishable from such physics in the absence of gravity. Einstein then proceeded to show that if this were true for electrodynamic physics, which was

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 195–212, 2001.

all fundamental physics at the time, light propagating out of a gravitational potential well must suffer a redshift.

The following modern formulation of the EEP expresses the possibility suggested by Einstein in 1907. It states that the outcome of any local, nongravitational test experiment is independent of the experimental apparatus' velocity relative to the gravitational field and is independent of where and when in the gravitational field the experiment is performed. This captures Einstein's suggestion because, in principle, local nongravitational test experiments can be performed in spacetime regions where gravity is negligible. The two conditions that the EEP imposes are referred to as local Lorentz and local position invariance, respectively. Note that a *local* experiment is one performed within a spacetime region so small that the experimental apparatus can detect no tidal effects. A *test* experiment is one performed with an apparatus having a mass so small that the apparatus can detect no effect of the perturbation it induces in the gravitational field.

From the perspective of 1907, the EEP is a striking generalization from the observed equality of test body accelerations in a gravitational field, and, as a matter of history, the gravitational redshift was the first physical consequence to be derived on the assumption that this generalization is valid. Given this history, it is not surprising that refinements of experimental tests of the universality of freefall acceleration and of measurements of the gravitational redshift remain among the most widely recognized tests of the validity of the EEP. Efforts to refine both kinds of test continue today. Readers can refer to Lute Maleki's article [3] in this volume for details on a proposed space–based variation on gravitational redshift measurements.

Tests of the universality of free fall acceleration are often referred to as Eötvös experiments because of the classic torsion–balance version performed by Baron von Eötvös and collaborators early in this century [4]. As noted above, the process of refining such tests continues today. For example, the group of Adelberger at the University of Washington [5] recently reported results of a torsion–balance experiment that include the conclusion that the gravitational accelerations of beryllium and copper test bodies toward the Earth are equal to better than 2.5 parts in 10^{12}. The ultimate refinement of tests of the equality of such accelerations may well be represented by the space–based STEP experiment under development at Stanford University. Readers can refer to an article in this volume for details on STEP [6].

Interest in such experiments remains high because our understanding of their significance as tests of the EEP has evolved significantly since 1907. The clear distinction Einstein made at that time between particle mechanics and other nongravitational physics, specifically electromagnetic physics, is no longer viable. We now understand how test bodies are composed of atoms and that they, in turn, are composites of the mass–energy of nucleons and electrons and of the electromagnetic, weak– and strong–interaction binding energies of these particles. Consequently, the freefall acceleration of test bodies can be influenced by many, if not all, aspects of nongravitational physics in an external gravitational

field, and experiments which test the universality of such accelerations turn out to be more profound tests of the EEP than one could have realized in 1907. We will return to this point later in section 2.1.

Set aside, for a moment, the issue of whether or not the EEP is valid and questions regarding the precision to which we may be able to establish experimentally that it is valid. Is it possible to formulate a gravitation field theory which predicts that it is valid? It took Einstein almost ten years from the time of his 1907 insight to establish that the answer to this question is yes. He did so by formulating general relativity [7]. Interested readers can refer to the recent review by Norton [8] for a discussion of the fascinating and in many ways still controversial history of Einstein's development of general relativity and of others' early attempts to understand the theory. This story is plagued by many formulations of the equivalence principle, confusion regarding the significance of coordinates and covariance and so on. We will not delve further into it here. Instead, we discuss the kind of analysis one must do to determine whether or not any given gravitation field theory predicts the validity of the EEP.

1.2 Theoretical Contexts for Analyses of the EEP

If an analysis reveals outcomes of some local nongravitational test experiment that depend on the velocity of the experimental apparatus relative to an external gravitational field or on where or when in that field the experiment is performed, it is clear that the underlying theory is nonmetric, that is, it violates the EEP. Significantly, such an analytical result also provides the basis for actual experiments that search for the specific preferred–frame or preferred–location effect revealed by the analysis. This approach has led to the development of many stringent new tests of the EEP that are quite different from familiar Eötvös experiments and gravitational redshift measurements. It has also clarified which structural features of gravitation field theories are constrained by experimental evidence that the EEP is valid to some level of precision.

Lagrangian field theory provides a natural setting for a general discussion of gravitation field theories. However, we note that much of what follows can be discussed in terms of gravitation field equations and matter–field equations of motion. Indeed, a number of recent papers exploit this latter approach to consider modifications of the Maxwell equations caused by quantum gravity effects. Papers by Gambini and Pullin [9] and by Ellis *et al.* [10] are examples. The paper of Haugan and Lämmerzahl [11] begins the analysis of physical consequences of a broad range of conceivable Maxwell equation modifications of this kind and the consideration of experiments that could detect them or constrain their magnitudes.

Returning to Lagrangian–based gravitation theories, we recall that each admits a formulation via an action principle,

$$\delta \int \mathcal{L}(\psi_g, \psi_m)\, d^4x = 0. \tag{1}$$

Here, ψ_g denotes dependence of the Lagrangian density on gravitational potentials and their derivatives, and ψ_m denotes dependence on matter fields and their derivatives. Not that long ago, one would have restricted attention to theories in which field derivatives appear only in conventional "kinetic" terms in $\mathcal{L}$, but attitudes have changed so that field theories tend to be viewed as effective rather than fundamental theories. Consequently, there is now a greater willingness to consider dependence on higher–order derivatives and the presence of derivative couplings between fields. Such things can make theories nonrenormalizable, but this is not the issue for effective theories that it is for fundamental ones.

A theory's Lagrangian density $\mathcal{L}$ can be split into a purely gravitational part and "nongravitational" remainder, $\mathcal{L} = \mathcal{L}_g + \mathcal{L}_{ng}$. The gravitational part $\mathcal{L}_g$ depends only on gravitational potentials and their derivatives. Its form specifies the dynamics of free gravitational fields in the theory. The nongravitational part $\mathcal{L}_{ng}$ depends on gravitational potentials and their derivatives and on matter fields and their derivatives. Its form specifies the coupling between matter and gravity in the theory. Its form determines both how matter responds to gravity and how matter acts as a source of gravity.

1.3 The Role of Locality

The matter fields involved in a local, nongravitational test experiment do not perturb gravitational potentials to a degree that the experiment can detect. It follows that when using a theory to predict the outcome of such an experiment we can treat gravitational potentials as specified functions which represent the gravitational environment generated by some source. They are external potentials. To predict the outcome of the experiment one needs only a theory's equations which govern the evolution of matter fields in the relevant gravitational environment. We derive these gravitationally–modified equations of motion from the theory's action principle (1) by considering variations of the matter fields while keeping the external gravitational potential functions fixed in the appropriate form. Consequently, the outcomes of experiments that directly test the EEP depend on the form of a theory's nongravitational Lagrangian density $\mathcal{L}_{ng}$ alone. Clearly, evidence that the EEP is valid to some precision can constrain only the manner in which matter couples to gravity.

Since experiments that directly test the EEP are local as well as test experiments, their outcomes are insensitive to the global form of external gravitational potentials. It is sufficient to consider initial terms of the Taylor–series expansions of external potentials when using a theory's action principle to predict a local experiment's outcome. The expansions should be made about an event inside the experimental apparatus during the course of the experiment. Keeping terms through first order is sufficient to predict the outcome of any local experiment. However, in some important cases it is sufficient to keep only zeroth–order terms. In particular, this can be the case when an experiment is completed quickly enough that the experimental apparatus can detect no effect of external potential time dependence and no effect of accelerations induced by external potential spatial dependence. This is generally true of realistic local

experiments which measure atomic transition frequencies, for example, and explains why atomic clocks may generally be treated as realizations of ideal clocks in the sense defined by theories of relativity.

To make the remainder of this discussion a bit more concrete focus on experiments involving such atomic transitions. To this point, we have established that a theory's predictions of their outcomes follows from the form that its nongravitational Lagrangian density $\mathcal{L}_{ng}$ takes when values of the external gravitational potentials and, if there are derivative matter–gravity couplings, values of their derivatives at an event inside the experimental apparatus during the experiment are plugged in. The resulting Lagrangian density has no explicit dependence on the spacetime coordinates and involves only matter fields. It determines the gravitationally–modified equations of motion which govern the structure of atoms treated as local test bodies. Note, however, that if we are interested only in atomic transition frequencies, we need not deal with these equations. We can, instead, compute energies of atomic states directly.

1.4 Relevant Observables

An expression for such energies follows from the form of the Lagrangian density introduced in the preceding paragraph because it is time independent. If *natural* coordinates are used in representing a theory's action principle (1), the energy expression's form will be that of the Standard Model Hamiltonian plus perturbating terms. In this context, natural coordinates are ones in which the form of the representation of the theory's nongravitational Lagrangian density $\mathcal{L}_{ng}$ reduces to the familiar representation of the Standard Model Lagrangian density as gravity is "turned off." Schwarzschild coordinates provide a familiar example of natural coordinates in the context of general relativity and situations in the external gravitational potential is static and spherically symmetric.

Accurate estimates of the energies of atomic states are easily computed when the gravitationally–modified Hamiltonian is a perturbed Standard Model Hamiltonian. In general, the results depend on an atom's velocity through and location in the external gravitational potentials. This is the case because the perturbing Hamiltonian terms reflect not only the form of the gravitation theory's nongravitational Lagrangian density $\mathcal{L}_{ng}$ but also the atom's gravitational environment. This environment is represented by the values in the atom's neighborhood of the external gravitational potentials and, if there are derivative matter–gravity couplings, the values of their derivatives.

Despite their velocity and location dependence these computed atomic state energies may not represent preferred–frame or preferred–location effects that signal violation of the Einstein equivalence principle. Even when we are using natural coordinates they are merely *coordinate* energies. Only velocity or location dependence of an experimentally measured atomic state energy, or frequency of a transition between such states, would constitute a genuine preferred–frame or preferred–location effect.

The distinction between coordinate energies or frequencies and measured energies or frequencies is, in some respects, subtle. However, it is not difficult to

appreciate if one remembers that, fundamentally, any measurement is simply the comparison of a property of some system of interest to the corresponding property of a chosen standard system. Thus, a measurement of the frequency of a transition between some pair of atomic states is simply a comparison of its coordinate frequency to the coordinate frequency of a selected standard transition. For example, one can imagine locking a laser to the transition whose frequency is to be measured and a second laser to the frequency of the standard transition. The coordinate frequency of each laser can depend on velocity through or location in the external gravitational potential, but the relative or beat frequency between them, which is the measured frequency, may not. While it may seem far–fetched to imagine cases in which coordinate energies of all atomic states depend on velocity through or location in external gravitational potentials in precisely the same way, thus, causing such dependence to cancel from measured energies, this is precisely what metric theories of gravity like general relativity predict. The preferred–frame or preferred–location effects predicted by nonmetric theories of gravity occur because their matter–gravity couplings distinguish between the contributions of rest–mass and of different types of interaction–energy to the energies of atomic states and, so, prevent such universal cancellations of coordinate effects. If such couplings are present, the observable frequency ratio for a pair of atomic clocks whose ticking rates are governed by different atomic transition and which move together through an external gravitational potential can depend on the clocks' location in and velocity through the potential. In the limit of slow motion and weak gravitation this frequency ratio takes the form

$$\frac{\nu_1(\boldsymbol{x},\boldsymbol{v})}{\nu_2(\boldsymbol{x},\boldsymbol{v})} = \frac{\nu_1^0}{\nu_2^0}\left(1 + \alpha_{ij}v^iv^j + \beta_{ij}U^{ij}(\boldsymbol{x})\right), \tag{2}$$

where the x^i denote natural spatial coordinates that reduce to Cartesian ones as gravity is turned off, the v^i denote corresponding components of the clocks' coordinate velocity and $U^{ij}(\boldsymbol{x})$ denotes the usual Newtonian gravitational potential tensor at the clocks' location. The parameters α_{ij} and β_{ij} depend on the particular transitions controlling the atomic clock rates, except in metric theories of gravity which predict that they all vanish. Their tensor character reflects the fact that the orientation of the atoms whose transitions govern the atomic clock rates can affect the observed frequency ratio (2). Nonvanishing α_{ij} and β_{ij} parameters characterize preferred–frame and preferred–location effects, respectively.

To conclude this discussion of the way in which Lagrangian–based gravitation theories predict outcomes of local, nongravitational test experiments, briefly consider the acceleration of test bodies in an external gravitational field once more. We focus on the freefall of atoms since, as we noted earlier, realistic test bodies are simply assemblages of them.

The analysis of atomic systems outlined above yields an expression for the coordinate energy of any atom in any state of interest. This energy is a function of

the atom's velocity relative to and location in an external gravitational potential

$$E = mc^2 + \frac{1}{2}m\left(\delta_{ij} + \frac{\delta m_{\mathrm{i}ij}}{m}\right)v^i v^j + m\left(\delta_{ij} + \frac{\delta m_{\mathrm{g}ij}}{m}\right)U^{ij}(\boldsymbol{x})\,, \tag{3}$$

where $\delta m_{\mathrm{i}ij}$ and $\delta m_{\mathrm{g}ij}$ are the anomalous inertial and gravitational mass tensors. They depend on the particular state of the particular atom under consideration, except in metric theories of gravity which predict that they vanish. The α_{ij} and β_{ij} parameters appearing in (2) are determined by the anomalous mass tensors of the states involved in the atomic clock transitions considered above.

When the external gravitational potential is time independent, the coordinate energy function (3) is globally conserved and its dependence on atomic velocity and location determines the atom's coordinate acceleration via familiar energy conservation arguments,

$$a^i = \delta^{ij}\partial_j U + \frac{\delta m_{\mathrm{i}}^{ij}}{m}\partial_j U + \delta^{ij}\frac{\delta m_{\mathrm{g}kl}}{m}\partial_j U^{kl}(\boldsymbol{x}), \tag{4}$$

(here $\delta m_{\mathrm{i}}^{ij} = \delta m_{\mathrm{i}ij}$). Kenneth Nordtvedt [12] and Mark Haugan [13] exploit such arguments to relate the outcomes Eötvös experiments to the outcomes of gravitational redshift measurements and other tests of the EEP.

In the end, the preceding overview of the kind of analysis one must do to determine the outcomes of local nongravitational test experiments predicted by Lagrangian–based gravitation field theories brings one full circle. We have come back to the most familiar experimental tests of the EEP, but with a deeper appreciation of the significance of their results. In the next section we consider examples of nonmetric theories and formalisms encompassing whole classes of such theories within which preferred–frame and preferred–location effects have been analyzed to provide a basis for testing the EEP. Much of the work on tests of the EEP done before 1993 is thoroughly reviewed in the early chapters of Clifford Will's *Theory and Experiment in Gravitation Physics* [14], for an update see [15], see also [16].

2 Theoretical Frameworks for the Analysis of EEP Tests

The approach outlined in the preceding section can be used to determine preferred–frame and preferred–location effects predicted by any nonmetric theory of gravity. Such effects reflect the form of the theory's nongravitational Lagrangian density $\mathcal{L}_{ng}$ once the external gravitational potential in which a local nongravitational test experiment is performed has been plugged in. Proceeding in this fashion, we would have to analyze and reanalyze any given experiment to determine its outcome as predicted by competing theories.

It is, instead, more efficient to analyze local nongravitational test experiments once and for all within the context of more general theoretical frameworks that encompass broad classes of nonmetric gravitation theories and gravitational

environments. Such frameworks are based on models of the nongravitational Lagrangian or corresponding matter field equations that depend on phenomenological gravitational potentials in ways that encompass the forms of these structures in many nonmetric theories and environments. The outcome of an experiment predicted within such a framework immediately yields the outcome predicted by any nonmetric theory it encompasses when the framework's phenomenological gravitational potentials are expressed in terms of the particular theory's potentials. The efficiency of this approach is somewhat like that provided by the PPN formalism [14] when dealing with the gravitational dynamics of metric theories.

Analyses carried out in the general nonmetric frameworks discussed below have the additional benefit of identifying mechanisms that lead to preferred–frame or preferred–location effects in entire classes of nonmetric theories and of providing theory–independent parametrizations of such effects that are useful in discussing the results of experiments designed to search for them.

2.1 The $TH\epsilon\mu$–Formalism

The $TH\epsilon\mu$–formalism (see [17,14]) is based on the form of the Lagrangian governing the dynamics of point particles with mass m_i and charge q_i and of the electromagnetic field in a static, spherically symmetric background gravitational field described by the phenomenological gravitational potentials T, H, ϵ and μ:

$$L = -\sum_i m_i \int \sqrt{T - H\dot{x}_i^2}\, dt + \sum_i q_i \int A_a \dot{x}_i\, dx^a + \frac{1}{8\pi}\int \left(\epsilon \boldsymbol{E}^2 - \frac{1}{\mu}\boldsymbol{B}^2\right) d^4x. \tag{5}$$

A striking feature of this framework and the nonmetric theories it encompasses is that the limiting speed of massive particles in the neighborhood of some point in the gravitational field can differ from the speed of light there. These coordinate speeds are given, respectively, by the values of $\sqrt{T/H}$ and $1/\sqrt{\epsilon\mu}$ at the point of interest. Preferred–frame effects result when the ratio of these speeds is not unity. Variation of the relative values of T, H, ϵ and μ with position in the gravitational field can also lead to preferred–location effects. Computations of the energies of atomic states using natural quantum mechanical extensions of the classical $TH\epsilon\mu$ Lagrangian reveal both kinds of effects and yield predictions for anomalous inertial and gravitational mass tensors (compare Eq.(3)) [13,18,19]. A quantum field theoretic extension of the formalism reveals EEP violations discernable in measurements of the Lamb shift, the anomalous magnetic moment of the electron and related phenomena [21].

This test theory has been widely used to interpret the results of experimental tests of the EEP. For example, its predictions of the energies of atomic states [13], [18] and [19] have be used to interpret Hughes–Drever type experiments as well as the Vessot–Levine rocket redshift experiment [20,14].

Originally conceived as a framework for analyzing the physics of charged particles and electromagnetic fields in an external gravitational field, the $TH\epsilon\mu$–formalism has also been extended in a natural way to cover the other sectors of nongravitational physics comprising the Standard Model [22].

2.2 The χg–Formalism

Like the $TH\epsilon\mu$–formalism the χg–formalism introduced by W.–T. Ni [23] originally provided a framework for the analysis of electrodynamic physics in a background gravitational field and has subsequently been extended to cover other sectors of the Standard Model. Unlike the $TH\epsilon\mu$–formalism the χg–formalism is not restricted to static, spherically symmetric gravitational environments. The χ of its name refers to a tensor field appearing in the electromagnetic part of the nongravitational Lagrangian density upon which the formalism is based,

$$\mathcal{L}_{\rm em} = -\frac{1}{16\pi}\chi^{\alpha\beta\gamma\delta}F_{\alpha\beta}F_{\gamma\delta}. \tag{6}$$

The independent components of this tensor comprise twenty–one phenomenological gravitational potentials capable of representing gravitational fields in a very broad class of nonmetric gravitation theories.

The coupling of one particular phenomenological potential to the electromagnetic field is interesting because it can be expressed as a purely derivative coupling to a pseudoscalar field φ. A particle physicist would describe it as an axion coupling. A relativist would describe it as a coupling to axial torsion [24]. The Hojmann–Rosenbaum–Ryan–Shepley theory [25] is but one example of a theory encompassed by Ni's χg–formalism. It features a novel torsion coupling that has been shown to predict effects inconsistent with the results of experimental tests of the weak equivalence principle.

2.3 The Kostelecky Formalism

String theory has the potential to provide a quantum theory of gravity that is unified with other fundamental theories of matter and interactions. Recently Colladay and Kostelecky have introduced a framework for treating the possibility of spontaneous breakdown of Lorentz symmetry in the context of string theory [31,32]. While somewhat different from the sources of preferred–frame effects considered to this point, these string induced effects are considered here because they lead to modifications of the Dirac and Maxwell equations like those considered in the next subsection.

2.4 Formalisms Based on Matter–Field Equations of Motion

The effects an external gravitational field on the dynamics of matter fields can be dealt with at the level of equations of motion rather than Lagrangians. A formalism based on forms of the equations of motion has the advantage of directly addressing the following natural requirements one would demand of the dynamics of any quantum field (i) deterministic evolution, (ii) the superposition principle, (iii) a finite propagation speed (whose maximum value, since it need not be isotropic, we call c_D) and (iv) the conservation of probability.

The equations governing the motion of a Dirac field which satisfy these requirements are a first–order hyperbolic system of the form

$$0 = i\widetilde{\gamma}^{\mu}\partial_{\mu}\varphi + M\varphi\,, \tag{7}$$

or, in $3+1$–form ($\hat{\mu} = 1,2,3$)

$$i\partial_0\varphi = c_D\widetilde{\alpha}^{\hat{\mu}}c\partial_{\hat{\mu}}\varphi + c_D\widetilde{\Gamma}\varphi + mc_D^2\widetilde{\beta}\varphi \tag{8}$$

which we call a generalized Dirac equation ($\widetilde{\alpha}^{\hat{\mu}} = (\widetilde{\gamma}^0)^{-1}\widetilde{\gamma}^{\hat{\mu}}$). The matrices $\widetilde{\gamma}^{\mu}$ are not assumed to define a Clifford algebra, instead they satisfy $\widetilde{\gamma}^{\mu}\widetilde{\gamma}^{\nu} + \widetilde{\gamma}^{\nu}\widetilde{\gamma}^{\mu} = 2g^{\mu\nu} + X^{\mu\nu}$ where $g^{\mu\nu} = \frac{1}{4}\mathrm{tr}(\widetilde{\gamma}^{\mu}\widetilde{\gamma}^{\nu})$ and $X^{\mu\nu}$ is a matrix. In general, M is also a matrix. A distinctive feature of this generalized Dirac equation is that it predicts a splitting of the null cones and mass shells. (For another modification of the Dirac equation see [26] in this volume.)

Taking the non–relativistic limit and specifying a general position–dependence of the matrices $\widetilde{\gamma}^{\mu}$ and M, one derives the generalized Pauli equation [27]

$$\begin{aligned} i\frac{\partial}{\partial t}\varphi = &-\frac{1}{2m}\left(\delta^{ij} - \frac{\delta m_{\mathrm{i}}^{ij}}{m} - \frac{\delta\bar{m}_{\mathrm{i}k}^{ij}\sigma^k}{m}\right)\partial_i\partial_j\varphi + \left(c_D A_j^i + \frac{1}{m}a_j^i\right)\sigma^j i\partial_i\varphi \\ &+\Big[m\,U(\boldsymbol{x}) + \boldsymbol{C}\cdot\boldsymbol{\sigma}\,m\,U(\boldsymbol{x}) + \delta m_{\mathrm{g}ij}U^{ij}(\boldsymbol{x}) + c_D\,\boldsymbol{T}\cdot\boldsymbol{\sigma} + mc_D^2\boldsymbol{B}\cdot\boldsymbol{\sigma}\Big]\varphi \end{aligned} \tag{9}$$

where the anomalous coefficients $\delta m_{\mathrm{i}}^{ij}$, $\delta\bar{m}_{\mathrm{i}k}^{ij}$, A_j^i, a_j^i, $\boldsymbol{C}$, $\delta m_{\mathrm{i}}^{ij}$, $\boldsymbol{B}$ stem from those parts of the $\widetilde{\gamma}$–matrices which prevent them from defining a Clifford algebra and from anomalous terms in the mass matrix M, for example, $\frac{\delta m_{\mathrm{i}}^{ij}}{m} + \frac{\delta\bar{m}_{\mathrm{i}k}^{ij}}{m}\sigma^k = \frac{1}{2}(1+\beta)\widetilde{\alpha}^{(i}\widetilde{\alpha}^{j)}$ where β is the usual Dirac γ^0 and σ^i are the usual Pauli matrices. The generalized Pauli equation predicts preferred–frame and preferred–location effects. Terms like those representing couplings between spin and the Newtonian gravitational potential were first introduced in references [28,29] and [30].

The Pauli equation (9) is a generalization of Schroedinger equation provided by M . Haugan's approach [13] to the dynamics of scalar matter. As in the case of preceding formalisms, this Pauli equation provides a basis for broad range of experimental tests of the EEP, including experiments exploiting matter–wave interferometry. The classical limit of the generalized Pauli equation describes the free fall of classical spin–polarized bodies,

$$a^i = \delta^{ij}\partial_j U + \left[\frac{\delta m_{\mathrm{i}}^{ij}}{m} + 2\left(\frac{\delta\bar{m}_{\mathrm{i}k}^{ij}}{m} + \delta^{ij}C_k\right)S^k\right]\partial_j U + \delta^{ij}\,\frac{\delta m_{\mathrm{g}kl}}{m}\,\partial_j U^{kl}(\boldsymbol{x})\,. \tag{10}$$

Notice that not all of the anomalous parameters appearing in the quantum equation (9) survive in the classical freefall acceleration. Only by considering the evolution of the spin as well can one design experiments in the classical limit that are sensitive to all possible anomalies.

Once a generalized Dirac equation (8) is available we can address the dynamics of the electromagnetic field in an analogous way. The electromagnetic field

can be defined operationally by considering the phase shifts in charged particle interferometry. Assuming that the dynamics of electromagnetic fields satisfies the same requirements as we demanded for the dynamics of the Dirac field, this leads to generalized Maxwell equations of the form,

$$\partial_{[\mu}F_{\nu\rho]} = 0\,, \qquad 4\pi j^{\mu} = \lambda^{\mu\nu\rho\sigma}\partial_{\nu}F_{\rho\sigma} + \bar{\lambda}^{\mu\rho\sigma}F_{\rho\sigma}\,. \tag{11}$$

In the case of small deviations from minimal coupling to the Riemannian space–time metric $g_{\mu\nu}$, we have $\lambda^{\mu\nu\rho\sigma} = \delta^{\mu[\rho}g^{\sigma]\nu} + \delta\lambda^{\mu\nu\rho\sigma}$ with small values of $\delta\lambda^{\mu\nu\rho\sigma}$ and $\bar{\lambda}^{\mu\rho\sigma}$. Clearly, $\delta\lambda^{\mu\nu\rho\sigma}$ can induce anistropic propagation of light and birefringence. The $\bar{\lambda}^{\mu\rho\sigma}$ can also modify propagation, in some cases leading to a damping of electromagnetic waves.

The generalized Dirac equation (8) and Maxwell equations (11) can be used just as the corresponding equations that emerge from the $TH\epsilon\mu$–formalism or the χg–formalism, respectively, to analyze to properties of atoms in background gravitational fields. They do, however, encompass a wider range of nonmetric couplings that influence spin and polarization. Consequently, they provide the broadest possible basis for the interpretation of experimental tests of the EEP.

3 Motivations for Continued Testing of the EEP

Although all tests of the EEP, including some of remarkable precision, have so far failed to detect any hint of a violation, recent theoretical developments continue to suggest the EEP must be violated at some level. All approaches to quantizing gravity and to unifying it with the other fundamental interactions currently under study are capable of predicting such violations.

3.1 String Theory

Today, string theory is among the most promising candidate for a theory of quantum gravity fully unified with other fundamental interactions, and it has been shown to predict a variety of EEP violations.

For example, departures from universal free fall accelerations have been computed in references [33,34]. The composition–dependent component of test–body acceleration is estimated to be as large as a part in 10^{15} of the mean gravitational acceleration. The proposed STEP experiment could easily detect such an anomaly (see [6]).

String theory can also predict a time–varying fine structure constant because of couplings to scalar (dilatonic) fields, for example, see [35].

In the latest versions of string theory, physical particles and fields are confined to the neighborhoods of D–branes and their propagation may be affected by recoil of the branes caused by that propagation [36]. The effect of this recoil can be accounted for via an energy–dependent effective metric. This leads to modifications of the Maxwell equations,

$$\boldsymbol{\nabla}\cdot\boldsymbol{E} + \bar{\boldsymbol{u}}\cdot\partial_t\boldsymbol{E} = 0 \qquad \boldsymbol{\nabla}\cdot\boldsymbol{B} = 0 \tag{12}$$

$$\boldsymbol{\nabla}\times\boldsymbol{B} - (1-\bar{u}^2)\partial_t\boldsymbol{E} + \bar{\boldsymbol{u}}\times\partial_t\boldsymbol{B} + (\bar{\boldsymbol{u}}\cdot\boldsymbol{\nabla})\boldsymbol{E} = 0 \qquad \boldsymbol{\nabla}\times\boldsymbol{E} = -\partial_t\boldsymbol{B}\,, \tag{13}$$

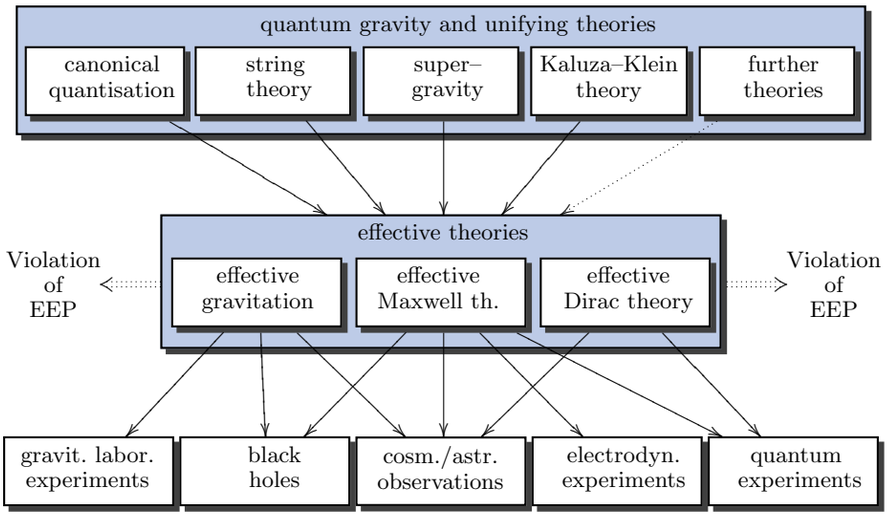


Fig. 1. Sources of violations of the EEP and classes of experiments or observations that are sensitive to them.

which predict dispersive light propagation. Analogous modifications of the Dirac equation account for the effect of brane recoil on the propagation of neutrinos and other fermions [37],

$$\gamma^a \partial_a \psi - m\psi + \gamma^0 (\bar{\boldsymbol{u}} \cdot \boldsymbol{\nabla})\psi = 0\,. \tag{14}$$

3.2 Loop Quantum Gravity

In the nonperturbative approach to quantum gravity based on observables analogous to Wilson loops, the semi–classical gravitational field is described via expectation values in so–called "weave"states. Gambini and Pullin [9] discuss the propagation of light through a gravitational field represented by a parity–violating weave state and find a polarization dependence of light propagation inconsistent with the EEP. The weave state is characterized by the length scale L, and gives rise to effective Maxwell equations [9] of the form

$$\partial_t \boldsymbol{E} = -\boldsymbol{\nabla} \times \boldsymbol{B} + 2\chi \ell_{\mathrm{P}} \boldsymbol{\nabla}^2 \boldsymbol{B} \tag{15}$$

$$\partial_t \boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{E} - 2\chi \ell_{\mathrm{P}} \boldsymbol{\nabla}^2 \boldsymbol{E}\,. \tag{16}$$

A corresponding effective Dirac equation [38] has the form

$$\left(i\widetilde{\gamma}^a \partial_a - \widetilde{m} + \widetilde{\gamma}^{ab} \partial_a \partial_b\right) \psi = 0 \tag{17}$$

with $\widetilde{\gamma}^a = \gamma^a + \kappa_1 \frac{\ell_{\mathrm{P}}}{L} G_1^a + \kappa_2 \left(\frac{\ell_{\mathrm{P}}}{L}\right)^2 G_2^a + \ldots$, $\widetilde{m} = m + \lambda_1 \frac{\ell_{\mathrm{P}}}{L} M_1 + \lambda_2 \left(\frac{\ell_{\mathrm{P}}}{L}\right)^2 M_2 + \ldots$, and $\widetilde{\gamma}^{ab} = \mu_1 \frac{\ell_{\mathrm{P}}}{L} G_1^{ab} + \mu_2 \left(\frac{\ell_{\mathrm{P}}}{L}\right)^2 G_2^{ab} + \ldots$, where γ^a are the usual Dirac γ–matrices,

m is the usual mass of the Dirac particle, G_i^a, M_i, and G_i^{ab} are arbitrary matrices, $\ell_{\rm P}$ is the Planck length and the coefficients κ_i, λ_i and μ_i are of the order unity $(i = 1, 2, \dots)$.

Since these field equations feature second–order spatial derivatives they are no longer hyperbolic and clearly single out a preferred frame. In addition, note that equation (16) modifies the homogeneous Maxwell equations, disrupting the relationship between field and 4–vector potential!

3.3 Gauge Theories of Gravity and Other Possibilities

Gauge theories of gravity like the Poincaré gauge theory [39] that leads to a Riemann–Cartan geometry, or the gauge theory of a linear group that leads to a metric–affine theory [40], gives rise to additional gravitational fields like torsion and, in the latter case, to nonmetricity. If these additional fields couple directly to matter, they can break local Lorentz invariance by singling out a preferred frame as well as breaking local position invariance.

In supergravity theories, which gauge the super–Poincaré group, torsion emerges as a bilinear combination of fundamental spin–$\frac{3}{2}$–field, see reference [41], for example.

Though invented long ago, Kaluza–Klein theories arise as a low–energy limits of string theory, with all that that entails regarding the validity of the EEP, see reference [42], for example.

Finally, we note that nonsymmetric theories of gravity, like those devised by John Moffat, have been shown to predict departures from universal free fall and violations of local Lorentz invariance in the electromagnetic sector [43] and [44].

4 Experimental and Observational Tests of the EEP

In principle, the outcomes of almost any experiment of observation conducted in different gravitational environments could yield evidence of the breakdown of the EEP. There are, however, certain classes of experiments and observations that are sensitive to characteristic violations of the EEP revealed by analyses within the theoretical frameworks discussed in section 2.

4.1 Tests of the Universality of Freefall

Tests with bulk matter. Experiments that search for composition–dependence of the freefall acceleration of macroscopic samples of matter are direct tests of the weak equivalence prinicple, one consequence of the EEP. It can be tested in traditional Eötvös fashion using torsion balance technology as in reference [5] or by monitoring the relative motion of freely falling bodies as in the Bremen drop tower experiment [45] and in the proposed MICROSCOPE [46] and STEP [6] space–based experiments. To date, the highest precision, of order 10^{-12}, has been achieved by torsion balance experiments, but the STEP experiment is designed to reach a precision of 10^{-18}.

Equation (10) shows that macroscopic samples of spin–polarized matter may experience different gravitational accelerations than unpolarized matter. Torsion balance experiments looking for such differential accelerations have been conducted by by Ritter, Gillies and coworkers [47,48]. They find no evidence of new spin–dependent forces.

Tests with quantum particles. As we saw in section 2, nonmetric theories of gravity can predict that quanta of different kinds fall with different accelerations in a gravitational field. Historically, there has been a great deal of interest in direct searches for such effects, especially in comparing the free fall acceleration of particles and antiparticles.

The first test of this type was performed by Witteborn and Fairbank [49] who tried to measure the gravitational acceleration of charged particles. A little later, Koester [50] showed that neutrons fall with the same way as classical bulk matter to an accuracy of a few percent. This result has been verified by means of neutron interferometry [51].

The potential for future matter interferometry tests of the EEP seems bright. Atomic interferometers have recently determined the gravitational acceleration toward the Earth to a part in 10^9 and yield results consistent with the measured acceleration of bulk matter. Refinements of these devices are expected to produce still more precise results and can be used to search for spin–dependent accelerations like those in (10).

4.2 Spectroscopic and Atomic Clock Tests of the EEP

As noted in preceding sections atoms are composites of the mass–energy of nucleons and electrons as well as of their electromagnetic and weak– and strong–interaction binding energies. Nonmetric theories whose matter–gravity couplings distinguish between these various contributions can cause not only gravitational accelerations that differ from atom to atom but also shifts in the energy spacings of atomic states that depend on an atom's velocity through or location in its gravitational environment. Spectroscopic and atomic clock experiments can search directly for these kinds of preferred–frame and preferred–location effects.

In Hughes–Drever–like experiments, for example, see Ref. [52], one searches for relative shifts between the frequencies of ground–state hyperfine transitions depending on atomic orientation in the gravitational environment. The interpretation of this type of experiment in the context of the $TH\epsilon\mu$–formalism is discussed in reference [14]. The interpretation in the context of the test theory of section 2.4 is discussed in reference [27].

Spectroscopic methods can also be used to search for the effects of spin–dependent and other EEP–violating effects predicted by equation (9), for example, see [53,27]. These techniques have been used to verify the spin–rotation coupling [54,55] in a search for anomalous spin–couplings [56].

Atomic clock technology is a particular, refined application of spectroscopic technique. Experiments that monitor the relative rates of different types of

atomic clock for dependence on the clocks' velocity through or location in a gravitational field provide another kind of spectroscopic test of the EEP. In essence, such tests are either tests of relativistic Doppler shifts that are sensitive to the parameters α_{ij} in (2) or tests of gravitational redshifts that are sensitive to the parameters β_{ij}. The difficulty of moving clocks through large changes in gravitational potential or at speeds approaching that of light limits the precision of such experiments. However, the Gravity Probe A experiment [57] succeeded in imposing the constraint $|\beta_{ij}| \leq 10^{-4}$. A recent experiment employing trapped Lithium atoms moving at 6.4% of the speed of light [58] was able to impose the constraint $|\alpha_{ij}| \leq 10^{-6}$. Atomic clock technology has also been used to constrain EEP–violating time–dependence of the fine–structure constant [59].

4.3 EEP Tests Involving Observations of Wave Propagation

Observations of the propagation of electromagnetic waves or other fields through a gravitational field are, in sense, a kind of experiment examing the effects of freefall. We discuss them separately, however, because of the distinctive way in which local effects that we think of as directly signalling violations of the EEP are allowed to build up as waves propagate over long distances.

Departures of the form of the Maxwell equations from their usual metric form induced by nonmetric couplings to gravity can lead to dispersive wave propagation or birefringence. Analogous departures of the Dirac equation from its usual metric form can also lead to dispersive propagation and make measurements of the arrival times of photons and neutrinos emitted from the same astrophysical event a test of EEP. The existence of very short duration events like supernova explosions and gamma ray bursts in combinations with the build up of gravitational delays over very great distances makes sharp tests possible.

Recently, limits on gravity–induced dispersion of electromagnetic wave propagation have been inferred from observations of quasars and gamma ray bursters [60]. They constrain

$$\frac{c_\gamma(\omega) - c_\gamma(\omega_0)}{c_\gamma(\omega_0)}\,, \quad \frac{c_\gamma^+ - c_\gamma^-}{c_\gamma^+}\,, \tag{18}$$

where $c_\gamma(\omega_0)$ is the velocity of the photon for a given frequency ω or polarization $\pm$. Exploiting rapid time variation of gamma ray bursters, Schaefer [60] is able to to impose sharp constraints on gravity–induced dispersion, $|(c_\gamma(\omega) - c_\gamma(\omega_0))/c_\gamma(\omega_0)| \leq 6 \times 10^{-21}$ for $\omega \sim 10^{18}$ Hz and $\omega_0 \sim 10^{19}$ Hz. See also [62] for implications of such data for quantum gravity models. Data constraining gravity–induced differences between the speed at which photons and neutrinos propagate are also imposed,

$$\frac{c_\gamma^\pm(\omega) - c_\nu^\pm(\omega)}{c_\gamma^\pm(\omega)}\,, \tag{19}$$

is also available.

Finally, observations that constrain differences between the speeds with which light with different polarizations propagates through a gravitational field have been analyzed [61] leading to the constraint $|(c_\gamma^+ - c_\gamma^-)/c_\gamma^+| \leq 10^{-28}$.

References

1. I. Newton: *Philosophiae Naturalis Principia Mathematica* (London 1686).
2. A. Einstein: Jahrb. Radioact. Elect. **4**, 411 (1907).
3. L. Maleki: SpaceTime Mission: Clock Test of Relativity at four Solar Radii, this volume.
4. R.V. Eötvös, V. Pekár and E. Fekete: Beiträge zum Gesetz der Proportionalität von Trägheit und Gravität, *Ann. Physik* **68**, 11 (1922).
5. Y. Su, B.R. Heckel, E.G. Adelberger, J.H. Gundlach, M. Harris, G.L. Smith, and H.E. Swanson: New test of the universality of free fall, *Phys. Rev.* **D 50**, 3614 (1994).
6. N. Lockerbie, J. Mester, R. Torii, S. Vitale, and P. Worden: STEP: A Status Report, this volume, p.213.
7. A. Einstein: *Preuss. Akad. Wiss Berlin Sitzber.* 688 (1916).
8. J. Norton: General covariance and the foundations of general relativity: eight decades of dispute, *Rep. Prog. Phys.* **56**, 791 (1993).
9. R. Gambini and J. Pullin: Nonstandard optics from quantum space–time, *Phys. Rev.* **D 59**, 124021 (1999).
10. J. Ellis, N.E. Mavromatos and D.V. Nanopoulos: Probing models of quantum space–time foam, preprint gr-qc/9909085.
11. M.P. Haugan and C. Lämmerzahl: On the experimental foundations of the Maxwell equations, *Ann. Phys.* (Leipzig) **9**, 119 (2000).
12. K. Nordtvedt: Qualitative relationship between clocks gravitational "red–shift" violatings and nonuniversality of free–fall rates in nonmetric theories of gravity, *Phys. Rev.* **D 11**, 245 (1975).
13. M.P. Haugan: Energy Conservation and the Principle of Equivalence, *Ann. Phys.* (N.Y.) **118**, 156 (1979).
14. C.M. Will: *Theory and Experiment in Gravitation Physics*, revised edition (Cambridge University Press, Cambridge 1993).
15. C.M. Will: The Confrontation between General Relativity and Experiment: A 1998 Update (Lecture notes from the 1998 SLAC Summer Institute on Particle Physics), gr-qc/9811036.
16. M.P. Haugan and C.M. Will: Modern tests of special relativity, *Physics Today*, May 1987, p. 69.
17. A.P. Lightman and D.L. Lee: Restricted Proof that the Weak Equivalence Principle Implies the Einstein Equivalence Principle, *Phys. Rev.* **D 8**, 364 (1973).
18. C.M. Will: Gravitational red–shift measurements as tests of nonmetric theories of gravity, *Phys. Rev.* **D 10**, 2330 (1974).
19. M.D. Gabriel and M.P. Haugan: Testing the Einstein Equivalence Principle: Atomic clocks and local Lorentz invariance, *Phys. Rev.* **D 41**, 2943 (1990).
20. J.P. Turneaure, C.M. Will, B.F. Farrel, E.M. Mattison, and R.F.C. Vessot: Test of the principle of equivalence by a null gravitational red–shift experiment, *Phys. Rev.* **27**, 1705 (1983).
21. C. Alvarez and R.B. Mann: Testing the Equivalence Principle in the Quantum Regime, preprint (1996), Honorable mention in the Gravity Research Foundation Essay Contest (and references cited therein).
22. J.E. Horvath, E.A. Logiudice, C. Riveros, and H. Vucetich: Einstein equivalence principle and theories of gravitation: A gravitationally modified standard model, *Phys. Rev.* **D 38**, 1754 (1988).

23. W.–T. Ni: Equivalence Principles and Electromagnetism, *Phys. Rev. Lett.* **38**, 301 (1977); *Bull. Am. Phys. Soc.* **19**, 655 (1974); A Nonmetric Theory of Gravity, preprint, Montana State University, Bozeman, Montana, USA (1973), http://gravity5.phys.nthu.edu.tw.
24. R.A. Puntigam, C. Lämmerzahl, and F.W. Hehl: Maxwell's theory on a post–Riemannian spacetime and the equivalence principle, *Class. Qaunt. Grav.* **14**, 1347 (1997).
25. S. Hojman, M.P. Rosenbaum, and L.C. Shepley: Gauge invariance, minimal coupling, and torsion, *Phys. Rev.* **D 17**, 3141 (1978).
26. C. Lämmerzahl, Ch.J. Bordé: Testing the Dirac equation, this volume, p. 463
27. C. Lämmerzahl: Quantum Tests of Foundations of General Relativity, *Class. Quantum Grav.* **14**, 13 (1998).
28. N.D. Hari Dass: Test for C, P, and T Nonconservation in Gravitaton, *Phys. Rev. Lett.* **36**, 393 (1976).
29. N.D. Hari Dass: Experimental Tests for Some Quantum Effects in Gravitation, *Ann. Physics (N.Y.)* **107**, 337 (1977).
30. A. Peres: Test of the equivalence principle with spin, *Phys. Rev.* **D 18**, 2739 (1978).
31. D. Colloday and V.A. Kostelecky: CPT–violation and the standard model, *Phys. Rev.* **D 55**, 6760 (1997).
32. D. Colloday and V.A. Kostelecky: Lorentz–violating extension of the standard model, *Phys. Rev.* **D 55**, 6760 (1997).
33. T. Damour and A.M. Polyakov: The string dilaton and a least action principle, *Nucl. Physics* **B 423**, 532 (1994).
34. T. Damour and A.M. Polyakov: String Theory and Gravity, *Gen. Rel. Grav.* **12**, 1171 (1996).
35. T. Damour: Equivalence Principle and Clocks, to appear in the *Proceedings of the 34th Rencontres de Moriond, "Gravitational Waves and Experimental Gravity"*, January 1999, gr-qc/9904032.
36. J. Ellis, N.E. Mavromatos, and D.V. Nanopoulos: Probing models of quantum space–time foam, gr-qc/9909085.
37. J. Ellis, N.E. Mavromatos, D.V. Nanopoulos, and G. Volkov: Gravitational–Recoil Effects on Fermion Propagation in Space-Time Foam, gr-qc/9911055.
38. J. Alfaro, H.A. Morales–Tecotl, and L.F. Urrutia: Quantum gravity corrections to neutrino propagation, *Phys. Rev. Lett.* **84**, to appear (2000).
39. F.W. Hehl, P. von der Heyde, G.D. Kerlick, and J.M. Nester: General relativity with spin and torsion: Foundations and prospects, *Rev. Mod. Phys.* **48**, 393 (1976).
40. F.W. Hehl, J.D. McCrea, E.W. Mielke, and Y. Ne'eman: Metric–affine gauge theory of gravity: Field Equations, Noether Identities, World Spinors, and Breaking of Dilation Invariance, *Phys. Rep.* **258**, 1 (1995).
41. P. van Nieuwenhuizen: Supergravity, *Phys. Rep.* **68**, 189 (1982).
42. M.J. Duff: Kaluza–Klein theory in perspective, hep-th/9410046.
43. C.M. Will, C.M.: Violation of the Weak Equivalence Principle in Theories of Gravity with a Nonsymmetric Metric, *Phys. Rev. Lett.* **62**, 369 (1989).
44. M.D. Gabriel, M.P. Haugan, R.B. Mann, and J.H. Palmer: Nonsymmetric gravitation theories and local Lorentz invariance, *Phys. Rev.* **D 91**, 2465 (1991).
45. W. Vodel, H. Dittus, S. Nietzsche, H. Koch, J. v. Zameck Glyscinski, R. Neubert, S. Lochmann, C. Mehls, D. Lockowandt: High Sensitive DC SQUID Based Position Detectors for Application in Gravitational Experiments at the Drop Tower Bremen, this volume.
46. P. Touboul: Space Accelerometers Present Status, this volume, p.273.

47. R.C. Ritter, C.E. Goldblum, W.-T. Ni, G.T. Gillies, and C.C. Speake: Experimental test of equivalence principle with polarized masses, *Phys. Rev.* **D 42**, 977 (1990).
48. R.C. Ritter, L.I. Winkler, and G.T. Gillies: Search for Anomalous Spin–Dependent Forces with a Polarized–Mass Torsion Pendulum, *Phys. Rev. Lett.* **70**, 701 (1993).
49. F.C. Witteborn and W.M. Fairbank: Experimental Comparison of the Graviational Force on Freely Falling Electrons and Metallic Electrons, *Phys. Rev. Lett.* **19**, 1049 (1967).
50. L. Koester: Verification of the equivalence principle of gravitational and inertial mass for the neutron, *Phys. Rev.* **D 14**, 907 (1976).
51. J.–L. Staudenmann, S.A. Werner, R. Colella, A.W. Overhauser: Gravity and Inertia in Quantum Mechanics, *Phys. Rev.* **A 21** 1419 (1980).
52. T.E. Chupp, R.J. Hoara, R.A. Loveman, E.R. Oteiza, J.M. Richardson, and M.E. Wagshul: Results of a New Test of Local Lorentz Invariance: A Search for Mass Anisotropy in ^{21}Ne, *Phys. Rev. Lett.* **63**, 1541 (1989).
53. C. Lämmerzahl: Constraints on space–time torsion from Hughes–Drever experiments, *Phys. Lett.* **A 228**, 223 (1997).
54. B. Mashhoon: Neutron Interferometry in a Rotating Frame of Reference, *Phys. Rev. Lett.* **61**, 2639 (1988).
55. B. Mashhoon: On the coupling of intrinsic spin with the rotation of the earth, *Phys. Lett.* **A 198**, 9 (1995).
56. B.J. Venema, P.K. Majumder, S.K. Lamoreaux, B.R. Heckel, and E.N. Fortson: Search for a Coupling of the Earth's Gravitational Field to Nuclear Spins in Atomic Mercury, *Phys. Rev. Lett.* **68**, 135 (1992).
57. R.F.C. Vessot, M.W. Levine, E.M. Mattison, E.L. Blomberg, T.E. Hoffmann, G.U. Nystrom, B.F. Farrel, R. Decher, P.B. Eby, C.R. Baughter, J.W. Watts, D.L. Teuber, and F.D. Wills: Test of Relativistic Gravitation with a Space–Borne Hydrogen Maser, *Phys. Rev. Lett.* **45**, 2081 (1980).
58. R. Grieser, R. Klein, G. Huber, S. Dickopf, I. Klaft, P. Knobloch, P. Merz, F. Albrecht, M. Grieser, D. Habs, D. Schwalm, and T. Kühl: A test of special relativity with stored lithium ions, *Appl. Phys.* **B 59**, 127 (1994).
59. J.D. Prestage, R.L. Tjoelker, and L. Maleki: Atomic clocks and variations of the fine structure constant, *Phys. Rev. Lett.* **74**, 3511 (1995).
60. B.E. Schaefer: Severe limits on variations of the speed of light with frequency, *Phys. Rev. Lett.* **82**, 4964 (1999).
61. M.P. Haugan and T.F. Kauffmann: A New Test of the Einstein Equivalence Principle and the Isotropy of Space, *Phys. Rev.* **D 52**, 3168 (1995).
62. S.D. Billet *et al.*: Limits to quantum gravity effects on energy dependence of the speed of light from observations of TeV flares in active galaxies, *Phys. Rev. Lett.* **83**, 2108 (1999).

STEP: A Status Report

Nick Lockerbie[1], John C. Mester[2], Rodney Torii[2],
Stefano Vitale[3], and Paul W. Worden[2]

[1] Department of Physics and Applied Physics, University of Strathclyde, Glasgow G4 0NG, UK
[2] W.W. Hansen Experim. Physics Labs, Stanford University, Stanford, California 94305, USA
[3] Department of Physics, Universitá Trento, 38050 Provo, Italy

Abstract. This paper presents an overview of the current technical status of STEP, the Satellite Test of the Equivalence Principle. STEP was originally presented as a candidate for ESA's M2 mission as a joint mission with NASA, and has since been studied as an M3 candidate, and under NASA as QuickSTEP and MiniSTEP. Studies especially during the last two years have resolved some long standing issues such as control of helium tide, improved the mission definition and error analysis, and have resulted in an improved baseline design which should be capable of comparing rates of fall to an accuracy approaching 10^{-18}.

1 Background

When two bodies are allowed to fall freely in a uniform gravitational field, they fall with the same acceleration regardless of their composition. This perplexing observation, known as the Universality of Free Fall, is both consequence and proof of the Equivalence Principle (EP), the fundamental concept that inertia and gravitational mass are the same physical property. This principle has been confirmed in both ground–based and lunar–ranging experiments with an accuracy of about a part in 10^{12}. The Satellite Test of the Equivalence Principle (STEP) is intended to test the EP in earth orbit to an accuracy of 1 part in 10^{18}. Rarely is it possible in science to advance in one leap the testing of an already well–established principle by six orders of magnitude, and do so in a range that is scientifically significant.

The Universality of Free Fall goes back in essence to Galileo's observation that two bodies of different composition dropped from the Leaning Tower of Pisa fall with the same acceleration. The deep significance of this discovery was first demonstrated by Newton. He distinguished two properties of a body, its 'weight' and the 'quantity of matter' in it, or as we would say, its gravitational mass m_g and inertial mass m_i, and he demonstrated that Universality of Free Fall makes the ratio m_g/m_i identical for all bodies regardless of composition. Even more importantly, he emphasized that it is exactly this property that marks off gravity from all the other forces in nature. Specifically, he pointed out that nothing like it applies to magnetism.

The term EP was coined by Einstein in 1907 in his famous 'falling elevator' argument. For Einstein, Equivalence was a more far–reaching concept than sim-

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 213–247, 2001.

ply the Universality of Free Fall. His conjecture was that not just falling bodies but all the laws of physics respond in the same way to either a gravitational acceleration or an inertial acceleration. He made the EP the basis for his idea that gravity is a consequence of the geometry of spacetime. Equivalence is therefore both the historical foundation and the best–established fact of General Relativity. Hence any breakdown of Universality of Free Fall would be in fundamental contradiction with Einstein's theory.

General Relativity provides the theoretical basis for our present description of the macroscopic world: the Big Bang, the cosmological expansion, the large–scale structure of the Universe, the end points of stellar evolution, gravitational collapse, neutron stars, black holes, gravitational waves, relativistic celestial mechanics of the solar system, the high–precision description of the motion of natural and artificial satellites, the definition of the international atomic time, etc. Nevertheless, General Relativity faces two obvious problems: it leaves the gravitational interaction apparently disconnected from the other interactions in nature, and it resists all traditional attempts at quantization. Historically, numerous alternative theories of gravity, have been proposed, including the Kaluza–Klein–Jordan theory, Brans–Dicke theory, Fierz–Brans–Dicke theory, and others, all of which predict violations of Equivalence at some level. Modern theories, including the Standard Model and string theory, either have internal deficiencies which require new fields to resolve —leading to a violation of Equivalence — or predict new fields such as the dilaton of the order of the strength of gravity, with the same result. Damour and Polyakov [1] devised a mechanism in which the dilaton field is attracted toward values that nearly decouple from matter, resulting in the prediction of an EP violation in the range between 10^{-14} and 10^{-23}, much of which is accessible to STEP.

Historically, there have been four important methods of testing Equivalence: (1) Galileo's free fall method, (2) Newton's pendulum method, (3) Newton's celestial method (using observations either of the earth–moon system or of the moons of Jupiter in the sun's field), (4) Eötvös's torsion–balance method. Curiously, the pendulum test, originally the most precise, has been the least satisfactory. The best was H.H. Potter's in 1926, who obtained an accuracy of a part in 10^6. Modern free fall experiments by groups from the US, Italy, and Japan give results between 1 and 3 parts in 10^{10}. Much the most accurate tests have been by torsion–balance experiments and the celestial method. The best torsion–balance results to date by Adelberger et al. have reached 2.4 parts in 10^{12} for a variety of test bodies. The celestial method was refined by Laplace in 1798 and then dramatically reinvented in the 1960s in the context of lunar laser–ranging. Its current limit reported by Dickey et al. is 5 parts in 10^{13}.

STEP returns to what is essentially Galileo's free fall method, without the sudden stop at the end, rather than attempting to orbit a torsion balance. The reason for this choice is in the disturbing effect of gravity gradients on the experiment. Ever since the work of Roll, Krotkov, and Dicke (and to some extent since Eötvös) it has been recognized that the two dominant practical limitations on torsion–balance experiments are seismic disturbance and the effects of grav-

ity gradients on the mass moments of the balance. In STEP, the seismic effects are greatly reduced in comparison with ground–based experiments through the use of a drag–free satellite. The gravity gradient effects remain; with a torsion balance, they would overwhelm the experiment. Already in the ground–based tests the experimenters have had to summon up very considerable ingenuity to reduce gravity gradient disturbances. To advance the precision six more orders of magnitude another approach has to be found.

The approach used in STEP depends on recognizing that gravity gradient effects enter the experiment in two ways, disturbances originating in the spacecraft and disturbances originating in gradients from the earth's gravity. In STEP it is possible to exploit the earth–gradient effects to our advantage. With two nearly concentric test masses in free fall, the earth–gradients produce a difference in acceleration, doubly periodic in the orbit, from which the displacement between the mass centers can be determined. The masses can be centered on each other very accurately using this acceleration as an error signal. The differential acceleration measurement is then insensitive to the earth's gradient, and even more important, it is insensitive to first–order gradient disturbances from the spacecraft. One disturbing gravity gradient effect has been used to eliminate both.

2 STEP Concepts

The simplest concept of STEP is Galileo's leaning tower experiment, with the difference that the test masses fall all the way around the Earth (Fig.1). If there is a difference in acceleration, the masses will go into separate orbits, and over short times they will separate in proportion to the square of the time of fall. For times of the order of one orbit or more, they tend to return periodically to their initial positions, with an added secular drift. The net gain is as if the experiment were done from a tower several thousand kilometers high, with the added advantages that the experiment repeats periodically and is performed in a relatively benign environment.

A practical difficulty with a tower experiment is releasing the masses without disturbance, at exactly the same height and exactly the same instant. This is overcome in STEP by observing the difference in the masses' motion. If the masses do not start from exactly the same initial conditions, they will likewise follow different orbits, but this difference is readily distinguishable from an EP violation by being at twice the frequency. This gravity gradient acceleration is used to precisely center the masses on one another, as described below. The STEP test masses are made in the shape of hollow cylinders so that one fits inside the other. This permits their centers of gravity to be precisely aligned, eliminating first order disturbances from gravity gradients which would result from miscentering.

Ambiguities in the measurement would result if the masses were allowed to move freely in all degrees of freedom. In a practical experiment the masses are not completely free, but are constrained to move along a line, the sensitive direc-

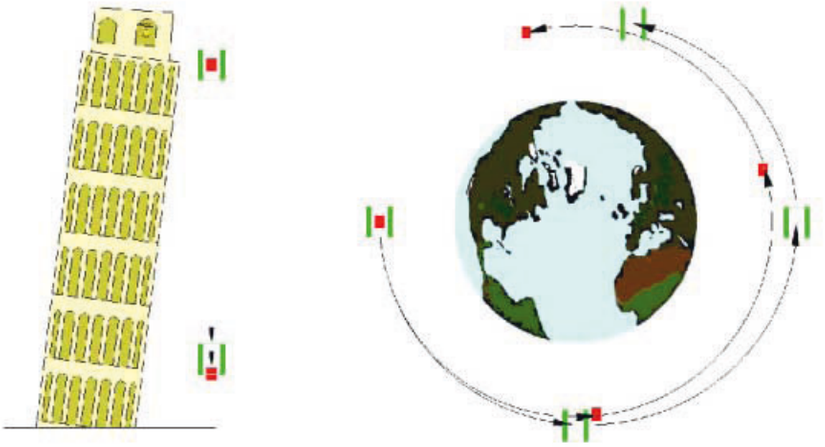


Fig. 1. Concept of STEP

tion, which coincides with the cylinder axis. Superconducting magnetic bearings with exceptional stability and extremely low losses provide the constraint mechanism. The only constraint along the sensitive direction comes from the position measurement. This gives the masses periods of the order of 1000 seconds, slightly reducing the amplitude of motion expected from a violation.

The position measurement is made by SQUID magnetometers which precisely measure currents from sensing coils near each test mass. The sensor makes a differential measurement between pairs of masses which is not much disturbed by satellite motion. Radial and angular vibrational modes of the masses are monitored by supplemental electrodes surrounding each test mass, which are also used to manipulate the masses by small applied voltages, and to measure their electrical charge.

Air drag would seriously degrade the STEP measurement (as in Galileo's experiment) if it were not compensated by a "drag–free" controller for the satellite. The motion of the satellite in response to drag is measured relative to the test masses, which are protected from the external drag by the satellite itself. A servomechanism fires small linear thrusters to correct the motion, cancelling the effect of the external drag or other disturbance. Reaction mass comes from boiloff of the liquid helium used to refrigerate the experiment, which would otherwise be yet another disturbance.

In orbit most disturbances tend to recur at the frequency of the orbit and its harmonics. The characteristic signature of a violation occurs at orbit frequency unless the satellite is rotated. In a rotating satellite, the frequency of the violation

signal is at the difference frequency between the orbit and rotation frequency of the satellite. STEP uses this fact to separate any EP violation from most external disturbances. Some disturbances may track the signal because they also depend on this difference frequency. We have prepared a program of measurements which allows identification of these systematic disturbances by measuring at different rotation frequencies and with varied experimental conditions. Details of some of these measurement procedures are below.

3 STEP Instrument Configuration

The STEP instrument comprises four differential accelerometers, which will be operated simultaneously to maximize the quantity and quality of the data returned. These accelerometers operate at a nominal temperature of approximately 2 K. Each accelerometer contains two cylindrical test masses which are constrained from moving radially by superconducting magnetic bearings that leave them free to move along the axis of the instrument. Motion along the cylinder axis is measured precisely by SQUID magnetometers, and a capacitance pickoff measures motion in all degrees of freedom at lower resolution. Magnetic force from the SQUID sensing coils constrains the masses in the axial direction. The SQUID outputs measure both differential mode acceleration and common mode acceleration along the cylinder axis. The masses can be manipulated by voltages applied to capacitance sensor electrodes surrounding them. The charge on the masses can be estimated from the response of the masses to these voltages, and it can be controlled by an ultraviolet discharging mechanism. The masses will be caged during launch by a hydraulic actuator, similar to the GP-B caging mechanism, using pressurized liquid helium as the working fluid.

Each differential accelerometer (Fig.2) contains two concentric test masses with cylindrical symmetry, each coated with a thin film of superconductor, which are independently constrained by superconducting magnetic bearings and which move freely along the symmetry axis. The "belted–cylinder" shapes of the test masses are picked to minimize higher order gravitational coupling to possibly moving masses in the satellite; for example each mass has the same moment of inertia about any axis. This makes the masses respond to gravity like a sphere to sixth order.

The SQUID displacement sensors will be able to detect displacements corresponding to about 3×10^{-19} g of differential acceleration in an averaging time of about 20 orbits. SQUIDs which measure the common mode will be sensitive to about 1×10^{-18} g over the same time span. Additional details of the SQUIDs are below.

The test masses must be prevented from moving in other than the sensitive direction to avoid ambiguity in the acceleration being measured; the sense coils necessarily have some response to motion in other modes than along the cylinder axis. Superconducting magnetic bearings perform this function. A meander pattern niobium circuit on a quartz cylinder provides the maximum radial constraint of each test mass with minimum force along the sensitive axis. This

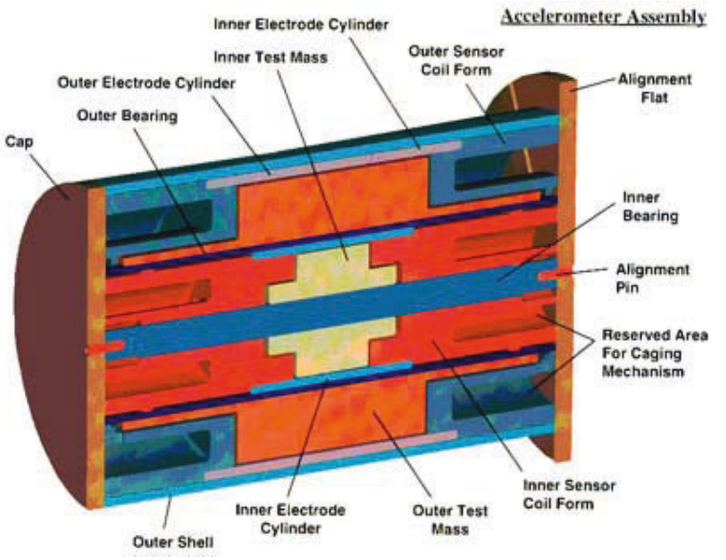


Fig. 2. STEP Accelerometer

type of circuit produces stiff radical force on the test mass with no net magnetic moment to disturb the SQUID. These circuits are divided in quadrants to allow adjustment of the test masses' center of mass positions during setup.

The accelerometer configuration (Fig.3) is best described by reference to spacecraft coordinates which have X and Y axes in the plane of the orbit and the Z axis roughly toward the Sun. The accelerometers have their sensitive axes alternately along the X and Y directions, and are stacked along the Z axis. They will be separated by about 15 cm in the Z direction. A quartz block supports and aligns the accelerometers and associated parts in this configuration. This configuration allows all the accelerometers to be centered on the roll axis of the satellite which is normally horizontal and perpendicular to the in–track direction.

In this attitude (Fig.4) the accelerometers can be operated simultaneously with minimum disturbance from the Earth's gravity gradient. Attitude errors will be estimated by using the common mode of the accelerometers to measure the Earth's gradient, and used to orient the satellite. This ensures that any disturbance from the Earth's gradient will be minimized.

The STEP accelerometers are contained in a dewar, part of a "drag–free" satellite (Fig.5) which completely surrounds the entire instrument, protecting it from disturbances such as air drag, magnetic field, and solar pressure. The satellite will follow the masses by using linear thrusters which compensate for the drag and cause the satellite to precisely follow the masses. The satellite will include a radiation sensor to improve the estimate of the masses' charging rate, an external magnetic field sensor, electronics to control the payload, a superfluid

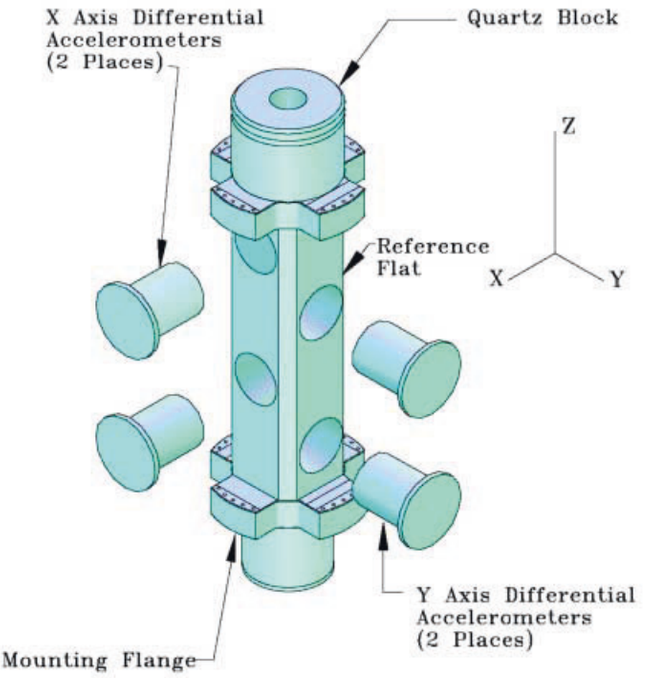


Fig. 3. Quartz Block Concept

helium dewar incorporating a superconducting magnetic shield and ultrahigh vacuum chamber that surrounds the instrument, communications equipment, power supply, computer, and other support equipment.

The essential requirement of the design is to eliminate all disturbances that could imitate the signal of a difference in rate of fall. The inherent stability and low losses in superconductors reduce the intrinsic sensor noise. Superconducting shielding eliminates external magnetic and electrical disturbances. Likewise the very low temperature ($\sim$ 2 K), temperature stability and uniformity ($<$ 0.5 mK per orbit), and pressure ($< 10^{-14}$ torr) reduce most thermal disturbances to insignificant levels. The shapes of the test masses are chosen to minimize disturbing gravity forces from the satellite, and the masses will be centered within 10^{-10} m to remove the effect of the Earth's gravity gradient. A charge control system reduces disturbances from particle radiation and the resultant electrical charging.

Active drag compensation ("Drag Free Control") is required to reduce satellite acceleration by seven orders of magnitude within the measurement bandwidth. The helium gas boiled off from the cryostat, which is otherwise a disturbance, is used as the propellant. This gas is vented through continuously variable thrusters developed for the GP-B program. Temperature control of the dewar is

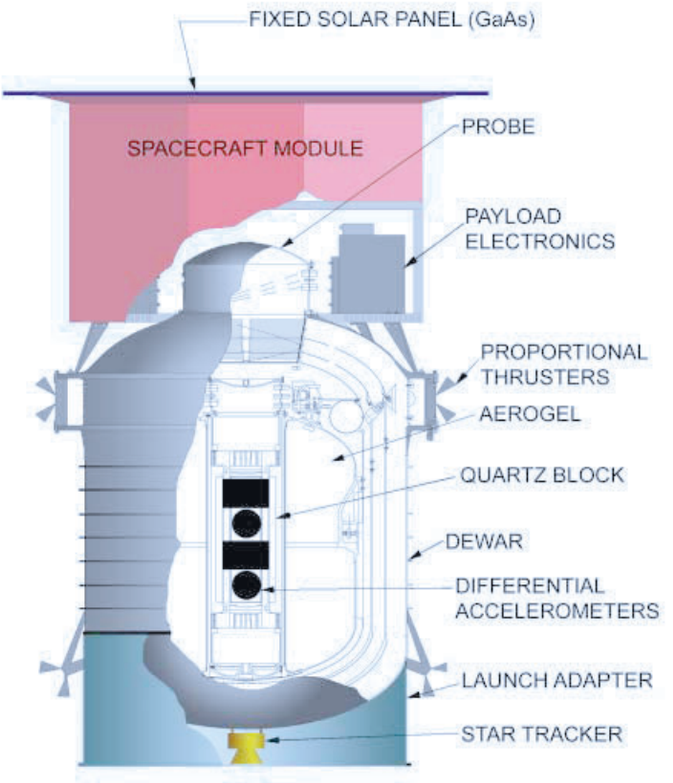


Fig. 4. STEP Satellite Concept

also maintained by the drag free system which will "null–dump" excess helium for cooling without changing the net thrust. A sun–synchronous orbit, about 500 km high, minimizes thermal changes to the satellite.

To distinguish a EP violation from most remaining disturbances, the satellite must rotate about an axis (the spacecraft Z–axis) normal to a plane parallel to the sensitive directions (or cylinder axis) of the accelerometers (Fig.5). The rotation axis must be horizontal to prevent the accelerometers being disturbed by common–mode gravity gradient accelerations from the Earth. The EP signal appears at the difference in frequency between the orbit frequency and the rotation frequency. The rotation frequency will be changed between measurements to move the signal frequency away from fixed frequency disturbances. The satellite's rotation axis and attitude will be chosen and controlled to minimize disturbances from the Earth's gravity gradients, by observation of the test masses. Experimental parameters, for example the electric charge on the test

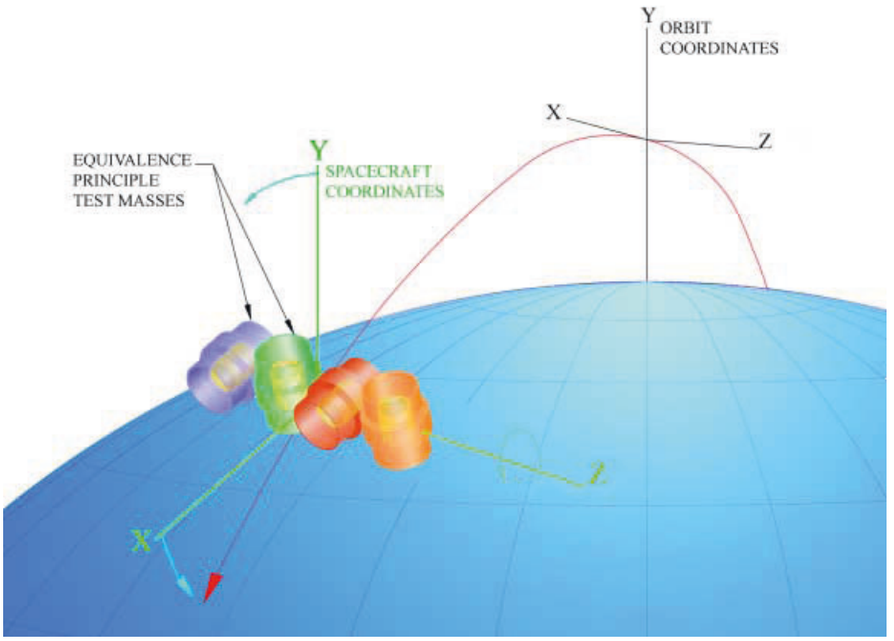


Fig. 5. STEP Orientation and Coordinate Systems

mass, or the center of mass displacement, can be varied to distinguish a possible violation from systematic disturbances.

Test Mass Choice

Extensive discussions have led to a baseline design choice of three materials, Pt-Ir, Nb, Be, in a cyclic condition, with the Pt-Ir/Be pair duplicated in an accelerometer with somewhat different design. This choice presumes that sensitive comparisons of two materials will be possible only within a single differential accelerometer, and that materials in different accelerometers cannot be directly compared. The choice of test mass materials was made to maximize information return. There is a tradeoff between two competing ideas, first to test as many different materials as possible, so as not to miss anything, and second to make certain that anything measured is a real effect. The choice is complicated by issues of manufacturability, that is, there are only a limited number of scientifically interesting materials which can be made into a test mass. Uncertainty about what might cause an EP violation suggests testing a multiplicity of materials. From this viewpoint a number of materials should be chosen from the low, middle and high portions of the periodic table. Though theoretical ideas are at present barely suggestive, it is expected that those properties which might lead to an EP violation will vary more or less smoothly across the periodic table; for

example mass deficit or neutron number. It is probably not too critical which ones are chosen, so long as they have linearly independent properties.

Testing whether a measured result is real or spurious involves both repetition of the experiment (to test whether the result is accidental) and variation of experimental conditions (to test that the result is not a systematic effect). Ideally one would like to test the same materials under many different conditions. In that case, if the result is unchanged, it is clear that the differing conditions have nothing to do with the outcome. This suggests testing a minimum number of materials very carefully. An important verification technique to use with multiple materials is the cyclic condition, in which the total acceleration difference between three pairs of masses AB, BC, CA must add to zero. The baseline choice shown in Table 1 incorporates a cyclic condition and the duplication of a pair of materials with a different accelerometer design. This baseline test–mass scheme

Table 1. STEP Satellite Roll Rates

Inner mass material	density (g/cm^3)	Outer mass material	density (g/cm^3)	mass ratio
Pt/Ir	21.68	Nb	9.57	3.97
Nb	9.57	Be	1.84	1.73
Pt/Ir	21.68	Be	1.84	0.76
Pt/Ir	21.68	Be	1.84	0.76

incorporates two model–independent validity checks: 1) ab, bc, ac; and 2) ab, (ab)'. The prime indicates that this accelerometer will include certain variations in experimental parameters as a further check for systematic effects. For example the ab and (ab)' accelerometers will be orientated at 90° relative to each other and will be physically separated by the length of the quartz block support structure. Radial and longitudinal spring constants will be made non–identical by setting bearing and readout circuit currents. Similarly, electrostatic position sensing voltages and electrode gap spacings will also differ.

We have verified through discussions with Speedring Inc. and BIPM that the three chosen materials can be machined to sub–micron tolerances. Their cryogenic properties are known sufficiently well to enable mechanical design and thermal analysis to proceed. We have conducted preliminary studies of their Nb thin film coating properties and foresee no obstacles in this regard.

3.1 Test Mass Shapes and Configuration

To ensure that non–uniform gravitational fields do not cause differences in acceleration of the test masses, their centers of mass must coincide, and their shapes and dimensions must be chosen so as to minimize the coupling of higher order gravitational gradients to each mass. This approach greatly reduces spurious

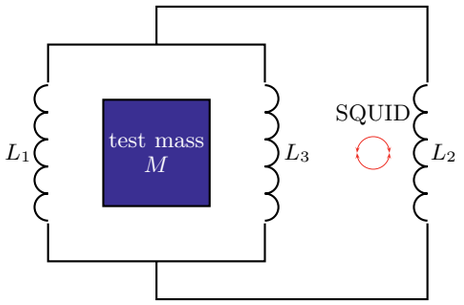


Fig. 6. Basic Position Sensing Circuit

gravitational effects due to helium tides in spacecraft's dewar, deformation of the spacecraft, etc. An ideal shape for the masses would be spherical, but this gives no good way to measure the inner mass, and although spheres can be very accurately made, any gravitational advantage is most likely nullified by density variations. The STEP test masses will be belted cylinders with their dimensions chosen to make the masses act as much like spheres as possible gravitationally. Where it is not possible to give the masses zero multipole moments, corresponding moments of inner and outer masses will be made identical. The shapes of the 2 masses in each pair separately make the quadrupole, hexadecapole, and 64–pole terms close to zero, while minimising the 256–pole gravitational coupling term [4]. Therefore the masses 'look like' gravitational monopoles to a few ppm for gravitational sources as close as 250 mm away [5].

3.2 Differential Accelerometer Operation

The differential accelerometers each contain a pair of test masses constrained by magnetic springs. The differential displacement along their common 'sensitive' axis is measured to very high precision ($\approx 7 \times 10^{-14}$ m in 10^5 s integration time) using superconducting pick–up coils connected to SQUID detectors. The differential displacement is therefore a measure of the difference in acceleration experienced by the masses in each pair. Each mass is guided along the sensitive axis by a superconducting magnetic bearing. Its equilibrium position may be trimmed by adjustment of supercurrents in the bearings and SQUID coils, or by a voltage applied to surrounding electrodes.

SQUIDs (Superconducting Quantum Interference Devices) are chosen for STEP because of their extremely high sensitivity and low intrinsic noise (approx. 2.6×10^{-28} J/Hz at the signal frequency). An ordinary non–differential accelerometer is made by placing one test mass between two superconducting pickup coils wired in series, with a persistent supercurrent flowing around the circuit (Fig.6). This circuit is the superconducting equivalent of a common device, the linear variable displacement transformer (LVDT).

Movement of the superconductor–coated test mass towards one pickup coil, and away from the other one, changes the inductance values. Because of flux conservation, this forces a current through a third coil in parallel with the first two. This signal current is roughly proportional to the displacement, and the third coil is the input coil of a SQUID. STEP's differential accelerometers use two such circuits—one for each test mass—connected back–to–back, so that they share the same SQUID input coil (Fig.7). With proper setup currents, the SQUID will sense a current proportional to the difference in displacement of the masses (and hence to their differential acceleration).

The accelerometers comprising each test mass pair must be balanced against each other to give a true differential signal. This is done by adjusting the magnitude of their respective circulating supercurrents. A supercurrent can be changed by destroying the superconductivity of part of the circuit by a heater and then restoring it after the current is changed, or indirectly by coupling to a transformer which incorporates a superconducting loop as its primary. These are standard cryogenic techniques which can be performed with high precision. The satellite will be accelerated to provide a dither acceleration, and the supercurrents adjusted to minimize the resulting differential mode response. The differential accelerometers are designed to have have excellent rejection of common–mode accelerations combined with exceptionally high sensitivity to differential accelerations. Moreover, they will remain balanced due to the unparalleled stability of the persistent supercurrents.

As previously mentioned, the test masses must be prevented from moving in other than the sensitive direction to avoid ambiguity in the acceleration being measured. The superconducting bearing circuits which constrain the masses are divided in quadrants to allow adjustment of the center of mass positions. This setup adjustment prevents coupling of first order gravity gradient forces into the measurement.

The center of mass offset is measured using the gravity gradient disturbance itself as an error signal, and the bearings are actively adjusted until the disturbance vanishes. The extremely stable supercurrent used for the adjustment is

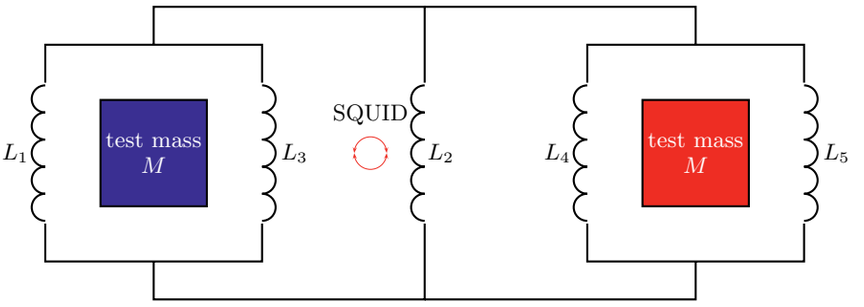


Fig. 7. Differential Position Sensing Circuit

"frozen" at the end of this setup procedure and then needs no further adjustment. The controller for this adjustment will be integrated with the drag–free controller for the satellite. The residual disturbance from drag–free control limits the measurement accuracy of the centering to perhaps 3×10^{-13} m, and will cause the masses to vibrate with an amplitude about 1.3×10^{-10} m, averaged over one signal period. This amplitude is nearly eight times smaller than the requirement of 10^{-9} m, and is at the ~ 0.01 Hz resonant frequency of the bearing, well removed from the signal frequency.

It will be necessary to control the masses' motion both before and after the magnetic bearing setup procedure. This manipulation will be done by an electrostatic positioning system. Several electrodes will be arranged around each test mass, and control voltages applied depending on its position and velocity. The positioning system will also capacitively measure the position of each test mass in six degrees of freedom, plus a charge measurement. During initialization procedures the electrostatic system controls the position of the mass (which must be fixed for SQUID setup) and provides positioning information to the drag free controller in those modes that the SQUID does not measure. During instrument operation the electrostatic system provides two additional functions. It provides damping of the motion of the test masses, thereby reducing intermode coupling. It also performs charge estimation and control, to reduce disturbances from electric charging during SAA passages. The charge measurement is conceptually made by applying a dither voltage to which the mass responds only if it is charged. This measurement is then used to drive the charge to zero in an active control loop. A UV light source, derived from GP-B heritage, allows charge to be moved on or off the mass without contacting it.

During launch the test masses will be held securely by a specially designed caging mechanism which will disengage before measurements begin, but can be re–engaged at any time should the masses need to be clamped for any reason.

4 Experiment Operations and Timeline

We have produced a detailed description of the operations sequence and schedule for the experiment, which is summarized below and in Fig.8. Beyond the setup and calibration stage of the experiment, a completely predetermined schedule of measurement operations for STEP is not possible. A flexible operations strategy is necessary to meet contingencies. This is because the optimum schedule, and which operations are to be performed, depends on the conditions actually found in orbit. These conditions cannot be known in advance because they depend on things such as the degree to which requirements have been met, environmental factors which cannot be predicted (or have never been measured), and possibly overlooked error sources. There is not time enough in the mission to perform all possible measurements, so a selection is necessary based on the conditions actually discovered in orbit. The strategy is therefore to prepare a set of measurements to be performed when predetermined conditions are met. Once preliminary data on the actual conditions are available, these "canned" proce-

dures will be used as needed to identify and calibrate possible systematic effects and optimize the system.

The rough outline of the experiment is therefore as follows. After an initial period of satellite commissioning, the drag free system is initialized in a "coarse drag–free" mode using capacitance sensor measurements for the reference. The differential accelerometers are initialized and calibrated. Really precise calibration is not generally required for STEP; the primary calibration reference is the capacitance sensor, which depends on the dimensional stability of the electrodes for its accuracy. This and the SQUID sensor will be precalibrated to $< 1\%$ on the ground. The accelerometers will be launched in a shutdown condition; the SQUIDs will be set up and recalibrated in orbit by reference to the capacitance sensor. The setups which need to be precise are the common mode rejection adjustment and the center of mass adjustment. These adjust internal ratios rather than absolute values and are self referenced by nulling or minimizing an output.

The common mode rejection adjustment changes the scale factors for each mass in a given differential accelerometer. The satellite will be accelerated at a known low frequency (much less than the longitudinal frequency of the test masses) to provide a dither signal in the sensitive direction. The supercurrents will be adjusted to produce a null response to this acceleration in the differential mode. The supercurrents can be changed in a few seconds, but the time required for the adjustment will need to be at least several times the test mass period ($\sim$1000 seconds) to allow the system to settle. All the differential accelerometers can be adjusted simultaneously, so this adjustment can be completed and checked in a few (2–5) days. Following the differential mode adjustment the satellite will be accelerated at right angles to the sensitive direction, to calibrate the coupling due to accelerometer misalignment. If this coupling is known to a few percent or better, it can be used in a software correction to improve the overall common mode rejection. This completes the adjustment and calibration of the common mode rejection ratio.

The procedure for centering the test masses on one another was described briefly above. In summary the gravity gradient acceleration in the differential mode is used as an error signal for the center of mass displacement. The center of mass displacement is adjusted by changes to supercurrents, and once the masses' equilibrium positions are centered on each other, they should remain centered. The masses are damped in the radial modes by the electrostatic positioning system, and are excited only by residual spacecraft motion, so that the masses never move very far from their equilibrium positions. No on–orbit calibration is necessary since both error signal and displacement go to zero together.

An initial measurement is made with the satellite inertially oriented. This is a reference measurement which can be taken immediately after setup. The signal frequency is equal to the orbit frequency which contains many disturbances that may mask an EP signal. The information is primarily for characterization of the disturbances, so that it will be possible to make a rational choice of the roll rate for subsequent measurements.

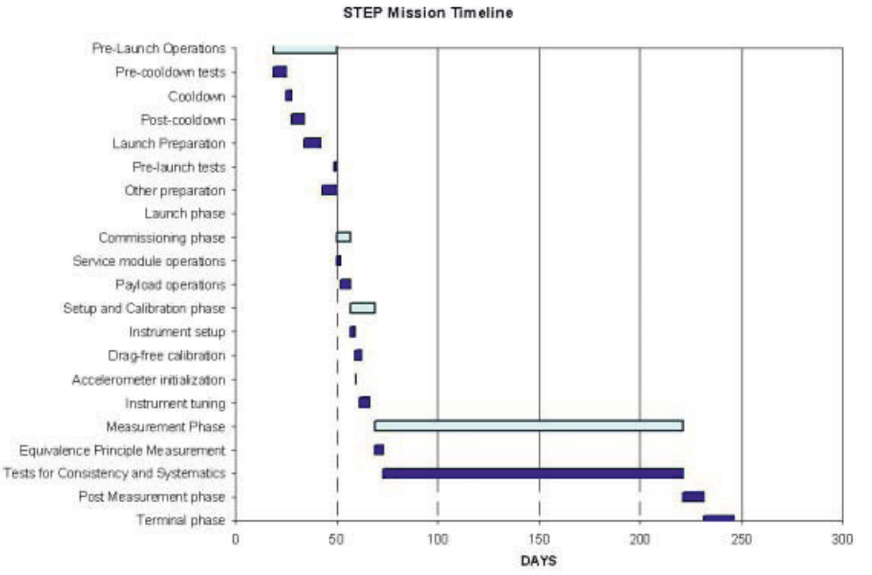


Fig. 8. STEP Mission Timeline

The EP signal appears at the difference frequency between the roll and orbit frequencies. Its phase is such that zero amplitude occurs when an accelerometer's sensitive axis is horizontal. Its amplitude (measured as an acceleration) is independent of roll rate. For an EP measurement, data will be taken for periods of 20,000 to 500,000 seconds, and at roll rates from -3 to $+3$ times orbit frequency ($\sim 1.8\times10^{-4}$Hz), or at signal frequency from -2 to $+4$ times orbit rate. EP measurements will be repeated with different roll rates and phases throughout the mission, at rates which avoid the frequencies of large disturbances. These tests will be intermixed with tests for systematic disturbances, which are discussed in the next section.

When the remaining helium is down to about ten day's supply, the experiment will enter a post–measurement phase which remeasures some of the items measured during calibration to see if they have changed, and adds a few tests which could not be done earlier because they might permanently change the setup. In the terminal phase, after the helium is gone, the satellite loses drag–free and attitude control and rapidly warms to a temperature where no further tests are possible.

We estimate about a year will be necessary to completely analyze the data. The analysis will be performed by two teams of scientists in Europe and America.

4.1 Test for Systematic Effects

Whether or not there is an EP signal it will be necessary to make tests that give confidence in the results. There are two possibilities for judging whether an apparent EP signal may be due to a disturbance, first by attempting to make the disturbance bigger, so that its effect becomes clear, or second, if this is not possible, measuring the disturbance and correlating it with the measurement output. In some cases the effect of a disturbance may be defined well enough to allow it to be calibrated out of the data, possibly giving an improvement in performance.

We can design tests for disturbances that can be anticipated, such as temperature change. This includes most if not all of the disturbances in the error analysis. The plan is twofold: we will have a prepared list of tests to be performed if given conditions exist, and in any case we will perform as many different tests as the schedule allows. Decisions on which tests to perform must be made in real time, picking from the prepared lists. In all cases as many experiment variables as possible should be measured simultaneously.

What follows is a selection from the list of tests we can perform. Except in case of the bearing force determinations (which map background forces) the tests are performed like an EP measurement, but with variation of one or more of the indicated parameters. That is, the indicated parameters are changed but the rest of the procedure is identical to an EP measurement. Note that many of these test for an entire class of disturbance, rather than any specific one.

Bearing Force Determination

The force from the magnetic bearings and environment can produce disturbances by varying in time, and by strong nonlinearities converting the drag free residual into the measurement bandwidth. The force is measured by balancing the unknown forces from the bearings and environment against the electrostatic positioning system. Magnetic and electric forces can be separated to a large extent by varying the bearing setup currents and charge on the mass. The purpose of this measurement is to verify that the forces on the test masses meet requirements, to measure how much of the total force is magnetic or electric, and to see if the forces have changed. This measurement should be made at least once at the beginning of the mission and again at the end. It should be repeated if there is any evidence of change, for example if the masses' resonant periods change for some reason.

Gravity Gradient Sensitivity Measurements

Gravity gradients are a source of concern because they are potentially large relative to the effect sought, and cannot be shielded. They are measured by offsetting the mass centers of the masses or by offsetting the common mass center and correlating the resulting acceleration with environmental factors. Also, comparisons between accelerometers will provide very sensitive measures of both local and

Earth gradients because of the long fixed baseline. These measurements test sensitivity of the experiment to changes in satellite and/or earth gravity gradients. They should be performed if there is an irreducible, significant second harmonic component to the differential acceleration, or if correlations between the common modes of separate accelerometers suggest that the satellite is changing shape.

Electric Charge Sensitivity Measurements

These measurements test that electrical forces are within requirements, how much of the total force is electrical, measure how much the charge contributes to the noise, and how much disturbance could be caused by the charging effect of particle radiation. For these tests the charge on the mass is controlled to a relatively large value (> 0.5 volts) and the effect on the acceleration measurement measured. They should be performed if there is any correlation between the differential signals and the measured residual charge or radiation dose.

Electrical Potential Sensitivity Measurements

The voltages on the sense electrodes for the electrostatic measurement and charge control system are potentially a source of disturbance, although the electrodes and measurement system are designed to minimize this. The disturbance results from both the magnitude of the voltages and voltage noise present on them. These tests increase the voltages and vary them while measuring the resulting accelerations. These measurements should be performed at least once during the mission.

Temperature Variation

Classically, changes in temperature are a major cause of disturbances. The instrument and satellite contain heaters for applying thermal gradients and changing local and global temperatures. These measurements test the coupling of temperature to the experiment through the mechanisms of gas pressure effects, penetration depth changes, thermal expansion, electronic drift, etc. It may not be possible in all cases to get a clean separation of the different coupling mechanisms. These measurements should be made at least once during the mission, or if there is any correlation between measured temperatures and the differential acceleration.

Test Mass Temperature

The masses are thermally isolated from their housings and heated by particle radiation, so that their temperature may rise a few Kelvin during the mission duration. We have not so far found any disturbances caused by a uniform temperature elevation of the test masses; the known disturbances are driven by thermal gradients at or near the signal frequency. The particle radiation is highly

penetrating and heats the masses uniformly. The masses will have very short thermal relaxation times of a few tens of milliseconds or less, so any thermal gradients caused by this small uniform heat input will be both insignificant and transient. Because the temperature of the masses is not directly monitored, if a disturbance from this source is suspected, the masses can be recaged and the temperature rise of the caging mechanism measured to estimate the temperature of the masses. A direct noncontacting measurement of the mass temperature, although of interest, is expected to be very expensive.

Satellite Motion

Motion of the satellite relative to the test masses is a major determinant of the performance of the experiment, since it directly determines the ultimate limit to all displacement measurements and directly affects many subsystems. Indirect effects include gravitational coupling of the satellite to the differential measurement. Two types of satellite motion disturbance may be distinguished. The most important motion disturbance is at low frequencies (< 2 Hz), which is due to the error in response of the drag free system to variations in residual drag. Higher frequency acoustic noise might be generated by overlooked internal sources (monitored by a microphone) or micrometeor impacts.

Sensitivity to satellite vibration can be directly checked by applying a known vibration with a PZT or other acoustic source. Measurements testing the masses' coupling to satellite motion will be performed at least once at the beginning of the mission during the common mode rejection setup. They will be repeated after SQUID setup changes, or if there is any correlation of the differential mode with the drag free error signal, or after changes in satellite rotation if the resonant frequencies of the masses differ significantly (in which case changes in rotation affect their equilibrium positions). These measurements are made by accelerating the satellite in a known way and measuring the resulting output from the accelerometers. For example, the quality of common mode rejection will be measured by accelerating the satellite at a frequency near the signal frequency. The common mode will then be readjusted if required, by changing the setup currents in the position sensors and magnetic bearings.

Other important tests related to satellite motion include switching to different drag free reference masses, applying bias accelerations, and changing the form of the control law. These will measure the sensitivity of the experiment to assumptions about the drag free control and its interaction with the environment.

Magnetic Shielding

External magnetic fields can penetrate into the instrument through flaws in the superconducting shield. This directly affects the SQUID position sensor measurement, adding a fictitious displacement to the real one. Large field leakage can also apply forces to the test masses. We will test the quality of magnetic shielding at least once during the mission, or if there is any correlation with the

external magnetometer output. The external field can be changed by exercising the magnetorquers, or the SQUID outputs can be correlated with the measured external field.

Particle Radiation

Particle radiation has three effects: charging the masses, heating them, and pushing them around by momentum transfer.

Measurements of the charge are planned continuously during the mission; these measurements test the integrity and reliability of the charge estimate from the measured dose. Charge measurements will be performed automatically as part of the standard data set, and will be automatically compared with the output of a radiation sensor to estimate the charge delivered by a given radiation dose. This estimate will be updated continuously in a convergent feedback loop. The effectiveness of this procedure for determining sensitivity to charging from particle radiation can be measured by passing through the South Atlantic Anomaly (SAA) in different orientations, which changes the shielding factor. Note that the charge on the test masses will be systematically and deliberately varied as part of the electric charge sensitivity measurements.

Particle radiation is highly penetrating and heats the masses uniformly. They are highly conducting and have a significant thermal mass relative to the small power in this source. A temperature rise of the order of 1 mK for each pass through the SAA is expected. The effect of this was discussed in the paragraph on test mass temperature.

Particles actually stopping in the test masses will transfer their momentum to them, and since the proton flux may be expected to be predominantly downward, there will be a net downward acceleration during SAA passage of the order of 10^{-19} m/s^2. This small disturbance will be correlated with the proton flux, and not with the signal which is at a different frequency. The corresponding disturbance from cosmic rays is at signal frequency, but is about a thousand times smaller.

Optical and Thermal Radiation

A UV light is used for charge control, and the momentum and heating of the light can disturb the measurement. The sensitivity of the instrument to UV light stability and power will be measured at least once near the beginning of the mission.

Helium Tide

The EP experiment is extremely sensitive to disturbance from gravity gradients from moving masses. The largest potentially movable mass in the satellite is the superfluid liquid helium used for refrigeration. If free to move, its surface would be excited by gravity gradients which have a harmonic relationship to the signal

frequency, resulting in a static distortion of the order of a centimeter, and possibly dynamic (wave) distortion as well. The helium is planned to be constrained by trapping it in a foam or aerogel. This restricts the motion by surface forces (or fountain pressure) much larger than the gravity gradients (See the results described in Section 5). The remaining motion will cause negligible disturbance. Constrained or not, it is important to be able to confirm that the helium is not disturbing the accelerometers. These measurements test the sensitivity of the instrument to motion of the liquid helium.

The amplitude of the disturbance from helium tide is expected to vary strongly depending on the amount of helium in the dewar and its distribution. The signature of a helium tide is expected to include significant fourth and higher harmonics of the signal frequency, and those harmonics will vary systematically in amplitude and phase relative to the signal frequency. This would be accompanied by a difference in common mode acceleration between accelerometers indicating a mass shift on board the satellite that doesn't correlate with satellite temperature.

Several sorts of tests can be performed. The first sort of test is to change the satellite rotation frequency. Because the helium tide has a (largely unknown) time–dependent response to Earth's gravity gradient, this will in general introduce a measureable phase and amplitude shift in the data. This response will vary systematically with helium level, and change slowly during the measurement. The next sort of tests would compare the common mode outputs of different accelerometers. The comparison would be highly sensitive to motion of nearby masses including helium tide. Third, introducing a static displacement between the masses in each accelerometer will enhance the test, possibly allowing some localization of the disturbance by measuring higher gradients. Finally, tests which apply heat or thermal gradients to the helium bath may change the distribution of helium within the constraining foam, by fountain effect, and therefore will change the helium tide response.

Mass Dynamics

Similar to the case of satellite motion, movement of the test masses relative to their housings can cause various disturbances. These disturbances include calibration changes, real and apparent coupling between modes of motion, and saturation of measurement systems. The masses' motion will normally be electronically damped to very low amplitudes in all degrees of freedom, using the electrostatic positioning system. A particular concern is rotation about the cylinder axis, which might be difficult to measure and control. We have shown that test mass rotation can be sensed and controlled using the residual out of roundness of the cylinders ($\sim 1\ \mu$m).

Measurements involving mass dynamics test the accelerometers' sensitivity to motion in modes other than the sensitive axis, including cross–coupling and frequency conversion. The masses will be displaced and excited in different modes by the electrostatic positioning system and magnetic bearings, while observing the accelerometer outputs.

5 Error Analysis and Mission Tradeoffs

STEP is necessarily a highly optimized experiment with many tradeoffs, and sections cannot simply be split off for separate analysis without the risk of inconsistency. Most error sources were evaluated analytically for the purpose of setting requirements in the original STEP Science Requirements Document (May 1990) and the analysis has been continually updated to the present day in an increasingly systematic and consistent way.

Early error calculations, sometimes performed by independent workers, tended to use independent assumptions for each error source resulting in requirements which were not entirely self–consistent. To eliminate these inconsistencies in assumptions and requirements we have developed a computer program which automatically calculates the errors and resulting tradeoffs from consistent starting assumptions.

The starting point for all the computer programs is the STEP Error Analysis, a text document actively maintained in our database. The STEP Error Analysis includes estimates of specific disturbances to the masses in the general categories of thermal noise, gas pressure forces, electrical forces, magnetic forces, gravitational forces, radiation pressure, and vibration. It also includes estimates of disturbances to the measurement system, including measurement noise, changes to the superconductors, and thermal and mechanical stability. Specific disturbances considered include electric potentials in the housing surrounding the mass (including patch effect), the radiometer effect, losses from eddy current damping, gravitational coupling from helium tide, drag free residual vibration coupled to the differential mode, SQUID noise, penetration depth changes in superconductors, and momentum transfer from penetrating particle radiation. The Error Analysis also lists interactions and tradeoffs within the experiment.

We also maintain an error analysis program, broadly based on the STEP Error Analysis. The goal is to embody a model of the experiment and apparatus (especially including tradeoffs) as a self–consistent whole. The extensive model of the accelerometers automatically calculates the common mode balance procedure, and the drag free and charge control laws are physically realizable models which will be updated with the actual control laws when those are finalized. The program systematically evaluates the known errors starting from a relatively small and consistent set of low level assumptions, and because it efficiently and explicitly incorporates the major tradeoffs it can be used to set requirements. Many if not most experiment parameters are explicitly dependent on other parameters or requirements; some simple examples are the mass of a test body from its dimensions and material, or the signal frequency from orbit height and satellite spin rate, or less directly the common mode rejection ratio from the setup currents and coil dimensions in the SQUID position sensor. The program automatically calculates these, preventing much of the inconsistency that results from independent assumptions. Two examples of the utility of this approach are given below, for the effects of orbit height and satellite rotation on the experiment.

5.1 Orbit Height Effects in the Model

Figs. 9 and 10 show sample results for two different orbit heights, 500 and 400 km respectively, with no other changes. Fig.9 is close to the STEP baseline mission. In this example the performance is dominated by two noise noise sources, SQUID noise and Nyquist noise. SQUID noise is the inherent flux noise of the SQUID sensor, converted into an equivalent acceleration. Nyquist noise is the fundamental thermal noise limit, likewise converted.

Some differences due to orbit height are readily apparent, especially in the disturbance *"dynamic CM offset"* disturbance which is greatly increased at the lower altitude relative to the other disturbances. Vibrations of the test masses about their equilibrium positions result in changing center of mass offsets that can couple gravity gradients into the signal frequency. These vibrations are caused by residual satellite motion. The drag free residual sensed in the differential mode of the accelerometers (*drag free residual in diff. mode*) is a direct limitation to the mass centering performance, and this residual is increased by the overall increase in drag in the lower orbit. Consequently the centering system cannot work so well in a lower orbit, and the disturbance increases rapidly with the centering error. There is a similar cause for the much smaller increase in the disturbance from electrical charge. The charge control system's performance (*Electric charge*) is also determined by the overall drag free residual, so that the charge cannot be controlled so well in a low orbit. This cancels the expectation that the charging problem would be less because the radiation decreases in lower orbits. The smallest charge that can be measured depends on the residual acceleration noise from drag, and the charge cannot be controlled better than it can be measured. In the model used for the radiation flux and its interaction with the test masses, the increase in drag noise is proportionally somewhat larger than the decrease in the disturbing flux, within the range of likely altitudes for STEP.

5.2 Satellite Rotation and EP Signal Detection in the Error Model

The ability to shift the signal frequency away from the frequency of disturbances is extremely important. Many disturbances, for example that due to particle radiation, occur at the frequency of the orbit or its harmonics. Moving the signal away from these frequencies may result in a significant improvement in signal to noise.

The disturbances of most concern occur at the signal frequency, the difference between orbit rate and rotation rate. For these disturbances, the rotation rate cannot have a direct effect because the disturbance tracks it. It may happen that indirect effects of rotation are important. One such case is when the centrifugal force of rotation is comparable to the spring constant restraining the test masses. Centrifugal force is a direct effect of rotation, and adds directly to the spring constants. However, the spring constants are determined by the set–up conditions, which tend to cancel the direct effects of the rotation. It is

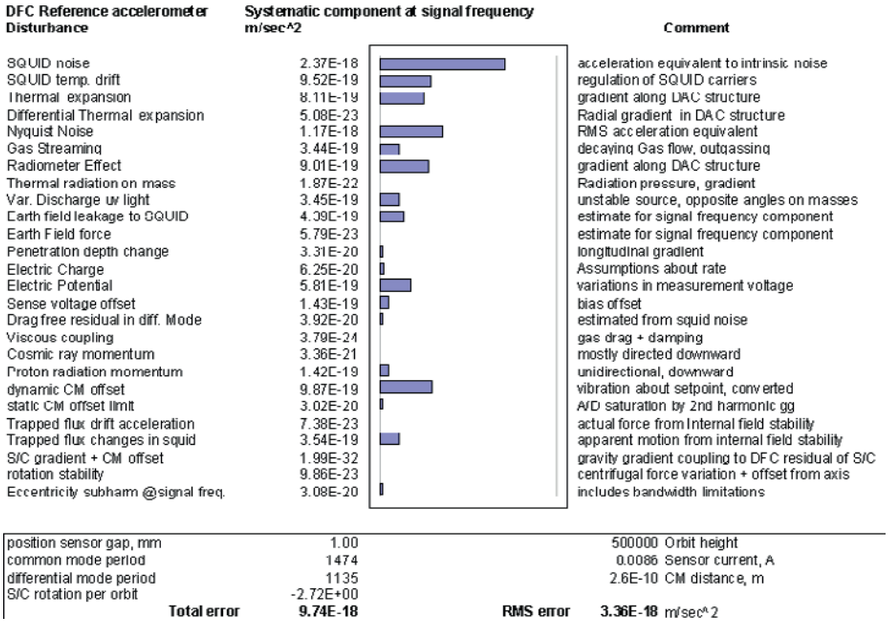


Fig. 9. Typical Errors for a 500 km orbit. Noise sources known to be incoherent, e.g. thermal noise, are converted to an expectation value at the signal frequency by RMS averaging. Coherent errors (e.g. some temperature driven effects) are calculated at signal frequency. A worst case estimate is made by simply summing the errors. For a statistical estimate, the errors are combined in an RMS sense. The bar chart is automatically rescaled in proportion to the total disturbance.

the changed set–up conditions which affect the sensitivity of the SQUID sensors (and the experiment) rather than the rotation directly.

The centrifugal force from the rotation has a tendency to destabilize the magnetic springs (from the position sensor) that hold the masses in place. This is the reason for holding the rotation to less than three times orbit rate; the springs are required to give a period of 1500 seconds or less to provide some margin of stability. In the frame of the satellite the centrifugal force is constant except possibly for small changes due to variation of rotation rate. The constant centrifugal force is easy to distinguish from EP signals. The centrifugal force disturbance due to changes in rotation rate w is $(\omega\delta(\omega) - \Delta X + \omega^2\delta(\Delta X))C_{\mathrm{mrr}}$, where ΔX is the distance from rotation axis and $\delta()$ indicates the variation in a quantity. $\delta(\Delta X)$ represents the amplitude of mass motion about ΔX. Variations in rotation rate $\delta(\omega)$ change the centrifugal force for both masses, and are reduced in the differential mode by the common mode rejection ratio C_{mrr}. For $\omega = -0.001$ rad/s ($\approx$ -orbit frequency), $\delta(\omega) \approx 0.01\omega$, $\Delta X \approx 7.5 \times 10^{-9}$ m, and $\delta(\Delta X) \approx 1.3 \times 10^{-10}$ m at signal frequency, this disturbance is about 3×10^{-21} m/s^2.

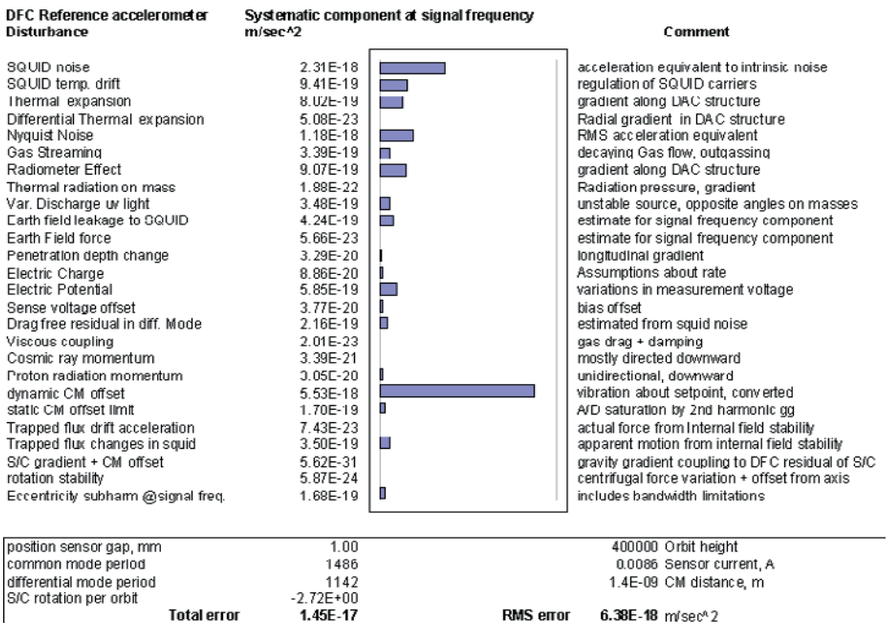


Fig. 10. Typical Errors for a 400 km Orbit

The drag–free and attitude control system controls all six degrees of freedom, including rotation. If uncontrolled, the satellite will rotate about some line through its center of mass rather than about the line joining the centers of the test masses. The attitude control system forces the satellite to rotate about the line through the centers of the test masses (the "rotation axis"), using linear thrusters to provide the necessary forces and torques. To do this, the system must apply a constant thrust in the rotating frame (with some perturbation from disturbances), directed toward the axis and proportional to the distance between the satellite's center of mass and the rotation axis. Enough thrust is available for this if the satellite center of mass is within 10 cm of the rotation axis, but some thrust must be reserved for translational control. The satellite is reasonably symmetric about the rotation axis, and there will be a requirement to passively trim the center of mass to within less than 1 cm of the axis. This leaves 90% of the thrust available for translation. Lockheed Martin has facilities for measuring the center of mass of satellites (specifically GP-B) to within 2.25 mm about the roll axis. Trimming to reduce the satellite's products of inertia may also be needed. These can be measured in the same facility to within 5 kg-m^2.

For many disturbances, neither the rotation rate nor the signal frequency is a direct input to the result of the error analysis, or it has a small effect. For disturbances which depend directly on the orbit rate, the error analysis assumes

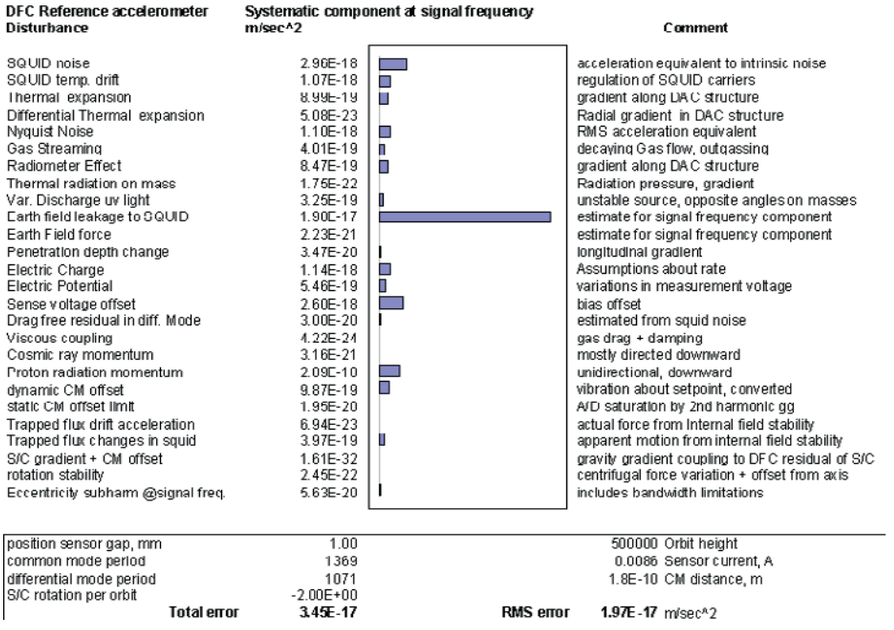


Fig. 11. Effect of Rotation Choice

they have harmonics at multiples of the orbit frequency; in many cases the amplitudes of these harmonics can be explicitly calculated, and in the others a conservative estimate is used. The frequency of disturbances and their harmonics influences the choice of rotation rates, which should not be multiples of the orbit frequency.

The expected effects of a change in rotation from -2.718 times orbit frequency (the previous examples) to -2.000 times orbit frequency are shown in Fig.11. The most apparent changes are a large increase in the disturbance from magnetic field penetration and smaller increases in electric charge effects and the momentum transfer from proton momentum. The magnetic field of the Earth changes direction at twice orbital frequency in orbit. Its fourier spectrum is made up of narrow lines at harmonics of the orbit frequency, and this particular choice of signal frequency coincides with one of them. Note that the narrow line width of regular disturbances such as magnetic field penetration is determined by the observation time, while their own structure determines the relative amplitude of their harmonics. Regular disturbances with short durations relative to the orbit, such as proton momentum transfer, are spread out over more harmonics, but have little energy away from multiples of the orbit frequency. Thus, small changes away from orbit harmonic frequencies may make a large difference in a disturbance.

The changes in electrical disturbances are related to the regular passage through the South Atlantic Anomaly. In this case the rotation rate of -2 times orbit frequency means that every third passage through the anomaly comes with the same phase. Because passage through the anomaly is short, its energy is spread over many harmonics, and the third harmonic changes the charge in time with the signal frequency. The relatively slow operating charge control system cannot correct the charge quickly enough to reduce the disturbance.

A selection of possible roll rates, and reasons for selecting them or not based on our analysis, is given in Table 2.

These two examples illustrate the importance of having an integrated model of the full system, in understanding the results. The effect of other errors can be similarly understood. Several of the most significant error sources are discussed below.

Table 2. STEP Satellite Roll Rates

SQUID Noise

A well designed experiment will be limited by sensor noise if not by a more fundamental disturbance. This is the case in STEP, where the very tiny intrinsic flux noise in the SQUID of 1×10^{-30} J/Hz is equivalent to an acceleration noise δa of about 5×10^{-13} m/s$^2/\sqrt{\text{Hz}}$ depending on setup conditions. This then is the limit of performance of the drag free control system, which cannot distinguish between the sensor noise and an actual drag variation. To compare the accelerations of two test masses to 10^{-17} m/s^2 in the presence of this (relatively) large disturbance it is necessary to make a differential measurement with common mode rejection C_{mrr} no bigger than 10^{-4}. This high common mode rejection is achieved by accurate manufacture of the accelerometers combined with the common mode adjustment procedure previously described.

Measured STEP SQUID noise characteristics are used to model the acceleration sensitivity in the STEP error analysis. The transfer function relating SQUID noise to acceleration sensitivity depends on readout circuit inductances, inductance derivatives, circuit setup currents, and test mass masses. See [6] for details of the derivation. The measured SQUID noise and analysis based on the baseline instrument design yield an acceleration sensitivity that meets the resolution requirement of 2.7×10^{-18} m s^{-2} in a bandwidth of 8×10^{-6} Hz. Figs. 12–14 show the measured properties of GP-B SQUIDS, which are identical with the STEP baseline choice.

The SQUID noise has consequences beyond the immediate acceleration disturbance. Two other systems depend on the acceleration performance of the drag–free system for their operation. These are the charge control system and the mass centering system. Because of the close dependence of the performance of these systems on acceleration noise, the overall performance of STEP may be degraded as a high power of the residual acceleration.

As mentioned above, the smallest measureable charge in the proposed system depends on the drag free residual acceleration. The charge measurement system works (conceptually) by measuring the acceleration of each test mass in response to a dither voltage. Any residual satellite motion appears as an apparent charge $M\delta a/E$ for a mass M, acceleration noise δa, and average applied electric field E. E is limited to small values (≈ 0.1 volt per mm) by the requirement to reduce disturbances from variations in the applied potentials and induced charges on the test masses, so the acceleration noise is a limiting factor to the performance of the charge measurement. Charge on the test masses is controlled by a system derived from GP-B which uses ultraviolet light and a bias electrode to steer electrons on or off of each mass. Incorporated into an appropriate control loop, this provides the best possible control of the charge, limited by the charge measurement noise $q = M\delta a/E$. This produces two sorts of disturbance to the EP measurement, one proportional to q^2 (or δa^2), and one proportional to q times any electric field that may be present.

The ability to center the masses in each accelerometer depends on the acceleration noise in the differential mode. The differential mode acceleration is used to monitor the center of mass separation, with the Earth's gravity gradient

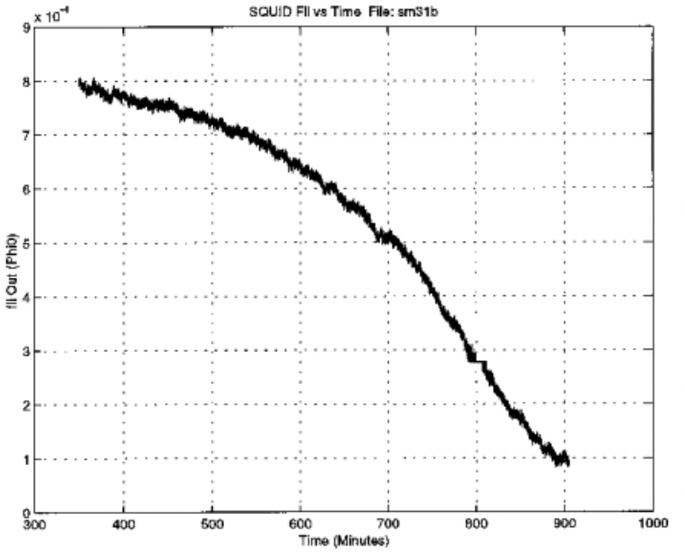


Fig. 12. Measured Output, in units of Φ_0, GPG Science Mission Qualified SQUID sm31b vs. time.

driving it. The centering can be no better than when the error signal from gravity gradient is the same size as the noise in the differential mode, which is the residual satellite motion times the common mode rejection. The static miscentering measureable by this technique in principle is quite small; at orbit radius $R = 6870$ km, g is about 8.43 m/s^2, and for observation time $T_0 = 100,000$ s, and $C_{\rm mrr} = 10^{-4}$, the measureable miscentering is $((T_0/2\pi)^{1/2} R\delta a C_{\rm mrr})/g \approx 3\times 10^{-13}$ m. This is not likely to be achieved because the masses' common modes will be excited by the acceleration noise, to an average amplitude $\delta a/(2\pi/T_b)^2 \approx 10^{-10}$ m, where T_b is the radial period of the magnetic bearings. It is this latter displacement which is converted to signal frequency by mixing with the Earth's gravity gradient at twice signal frequency, and causes the disturbance *dynamic CM offset* mentioned above.

Thermal Expansion

Thermal expansion in the instrument has two effects. It can change the shape of the test masses, causing them to couple more strongly to gravity gradients, and it can directly affect the measurement.

Thermal contraction coefficients of the test masses must be known to ∼10% in order that they will come out the right size at low temperature. Great accuracy of knowledge is not necessary because the contraction is a small part of the

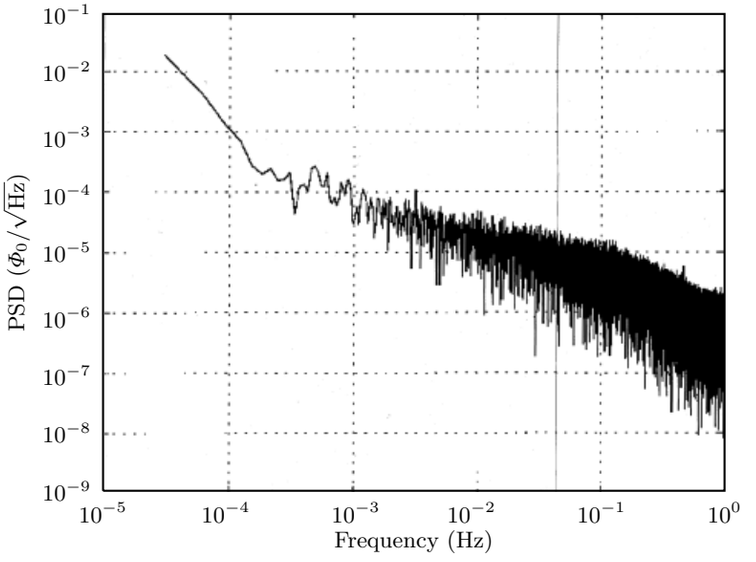


Fig. 13. Power Spectral Density of Trace in Fig.12, in units of $\Phi_0/\sqrt{\text{Hz}}$.

total size. The important thing is that the thermal contraction be uniform for the whole mass. Differences in thermal contraction can mechanically warp the shape of the mass much more than the actual thermal contraction would suggest. Current thinking is that the test mass thermal contraction needs to be uniform to about 10^{-4} to prevent warping. If this is the case, apparent changes in position due to nonuniform expansion coefficient will be limited to less than 1.4×10^{-16} m, which would not be expected to recur at signal frequency. This offset is insignificant for the acceleration measurement.

The concern then is about the warping of the test masses. This might change the carefully minimized gravitational coupling terms. The total thermal expansion of beryllium (for example) over the range 0–293 K is about $\alpha_{Be} = 0.00131$; the masses must be oversized by this fraction to come to the correct size when cold. If the expansion coefficient is anisotropic, and differs by 10^{-4} in perpendicular directions, the diameter/length ratio will change by about 10^{-4} $\alpha_{Be} = 1.3 \times 10^{-7}$ upon cooling from room temperature, and the length (or diameter) will differ from its intended value by about 2×10^{-8} meters. This is comfortably smaller than the smallest relevant tolerance in the manufacture, 0.3 μ or 30×10^{-8} meters.

Much larger dimensional changes could occur if there are gradients in the expansion coefficient over scales about equal to the size of the mass, because of thermal bending. The worst case thermal bending would be the case of a bender comprised of two strips of material with differing expansion coefficients welded to each other. The deflection of such a bender is roughly $L^2\delta\alpha/S$ where L is the

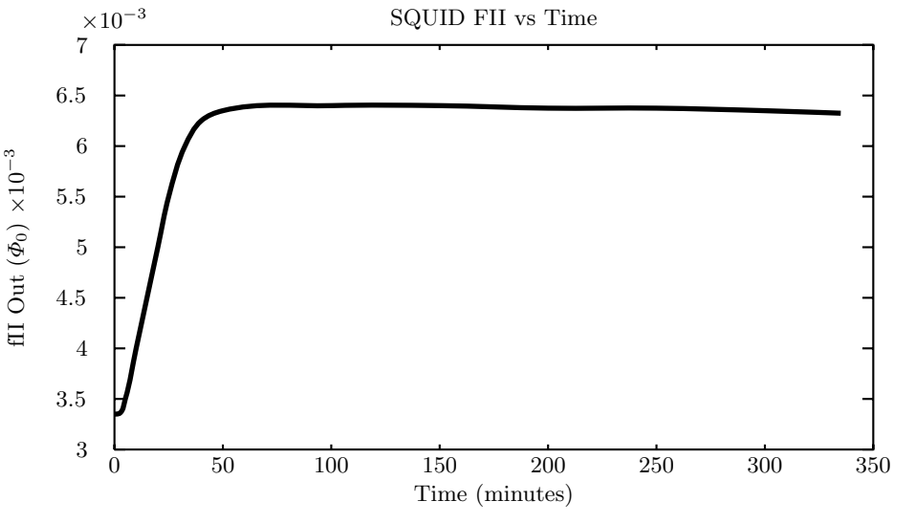


Fig. 14. Measured Output, in units of Φ_0, GP-B SQUID with Temperature Controlled Electronics.

length, $\delta\alpha$ the difference in thermal expansion, and S the thickness; Substituting worst case values from our test mass dimensions, $L = 0.069415$ m, $S = 0.0025$ m, $\delta\alpha = 10^{-4}\ \alpha_{Be}$, the deflection will not exceed 25×10^{-8} meters. This is also less than our smallest required tolerance. Coupon samples will be taken from the material during machining to verify that the expansion differs by no more than a specified amount.

The most stringent mechanical stability requirement comes from considering the case that one position sensing coil substrate undergoes a dimensional change. Then the resulting change in the radius of the thin film pickup coil will change the pickup coil inductance and could thereby mimic a change in position of the test mass.

The mechanical stability requirement on the accelerometers is set to limit the possibility of false differential acceleration signals at signal frequency. Ultimately, the acceleration measurement is a measurement of test mass position. The natural period of the longitudinal mode of test mass oscillation is ∼1000 seconds, so $\omega_n \approx 2\pi \times 10^{-3}$ rad/s. Therefore the readout system requirement of differential acceleration resolution of less than 4×10^{-18} m s^{-2} in a 6×10^{-6} Hz bandwidth implies the differential position measurement resolution in this same bandwidth must be 4×10^{-8} m/s^{-2}/$(2\pi \times 10^{-3}$ s$^{-1})^2 \approx 10^{-13}$ m. The position of each test mass is measured by a differential circuit with a pickup coil near both ends of each test mass belt. This makes the measurement relatively insensitive to changes in the test mass size and accelerometer length, and also greatly eases the dimensional stability requirement of the accelerometer housing along the direction of the symmetry axis.

Using the STEP baseline pickup coil design parameters we have modeled the dependence of inductance on coil radius and thin–film trace width and compared this to the dependence of inductance on test mass position. We require that the dimensional changes in the sensing coil contribute at most 10% of the required test mass position resolution and obtain a mechanical stability requirement of 1 part in 10^{-11} for changes modulated at signal frequency.

Given the systematic temperature difference requirement of 500 μK, the mechanical stability requirement will be easily met using either quartz or sapphire material to construct the accelerometer housing, bearing, sense coil, and electrode structures. For example, Braginski and co–workers [3] report measurements of the linear expansion of sapphire at 4.5 K to be $\sim 5\times 10^{-11}$ K^{-1} with a good fit to the predicted T^3 law in there measurement range of 4.5 - 15 K. Taylor and co–workers [7] obtained similar results for a sapphire Fabrey–Perot optical cavity between 5 K and 77 K. Measurements of Heraeus engineering grade quartz for the Gravity Probe B program indicate a linear expansion of below 1×10^{-10} K^{-1} at 77 K while the higher grade Herasil 1 Top has an even smaller thermal expansion.

Helium Tide

If the liquid helium used to refrigerate the experiment has a free surface, it can move under small influences such as the gravity gradient of the Earth ($\sim 10^{-7}$ g), which are harmonically related to the signal frequency. This moving mass would produce significant gravitational disturbances to the experiment, which by frequency conversion may appear in the signal bandwidth. This is a concern because the estimated disturbance might be several thousand times the proposed sensitivity and is essentially impossible to calculate reliably. Moreover, separating this disturbance by rotating the spacecraft may not work well since it is locked to the signal frequency, although its amplitude and phase could be varied. The disturbance would be expected to change its signature significantly during the mission, as the helium boiled away and the free surface changed its size and response characteristics.

We have observed that the liquid–vapor interface in an aerogel nearly filled with superfluid helium II takes a shape that is independent of gravity over times up to 3 hours (Fig.15). The most likely reason that the interface is fixed is that its position is determined by capillary forces. The studies were for nearly filled aerogel, since this is when capillary forces are weakest and least understood, and the distribution of helium in the aerogel is then most susceptible to gravity. Apparently, therefore, the helium II liquid–vapor interface will not change shape in response to gravitational forces of 1 g or less for any amount of helium in the aerogel. This makes aerogel a very promising candidate for helium control in STEP. Aerogel is expected to eliminate this disturbance, so that elimination by experimental procedures or data analysis is no longer necessary. Studies are continuing on aerogels and other materials to determine their suitability for this purpose.

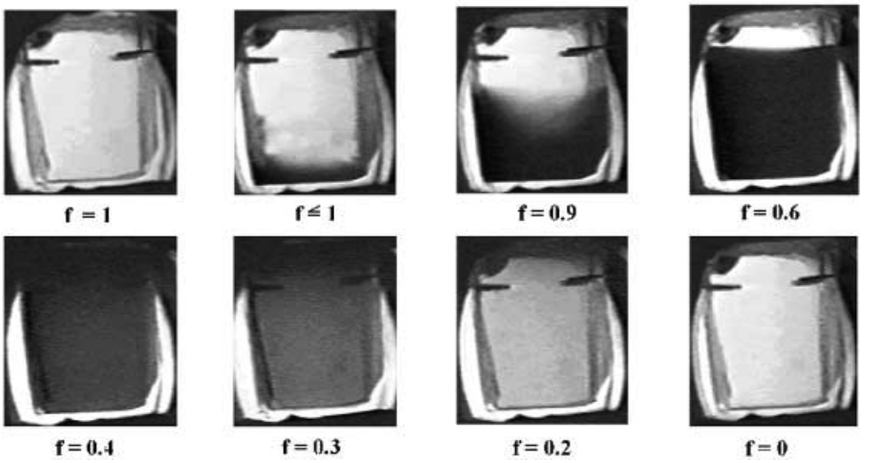


Fig. 15. Images of an aerogel sample as a function of the filling fraction, at 1.9 K. The dark regions are caused by an increase in light scattering in regions from which helium has been removed. The lower density region is below the higher density region, indicating a weak influence of gravity on the helium distribution. Tilting the cryostat 11 degrees had no noticeable effect.

Patch Effect

The patch effect is the variation of electrical potential on conducting surfaces, which can give rise to spurious couplings between the test masses and their housing. Work function differences can occur in polycrystalline metals due to the surface exhibiting random crystal orientations or due to non–uniformly adsorbed surface layers. We have recently re–evaluated this potentially disturbing effect and performed a trade between it and the mass—housing gap, an important dimension in the baseline design.

We have performed a theoretical analysis of the expected patch effect forces, force gradients, and dissipation and noise effects, and characterised the grain structure of sputtered Nb films which will coat the test masses, and which form the bearing and position sensing circuits. A theoretical study of the forces and gradients of forces due to variations of electrostatic potential over the surfaces of two adjacent metallic surfaces has been given for the LISA gravitational wave experiment in [8]. Based on these studies we conclude that the patch effect will not present a significant disturbance to the STEP experiment. Therefore, the patch effect does not determine gap spacing tradeoff in the baseline accelerometer design, although it could become important for significantly smaller gaps than our current baseline of 100 μm. The error analysis includes the following effects to determine design tradeoffs: bearing magnetic fields, effects of trapped flux, capacitive forces and dissipation (in particular with non–zero charge on test

masses), the variation of inductance with distance of position sensing circuits, including its first and second derivatives, and gravity gradient couplings.

A basic result for patch effect forces is as follows: If there is significant coherence between the 2–dimensional potential distributions on both surfaces *and* there is significant power in the spatial spectrum of potential variations at wavelengths comparable with the spacing between the surfaces, then the force gradients are approximately $\epsilon_0 A V^2/a^3$ where a and A are the separation and common area of the electrodes respectively, and V is the amplitude of the variation in potential. If all the spectral power lies at wavelengths much smaller than the separation, the forces and force gradients are reduced exponentially. If there is no correlation between the two distributions of potential then there will be no force or force gradient. Correlations between the distributions for clean surfaces will in general be determined by the distribution of surface grains, discussed below for Nb films similar to those for STEP.

Differences in the parasitic stiffness coupling the test–masses to the spacecraft, caused by patch effect, will not lead to a loss of common–mode rejection because they will be trimmed out during setup by adjustment of the detector supercurrents. The stiffness due to the patch–potential correlations must not change the resonant frequency by more than a factor 10^{-4}. STEP can tolerate potential variations of 0.3V and separations of 0.1 mm. The drag–free (or common–mode acceleration noise) requirement on the patch–field force gradients should be approximately satisfied by the current STEP accelerometer design irrespective of the spatial distribution and correlation of the patch–potentials.

Any variation in time of the forces and force gradients due to patch–fields will necessarily produce energy losses and will affect the quality factor of the accelerometer. We have shown that, provided the thermal noise is sufficiently low to satisfy the STEP science goals, the patch–potential variations will not produce significant differential forces on the test–masses.

Perhaps the most likely way that the surface potentials can be modified over time and over distances large enough to cause a correlation is through the charge management system. The UV radiation will liberate surface contaminants as well as electrons. The work functions of both the test–mass and the levitation coil will be modified by the UV exposure (and the high energy protons in the SAA). Some of these free ions will have low energies and will move in the electric fields generated by the patch fields near the metallic surfaces and presumably will occupy potential wells due to the patch fields. Changes in the electric field experienced at one surface due to the presence of the other will lead to a change in the population of adsorbate atoms and corresponding losses. The time taken for an adsorbate to diffuse will depend on the depth of the potential well in which it sits [2]. Such motion of adsorbates will give rise to losses in energy. Depending on the number density distribution of wells as a function of energy this will give rise to 1/f noise in a fashion similar to anelasticity in mechanical suspensions.

If the patch fields are correlated initially, charge motion will take place in order to anti–correlate them as much as possible. After a time the stiffness will reduce by an amount ΔK_p which is analogous to the lossy component of a

mechanical spring. We can put an upper limit on ΔK_p using the requirement that the common–mode rejection ratio be stable to one part in 10^4, $\Delta K_p \approx 10^{-4} m\omega^2$, where ω is the resonant angular frequency of the differential mode of the accelerometer with equivalent mass m. Estimates for $\Delta \mathrm{K}_p$ lead to a condition on ω. The spectral power density of thermal noise forces produced by fluctuations in the surface contaminants is expected to be $F^2 = 4kT\Delta K_p/\omega$, which must be less than 10^{-30} $\mathrm{N}^2/\mathrm{Hz}^{1/2}$ to achieve the STEP goal. In the worst case the thermal noise is generated entirely by motion of surface contaminants. Thus ΔK_p must be less than $10^{-30}\omega/4kT$. Thus the common–mode rejection ratio will be stable enough provided $10^{-4}\omega^2 > 10^{-30}\omega/4kT$. Using an angular frequency of $\omega = 10^{-3}$, we find that this inequality is satisfied with an order of magnitude to spare.

Niobium thin films grown by sputtering tend to be very fine grained polycrystalline materials. Niobium films of order a micrometer or larger in thickness become quite rough. These films and the oxide layer on them have been well characterized. One point to consider is that any niobium surface which has been exposed to the atmosphere will develop an oxide coating consisting of amorphous $\mathrm{Nb}_2\ \mathrm{O}_5$, which is an insulator, mechanically hard and stable, dense and well bonded to niobium. In STEP these films will have to be covered with an inert, thin, slightly conducting layer to prevent charge buildup.

Fig.16 below shows an example of a niobium film grown at Stanford for coating on quartz substrates. The small grain size suggests that any correlation of surface potentials will be due to overall texturing rather than specific details of the grains.

6 Conclusion

We have presented a simplified description of the STEP apparatus and procedures, followed by highlights of recent technical progress. The STEP accelerometer design has been advanced to a stage where we can build a prototype flight accelerometer based on previous working models. Particularly important progress has been made in a unified error model for the experiment and major disturbances, and in practical solutions to some longstanding problems such as helium tide and patch effect disturbances.

References

1. T. Damour and A.M. Polyakov: The string dilaton and a least action principle, *Nucl. Physics* **B 423**, 532 (1994); String Theory and Gravity, *Gen. Rel. Grav.* **12**, 1171 (1996).
2. C.C. Speake, T.J. Quinn, R.S. Davies, and S.J. Richman: Experiment and theory in anelasticity, *J. Meas. Sci. and Technol.* **10**, 430 (1999).
3. V.B. Braginskii, S.I. Vasil'ev, and V.I. Panov: Thermal expansion of aluminum oxide crystals at low temperatures, *Sov. Tech. Phys. Lett.* **6**, 665 (1980)

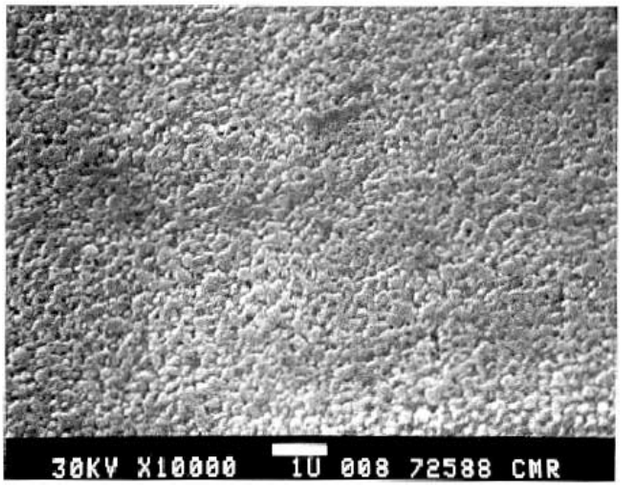


Fig. 16. A SEM image of a 1.4 μm thick sputtered niobium film deposited on a 3.8 cm diameter fused quartz sphere. Deposition conditions were as follows, RF Diode at 600 W, 3.0 – 3.2 kV peak, argon pressure 9.9 mTorr. The coating was deposited in 2 layers, with several days air exposure between layers. Base pressure below 1×10^{-6} Torr. By comparison with the 1 μm scale bar, the largest grains are about 0.25 μm in diameter and the average grain size is closer to 0.1 μm.

4. N. A. Lockerbie, A.V. Veryaskin, and X. Xu: Differential Gravitational Coupling Between Cylindrically–Symmetric, Concentric Test Masses and an Arbitrary Gravitational Source: Relevance to the STEP Experiment, *Class. Quantum Grav.* **10**, 2419 (1993).
5. N.A. Lockerbie, X. Xu, A.V. Veryaskin, and M.A. Hosey: Optimization of Immunity to Helium Tidal Influences for the STEP Experiment Test Masses, *Class. Quantum Grav.* **11**, 1575 (1994).
6. H.A. Chan and H.J. Paik: *Phys. Rev.* **D 35**, 3551 (1987).
7. C.T. Taylor, M. Norcutt, E.K. Wong, A.G. Mann, and D. Blair: Measurement of the coefficient of thermal expansion of a cryogenic, all–sapphire, Fabry–Perot optical cavity, *Optics Communications* **131**, 311 (1996).
8. C.C. Speake: *Class. Quantum Grav.* **13**, A291 (1996).

High Sensitive DC SQUID Based Position Detectors for Application in Gravitational Experiments at the Drop Tower Bremen

Wolfgang Vodel[1], Hansjörg Dittus[2], Sandor Nietzsche[1], Helmar Koch[1], J.v. Zameck Glyscinski[1], Ralf Neubert[1], Stephan Lochmann[2], Carsten Mehls[2], and D. Lockowandt[2]

[1] Institut für Festkörperphysik, Friedrich–Schiller–Universität Jena
[2] Zentrum für angewandte Raumfahrttechnologie und Mikrogravitation (ZARM), Universität Bremen

Abstract. Free fall tests for proving the Weak Equivalence Principle (WEP) have been rarely be done in history. Although they seem to be the natural experiments to test the equivalence of inertial and gravitational mass, best results for proofs of the WEP could be attained with torsion pendulum tests to an accuracy of 10^{-12}. Pendulum tests are long term periodic experiments, whereas free fall tests on Earth can be carried out only for seconds causing certain limitations in principle. Nevertheless, very precise fall tests in the 10^{-12} to 10^{-13} range are possible and under preparation to be carried out on the Drop Tower Bremen for a free fall over 110 m. These tests require position detectors with an extremely high resolution in order to measure tiny displacements of freely falling test masses. Using SQUID–based sensing technique, the displacements can be determined with an accuracy of 2×10^{-14} m/$\sqrt{\text{Hz}}$. The SQUID system, developed and manufactured at Jena University, provides high sensitivity and extremely low intrinsic noise, especially at low frequencies. Some recent results are discussed.

1 Introduction

The study of gravitational interactions is an important challenge for basic science. Many theories have been developed to describe the possible material dependent coupling of masses to gravitation, but up to now a corresponding experimental proof is missing. The evidence of such a dependence is indicated by a violation of the Weak Equivalence Principle (WEP). This principle requires an exact equality between gravitational mass m_{g} and inertial mass m_{i}. The experiments carried out to test the WEP may be classified into two groups:

1. Torsion balance experiments and
2. Galilean type (or drop tower) experiments.

Torsion balance experiments carried out by Dicke [1] and Su *et al.* [2] placed an upper limit of 5×10^{-12} on the so–called Eötvös–factor

$$\eta(A,B) = 2\frac{(m_{\text{g}}/m_{\text{i}})_A - (m_{\text{g}}/m_{\text{i}})_B}{(m_{\text{g}}/m_{\text{i}})_A + (m_{\text{g}}/m_{\text{i}})_B}\,. \tag{1}$$

This is still the best proof of the validity of the WEP.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 248–272, 2001.

Although Galilean tests seem to be the easiest way to prove the WEP, the best experiments as performed by Niebauer et al. [3] and Kuroda and Mio [4], e.g., allow an accuracy of 5×10^{-10} only. In these free–fall experiments absolute gravimeters were mainly used adapted especially to this application. Although in these experiments the difference signal is detected, the measurement is absolute in a sense that the unavoidable time delay between the release of both test masses at $t = 0$ will cause a displacement $\Delta s \approx g\,\Delta t\,t$ which can be larger by orders of magnitude than the expected signal due to a violation of the WEP in the aspired limit $\eta \leq 10^{-13}$. On the other hand, the short free–fall time of 0.3 s only yields a very small gravity dependent relative shift in the position of the test body. The Drop Tower Bremen as an earth bound microgravity facility with a free–fall distance of 110 m and a flight time of 4.7 sec offers the unique possibility of a pseudo–Galilean test of the WEP in which a true relative measurement can be performed. This can be realized by integrating the measuring device into the free falling system so that displacements caused by the time delay become very small. The experiments cover the interesting and rarely measured distance range of up to 100 m and are ideal tests for the planned space experiments like the Satellite Test of Equivalence Principle (STEP).

In this drop tower experiment as well as in the satellite STEP experiment of NASA/ESA, one wants to prove a possible violation of the WEP at a level of an accuracy up to one part in 10^{13} and 10^{18}, respectively. High performance SQUIDs will be used as ultra sensitive superconducting position detectors for measurements of tiny displacements of test masses [5,6].

2 Experiment Description

2.1 Drop Tower Bremen

The Drop Tower Bremen is a 146 m high building located at the campus of the University of Bremen, Germany. Inside capsules weighting up to 500 kg, experiments can be carried out under conditions of weightlessness during 4.74 ses in free fall inside a 100 m high drop tube. The residual disturbing acceleration is reduced to 10^{-5} to 10^{-4} m/sec^2 ($\approx 10^{-5}$ to $10^{-6} g$, where g is the local attractive acceleration of the Earth gravitational field) below the frequency range of 100 Hz. This is achieved essentially by pumping down the entire drop tube inside the concrete tower to a vacuum pressure of 1 to 10 Pa. By using a special release system for the drop capsule, the rotation of the capsule during free fall can be minimized down to 0.5°/sec. To minimize the transfer of forces induced by wind, the drop tube is free–standing on a 2 m–thick concrete ceiling without any link to the surrounding concrete shaft. The deceleration chamber, 10 m in height and 9 m in diameter, is integral with the tower foot. After free fall in the Earth's gravitational field under vacuum the capsule's velocity is about 46.5 m/sec; for breaking, the capsule falls into a tank 8 m high filled with 30 m^3 fine graded polystyrene. Although the impact seems to be a strong mechanical shock, the deceleration system in fact provides a very soft landing. The deceleration process

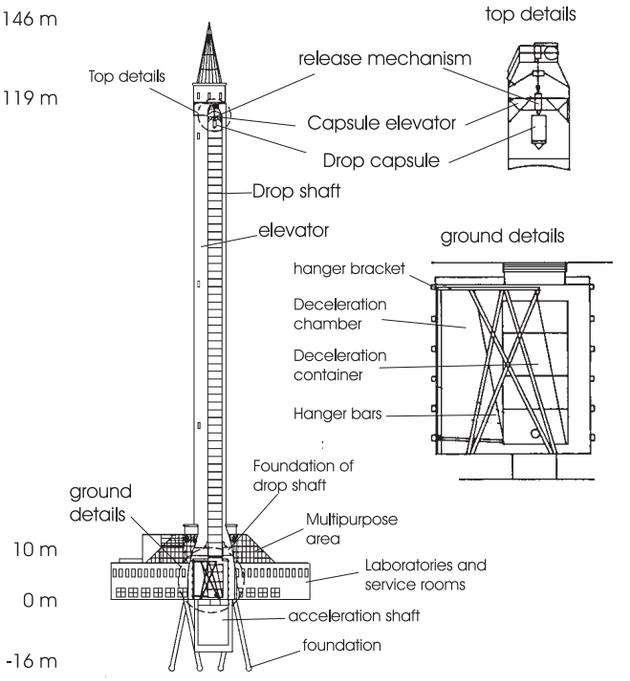


Fig. 1. Drop Tower Bremen.

lasts about 0.2 sec with an average deceleration rate of 250 m/sec^2, the maximum peaks do not exceed 500 m/sec^2.

Drop capsules are cylindrical in shape; 3 m high and 80 cm in diameter. At the front of each capsule, a cone 50 cm long is mounted to stabilize it during impact deceleration. The experiment is mounted on horizontal platforms inserted in a system of 4 vertical stringers. After assembly and final tests a vacuum–tight cover is slipped over the stringer structure. Normal pressure conditions inside the capsule are helpful, because additional provision for adaptation of the experimental equipment to vacuum conditions can be avoided. The experiment platforms are of "sandwich" construction, made from wood covered by thin aluminum plates, and have excellent damping characteristics. At the top of the tower, the capsule is connected by an umbilical cord and helium vent lines. Just before release, the umbilical is drawn and the vent lines become closed to avoid any disturbance to the freely falling capsule [7].

In optimizing the release process, not only the possible rotation of the falling capsule must be considered but also the disturbing structure vibrations induced during release. Rapid release causes oscillations of the structure along its cylindrical axis. These oscillations, as a result of relaxing the stretched structure, do not depend on any properties of the structure's material, which only influences the natural frequency of the structure, and must be damped within ade-

quate time (ca. 0.15 sec). Beside release process and air drag already mentioned, some other physical reasons exist which limit the the possibility of attaining real weightlessness in the sense of total compensation of the gravitational force and disappearance of any disturbing acceleration. In [8], disturbance effects are classified and listed. Nevertheless, carrying out a gravitational experiment at Drop Tower Bremen needs careful analysis of further disturbing effects which will be described below.

2.2 Experimental Set–Up and Timing

The drop tower experiment yields to compare the gravitational accelerations acting on two test masses falling freely in the Earth's gravitational field [9]. The test masses of different composition (e.g. niobium, aluminium, or lead) are hollow cylinders aligned along their vertical axes to place the centres of mass (CM) at the same point. Position differences between the test masses caused by, no matter what, will lead to a differential acceleration due to the Earth's gravitational field, which could mask a possible material dependent difference of accelerations by magnitudes. Therefore in the ideal cylindrical constellation (both CMs the same), both cylinders suffer the same acceleration by the Earth's gravitational field and their mutual gravitational attraction will become zero. The experiment is carried out in a small high vacuum container cooled down in a liquid helium Dewar vessel to a temperature of 4.2 K. Because the vacuum inside the tower's tube is insufficient to attain a residual acceleration level of $< 10^{-6}$ m/sec^2 ($\approx 10^{-7} g$), the Dewar is fixed on a free–flyer platform with a diameter of 460 mm and a height of 1190 mm. Handling the entire experimental apparatus inside the drop capsule as a free flyer additionally released just after the drop capsule's release, influences by air drag can be minimized.

Both test masses become adjusted with an accuracy of 1 μm by a position coil system at top and bottom of each test mass. After positioning which needs about 0.5 sec only, the test masses are released inside the vacuum container by shut–off of the position coil system. Vacuum is needed to minimize mechanical noise due to Brownian motion of the gas atoms surrounding the test masses; the resulting drift velocities of < 10 μ m/sec during free fall are small enough to neglect air drag effects inside the vacuum container. Cylindrical meander coils surrounding the test masses, which have a superconducting coating, levitate them along the two axes perpendicular to the cylindrical symmetry axis. Levitation is based on the phenomenon of magnetic flux exclusion from the bulk of a superconductor, called the Meissner–Ochsenfeld effect.

The tiny differential displacements and accelerations are measured by SQUID–based sensors. SQUIDs sense the relative displacement of the test masses as an inductance change with ultra high accuracy of 4×10^{-14} m/$\sqrt{\text{Hz}}$. In principle, a pair of pick–up coils, wound in series opposition, are placed on either side of a thin superconducting diaphragm covering each test mass. When the pick–up coil system is connected across the input coil of the SQUID any small movement of the test mass results in an output signal of the SQUID. Each test mass acts as a superconducting tuning slug changing the inductance of the pick–up coils placed

in front of the top and bottom area of the cylindrical test masses. In the initial position a persistent current has to be trapped into a superconducting input circuit with definite flux. A motion of the test masses modulates the inductivity of the pick–up coils and, since the flux in the superconducting circuit is constant, forces an additional screening current through the pick–up circuit which will be detected by the SQUID. Such high precision measurements need robust SQUIDs unaffected by repeated thermal cycling, with an extremely low noise especially at low frequencies, excellent long term stability of all parameters, and immunity to mechanical shocks, vibrations, and r.f. interference.

After a drop time of about 4.3 sec, the experiment is finished by clamping and locking the free flyer platform and re–caging the test masses to avoid any destruction of the experimental equipment.

2.3 Main Error Sources

The accuracy of the experiment depends directly on the precision of the test masses' alignment and orientation after release. Disturbances during release result in initial displacements ($\Delta z \neq 0$) and initial velocities ($\Delta v \neq 0$) causing an acceleration difference of the test masses Δa_{grad} due to the Earth's gravity gradient and the mutual gravitational attraction of both test masses, respectively. This relative acceleration may dominate the potential material dependent acceleration Δa_{WEP} by orders of magnitudes. To elude this problem by taking into account the desired accuracy of the apparatus, boundary conditions for the experimental set–up can be derived in the following manner:

1. Choose of proper cylinder materials (appropriate for machining), and
2. Optimizing the geometry of the cylinders by keeping the contribution of Δa_{grad} clearly below Δa_{WEP}.

The initial values of the maximum acceptable relative displacement and velocity of the test masses are directly connected to a proper choice of the cylinders' materials, with respect to their densities, and their geometry. With suitable values it is possible to compensate the differential acceleration of the inhomogeneous Earth's gravity field by the mutual gravitational attraction of the test masses. The motion of the free falling test masses is described by two coupled differential equations (assuming a motion only in z–direction), which have to be solved numerically:

$$\begin{aligned} a_1 &= -\partial_z \Phi_{\mathrm{earth}} - G\partial_z \int \frac{\rho_2}{z_1 - z_2} + \varepsilon, \\ a_2 &= -\partial_z \Phi_{\mathrm{earth}} - G\partial_z \int \frac{\rho_1}{z_2 - z_1}\,, \end{aligned} \tag{2}$$

where $a_{1,2}$ are the test masses accelerations, $z_{1,2}$ their coordinates in fall direction, and $\rho_{1,2}$ their densities, respectively. Φ_{earth} is the Earth's gravitational potential, and G Newton's gravitational constant. A material dependent acceleration of the test masses is taken into account by adding ε to the equation of

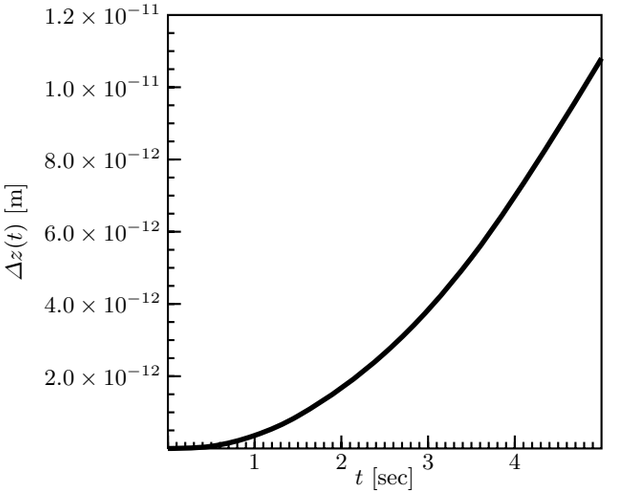


Fig. 2. Displacement $\Delta z(t)$ of the test masses in presence of a material dependent gravity force ($\varepsilon \neq 0$) taking into account an initial displacement $\Delta z_0 = 10\ \mu$m and an initial velocity $\delta v_0 = 10\ \mu$m/sec.

motion of one test mass. The second term of eq.(2) is given by:

$$\partial_z \Phi^i_{\rm cyl} = G\pi\rho_i\, f(L_j, z_i - z_j, r^j_{\rm in}, r^j_{\rm out})\,, \quad i = 1,2, \;\; i \neq j\,, \tag{3}$$

where L denotes the cylindrical length, $r_{\rm in}$ and $r_{\rm out}$ the inner and outer radius of the cylinders, respectively, and ρ the density. $\Phi^i_{\rm cyl}$ is the gravitational potential of the i–th cylinder. Eq. (3) shows that optimization is a pure geometrical problem and could be solved for each test mass pair individually. In practice, we choose only one geometrical configuration for all test mass pairs and try to compensate the incomplete compensation by perfect release and positioning of the test masses. For a detailed problem description, see [10]. If initial displacement and velocity do not exceed certain values of Δz_0 and Δv_0 (declined by the cylinder design), fictitious position sensitive detectors with a resolution capability of $\Delta z \approx 10^{-13}$ m show a motion of two test masses which do not interact gravitationally and move coincidently in a quasi–homogeneous gravity field. Hence, the differential motion looks linear and a WEP–violation should be observed as a relatively accelerated motion.

Assuming a material dependent acceleration difference ($\varepsilon \neq 0$) of $\Delta a/a = 10^{-13}$ an approximation of the additional ε–term is given by

$$\varepsilon = 10^{-13}\frac{GM_{\rm earth}}{z_1^2}\,, \tag{4}$$

where $M_{\rm earth}$ denotes the Earth's mass. The numerical solution of the differential equation system (2) neglecting the cylinder–cylinder gravitational interaction is shown in Fig.2. Acceleration difference and displacement are corrected as described above. To test the WEP with an accuracy of 10^{-13} by use of optimally

designed test masses[1] (see Fig.3), the effecting resolution of the position sensing SQUID–based system has to be $\Delta z \leq 10^{-12}$ m. Disturbances due to a non–momentum free release resulting in different initial velocities of the test masses may be acceptable if they do not exceed $\Delta v_0 = 10$ μm/sec. The maximum displacement tolerance of the CMs of the test masses is $\Delta z_0 = 10^{-5}$ m without any loss in performance. These requirements are in accordance with our experimental results for positioning the test masses.

Temperature drifts inside the vacuum container during the measurement would lead to a linear expansion of the test masses superposing the displacement measure, and temperature gradients in the vacuum container may cause a non–symmetric temperature gradient inside the test masses effecting a gravity gradient dependent influence on the free fall motion of the test masses. Considering the maximum acceptable tolerance of the CM displacement of 10 μm during free fall, temperature must be stable within an accuracy of ± 0.1 K. Measurements showed that the temperature is increased by an amount of < 0.05 K only during the drop time. The temperature increase is correlated by a small He–pressure increase inside the Dewar vessel, because the exhausting valve has to be closed during free fall to avoid a reduction of the quality of the microgravity. Only a slight increase of the helium level (ca. 8.5 %) is observed.

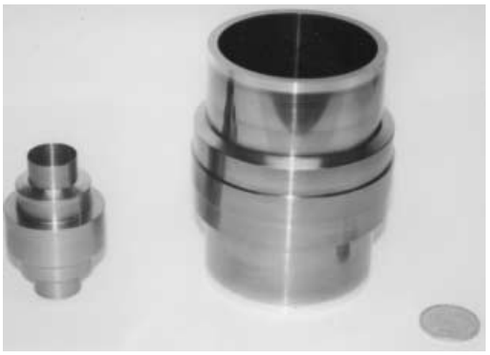


Fig. 3. Test masses: inner test body made of special lead alloy (left); outer test body made of aluminium coated with niobium (right). (For comparison: a coin of 1 Deutsche Mark at the lower right.)

In addition, several experiments have been performed to measure the relatively strong magnetic field components B_z (vertical), B_y (horizontal), and the vertical gradient $\mathrm{d}B_z/\mathrm{d}z$ inside the steel tube during free fall. The diameter of the pick–up coils connected across the SQUID input was 1 mm only. The magnetic field pattern observed (see Fig.4) was caused by the welding seams and changing magnetic remanence of the tube segments. After shielding the experiment perfectly by superconductors, the disturbing magnetic field signals disappear.

[1] In inhomogeneous gravitational fields the acceleration of a test body depends not only on the position of its center of mass but also on its shape. The geometry of both test bodies were numerically optimized in such a way that these test bodies move with nearly the same acceleration in the inhomogenous gravitational field of arbitrary disturbing masses nearby provided, of course, that the centers of mass of the test bodies are equal.

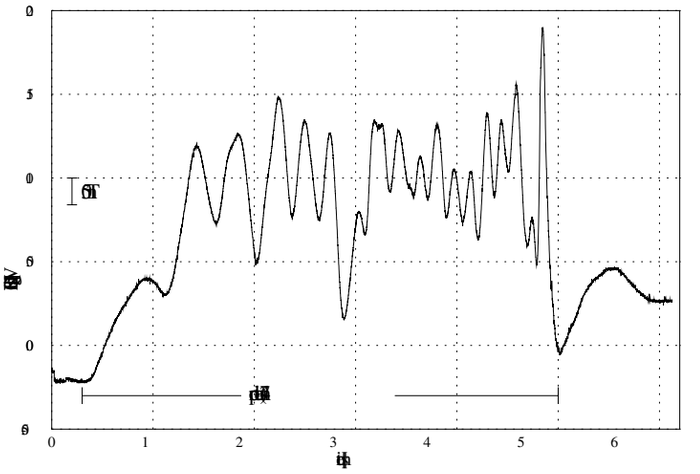


Fig. 4. Magnetic field inside the drop tube (measured during a free fall experiment using a SQUID–based magnetometer).

Principle limitations of the resolution and systematic errors of the SQUID–sensor and the sensing system are discussed separately below.

3 SQUID Based Position Detector

In order to realize the precision measurement of tiny displacements of free falling test bodies, highly sensitive SQUID–based measurement equipments are required. The Superconducting Quantum Interference Device (SQUID) is the most sensitive magnetic flux detector known today. The operation of SQUIDs is mainly based on two effects, observable only at low temperature in the presence of superconductivity:

- flux quantization in superconducting loops and
- Josephson effects.

SQUIDs has often been referred to as a cryogenic or superconducting magnetometer, and indeed the measurement of extremely small magnetic fields is one of the most important applications. However, the scope of the SQUID's usefulness extends far beyond simple field measurements. Almost any low–frequency signal that can be converted into a corresponding magnetic flux signal will be detected at a greater level of sensitivity with a SQUID than with any other instrument.

Applications of SQUID systems have ranged from the investigation of magnetic and electronic properties of materials to biomagnetic research, and from the measurement of millikelvin temperatures to the detection and characterization of subsurface magnetic structures on land and at sea. In recent years the

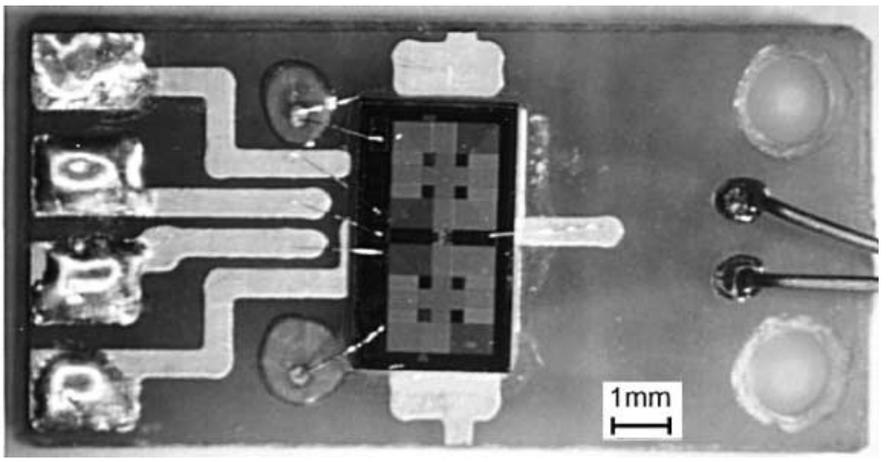


Fig. 5. DC SQUID *UJ 111*. Chip fixed on a printed circuit board with contact pads.

application of high performance SQUID measurement systems in fundamental experiments like the proof of the universal proportionality of inertial and gravitational mass is one of the most promising field in modern physics.

3.1 The DC SQUID

The magnetic flux sensor employed in this work is an eight–loop thin film DC SQUID *UJ 111* [11] based on Nb–NbO_x–Pb/In/Au window–type Josephson tunnel junctions developed and fabricated at the Department of Physics of Friedrich–Schiller–University Jena. In contrast to other sensors the SQUID *UJ 111* was designed for universal applications in precision measurement technique and works at an extremely low noise level also in a magnetically unshielded environment.

Each Josephson junction is resistively shunted by strips of Ag/In alloy having a resistivity of 2 Ω at 4.2 K. The integrated input transformer consists of two coils

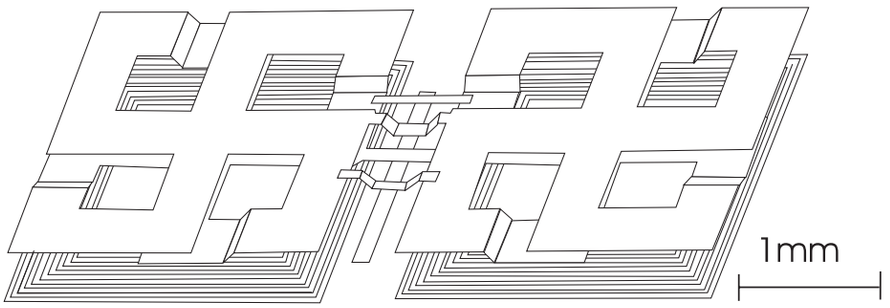


Fig. 6. Simplified layout of the thin film DC SQUID *UJ 111*. See text for details.

Table 1. Parameters of the DC SQUID *UJ 111* (flux quantum Φ_0=2.07×10^{-15} Vs).

Parameter	Value
SQUID inductance L_{SQ}	40 pH
Input inductance L_i	0.8 μH
Junction shunt resistance R	2 Ω
Junction capacitance C	$\approx$ 0.6 pF
Stewart–McCumber parameter β_c	< 0.5
SQUID parameter β_L	1
Junction critical current I_c	3 - 30 μA
Number of turns in input coil N_1	2×18
Coupling coefficient k between SQUID and input coil	0.9
Input current sensitivity $\Delta I_i/\Delta\Phi$	0.4 μA/Φ_0
Modulation current sensitivity $\Delta I_m/\Delta\Phi$	20 μA /Φ_0

of 18 turns each, connected in a gradiometric configuration, providing an input inductance of about 0.8 μH. The one turn flux modulation coil is inductively coupled only to one half of the gradiometric SQUID loop system. The contact pads for bias current, voltage measurement, and flux modulation are made of niobium and are connected with the printed circuit board of the sensor using 20 μm thick aluminium wire and bond techniques. Niobium wire (thickness: about 25 μm) is also used for the bonded interconnections between the input coil and the niobium input leads of the SQUID. The complete SQUID has been integrated on a 3 mm × 5 mm silicon chip and is placed on a printed circuit board with an overall size of 8 mm × 16 mm (see Fig.5).

A simplified structure of the integrated DC SQUID sensor, as described above, is shown schematically in Fig.6. An overview on some important electrical parameters of the applied SQUID type is given in Table 1.

The long term stability of the parameters given during a time period of several years is remarkable although the SQUID described is not encapsulated hermetically. According to our experience there is no influence on the SQUID parameters even after more than 100 cooling down cycles. It should be pointed out that the SQUID sensors are also very insensitive to mechanical shocks tested by free–fall experiments over a height of up to 110 m causing a deceleration of up to 500 m/sec^2 (50 g) at the end of the flight.

3.2 SQUID Control Unit

The SQUID electronics consist of the low noise preamplifier and the SQUID control and detector unit (see Fig.7). The low source impedance of the SQUID (about 1 Ω) is stepped up to the optimum impedance of the preamplifier with the help of a matching transformer. The first stage of the preamplifier consists of two low noise junction field effect transistors (i.e. Toshiba 2SK146) coupled in parallel with a noise optimum source impedance within the range of 1 and

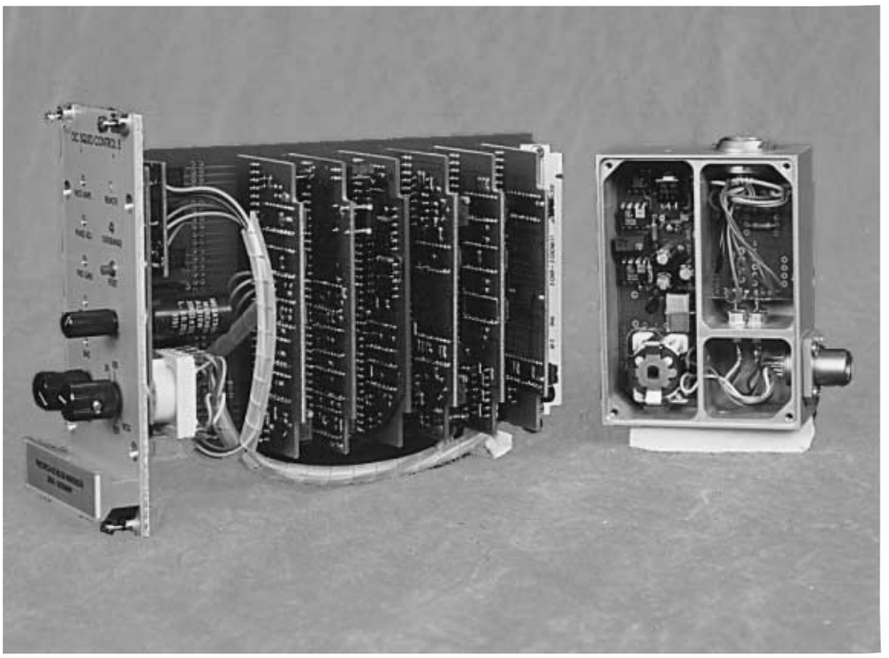


Fig. 7. Photo of the SQUID Control unit with preamplifier.

100 kΩ. Thus, the matching condition for low noise operation is well satisfied. The choice of the cables connecting the transformer with the SQUID and the preamplifier, respectively, is usually rather critical. Twisted pairs of teflon insulated copper wires were used to keep the contribution of current noise from the cables negligible. Each pair is installed in a grounded Cu–Ni–tube (diameter: 1 mm) preventing disturbances by RF interferences. The contribution of the SQUID preamplifier to the total voltage noise observed was negligible in all cases reported.

The d.c. bias and flux modulation current (f = 125 kHz) are fed into the SQUID via voltage–controlled current sources situated in the preamplifier case. The amplification and detection of the SQUID signal is achieved by the state–of–the–art design (Fig.8), i.e. the preamplifier is followed by an AC amplifier and phase sensitive detector with a PI–type integrator. The output signal returns via a resistor RF to the modulation coil to close the feedback loop. Furthermore, the SQUID control unit has three different modes of operation:

- the test mode,
- the small signal measurement mode, and
- the flux–locked–loop mode.

All experiments were usually performed with standard cryogenic equipment at liquid helium temperature. No special care was taken to stabilize the bath

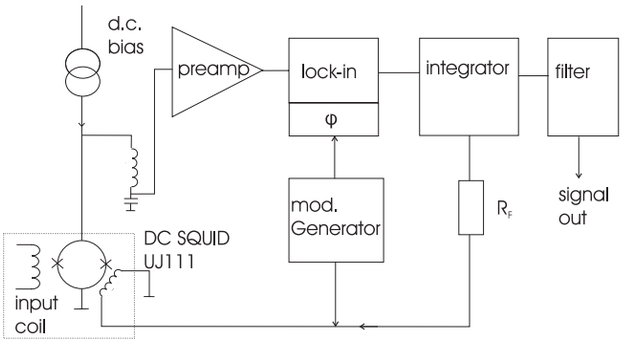


Fig. 8. Block diagram of the DC SQUID Controller.

temperature. The cryostat was especially developed for drop tower experiments and contains a maximum of 20 l liquid helium. All parts of the dewar were made of nonmagnetic stainless steel. As radiation shields, layers of 25 nm thin semireflecting aluminium coated Mylar foil were used. In this configuration an operating time of more than 10 hours is available.

The ground based measurements to test the performance of the SQUID measurement system and the position detectors were performed in a open laboratory at the Friedrich–Schiller–University Jena. The low frequency noise spectra (Fig.9) were measured with a Hewlett–Packard HP 3582A Fast Fourier Transformation (FFT) spectrum analyzer using the Hanning window. The white noise levels were obtained by calculating the FFT. To attain the result in units of the flux quantum, the spectral density of the voltage was divided by the change of the output voltage corresponding to one flux quantum Φ_0 measured before. In noise measurements the SQUID was operated with a peak–to–peak modulation of $\Phi_0/2$. In general, the optimum noise performance was obtained when the SQUID was biased slightly above the critical current.

For an optimum choice of bias and flux modulation point, a white noise flux spectral density of $2 \times 10^{-6}\,\Phi_0/\sqrt{\text{Hz}}$ for the SQUID system was found when the antenna was not connected but the SQUID input coil was shunted. This flux noise corresponds to an equivalent current noise through the input coil of 0.9 $\text{pA}/\sqrt{\text{Hz}}$, an effective energy factor of 543 h (h = Planck's constant), and an energy resolution of 3.6×10^{-31} $\text{J}/\sqrt{\text{Hz}}$ [11].

3.3 Detector Principle

The main components of a SQUID position detector consist in a superconducting pick–up coil and a superconducting diaphragm covering the test body. When this pick–up coil is connected across the input coil of the SQUID and a definite current is trapped in this closed superconducting loop any small movement of the test mass results in an output signal of the SQUID.

We started our investigations with a position detector using one pick–up coil only [12]-[14]. The basic scheme of an improved version of a position detector

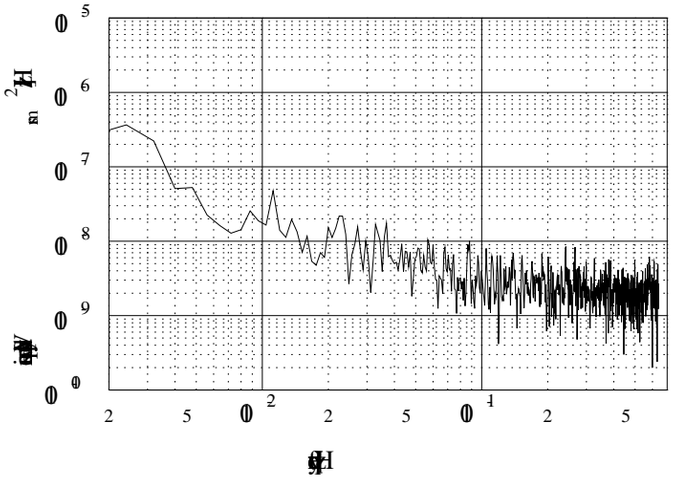


Fig. 9. Low frequency noise of the Jena DC SQUID system with the sensor UJ 111. For this measurement the sensitivity of the SQUID system was adjusted to 10 V/Φ_0.

with two pick–up coils is shown in Fig.10. The test mass changes the inductance of both pick–up coils L_1 and L_2. In the initial position a persistent current I_0 has to be fed across heat–switch S_1 into the superconducting pick–up circuit causing a definite magnetic flux. During this procedure the input coil of the SQUID is decoupled by heat–switch S_2.[2] The motion of the test mass modulates L_1 and L_2 and, since the magnetic flux in superconducting loops is constant, the variations of the inductance of both coils force a screening current I_i corresponding to the displacement of the test body Δx to flow in L_i.

The circuit can be described by the following equations:

$$I_2 = I_1 - I_i \tag{5}$$

$$\Phi_{\mathrm{II}} = L_2 I_2 - L_i I_i \tag{6}$$

$$\Phi_{\mathrm{I}} = L_1 I_1 + L_i I_i \tag{7}$$

Eqn. (5) comes from the current conservation. Equations (6) and (7) come from the flux conservation in two superconducting loops of the circuit, where Φ_{I} and Φ_{II} are the fluxes trapped in these loops. These equations correspond to the

[2] The heat switches consist of a small heater (a special non–magnetic resistor) and a thin superconducting wire (niobium) possessing a good thermal contact to the heater. The superconduting wire is connected with the remaining superconducting circuit whereas the heater is insulated electrically from the circuit and has its own power supplies (not shown in Fig.10). If a current flows through the heater, the temperature increases and the superconducting wire becomes normal conducting. Since the remaining circuit is superconducting, no current flows through the normal conducting part. If the power supply of the heater is disconnected, the switch becomes superconducting again, and the circuit preserves its magnetic flux distribution.

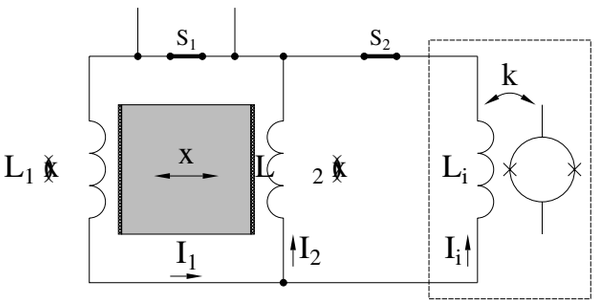


Fig. 10. Extended position detector circuit with DC SQUID.

equations for the voltage in room temperature circuits. The solution for the current I_i is

$$I_i = \frac{\Phi_{\mathrm{I}} L_2 - \Phi_{\mathrm{II}} L_1}{L_i(L_1 + L_2) + L_1 L_2} \quad . \tag{8}$$

The sensitivity of the detector depends not only on the magnitude of the fluxes but also on their sign. In order to get a maximal sensitivity for displacements in the x–direction and a minimal sensitivity perpendicular to x the fluxes Φ_{I} and Φ_{II} should have equal magnitudes and signs. Two switches are necessary to feed two defined fluxes into the circuit (Fig.10). The current should be supplied across switch S_1, when the test mass is centered ($L_1 = L_2$). Assuming a linear ansatz for the inductance (corresponding to the dependence of plunger type coils)

$$L_1(x) = L_0 + L'x \; , \qquad L_2(x) = L_0 - L'x \; , \tag{9}$$

and considering the initial conditions

$$\Phi_{\mathrm{I}} = L_0 I_0 \; , \quad \Phi_{\mathrm{II}} = L_0 I_0 \; , \tag{10}$$

we get for the current I_i:

$$I_i = \frac{-2\,L'\,x\,I_0}{2L_i + L_0 - (L'x)^2/L_0} \approx \frac{-2}{2L_i + L_0}\; L' I_0 x \; , \tag{11}$$

where I_0 is the current fed into the circuit at position $x = 0$. The dependence $I_i(x)$ is linear for small displacements x. The ansatz (9) describes approximately the inductance of plunger–type pick–up coils in their optimal working range (see also section 4.1). In order to get a detector with a linear dependence $I_i(x)$ within a larger range one has to use coils with a special dependence of the inductance:

$$L_1(x) = \frac{1}{u + vx} \; , \qquad L_2(x) = \frac{1}{u - vx} \; , \tag{12}$$

where u and v are arbitrary parameters. However, it is difficult to produce wire wound coils possessing such a dependence over a wide working range.

According Eqn. (11) the detector measures the absolute position from a fixed reference point, where $I_i = 0$. However, this reference point is not the centered position of the test body but depends on the position of the test body at the time the flux was fed into the circuit. Therefore, the reference point varies slightly each time a new current is fed into the detector circuit. Therefore we get a high accuracy of measurement of relative displacements whereas the accurracy of the absolute position depends on the reproducibility of the supplied fluxes and is much lower.

The current I_i is measured by the SQUID. As we described above we use a flux locked loop electronics for readout. This electronics react to the detector circuit by an integrated feedback coil. As long as the feedback loop of the SQUID electronics is closed the reaction can be taken into account by using a smaller effective inductance for L_i instead of the real inductance of the SQUID input coil. This corresponds to the behaviour of a transformer where the effective inductance of the primary coil depends on wether the secondary coil is open or bypassed. But note that the problem becomes more complicated if the feedback loop of the elektronics is opened for a moment.

The most important advantage of this detector, compared with a detector with a single pick–up coil, consists in the fact that the current I_0 is not limited by the critical current of the input coil of the SQUID [14], as long as x will be small. According to 13, a higher current I_0 will provide a much better resolution δx of the detector:

$$\delta x = \frac{2L_i + L_0}{2k\sqrt{L_i L_{\mathrm{SQ}}}} \, \frac{1}{I_0} \, \frac{1}{\alpha} \, \delta\Phi_{\mathrm{SQ}} \; , \tag{13}$$

where L_0 is the inductance of both pick–up coils in the initial position of the test body (middle position), k is the coupling factor between input coil and SQUID loop, L_{SQ} is the SQUID inductance, $\delta\Phi_{\mathrm{SQ}}$ is the flux resolution of the SQUID, and $\alpha = |dL_1/dx| = |dL_2/dx|$.

In order to characterize a detector or a converter it has to be distinguished between sensitivity and resolution. The sensitivity is the ratio of the output signal (e.g. the output voltage of the SQUID electronics) to the input signal (e.g. the displacement). The resolution is the smallest difference of two values which can just be distinguished. Often, the resolution is limited by the noise of the detector. In this case, the absolute resolution can be increased by averaging over a long time. Therefore, the so–called noise limited resolution is used to characterize these detectors. This is the product of absolute resolution and the square root of the averaging time (one divided by square root of measurement bandwidth). For instance a position resolution of $\delta x = 10^{-12}\, m/\sqrt{\mathrm{Hz}}$ means:

- If you measure the position of a test body at two different times by averaging the signal over 1 s per measurement point (or by using a low pass filter of 1 Hz), then you can dectect displacements of $\delta x \geq 10^{-12}$ m
- If you average over a time period of 10 ms per measurement point, the smallest detectable displacements is 10^{-11} m only.

If the signal changes during the averaging time or the noise of the detector is not white, the problem becomes more complicated.

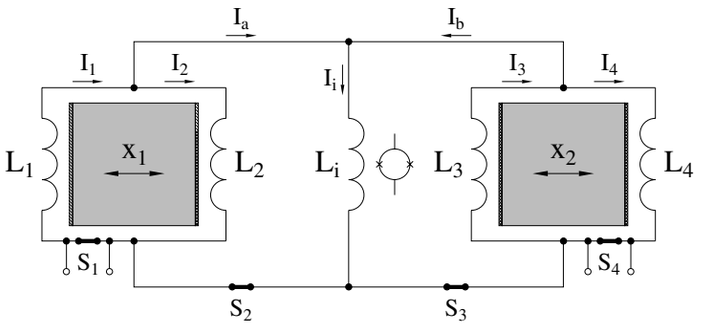


Fig. 11. Differential detector with two test masses.

The resolution of a detector depends on its sensitivity. A good resolution can only optained, if the sensitivity is sufficient high. However, an increasement of sensitivity (e.g. by amplification of the output voltage) not necessarily improves the resolution, because both signal and noise will be increased.

The position detector interacts mechanically with the test mass because each pick–up coil acts like a small magnet which repels the superconductive covered mass. For the centered position of the test mass both forces compensate each other. For this position we can calculate a spring constant by twice differentiating the energy in the circuit:

$$E = \frac{1}{2} \frac{(\Phi_{\mathrm{I}} + \Phi_{\mathrm{II}})^2 L_i + \Phi_{\mathrm{I}}^2 L_2 + \Phi_{\mathrm{II}}^2 L_1}{L_1 L_2 + L_1 L_i + L_2 L_i} \ . \tag{14}$$

For most coils the spring constant is positive but for certain geometries also negative spring constants are possible. For plunger–type coils negative spring constants can occur if both ends of the test mass are completely inside the two pick–up coils. After a small displacement one coil pushes the test mass further into the other coil. The gradient of the gravitational field of the earth causes an effect which can be described by a negative spring constant. The resulting total spring constant can be compensated by additional levitation coils.

A second important parameter of the detector, besides the resolution, is the dynamic range. This means the ratio of working range to the smallest signal which can be detected. The SQUID works up to its maximal input current of 2 mA (corresponding to about 5000 Φ_0) without any loss of sensitivity. Above this current some parts of the input coil become normal conducting. The usable working range is limited by the SQUID electronics. Because the SQUID is designed for precision measurement, it has only a small integrated feedback coil. The maximum working range of the SQUID electronics ($\pm 100\ \Phi_0$) is limited by the relatively high feedback current needed. Even smaller working ranges of 1 or 10 Φ_0 will be used in the final experiments due to certain advantages (faster, less noise). With a resolution of $2 \times 10^{-6}\ \Phi_0/\sqrt{\mathrm{Hz}}$ for our best SQUIDs we get a dynamic range of about 10^6. This corresponds to the dynamic range of our 24–bit sigma–delta A/D converter (121 dB at 500 Hz).

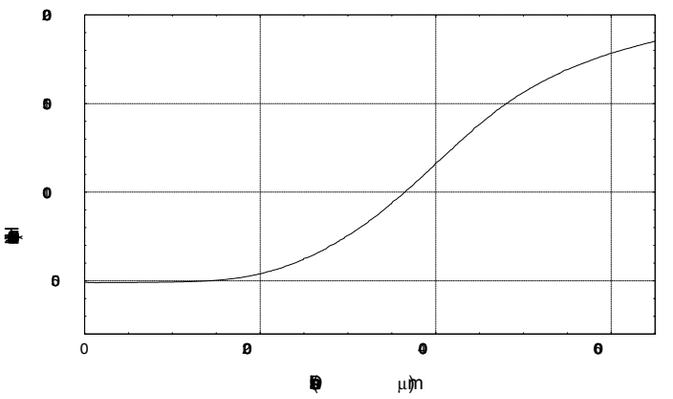


Fig. 12. Dependence of the inductance of a pick–up coil on the position of the test body. See text for details.

During the measurement time of about 4 sec the test mass should not leave the working range of the detector. For a detector with a resolution of 10^{-12} m the velocity of the test mass should not exceed 0.25 μm/sec. Before the release all parts of the capsule are tensed by their own weight. After the release the structure contracts and accelerates the test mass to a velocity of about 1 cm/sec. Therefore the test mass has to be fixed during the release of the capsule and after the vibrations have faded away it has to be placed in the centered position. This can be done by mechanical means (piezo elements) or with additional suspension coils. In our recent experiments we measure the displacement of two simultaneously free falling test masses. A violation of the WEP corresponding to an Eötvös factor $\eta \neq 0$ will cause a displacement Δx between the two test masses during the fall. If both test bodies start with the initial velocity of $v_0 = 0$ then this displacement is:

$$\Delta x = \eta \, s_0 \; , \tag{15}$$

where s_0 is the drop length.

The vertical gradient of the gravitational field of the earth causes a displacement of bodies if their centres of mass are not at the same height. Therefore the centres of mass of the two test bodies have to be in coincidence with an accuracy of < 1 μm. The position detectors of the two test bodies can be combined in a circuit [15], as shown in fig. 11. The current in the SQUID input coil is:

$$I_i = \frac{\Phi_1 L_1^{-1} - \Phi_2 L_2^{-1} + \Phi_3 L_3^{-1} - \Phi_4 L_4^{-1}}{L_i (L_1^{-1} + L_2^{-1} + L_3^{-1} + L_4^{-1}) + 1} \; , \tag{16}$$

where Φ_m $(m = 1, \ldots, 4)$ is the flux in the loop consisting of pick–up coil L_m and the SQUID input coil L_i. It is possible to balance the circuit by adjusting the ratio of the two currents fed into the circuit (across switch S_1 and S_4, respectively). For such a balanced detector the current I_i would be proportional

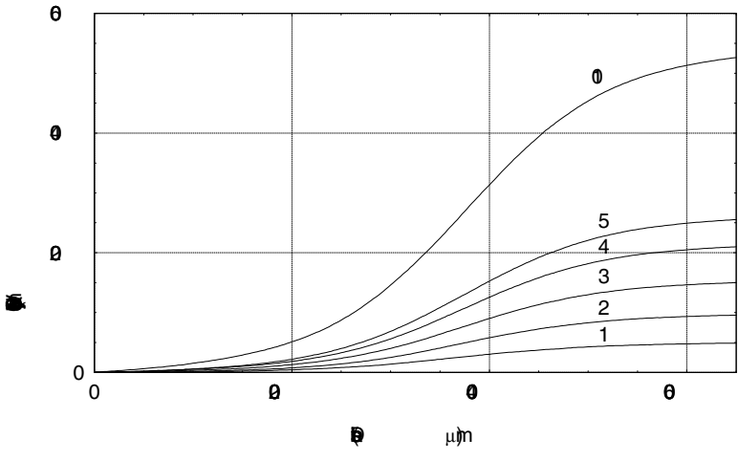


Fig. 13. SQUID system output $U(x)$ in dependence on the position of the test body for various currents I_0 in the pick–up circuit. The numbers at the curves indicate the amount of the current I_0 in mA.

to the displacement $(x_1 - x_2)$ of the two test masses, but only for small displacements x_1 and x_2. The detector can be balanced for all positions x_1 and x_2 using inductance characteristics corresponding to (7), but geometries having such dependencies in a large range are difficult to realize.

4 Experimental Results

4.1 Inductance Measurements

Beside theoretical determinations of L_0 and dL/dx we used a special experimental equipment to measure the inductance of a pick–up coil in dependence on the distance of a cylindrical test body with the help of a commercial LCR meter. An example of the characteristics measured is given in Fig.12. In this case the coil was hand–wired from a 0.3 mm niobium wire applying 8 turns. The geometry of the coil and the body was a plunger–type one. The inductance in the initial position ($x = 4000$ μm) is 1.2 μH and α is determined by differentiating the curve in Fig.12 to be 0.45 nH/μm. These results are in good agreement with the expected values.

4.2 Performance of the Detector

The performance of the DC SQUID position detector was investigated in ground–based experiments. Therefore a special measurement equipment allowing us to move the test body relative to the pick–up coils was used. It consists of a cylindrical test body with superconductive coated jacket and the two pick–up coils at each end. In order to determine the sensitivity $\beta = dU/dx$ of the detector numerous characteristics $U(x)$ for different currents I_0 in the pick–up circuit were

recorded. Fig. 13 shows an example for these characteristics. The unusual high voltages come from editing the curves to avoid the reset branches of the SQUID controller if the output voltage exceeds the maximum value by shifting segments of the curves in the voltage direction. According to Eq. (13), the sensitivity increases with higher currents in the pick–up circuit. Differentiating these curves one can get the maximum sensitivity which also increases with the current I_0. This dependence is shown in Fig.14.

On the basis of the measured sensitivity, the noise limited resolution of the detector (Fig.15) can be specified using the flux resolution $\delta\Phi_{SQ}$ of the SQUID and the transfer function $dU/d\Phi$ of the SQUID system. This correlation is expressed by the following equation:

$$\delta x = \frac{1}{\beta}\,\frac{dU}{d\Phi}\,\delta\Phi_{\mathrm{SQ}}\ . \tag{17}$$

Inserting the flux resolution of $6\times10^{-6}\,\Phi_0/\sqrt{\mathrm{Hz}}$ (including some environmental noise, in contrast to the value $2\times10^{-6}\,\Phi_0/\sqrt{\mathrm{Hz}}$ described in section 3.2) and the flux sensitivity of 2.5 V/Φ_0 of the SQUID system used one obtains a position resolution of

$$\delta x = 4\times10^{-14}\frac{\mathrm{m}}{\sqrt{\mathrm{Hz}}} \tag{18}$$

assuming a current $I = 200$ mA through the pick–up circuit. For comparison, the radius of a uranium nucleus is in the same order of magnitude.

4.3 Free Fall Measurement System

After performing many free–fall experiments with a simplified system having only one single test body, a new system with two test bodies was completed and

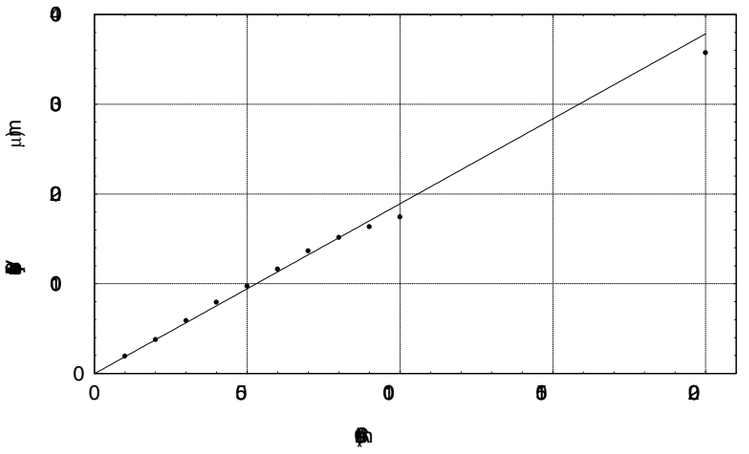


Fig. 14. Sensitivity of the SQUID position detector in dependence on the current I_0 in the pick–up circuit.

has been successfully tested in first free–fall tests. Fig.16 represents a simplified drawing of the new two–body system. The two concentric test bodies (a) and (b) are made of lead and aluminium (coated with niobium), respectively. They are stabilized in the radial direction by meander–shaped coils and can be controlled in the vertical direction by levitation coils. These coils are supported by special carriers (c), (d), (e). For instance, (f) denotes the upper levitation coil for the outer test body and (i) the lower levitation coil of the inner test body. The pick–up coils required for the position measurement are also attached close to the test bodies. For each body two pick–up coils are needed as described above. (h) shows the lower pick–up coil of the inner test body and (g) the upper pick–up coil of the outer test body. Furthermore all parts are shielded by superconducting screenings (j). The whole system is housed in a vacuum chamber immersed in liquid helium at 4.2 K.

The cryostat (20 l liquid helium), the SQUID electronics, the data processing unit, the control unit and the power supply are mounted in a special structure, the so–called free–flyer (see Fig.17), which resists the vibrations and shocks at the beginning and at the end of the drop. The free–flyer itself falls freely inside of the outer fall capsule, which contains a computer–controlled caging mechanism. Despite the weightlessness and the rapid deceleration of about 500 m/sec^2 (50 g) after numerous drop experiments neither the cryogenic equipment nor the SQUID electronics have been damaged.

4.4 Free Fall Tests of the Measurement System

Experiments were done to improve the test body position control unit. This unit fixes the test bodies at the beginning of the drop. Before release, the hanging

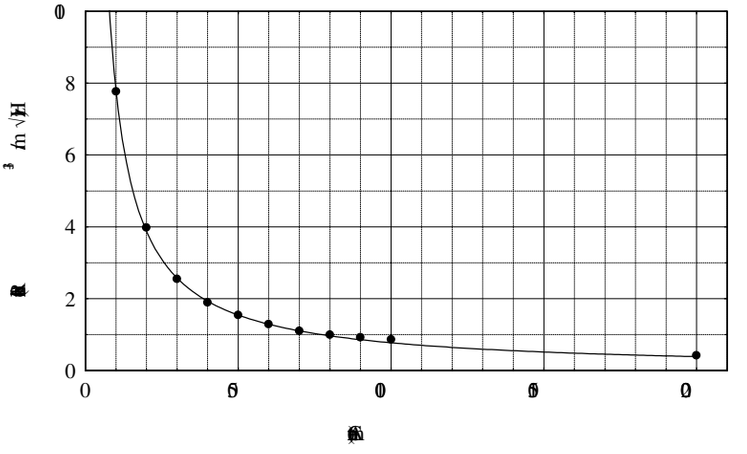


Fig. 15. Resolution of the position detector in dependence on the current I_0 in the pick–up circuit.

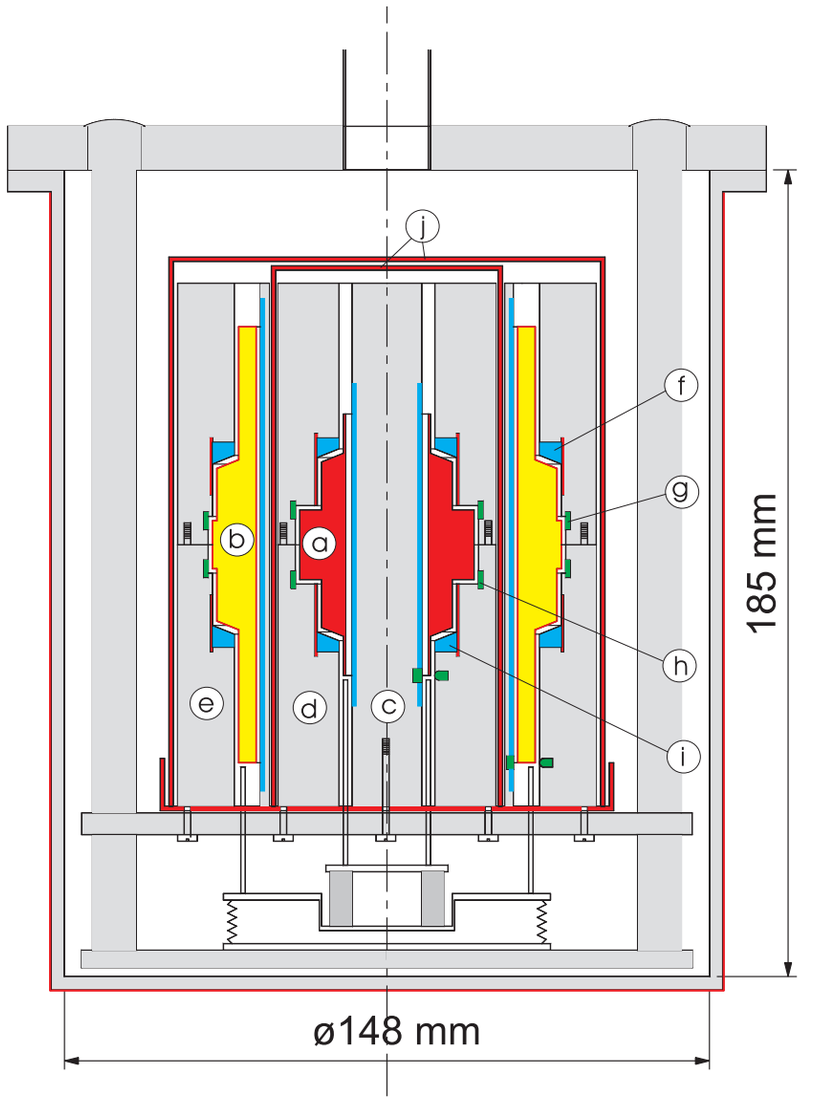


Fig. 16. Simplified assembly drawing of the cryogenic part of the measurement system for testing the Equivalence Principle including two test bodies, levitation coils, and pick–up coils for position measurement. See text for further explanation.

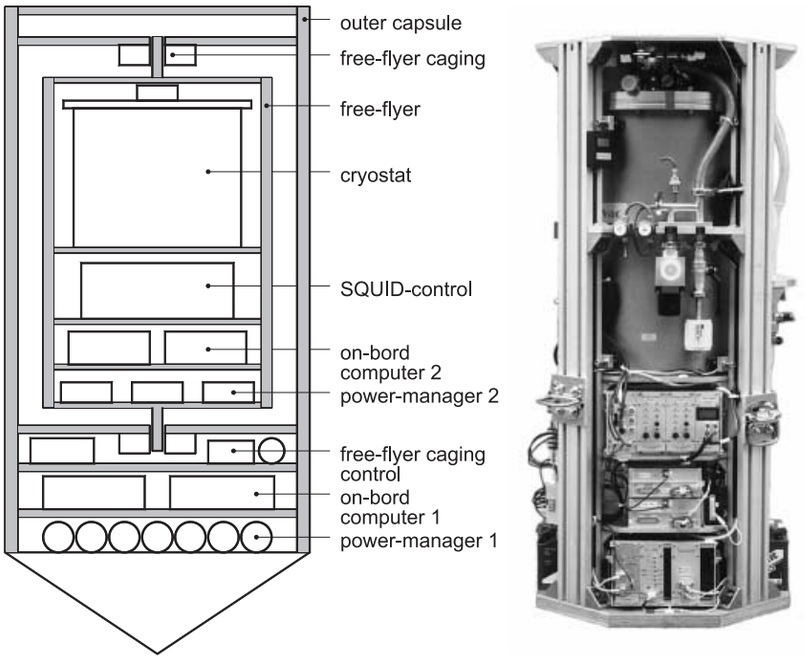


Fig. 17. Scheme of drop capsule (left) and photograph of free–flyer (right).

capsule is under tension from its own weight resulting in a contraction upon release. This would result in an upwards acceleration of the test bodies. After vibrations of the system have faded away, the control unit has to center the test bodies in their working position as fast as possible to attain a long measurement time for determination of the Eötvös coefficient. For this purpose the control unit measures the position and the velocity of the test bodies and sends short rectangular shaped current pulses of a calculated length and amplitude through the upper and the lower levitation coils, respectively. These coils situated at both ends of each test body can repel the superconducting test bodies. Theoretically, two pulses from opposite directions (upper and lower levitation coil, respectively) are sufficient to attain centering even given an initial velocity. However, after the two pulses there are small deviations from the aspired position and velocity due to errors in determination of the initial velocity and position or due to limitations of the control unit. Therefore, it will be necessary to repeat the procedure several times. For precision experiments the final velocity of the test bodies relative to the measurement system should be lower than 0.25 μm/sec to avoid exceeding the working range of the position detector (see section 3.3).

Fig.18 shows a preliminary result of a free–fall test for the position control of the inner test body. At the beginning of the drop the test body was magnetically fixed by a current of 1 A through the upper levitation coil. The following test sequence starts with a current pulse of 0.5 A through the lower levitation coil

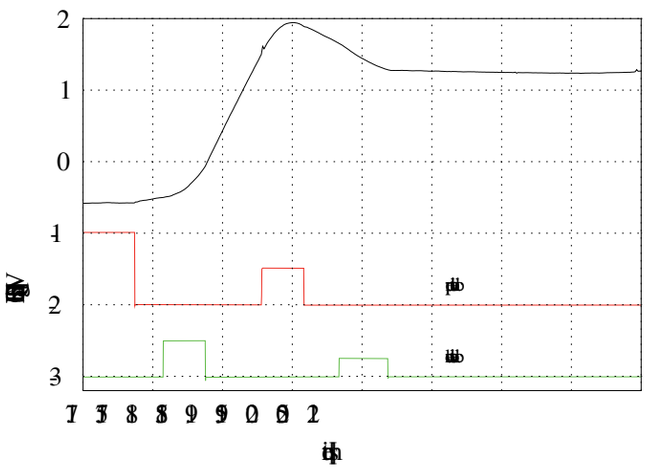


Fig. 18. Controlling of the inner test body by short pulses through the levitation coils during a drop; SQUID Signal (upper curve, 14 V ≡ 1 mm displacement); currents of the levitation coils (lower curves) measured by means of current–voltage converters (1 V ≡ 1 A).

accelerating the test body. After 40 msec the test body is decelerated again by a pulse through the upper coil. Both pulses have the same length (30 msec) and amplitude. Despite of same currents and simular geometry the upper and lower levitation coils cause slightly different forces on the test body. Therefore, the test body has not been stopped by the second pulse of the same length but has a residual velocity. This velocity is measured by the control unit and a short pulse is send to the lower levitation coil stopping the test body as it is clearly seen in Fig.17 ($t > 1.92\,\mathrm{sec}$).

After both test bodies have been stopped with a sufficient precision in the final experiments, the sensitivity of the detectors have to be switched to a higher sensitivity and resolution (but smaller working range) and the determination of Eötvös coefficient can start[3]. The position control unit is switched off for the rest of the drop.

If a higher sensitivity is used, one would see that the test bodies don't stay in rest after the controlling process, as it seems in Fig.18, but appear as small ($< 10^{-6}g$) acceleration caused by the residual acceleration of the drop capsule. This acceleration should be equal for both test bodies if $\eta = 0$. Beside the acceleration, the test bodies have a different initial velocity caused by limitation of the controlling process at the beginning and an unknown initial displacement.

In order to determine the Eötvös coefficient, the displacement of the two test bodies has to be measured from the end of the controlling process to the end of

[3] It is also possible to use two different sensors for each test body. One sensor possessing a large working range (0.5 mm) is used for controlling the test body. The other possessing a high resolution is used for the determination of the Eötvös coefficient.

the drop[4]. We expect a linear dependence of the displacement on time (caused by the different initial velocities) (if $\eta = 0$), or a parabolic dependence (if $\eta \neq 0$). Theoretically, three measurement points would be sufficient to determine η. But the measurement curves will be superimposed by the noise of the detector and the environment (magnetic disturbances, vibration of the capsule). Therefore, it will be better to measure much more points to get the highest possible resolution for η limited by the noise spectrum and the measurement time. The sampling rate of the measurement point needed to attain an optimal resolution of η depends on the bandwidth of the detector. Both parameters are chosen on the basis of the dynamical performance of the A/D–converter, the memory capacity of the data recording unit and the measurement time. In the final experiments we will use a 24–bit A/D converter[5] with a sampling rate of 2 kHz and a high–order lowpass filter with a cut–off frequency of 500 Hz.

5 Conclusions

Due to their exceptional sensitivity and universality, DC–SQUIDs made of classical low–T_c materials like niobium and lead can be applied in position detectors providing an unusually high position resolution. On the basis of a set of two wire–wound plunger–type pick–up coils, the position of a superconducting test body could be measured with a noise limited resolution of the order of 4×10^{-14} m/$\sqrt{\text{Hz}}$. Using this sensor technique, gravitational experiments can be performed such as the test of the validity of the WEP on earth by using drop tower facilities or in space like the current STEP project of NASA/ESA.

On the basis of this type of position detector, it would be possible to prove the validity of the WEP with an accuracy of about $\eta = 10^{-15}$ on earth by using the drop tower facility in Bremen. Because of disturbing effects (such as magnetic background fields, mechanical vibrations of the drop capsule after release, liquid helium tides, etc.) the accuracy will be limited to an estimated level of $\eta = 10^{-13}$. Nevertheless, this would be more precise by at least one order of magnitude than the most accurate ground based experiments performed by Roll, Krotkov, and Dicke [1,2] using modern torsion balance equipments.

On the other hand, drop tower facilities like the Bremen Drop Tower seem to be an appropriate and excellent tool for the ground based experimental test and characterisation of all components of SQUID based position detector systems as required, e.g., for the proof of the validity of the WEP in space.

Acknowledgement

The authors wish to thank G. Schmidt (Friedrich–Schiller–University Jena) for his expert assistance in the design and fabrication of the mechanical parts. This

[4] The displacement can be determined by measuring the position of each test body and calculate the difference after the drop or by using of a differential detector.

[5] The resolution of such a precision A/D converter depends on the sampling rate. It has only a resolution of 20 bit at a sampling rate of 2 kHz.

work was supported in part by the DARA GmbH/DLR e.V., Bonn, under contract No. 50WM94383 and 50OY98020.

References

1. P.G. Roll, R.V. Krotkov, R.H. Dicke: The equivalence of inertial and passive gravitational mass, *Ann. Phys.* (N.Y.) **26**, 442 (1964).
2. Y. Su, B.R. Heckel, E.G. Adelberger, J.H. Gundlach, M. Harris, G.L. Smith, and H.E. Swanson: New tests of the universality of free fall, *Phys. Rev.* **D 50**, 3614 (1994)
3. T.M.. Niebauer, M.P. McHugh, J.E. Faller: Galilean test for the fifth force, *Phys. Rev. Lett.* **59**, 609 (1987).
4. K. Kuroda, N. Mio: *IEEE Trans. Instr. Meas.* **38**, 196 (1989).
5. F. Barlier *et al.*: *STEP Assessment Study Report*, SCI(91)4, January 1991.
6. C.W.F. Everitt. P. Worden, L. Ryans–Culclager, R. Farnsworth: *STEP NASA Research Announcement*, NRA-96-HEDS-03, 19 March, 1997.
7. H. Dittus, Chr. Eigenbrod, U. Kaczmarczik, J. Middelberg, H.J. Rath, B. Schwark: High quality of microgravity environment during short time experiments on the Drop Tower Bremen, *Proc. 7th Europ. Symp. on Materials and Fluid Sciences in Microgravity*, European Space Agency (ESA-SP 295), p. 703 (Paris 1989).
8. H. Dittus: Drop Tower Bremen – A weightlessness laboratory on Earth, *Endeavour*, New Series, **15**, 72 (1991).
9. H. Dittus, R. Greger, St. Lochmann, W. Vodel, H. Koch, S. Nietzsche, J. v. Zameck Glyscinski, C. Mehls, P. Mazilu: Testing the Weak Equivalence Principle at the Bremen Drop Tower: Report on recent developments, *Class. Quantum Grav.* **13**, A43 (1996).
10. N. Lockerbie, X. Xu, A.V. Veryashkin. M.A. Hosey: The gravitational coupling between longitudinal segments of a hollow cylinder and an arbitrary gravitational source: Relevance to the STEP experiment, *Class. Quantum Grav.* **13**, 2041 (1996).
11. W. Vodel, K. Mäkiniemi: An ultra low noise SQUID system for biomagnetic research, *Measurement Science and Technology* **3**, 1155 (1992).
12. W. Vodel, H. Koch, J. von Zameck Glyscinski, C. Mehls, S. Nietzsche: Application of High Performance DC SQUID Systems in Fundamental Physics, *Proc. of 5th International Superconductive Electronics Conference*, Nagoya, Japan, 18-21 September, 1995.
13. W. Vodel, H. Koch, S. Nietzsche, J. von Zameck Glyscinski, H. Dittus, S. Lochmann, and C. Mehls: Low noise LTS SQUIDs for application in gravitational experiments, *IEEE Transactions on Applied Superconductivity* **7**, 3343 (1997).
14. S. Nietzsche: *Aufbau und Test von Grundanordnungen für Positionsdetektoren mit SQUID*, Diploma Thesis (University of Jena, 1996).
15. H.A. Chan and H.J. Paik: Superconducting gravity gradiometer for sensitive gravity measurements, *Phys. Rev.* **D 35**, 3551 (1987).

Space Accelerometers: Present Status

Pierre Touboul

Physics, Instrumentation and Sensing Department
Office National d'Etudes et de Recherches Aérospatiales, BP 72,
92322 Châtillon Cedex, France

Abstract. In view of space missions, for accurate recovery of the Earth gravity field, for the test of the equivalence principle and for the observation of gravity waves in particular, specific inertial sensors are developed exhibiting very high resolution and limited full scale range suited for in orbit operation. These sensors are constructed around a high density proof–mass with a very fine and stable silica gold coated core. The proof–mass position and attitude are measured with highly sensitive capacitive sensors and are controlled with electrostatic actuators. The configuration and the major design parameters of these instruments are described in relation to the expected performances. The present status of the development of these instruments is shown together with the associated space mission scientific objectives. The main experimental results obtained during the ground qualification of these accelerometers are also presented.

1 Introduction

Space missions are not only dedicated to the discovery and the observation of the universe, as in astrophysics or in planetary sciences, but space provides also a specific microgravity environment, the advantage of which is exploited in crystallography, biology, and physics experiments. Moreover, it provides a fine possibility of global and accurate observation of the Earth.

For the micro–gravity experiments, the acceleration levels of the space platforms have to be monitored especially in the very low frequency range from 1 Hz down to the orbital frequency of about 1.7×10^{-4} Hz, and with a sensitivity better than the micro-g. Such a sensitivity must even be increased for the control of dedicated isolated bench [1] or for the control of spacecraft. The DC acceleration and the vibration levels on board the Spacelab shuttle have been monitored [2] in orbit to less than 10 micro-g (1 g $\sim$ 9.8 m/s^2) mean value, depending on the orbit altitude between 300 km and 500 km and on the orientation of the shuttle with respect to its velocity vector. The fluctuations in the 1 Hz bandwidth reach 100 micro-g and time to time 1 milli-g while the local vibrations increase at upper frequencies up to a hundred of milli-g. This is far from the simple idea of punctual weightlessness. In fact, the spacecraft is not only submitted to the orbital gravity field but also to the atmospheric drag of the Earth (or the planet), the radiation pressures from the Sun and the Earth, the magnetic forces induced by the coupling between the planet's magnetic field and the spacecraft charge (Lorentz force), or its magnetic susceptibility or momentum. Furthermore, when not at the centre of mass, the effect of the gravity gradient tensor has to be taken into account as well as the angular and the centrifugal acceleration.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 273–291, 2001.

Table 1 provides evaluation of the amplitudes of these accelerations for two recently designed satellites, the first one at very low altitude for a geodesy mission, the second one at higher altitude for space fundamental physics. Obviously, at 36 thousand kilometres, the geosynchronous satellites are no more affected by the atmospheric drag and the Earth's gravity gradient.

For what concerns the International Space Station, it is designed to provide a space laboratory micro–gravity environment limited to 2 μg for frequency lower than 0.1 Hz and with a f law increase for upper frequencies [3].

Table 1. The Esa GOCE mission is devoted to the Earth gravity field recovery (launch foreseen in 2004) with a 800 kg satellite; the CNES μSCOPE mission aims at the Equivalence Principle test (launch foreseen in 2004) with a micro-satellite of 120 kg; both missions require a very fine orbit and altitude control.

	GOCE satellite	μSCOPE satellite
altitude	250 km	700 km
atmospheric drag	1.5×10^{-5} m s^{-2}	6×10^{-8} m s^{-2}
radiation pression	6.1×10^{-8} m s^{-2}	3.7×10^{-8} m s^{-2}
gravity gradient	4.1×10^{-6} m s^{-2}/m	3.4×10^{-6} m s^{-2}/m
angular acceleration	1.4×10^{-6} m s^{-2}/m	1.1×10^{-6} m s^{-2}/m
centrifugal acceleration	1.5×10^{-5} m s^{-2}/m	3.3×10^{-6} m s^{-2}/m

In order to finely measure these environments, specific sensors exhibiting very low bias and high resolution, hundred times better than the accelerometer used in aircraft for inertial navigation have to be developed.

For industrial applications, accelerometers exhibiting a full scale range of up to several hundreds of g and exhibiting parts of g sensitivity are available [4]. The piezoelectric accelerometers exploit the electrical potential variations of a piezoelectric device which internal constraints are linked to the relative motion of an inertial mass. These very robust and small instruments present very large bandwidth but no measurement at low frequencies.

For aircraft navigation, the accelerometer accuracy has to be better than one hundred micro-g, after data processing taking into account the instrument calibration and thermal sensitivity, in the important range of temperature from -50°C to $+90$°C [5]. Most of these accelerometers rely on a spring–mass or a pendulum-mass device with a position sensing of the mass. When the mass motion is servo–controlled through magnetic actuators for instance, the acceleration of the instrument frame is deduced from the measured applied force. The stability of the spring or of the hinge which sustains the mass is of peculiar importance in order to obtain the accelerometer bias stability [6]. The servo–loop

operation leads fortunately to one single point condition of operation (deviation and constraint) to the benefit of the behaviour stability.

New generation of vibrating beam accelerometers are now proposed for aircraft navigation. The natural mode of vibration of a silica or a quartz beam varies with the applied constraints which are induced by a mass fixed at one of its extremity. The vibrations are excited by piezoelectric or capacitive means and the oscillator frequency is measured with a counter. These accelerometers exhibit now better than one hundred micro-g accuracy in a large operating range of temperature while the resolution can be in the order of one micro-g. The main interests consist in their easy way of production, few cm^3 sizes and the already digital outputs [7].

All these accelerometers can be adapted to some space applications but present for the quasi–steady acceleration a non sufficient resolution because of their too important full range and bias, not suited to the in orbit acceleration levels.

Furthermore, scientific missions need much better performance. In the field of the Earth's observation, two main applications of the ultra–sensitive accelerometers merge. First, the measurement of the surface forces acting on low Earth orbit satellites allows to eliminate the non gravitational effects in the accurate determination of the orbit. This can be simply exploited for the autonomous navigation of the satellite, the orbit of which is then predictable. Associated to a fine trajectory recovery from GPS receiver or Doris on board system, this yields to the determination of the Earth's gravity field as in the CHAMP project or the DORIS D micro–satellite project [8]. For the determination of the gravity anomalies at smaller geographical scale, from wave lengths of several thousands down to one hundred kilometres (and so of the higher spherical harmonics of the gravity potential) two other techniques are now elaborated, the satellite to satellite tracking and the gravity gradiometry [9], [10]. For both, accelerometers with expected resolution up to the pico-g or better are necessary [11].

Future scientific missions in the domain of Fundamental Physics require also satellites flying drag–free, which means that all non-gravitational disturbances acting on the satellite are compensated by throttling the thrust with closed–loop controls exploiting acceleration sensors. This is the case of the LISA mission satellites for the observation of the gravity waves [12]. The in orbit test of the Equivalence Principle is also proposed by considering ultra–sensitive differential accelerometers including two test-masses made of different materials [13], [14]. For these missions, the objectives of sensitivity are much better than the femto–g.

These outstanding resolution cannot be achieved with instruments able to measure or even support in their nominal measurement mode one g. Specific space accelerometers must be conceived taking advantage of the in orbit environment that much reduces the applied acceleration to sustain especially on board future drag–free satellites.

2 Accelerometers Dedicated to Space

Aircraft pendulum accelerometers have been optimised for space applications; this is the case of the sensors of the SAMS system which has been used to characterise the shuttle microgravity [15]. In fact, the sensor core is quite not modified but the servo–loop electronics of the mass control can be changed in order to increase the accelerometer scale factor from 0.1 V/g to 100 V/g. The resolution is then modified from 0.5 μg/$\sqrt{\text{Hz}}$ to about 50 nano-g/$\sqrt{\text{Hz}}$ as seen in Fig.1.

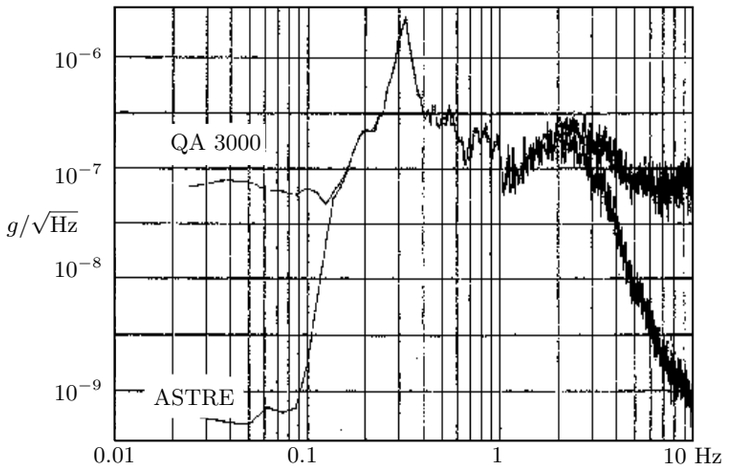


Fig. 1. The pendulum accelerometer is mounted on an antiseismic testing bench; the residual acceleration of the bench, which has a tilt natural mode around 0.3 Hz, is measured with the ASTRE space accelerometer (presented below) to less than 1nanog for frequency lower than 0.1 Hz; the intrinsic noise of the QA 3000 is run printed out.

To go further in resolution, electrostatic space accelerometers have been developed since the seventies. The Discos and the Cactus instruments have been respectively launched in 1972 and 1975 for drag compensation control or drag measurement [16], [17]. Both was designed around a high density proof–mass, the motion of which is measured with three capacitive sensors. The Cactus accelerometer (Capteur Accélérométrique Capacitif Triaxial Ultra Sensible) was designed to measure the surface forces exerted on a satellite by the aerodynamic drag and the solar radiation pressures. The spherical proof–mass position is controlled but not the rotation. This simple design is well suited for high sensitivity but incompatible with high accuracy: no matter how carefully the instrument is built, the residual geometric or electric defects of sphericity make the sensitivity of the accelerometer fluctuates as the proof–mass turns.

The Miniature ElectroStatic Accelerometer (MESA) includes a thin-walled cylinder as a proof–mass with a thin central flange and operates in various space

missions [18]. Initially developed by Bell Aerospace as a one–axis accelerometer for rotating accelerometer gravity gradiometre, the MESA is presently exploited as a three axis space accelerometer taking advantage of its five degrees of freedom electrostatic "suspension". It is in particular one of the sensors of the Orbital Acceleration Research Experiment (OARE) currently used by NASA to measure the space shuttles on-orbit linear acceleration within the sub–microgravity regime. The resolution of the sensor is better than 10 nanog with 25 s integration and its bias and scale factor are calibrated in orbit by rotating the MESA sensor mounted on a dual–gimbals platform [19]. The European QSAM (Quasi-Steady Acceleration Measurement) instrument follows the same approach with two pairs of QA 3000 pendulum accelerometers, each pair being mounted on a flipping device for calibration and associated with a static triad of accelerometers. Due to the use of QA3000, the resolution is of the order of 5×10^{-7} g [20]. The three instruments, QSAM, OARE and ASTRE have flown in 1996 during the same shuttle mission MSL-1.

Superconducting accelerometers, with cryogenic magnetic levitation instead of room temperature electrostatic one, have been developed for ultra–sensitive gravity gradiometry [21]: the cylindrical mass being maintained in radial direction with a blade, highly stable magnetic stiffness are induced between the mass and the instrument cage and along the axial direction by superconducting loops, associated with squid position sensing. Same type of instrument configuration are also being defined for three axis measurement but without any mechanical stiffness and with a three axis magnetic suspension [13].

3 Electrostatic Servo–Controlled Accelerometer Operation

The principle of operation of a servo–controlled electrostatic accelerometer is based on the measurement of the electrostatic force necessary to maintain the accelerometer proof–mass motionless with respect to the sensor cage. For space applications, the proof–mass can be fully suspended in the three directions, suppressing any mechanical contact to the benefit of the resolution and yielding to a three axis accelerometer.

The performance relies on the resolution of the proof–mass position sensor, the very limited stiffness that links the test–mass to the accelerometer cage and the very weak level of the test–mass motion disturbances, Γ_{dtb} due to the environment (magnetic, electric, thermal), and independent to the electrostatic suspension. Let us consider the expression of the test–mass motion x_{mass} inside the sensor cage:

$$s^2 x_{\mathrm{mass}} = g + \Gamma_{\mathrm{elec}} \pm \omega_p^2 \left(x_{\mathrm{mass}} - x_{\mathrm{cage}}\right) \quad \text{and} \quad s^2 x_{\mathrm{cage}} = \Gamma_{\mathrm{spcft}} \tag{1}$$

with s the Laplace derivative variable, Γ_{elec} the acceleration from the electrostatic suspension, g the gravity field and ω_p the angular frequency associated to the parasitic uncontrolled stiffness. Then the measurement of Γ_{elec} provided by

the accelerometer leads to the acceleration of the spacecraft Γ_{spcft}:

$$\Gamma_{\text{elec}} = \frac{G_{\text{Loop}}}{G_{\text{Loop}} + s^2 \pm \omega_p^2} \left(\Gamma_{\text{spcft}} - \Gamma_{\text{dtb}} + \left(s^2 \pm \omega_p^2\right) x_{\text{noise}} - \frac{m_{\text{g}}}{m_{\text{i}}} g \right) \tag{2}$$

with G_{Loop} the gain of the servo–loop which controls the proof–mass degree of freedom $x_{\text{mass}}, G_{\text{Loop}} \gg 1$ at low frequencies and with x_{noise}, the capacitive sensor noise expressed in displacement according to the geometry of the configuration electrodes–mass.

One of the advantages of the electrostatic suspension is the use of very weak acceleration with measurable level of applied voltages: electrostatic actuators with sensitivity lower than one nano-g per Volt can be achieved. Furthermore, when the position of the mass is servo–controlled, the difference $x_{\text{mass}} - x_{\text{cage}}$ is almost null in the frequency bandwidth of the suspension where the gain of the servo–loop is important. Then, the optimisation of the sensor–head configuration which aims at reducing Γ_{dtb} and at producing well identified Γ_{elec} is made easier by the steady geometrical configuration.

Because the electrostatic forces remain normal to the faces of the mass, assumed to be a perfect conductor (then an electrical equipotential), a cubic proof–mass with parallelism and perpendicularism deviations less than one arc second ensures very low coupling between the three pairs of servo–control channels that keep it motionless in position and attitude. And thus low coupling of the sensitive axis measurements are obtained.

The pair of electrodes corresponding to each loop can be used for both capacitive position sensing and electrostatic restoring force generation (Fig.2). From the measurement of the capacitive sensor, a corrector determines the opposite drive voltages to be applied on the opposite electrodes for the generation of the electrostatic field. A unique sine wave pumping signal V_d is applied to the proof–mass for the six capacitive sensors. A biasing reference voltage can be also applied to the proof–mass or to both electrodes in order to make linear the electrostatic actuators. Both electrodes attract the proof–mass with forces F_1 and F_2 proportional to the gradient of the capacitance and to the square of the electric potential differences between the proof–mass and the electrodes. The resulting force F ($F = F_1 + F_2$) is expressed by:

$$F = \frac{1}{2} \left(\nabla C_2 \left(V_2 - V_p\right)^2 + \nabla C_1 \left(V_1 - V_p\right)^2 \right) = m\Gamma_{\text{elec}} \tag{3}$$

In case of a perfect configuration symmetry with:

$$\Delta C_2 = -\Delta C_1 = \Delta C \qquad V_1 = -V_2 = V \tag{4}$$

the resultant force F is proportional to V:

$$\Gamma_{\text{elec}} = \frac{1}{m} F = 2\, \nabla C\, V_p\, V \tag{5}$$

The loop provides a cold damping of the proof–mass motion and a strong control at frequencies lower than the closed loop natural frequency, $f_{0_{\text{c}}}$. The

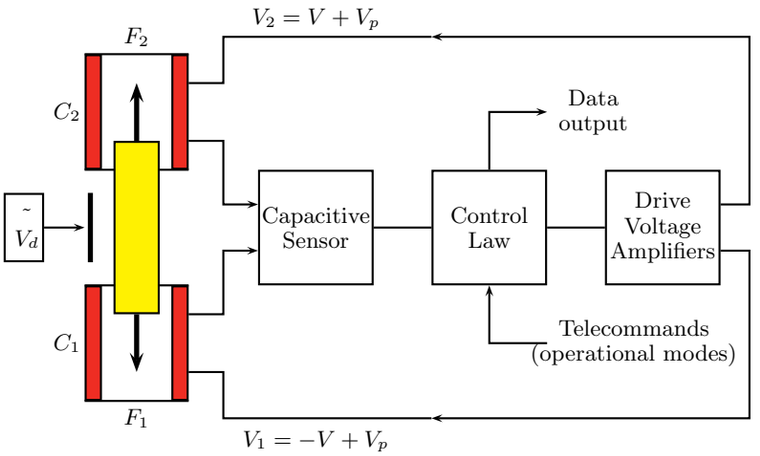


Fig. 2. Scheme of one servo–loop channel.

output network provides the measurement of the actual applied voltage V which constitutes the analogue output of the accelerometer. $H(s)$ being the linear relation provided by the electronics between V and the capacitive sensor output, the expression of the measure is:

$$\text{Measure} = \Gamma_{\text{elec}} = \frac{H(s)}{H(s) + s^2 \pm \omega_p^2} \left(\Gamma_{\text{spcft}} - \Gamma_{\text{dtb}} - g + \left(s^2 \pm \omega_p^2 \right) x_{\text{noise}} \right) \quad (6)$$

and the expression of the mass displacement:

$$x_{\text{mass}} = x_{\text{cage}} - \frac{H(s)}{H(s) + s^2 \pm \omega_p^2} x_{\text{noise}} + \frac{1}{H(s) + s^2 \pm \omega_p^2} \left(\Gamma_{\text{spcft}} - \Gamma_{\text{dtb}} - g \right) . \quad (7)$$

The electrostatic suspension bandwidth can be selected much larger than the measurement frequency domain without affecting the accelerometer resolution. Then, the accelerometer frequency response is flat (first term of the expression of the measure is almost one) and the proof–mass motion is very limited to the benefit of the accelerometer linearity and stability of characteristics.

4 The ASTRE and STAR Accelerometers

On the basis of the previously described concept, several configurations of the sensor-head with its associated electronics have been developed according to the mission requirements.

ASTRE (French acronyms of Accelerometre Spatial TRiaxial Electrostatique) has been optimised for the monitoring of the manned spacecraft environment and its configuration is well suited to realise a three axis accelerometer with the possibility to perform ground tests under normal gravity [22].

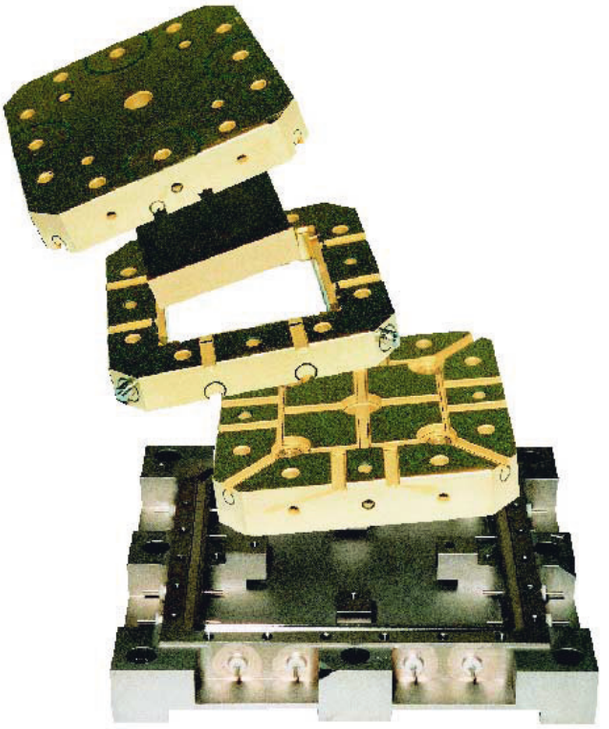


Fig. 3. ASTRE–STAR sensor–head.

The mechanical core of the sensor is defined with a silica core surrounding a 72 grams parallelepiped proof–mass made of titanium alloy, 4 cm side and 1 cm height (see Fig.3). The core is internally gold coated and the electrodes face the proof–mass by pair. The mass is ground with parallelism and perpendicularly deviations less than 10^{-5} rad, that ensures the high uncoupling between the three axes. The asymmetry of the configuration (not a cubic proof–mass) is required to perform the electrostatic suspension of the mass under one g, the electrode areas being larger along the vertical axis and the distance between the proof–mass and the electrodes being reduced to a few tens of microns. An electrostatic field as strong as 3×10^7 V/m can be necessary to sustain the proof–mass. The core is integrated inside a tight housing ensuring a secondary vacuum of 10^{-6} mbar and a magnetic shield to the benefit of the mass motion disturbances.

According to the mission, the ASTRE configuration can be slightly modified: the range of the accelerometer can be settled by changing the proof–mass material (the proof–mass density can vary from 2.2 g/cm^3 for Silica up to 20 g/cm^3 for Platinum) without modifying the silica core; the proof–mass sizes can also be modified and thus the gaps between the electrodes and the proof–mass can be

adjusted between 30 μm and 1 mm; the strength of the electrostatic actuators being reduced, the accelerometer sensitivity can be increased.

The analogue servo–loop electronics are designed around six capacitive sensors associated to six electrostatic actuators. The sensitivity of the capacitive position sensing must be settled to the electrode configuration and sizes. The areas of the two z electrodes are for instance 2 cm^2 and the distance to the mass is 75 μm leading to capacitance of 24 pF. With a sensitivity of 0.3 V/μm, the measured resolution of the position sensing is $2 \times 10^{-12}(1 + 0.4/f)^{1/2}$ m/$\sqrt{\text{Hz}}$ as shown in Fig.4 and with a negligible back–action of the electrical sensing signals on the mass acceleration.

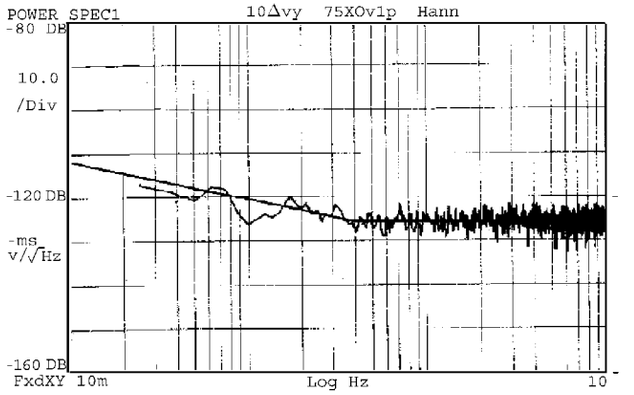


Fig. 4. Capacitive position sensing resolution with ASTRE accelerometer configuration and electronics; –120 dB equals 1 μV sensor output i.e. 3.3 picometre

In spite of the differences of the geometry, the characteristics of the three outputs of the ASTRE accelerometer are very similar with a resolution of 1 nano-g and a full scale range of one milli–g.The ASTRE accelerometer flew three times on board the Columbia shuttle in 96 and 97 for the missions STS-78, STS-81 and STS-81R. Integrated inside its host system, the Microgravity Measurement Assembly (MMA), ASTRE collected data during the whole missions. Fig.5 shows the statistic distribution of the measured acceleration during the third mission, demonstrating that the microgravity level is unfortunately not reached during a wide window [23].

The Fig.6 illustrates the impact on the Spacelab acceleration in translation of the actuation of the Columbia shuttle thrusters during the attitude control phase of the shuttle. It is to be noticed that the average level is very close to zero demonstrating the weak instrument bias.

From the ASTRE experience, the STAR accelerometer has been defined for the German CHAMP mission (CHAllenging Micro-satellite Payload for geophysical research and application) dedicated to Earth's observation. Five scientific

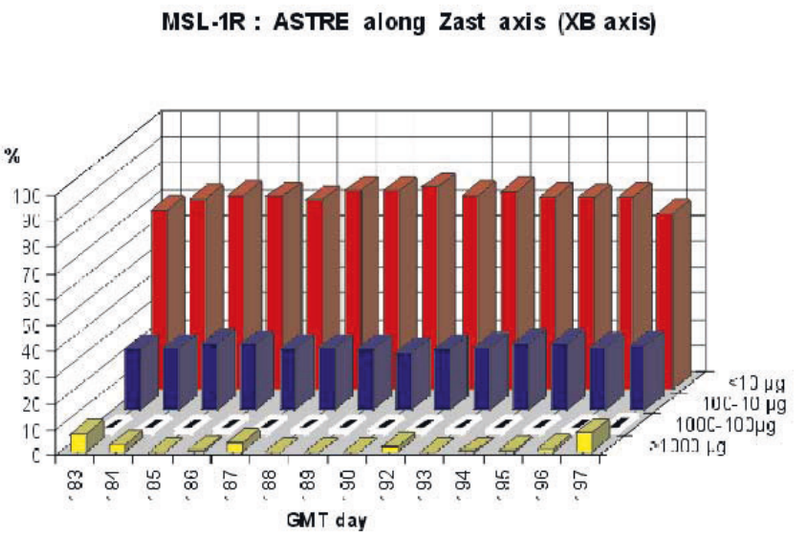


Fig. 5. Distribution of time percentage per mission day when the measured acceleration is lower than 10 μg, between 10 and 100 μg, 100 μg and 1000 μg or larger

instruments constitute the satellite payload: two magnetometers for the global and accurate Earth's magnetic field mapping, one digital ion drift meter to deliver complementary measurements of the Earth's electrical field, one GPS dual–frequency receiver contributing to the atmospheric analysis, density, pressure and temperature sounding but devoted first to the precise orbit determination with centimetre accuracy and one STAR (Space Three–axis Accelerometer for Research) accelerometer [24], [25].

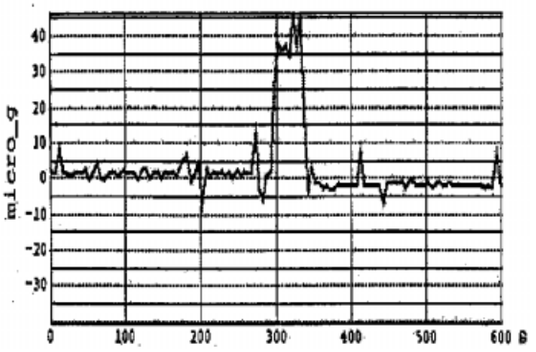


Fig. 6. Measured Spacelab acceleration during the throlling of the shuttle attitude control thrusters

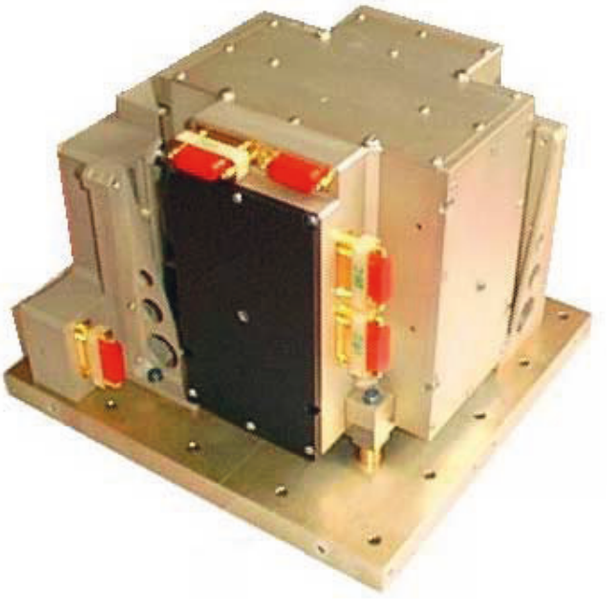


Fig. 7. STAR sensor unit before integration in the CHAMP satellite

The CHAMP satellite is scheduled to be launched in April 2000 by a Russian Kosmos rocket at an altitude of 550 km in a circular orbit with a 100 degree inclination, the eccentricity being less than 0.01. During the five years of the mission, the orbitography will be finely performed and the accelerometer, integrated at the centre of mass of the satellite with an accuracy of a few millimetres shall measure in a frequency bandwidth from DC to a few tenth of Hertz the air drag, the solar and Earth radiation pressures and the attitude manoeuvre effects aiming at the Earth's gravity field up to degrees and orders of about 50. A by–product of the accelerometer measurements will be the determination of the atmospheric density variations, in particular at the lower altitude of the mission.

STAR has an internal core configuration similar to ASTRE. This six–axis accelerometer provides the three linear accelerations along the instrument sensitive axes and the three angular accelerations about these axes. STAR presents a measurement range of 10^{-4} m s^{-2} and exhibits a resolution of better than 3×10^{-9} m s^{-2} for the y– and z–axes and 3×10^{-8} m s^{-2} for the x–axis within the measurement bandwidth from 10^{-4} Hz to 10^{-1} Hz. The measurements are integrated over one second before delivery to the satellite data bus. In orbit, STAR x–axis is vertical from the Earth, while the y–axis is normal to the orbit and the z–axis is along track. The Fig.7 shows STAR stand–alone instrument constituted by a sensor core depicted previously and integrated inside a tight housing with servo–loop electronics boxes around, all packaged in the same unit.

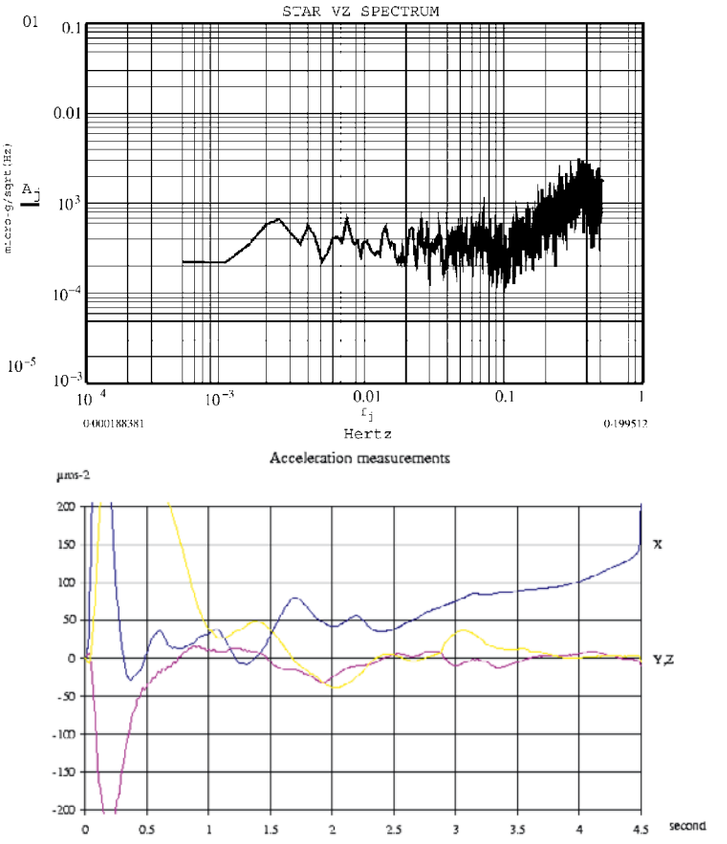


Fig. 8. STAR resolution as demonstrated with ground tests (left) and free fall tests in drop tower (right); the accelerometer is measuring in both case the residual level of vibrations of the testing bench maintained horizontal or of the falling capsule

Its total mass is 7.8 kg, inside a volume of $19.5 \times 18.5 \times 16.5$ cm^3 and its nominal power consumption is 1.7 W.

The STAR performance has been verified in laboratory as far as it is possible in presence of seismic vibrations and with a mass levitation under one g by specific booster electronics. The y and z accelerometer axes being maintained horizontal, measurements of residual low frequency vibrations of the testing platform have been performed with a resolution of 10^{-8} m s$^{-2}/\sqrt{\text{Hz}}$ at frequencies lower than 0.1 Hz corresponding to the requirement of 3×10^{-9} m s^{-2} rms in the measurement bandwidth (see Fig.8). The accelerometer low level of bias (less than 10^{-5} m s^{-2}) has been also verified in free fall in the Bremen drop tower [26].

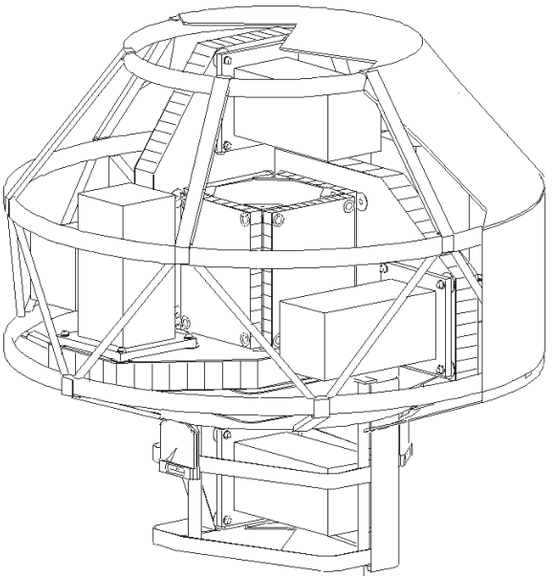


Fig. 9. Diamond configuration gradiometer.

5 From CHAMP to GRACE and GOCE Instrument

With an up–date of the STAR configuration by modifying the core geometry, the position sensing and the electrostatic actuator sensitivities, the SuperSTAR accelerometer has been defined for the Gravity Recovery And Climate Experiment (GRACE) mission [27]. This instrument presents two modes of operation with two different ranges: the first one is more robust ensuring the mass levitation with a strong control, the second one corresponds to a full measurement range of 5×10^{-5} m s^{-2} with a resolution of 10^{-10} m s$^{-2}/\sqrt{\text{Hz}}$. Two flight models of SuperSTAR accelerometer are being produced in order to be integrated in the two GRACE satellites in the first half of 2000. The mission aims at the recovery of the Earth's geoid with an accuracy of 3.5 mm and a geographical resolution of 200 km. The satellite to satellite tracking performed with a micro–wave link at altitudes between 500 km and 300 km provides the measurement of their relative velocity with an accuracy better than 1 μm s^{-1} depending on the over flight gravity anomalies and on the satellite surface forces measured by the accelerometers [28].

As a complement to the previous JPL mission, the Gravity and Ocean Circulation Explorer (GOCE) mission has been selected in 1999 by the European space agency in the frame of the Earth explorer program [29], [30]. On board a drag–free satellite, at an altitude as low as 250 km, the gravity gradiometer shall provide the accurate measurements of the Earth's gravity gradient tensor components for the determination of the upper harmonics of the potential [31]. This gradiometer is composed of six identical electrostatic accelerometers in a

diamond configuration and accommodated on a rigid and stable Carbon–Carbon structure inside a double chamber thermal enclosure. The theoretical models of the accelerometer operation at room temperature, which are based on the experimental results obtained with the STAR and SuperSTAR qualifications, have been exploited to demonstrate the adequacy of such an instrument for the conditions of the very soft environment of GOCE satellite. With maximum levels of acceleration less than 10^{-6} m s^{-2}, due for a large part to the Earth gravity gradient itself, and temperature variations lower than 10 mK over the orbit, the expected resolution of the accelerometers is 4×10^{-13} m s$^{-2}/\sqrt{\text{Hz}}$ in the frequency bandwidth from 5×10^{-3} Hz to 0.1 Hz. Taking into account the 10^{-5} rejection factor of the residual satellite acceleration in translation and attitude, the gradiometer resolution is evaluated to better than[1] 5 mE [32].

6 Electrostatic Accelerometers to Test the Equivalence Principle in Space

The objective of the μSCOPE space mission that we presently propose to CNES (French national Agency) is to improve by a factor one thousand the test of universality of free fall leading to an accuracy of 10^{-15} [33]. The laboratory experiments for the Equivalence Principle (EP) test exploit torsion pendulum and have to deal with the environmental instabilities and in particular the Earth gravity gradient fluctuations [34]. Recent results have been obtained by considering the Moon–Earth laser ranging data but the material composition of the two bodies is not sufficiently well known [35]. Space experiment takes advantage of the in orbit soft conditions and the possibilities of long term signal integration. The usual 'Galileo type' experiment consists of comparing the relative motion of two free falling test masses made of different materials. Another way to perform the EP test is to control at null the relative mass motion (the masses having a free common fall motion in orbit) so that any EP violation appears through the measured forces necessary to nullify this relative motion. Then a much better management of the environment disturbances is possible because of the steady configuration. This is why we propose to perform this experiment on board a drag–free satellite with two servo–controlled electrostatic accelerometers, one implemented inside the other one [36]. Thus, the two test–masses are centred and submitted to the analog Earth and satellite gravity field: the ratios $M_{\text{g}}/M_{\text{i}}$, gravitational mass M_{g} over the inertial mass M_{i}, of the test–masses made of different materials will be compared . The μSCOPE payload is in fact composed of two of these differential accelerometers operating at room temperature: two pairs of test masses are then compared, one with the same material, the second with two different materials leading to a double differential measurement to the benefit of the rejection of the common motion disturbances. The satellite spin about the perpendicular axis to the orbit varies the orientation of the Earth gravity field in the instrument reference frame leading to a research of the violation signal at a well known frequency and helps to the discrimination from other

[1] 1 E = 1 Eötvös = 10^{-9} s^{-2}.

effects like thermal or gravity gradient ones. With a signal integration over 20 orbits (10^5 seconds), the difference of the two test–mass accelerations shall be measured with a resolution of 10^{-12} m s$^{-2}/\sqrt{\text{Hz}}$ in the vicinity of the sum of the orbital and spinning frequencies, about 10^{-3} Hz [37].

The expected matching of the characteristics leads to a 10^{-4} rejection of the accelerations applied in common mode on the two concentric accelerometers and the necessary drag compensation system of the satellite must then reduce these accelerations to levels of less than 10^{-8} m s$^{-2}/\sqrt{\text{Hz}}$.

The cylindrical test masses present diagonal inertia matrices to limit the effects of the satellite and Earth gravity gradient and the common revolution symmetry axis is the EP test sensitive axis. The electrode configuration presented in Fig.10 allows the measurement and the control of the 6 degrees of freedom of the two test-masses. While the eight quadrant electrodes associated by pairs allow the control of the radial translations and rotations, the two cylindrical sensing electrodes at the ends of the test–masses are used to control motionless the test–masses along the axial direction quite independently to the mass relative position that is adjusted in orbit to a few microns.

The 10^{-12} m s$^{-2}/\sqrt{\text{Hz}}$ expected resolution of this instrument is deduced from the three major noise sources. At lower frequencies, the thermal instabilities induces radiation pressure and radiometer acceleration fluctuations due to residual gas for the latter. At higher frequencies, the position sensing resolution affects the resolution with a square frequency law. Between, the thermal noise of the mass motion is derived from the damping factor estimated from dedicated laboratory experiments and mainly due to the thin 5 μm wire used for the charge control of the mass [38], [39].

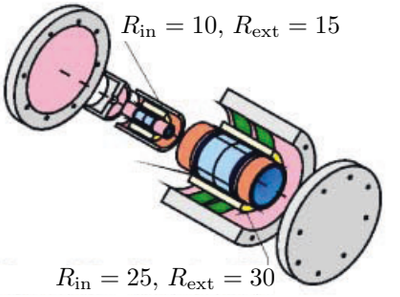


Fig. 10. Differential electrostatic accelerometer for the in orbit test of the Equivalence Principle.

7 Space Gravity Wave Antenna

The LISA space mission (Laser Interferometer Space Antenna), under study by US and European Space Agencies, aims at the observation of the low-frequency astrophysical gravitational radiation with cosmological and fundamental physics objectives [12]. The interferometer is realised with a triangle formation of three drag–free spacecraft in heliocentric orbit. At the centre of the satellite, the proof–masses of inertial sensors are the mirrors at each end of the 5×10^6 km interferometer arms. The optical bench of each satellite is highly stabilised in temperature (10^{-6} K/$\sqrt{\text{Hz}}$) and accommodated around the inertial sensors, the YAG laser beam being reflected directly on one side of masses. Then, the proof–masses constitute the inertial references, free of any acceleration disturbances, with an

expected level of a few 10^{-15} m $s^{-2}/\sqrt{Hz}$ in the very low frequency domain from 10^{-4} Hz up to 10^{-2} Hz . Furthermore, the inertial sensors are used by the drag–free satellite subsystem for the compensation of the drag due to radiation pressures. So, the satellite carries the optical bench and is a shield for the mass that the satellite follows accurately, at 10^{-9} m$/\sqrt{Hz}$, in the same low frequency bandwidth, in order to limit the mass disturbances, satellite gravity gradient for instance.

As in the case of μSCOPE configuration, this sensor presents a preferential axis (along the laser beam) and is optimised for a very weak range of operation thanks to the drag–free satellite. The accelerometer is built around a 1.5 kg bar test–mass made of Gold–Platinum alloy leading to a very weak magnetic susceptibility. Its cage is made of ultra low expansion (ULE) silica as well as its reference plate fixed on the optical bench taking advantage of the quite null coefficient of thermal expansion at the 25°C operating temperature [40].

The instrument housing is opened to the space vacuum in order to obtain a residual internal pressure less than 10^{-6} Pa. The principle of operation of the sensor is similar to the previous ones except that the satellite drag compensation system acts the electrical satellite thrusters in order to nullify the voltages that should be applied on the electrodes to keep the proof–mass motionless. In this way, no electrostatic forces will be applied to the proof–mass within the frequency bandwidth of the satellite control loop, larger than the interferometer measurement bandwidth. So, the inertial sensor has not to deliver an accurate measurement of acceleration with well known and steady scale factor but the configuration is focussed on the reduction of all parasitic forces. No discharging wire is envisaged to control the proof–mass potential for such a room temperature sensitive sensor. So an active control of the mass charge is necessary. The Fig.11 presents the silica core of the first laboratory model that has been in-

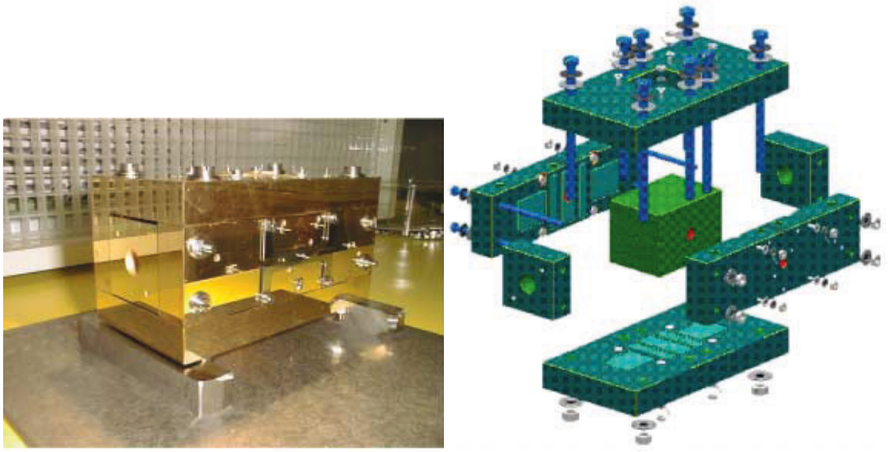


Fig. 11. Inertial sensor for gravity wave space interferometer: first laboratory model core with a parallelepipedic configuration.

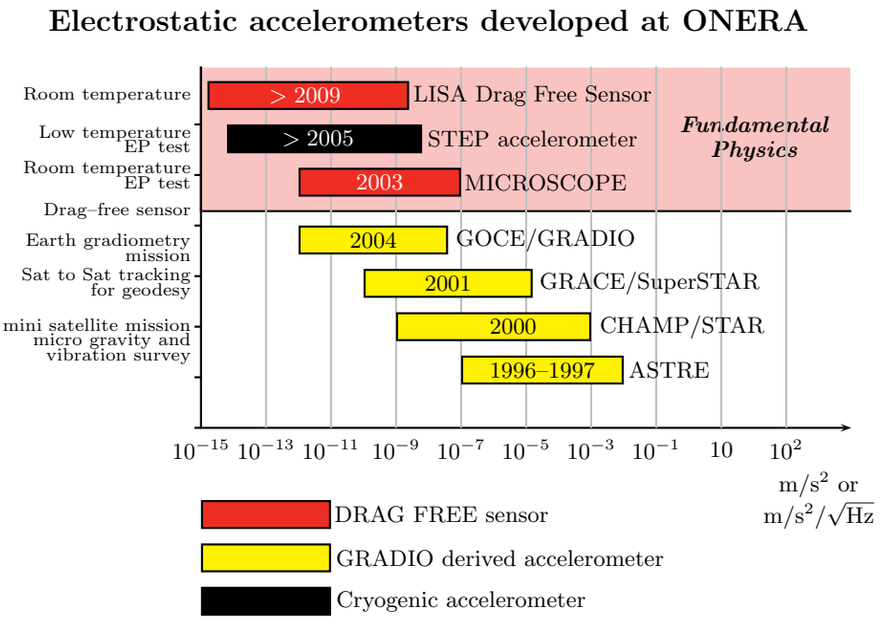


Fig. 12. Electrostatic space accelerometer range and performance.

tegrated with a silica mass in order to perform laboratory test under normal gravity: the levitation of the mass is performed and the electrostatic configuration is under test in order to demonstrate its compatibility with very weak induced stiffness, lower than 10^{-6} N/m.

8 Perspective

One of the major advantages of space environment is the low level of acceleration disturbances. On one hand, the survey of these levels require specific instrument, on the other hand, the dedicated space accelerometers take advantage of the specific range of operation and measurement. The concept of electrostatic servo–controlled accelerometer is well suited for these space applications: the electrostatic forces give the possibility to generate very weak but accurate accelerations while the capacitive sensing offers a high position resolution with negligible back reaction.

The Fig.12 summarises the different models that are under development or have been already produced for space missions. The large flexibility of configuration and operation allow many mechanical sensor configurations. Moreover, the present development of digital servo–loop electronics will permit to calibrate in orbit these instruments and to adjust the parameters of their operation [39]. This is now mandatory: the outstanding performance, needed by the future space mission s and expected because of the drag compensation system of the

satellite, are too far from the test accuracy in the laboratory, limited by human activities and Earth gravity field. The room temperature operation facilitates the integration on board the satellite. Nevertheless, cryogenic temperatures may be necessary to improve further the accelerometer resolution. This is why the operation of such an instrument at Helium temperature has been successively performed and prepare future physics investigations.

References

1. M. Nati, A. Bernard, B. Foulon, P. Touboul.: ASTRE, a highly performant accelerometer for the low frequency range of the microgravity environment, 24th Symposium on space environmental control systems, Friedrichshafen, Germany (1994).
2. K.M. McPherson, M. Nati, P. Touboul, A. Schütte, G. Sablon: A summary of the Quasi–Steady Acceleration Environment On–Board STS-94 (MSL-1), NASA/TM-1999-208853, AIAA-99-0574, January 1999.
3. H. Hamacher: Characterisation and μg improvement for the Columbus orbiting facility, ESA Report SP-385, p. 209 (1996).
4. R. Laver: PXE accelerometer, Technical Information 031, Philips, 1997.
5. J.C. Radix: Systèmes inertiels à composents liés strap-down, Cepadues Ed., Toulouse France, 1991.
6. S.A. Foots, D.B. Grindeland: Model QA 3000 K-FLEX Accelerometer High Performance Tests Results, IEEE-0-7803-0468, 1992.
7. D. Janiaud, O. Le Traon, S. Muller, P. Bouniol: The VIA accelerometer, an occurate low cost miniature sensor, Symposium Gyro-Technology, Stuttgart Germany, 1998.
8. A. Cazenave, D. Dédié: proposition de mission pour μsatellite, CNES (1999).
9. O.L. Colombo: Geodetic theory, *Rev. of Geophysics* **25**, 851 (1987).
10. R. Rummel, O.L. Colombo: Gravity field determination from satellite gradiometry, *Bull. Geod.* **59**, 233 (1985).
11. A. Bernard, P. Touboul: A spaceborne gravity gradiometer for the nineties, IAG annual meeting Edinburgh U.K., 1989.
12. K. Danzmann et al.: LISA, Interferometer Space Antenna, MPQ 208, Garching, Germany, 1996.
13. C.W.F. Everitt et al.: Satellite Test of the Equivalence Principle 'Quick STEP' phase A study, report JPL D-12453, 1995.
14. R. Bonneville: GEOSTEP: a Gravitation Experiment in earth Orbiting Satellite to Test the Equivalence Principle, proceedings COSPAR 31, Birmingham, 1995.
15. T.J. Kacpura and J.C. Acevedo: Space Acceleration Measurement System for Free Flyers (SAMS-FF) - Initial test results, AIAA-98-0454.
16. R.E. Jenkins: Performance in orbit of the TRIAD disturbance compensation system, *APL Technical Digest* **12**, 1973.
17. A. Bernard, M. Gay, R. Juillerat: The accelerometer CACTUS, AGARDograph, 254, 1982.
18. W.G. Lange, R.W. Dietrich: The MESA accelerometer for space application, NASA Conference Publication 3088, 1990.
19. J.E. Rice: OARE STS-78 (LMS-1). Final Report, Canopus Systems Inc., 1996.
20. H. Hammacher, R. Jilg, H.E. Richter: QSAM – An Approach to Detect Low Frequency Acceleration in Spacelab, 24th International Conference on Environmental Systems, Friedrichshafen Germany, SAE tech. paper 941362, 1994.

21. J.W. Parke: Null test of the gravitational inverse square law and the development of a superconducting six–axis accelerometer, Thesis, University of Maryland, 1990.
22. P. Touboul, B. Foulon, G.M. Le Clerc: The accelerometer ASTRE for the microgravity environment monitoring on board space laboratoires. Final Report 10/3825 PY, Onera Ed., Châtillon France, April 1996.
23. D. Horrière, B. Foulon, P. Touboul: ASTRE on MSL-1 Spacelab mission – Post mission analysis, Report 15/3825 DMPH/Y, ONERA Ed., Châtillon France, July 1998.
24. C. Reiberg: CHAMP a challenging micro-satellite payload for geophysical research and application, GFZ Final Report, Postdam, Germany, 1995.
25. R. Kasper: Satellite CHAMP, lecture given at 277^{th} Jenaer Carl–Zeiss–Optikkolloquium, Jena, Germany, 1998.
26. P. Touboul, B. Foulon: Space accelerometer developments and drop tower experiments, *Space Forum* **4**, 145 (1998).
27. B.D. Tapley: Gravity Recovery and Climate Experiment (GRACE), Proposal to NASA's Earth System Science Pathfinder Program, 1996.
28. Gravity Recovery and Climate Experiment (GRACE), Science and Mission Requirements Document, 327-200, JPL, Pasadena, CA, 1998.
29. Gravity field and steady state Ocean Circulation Mission, Earth Explorer mission selection report, ESA SP-1233(1), Noordwijk, Netherland, 1999.
30. R. Rummel, F. Sanso, M. Van Gelderen, M. Brovelli, R. Koop, F. Migliaccio, E.J.O. Shramma and F. Sacerdote: Sperical Harmonics Analysis of Satellite Gradiometry, *Neth. Geodetic Commission* N.39 (New Series), 1993.
31. G. Balmino, F. Perosanz, R. Rummel, N. Sneeuw, H. Sünkel, P. Woodworth: Some views on dedicated gravity field missions: GRACE and GOCE, European Space Agency document, 1998.
32. P. Touboul, E. Willemenot, B. Foulon and V. Josselin: Accelerometers for CHAMP, GRACE, GOCE, synergy and evolution, Proceedings 2nd Joint Meeting of the IGC and IGeC commissions, Trieste 1998, Bolletino di Geophisicae, in press.
33. P. Touboul and G. Metris: μSCOPE, μSatellite à Compensation de traînée pour l'Observation du Principe d'Equivalence, proposition de mission μsatellite, CNES 1998.
34. Y. Su, B.R. Heckel, E.G. Adelberger, J.H. Gundlach, M. Harris, G.L. Smith and H.E. Swanson: New tests of the universality of free fall, *Phys. Rev.* **D 50** 3614 (1994).
35. J.O. Dickey: Linear laser ranging: continuing legacy of the Apollo program, *Science* **265**, 482 (1994).
36. P. Touboul, M. Rodrigues, E. Willemenot, A. Bernard: Electrostatic accelerometers for the Equivalence Principle, Proc. Symposium of Fundamental Physics in Space, London UK (1995).
37. M. Rodrigues and P. Touboul: Specification de la charge utile pour le projet μSCOPE, Onera Châtillon France, Report RTS 27/3815 DMPH/Y (1999).
38. E. Willemenot: Pendule de torsion à suspension électrostatique, très hautes résolution des accéléromètres spatiaux pour la physique fondamentale, *Thèse* (Paris XI, 1997).
39. V. Josselin: Etalonnage en orbite des accéléromètres ultrasensibles pour le test du principe d'équivalence, *Thèse* (Paris XI, 1999).
40. M. Rodrigues, P. Touboul, G.M. Le Clerc: The Inertial Reference Sensor CAESAR for the Laser Interferometer Space Antenna Mission, Symposium on Fundamental Physics in Space, London, UK, 1995.

Searching for Extra Dimensions and New String-Inspired Forces in the Casimir Regime

Dennis E. Krause[1,2] and Ephraim Fischbach[2]

[1] Wabash College, Crawfordsville, IN 47933-0352, USA
[2] Purdue University, West Lafayette, IN 47907-1396, USA

Abstract. The appearance of new fundamental forces and extra-dimensional modifications to gravity in extensions of the Standard Model has motivated considerable interest in testing Newtonian gravity at short distances ($\lesssim 10^{-3}$ m). Presently a number of new gravity experiments are searching for non-Newtonian effects in the ranges $\sim 10^{-4}$–10^{-3} m. However, as challenging as these experiments are, formidable new obstacles await the next generation of experiments which will probe gravity at distances $\lesssim 10^{-4}$ m where Casimir/van der Waals forces become dominant. Here we will review the motivation for conducting such very short distance gravity experiments, and discuss some of the new problems that may arise in future experiments. Finally, we suggest schematic designs for null experiments which would address some of these problems using the "iso-electronic" and "finite-size" effects.

1 Introduction

When Isaac Newton formulated his law of universal gravity over 300 years ago, he provided the first mathematical description of one of the fundamental forces of nature. Yet, physicists have realized only relatively recently that tests of Newtonian gravity can still provide a unique window into new physics [1–7]. Within the past 20 years, experimentalists have put Newtonian gravity to the test for distance scales 10^{-3}–10^{15} m by searching for violations of the weak equivalence principle (WEP) and inverse square law (ISL). The fact that no such violations have been observed places stringent constraints on extensions of the Standard Model that would naturally lead to such effects [1]. Despite this effort, a number of authors have pointed out that very little is known of the validity of Newtonian gravity at distances $\lesssim 10^{-3}$ m [8–12]. Several experimental groups are currently attempting to extend these limits down to 10^{-4} m [9,13,14], which is near the point where Casimir/van der Waals forces overcome gravity to become the dominant force between neutral, non–magnetic bodies. This strong intermolecular force background will become a major challenge for experimentalists who attempt to probe gravity at much smaller distances. The purpose of this paper is to show that such experiments are worth the effort despite the new difficulties, and to suggest ideas which may be useful in detecting new forces of gravitational strength against a strong intermolecular force background.

We begin by examining the theoretical motivation for studying Newtonian gravity and the phenomenology used to characterize non-Newtonian effects which would be the signal of new physics in a gravity experiment. After briefly reviewing the current constraints on new forces achieved by longer distance gravity

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 292–309, 2001.

experiments ($\gtrsim 10^{-3}$ m), we will see that new problems arise when one attempts to set comparable limits in the Casimir/van der Waals regime ($\lesssim 10^{-4}$ m) where intermolecular forces become large. We then investigate these problems quantitatively by computing the forces between two parallel plates over a range of separations. Finally, we propose two schematic designs for null experiments designed to subtract out the unwanted intermolecular and gravitational backgrounds using the "iso-electronic" and "finite-size" effects. These will hopefully allow one to search for signals of new physics at very short distances in the presence of Casimir/van der Waals forces.

2 Theoretical Motivation and Phenomenology

2.1 Overview

The Standard Model currently provides an adequate description of the electromagnetic, weak, and strong interactions within the framework of quantum field theory. However, a consistent description of quantum gravity has yet to be formulated despite intense work over the past fifty years. The lack of a quantum theory of gravity currently provides much of the motivation for studying extensions of the Standard Model which would bring all the fundamental forces into the quantum realm. In fact, many believe that the present Standard Model is really only an effective theory which would be superseded at much higher energies by a more fundamental theory, such as string or M-theory [15,16]. One of the main problems with these more fundamental theories is that, despite their purported mathematical beauty, many of their principal consequences lay far beyond the reach of most foreseeable experiments. It is therefore vital to investigate the low-energy limits of these fundamental theories to allow experimentalists the opportunity to constrain the proliferation of models which would otherwise go unchecked.

It is against this backdrop that one should view recent and future experiments testing Newtonian gravity. Many extensions of the Standard Model, including string theory, contain new light bosons which would manifest themselves as new fundamental forces[1,6,7,12]. These new forces would compete with the other known forces, but they would most likely be revealed in a gravity experiment for several reasons. First, in many ways gravity remains the least understood of the fundamental forces and is relatively untested over a wide range of distance scales. Second, any new forces probably couple very weakly with matter—otherwise they would have been seen already. Since gravity is by far the weakest fundamental force, it sets a natural scale from which to measure new weak forces. Third, there are two signatures of gravity which help one extract a signal from the background of other forces: 1) Since the gravitational force couples to mass, it obeys the weak equivalence principle (WEP), so violations of the WEP would indicate the presence of a non-gravitational force. 2) The Newtonian gravitational force between point particles obeys an inverse square law (ISL), hence any departures from the ISL might be attributed to new forces. Finally, Newtonian gravity is the weak-field, non-relativistic limit of General Relativity, a

theory in which gravity is seen as a manifestation of spacetime that has been curved by mass-energy. Therefore, any theory that impacts our understanding of space and time must involve gravity. This is important because string theory requires that there exist more than three spatial dimensions, the new extra dimensions being rendered invisible to current experiments by some yet-to-be-understood mechanism. As will be discussed below, recent models suggest that these new dimensions may modify Newtonian gravity at short, but macroscopic, distances.

2.2 Yukawa Potentials

The form of the violations of Newton's law of gravity arising from new physics will be to some extent model dependent, but one finds in practice that most theories yield modifications that have similar generic features [1]. For example, suppose there exists a new vector field $A^\mu(x)$ which couples to fermions via the Lagrangian density

$$\mathcal{L}(x) = if\,\overline{\psi}(x)\gamma_\mu\psi(x)\,A^\mu(x). \tag{1}$$

Here f is the dimensionless vector-fermion coupling constant ($\hbar = c = 1$) and $\psi(x)$ is the fermion field operator. If two fermions 1 and 2 exchange a single vector boson with mass m via this coupling, the lowest order interaction in the non-relativistic limit yields a Yukawa potential

$$V_v(r) = \pm\frac{f_1 f_2}{4\pi}\,\frac{e^{-mr}}{r}, \tag{2}$$

where "+" ("−") indicates that the force is repulsive (attractive) between like charges. If this was electromagnetism, massless photons give $m = 0$, and for electrons $f = -e$. In the units we use, the range of the interaction is $\lambda \equiv 1/m$, so that $m \sim 10^{-5}$ eV gives $\lambda \sim 1$ cm, for example. If the exchanged bosons were scalars instead of vectors, one arrives at an attractive Yukawa potential between identical fermions:

$$V_s(r) = -\frac{f_1 f_2}{4\pi}\,\frac{e^{-mr}}{r}, \tag{3}$$

where f_i is now the scalar coupling constant.

If the fermions have masses m_1 and m_2, the total interaction potential including gravity and scalar/vector interactions can be written in the general form

$$V(r) = -\frac{Gm_1m_2}{r}\left(1 + \alpha_{12}\,e^{-r/\lambda}\right), \tag{4}$$

where

$$\alpha_{12} \equiv \mp\frac{f_1 f_2}{4\pi G m_1 m_2}. \tag{5}$$

The dimensionless constant α_{12} then characterizes the strength of the interaction relative to gravity, and its sign depends on type of boson exchanged. When $r \ll \lambda$, $|\alpha_{12}| = 1$ indicates a force of gravitational strength.

Yukawa potentials also arise naturally in models where new gravitational forces appear from extra spatial dimensions. String theory requires there to be more than 3 spatial dimensions, but until recently it was thought that all the extra dimensions were compactified on the Planck scale and thus invisible to any conceivable experiment. However, much attention recently has focused on a number of string-inspired models in which all Standard Model particles are confined to the usual 3 spatial dimensions (a 3-brane) while gravity can "see" all dimensions[17–21]. In such models, one would never expect to see the effects of extra dimensions in Standard Model physics, but their effects would appear in gravitational physics. Since so much of the parameter space of gravity remains unexplored, these effects could have easily escaped detection. For example, in models in which the extra dimensions are compact, it is possible that the compactification radius r_c could be as large as 10^{-3} m and thus would have not been seen in any experiment to date [17,18]. These models would produce dramatic deviations from Newtonian gravity at short distances since they imply that the r-dependence of the gravitational potential between point masses changes when the particle separation approaches r_c:

$$V_{\rm grav}(r) = \begin{cases} -\dfrac{G_4\, m_1 m_2}{r} & \text{for } r \gg r_c \\ -\dfrac{G_{4+n}\, m_1 m_2}{r^{1+n}} & \text{for } r \ll r_c. \end{cases} \tag{6}$$

Here n is the number of extra spatial dimensions, $G_4 = G$ is the usual macroscopic Newtonian gravitational constant, and G_{4+n} is the more fundamental gravitational constant for the total $4 + n$ dimensional spacetime. Thus, in this model, Newton's law of gravity is merely a projection of a more fundamental law of gravity onto 3 spatial dimensions, and the unusual weakness of gravity relative to the other fundamental forces is attributed to this projection. One of the striking features of some recent string models is that compactification can occur over scales much larger than the Planck scale ($1/M_{\rm Planck} \sim 10^{-35}$ m)[17,18,21]. For example, if in a $4 + n$ dimensional spacetime the fundamental mass scale $M_{\rm fund} \sim M_{\rm EW}$, where $M_{\rm EW} \sim 1$ TeV is the electroweak scale, then one expects the compactification scale r_c to be given by [17,18,21]

$$r_c \sim \frac{1}{M_{\rm fund}} \left(\frac{M_{\rm Planck}}{M_{\rm fund}} \right)^{2/n} \sim (10^{-19}\ {\rm m})(10^{16})^{2/n}, \tag{7}$$

which yields,

$$r_c \sim \begin{cases} 10^{13}\ {\rm m}, & n = 1, \\ 10^{-3}\ {\rm m}, & n = 2, \\ 10^{-9}\ {\rm m}, & n = 3. \end{cases} \tag{8}$$

Since no deviations from Newtonian gravity have been observed for $r \gtrsim 10^{-3}$ m, theories which suggest that $r_c \lesssim 10^{-3}$ m (e.g., $n \geq 2$) are compatible with current experimental limits. As one tests gravity over smaller distance scales, the effects of new extra dimensions would first appear as corrections to the usual

gravitational potential. It has been shown that these corrections for $r \gg r_c$ have a Yukawa form [18,22,23]:

$$V_{\rm grav}(r) \sim -\frac{G_4\, m_1 m_2}{r}\left(1+\alpha_n e^{-r/\lambda}\right). \tag{9}$$

Here α_n is a composition-independent universal constant that depends on n and the nature of the compactification, and the range of the interaction is $\lambda \sim r_c$, the compactification scale. For example, for n extra dimensions compactified on an n-torus, $\alpha_n = 2n$ [18,22,23]. It is thus conceivable that the first evidence supporting the existence of extra spatial dimensions (and string theory) could come from the detection of a composition-independent Yukawa modification of Newton's law of gravity.

2.3 Current Constraints on New Yukawa Forces

Let us now turn to the current laboratory constraints on new Yukawa forces which arise from a generic potential of the form,

$$V_Y(r) = -\alpha\, \frac{G m_1 m_2}{r}\, e^{-r/\lambda}. \tag{10}$$

(Here we assume that the interactions are attractive for positive α.) This potential will lead to a violation of the WEP in a gravity experiment if α is composition-dependent, as in the case of a vector or scalar interaction. In addition, even if α is independent of the composition of the test masses, as in the case of the extra dimension theories described earlier, small violations of the WEP will still be present in an experiment using different materials due to the "finite-size" effect [1]. This effect arises because a non-uniform Yukawa field will "capture" different fractions of two finite-sized objects having the same mass, but different densities, as will be the case in the null experiments considered below. Finally, in addition to violating the WEP, $V_Y(r)$ will also violate the ISL, and so constraints on α can be inferred from tests of the gravitational ISL.

As shown in Fig. 1, the current experimental constraints on the Yukawa coupling constant α as a function of range λ are quite stringent (allowed $|\alpha| \ll 1$) for 10^{-3} m $\lesssim \lambda \lesssim 10^{15}$ m, but they fall exponentially outside this region. Composition-dependent experiments have set strong limits for specific couplings (e.g., to baryon number) when $\lambda \gtrsim 10^{15}$ m [1], but also fall off exponentially for $\lambda \lesssim 10^{-3}$ m.

As discussed in more detail in Ref. [9], current constraints allow $\alpha \gtrsim 1$ for $\lambda \lesssim 10^{-3}$ m. For 10^{-4} m $\lesssim \lambda \lesssim 10^{-3}$ m, these limits were obtained from a test of the gravitational ISL by Mitrofanov and Ponomareva [24], but they still permit a new force with $\alpha \sim 10^4$ for $\lambda \sim 10^{-4}$ m. However, a new round of gravity experiments [9,13,14] should fill in much of this region of parameter space within the next few years. At shorter distances, Casimir/van der Waals forces dominate gravity so the current limits are set by Casimir force experiments [9,25–27] and are much less restrictive than those obtained from the longer ranged gravity experiments. [See also Refs. [28,29] for detailed discussions on extracting limits on

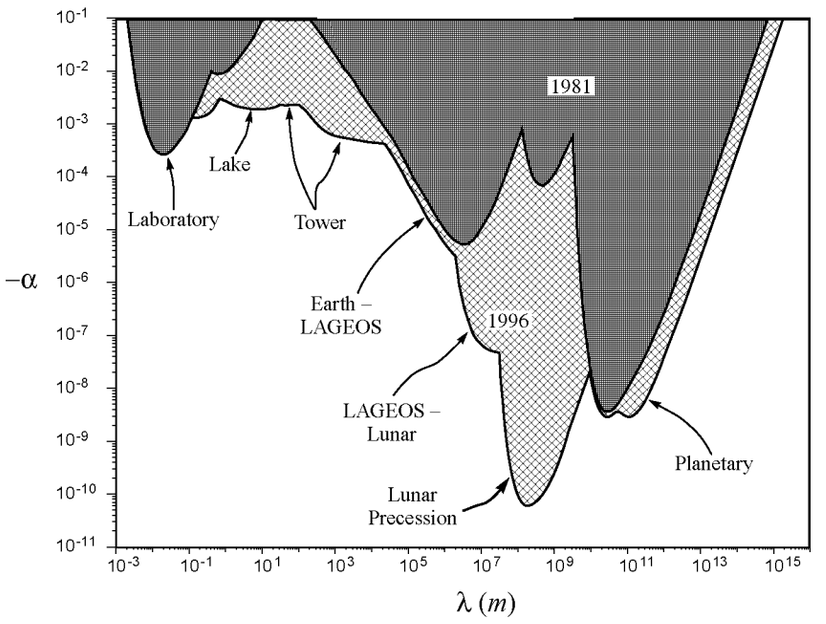


Fig. 1. 2σ constraints on the coupling constant α as a function of the range λ from composition-independent experiments [1]. The dark shaded area indicates the region excluded as of 1981, and the light hatched region gives the 1996 limits which remain current.

new forces from Casimir force experiments.] The region $\lambda \lesssim 10^{-4}$ m will remain essentially unexplored until the next generation of experiments specifically dedicated to search for new forces is designed and carried out. We turn next to a discussion of some of the difficulties likely to be encountered in developing this next round of experiments.

3 Problems in Testing Gravity at Very Short Distances

3.1 General Problems

As noted in the Introduction, a number of authors [8–12] have called attention to the huge gap in our understanding of gravity at very short distances, and to its potential to reveal new physics. The fact that short-distance gravity experiments can potentially expose the presence of extra spatial dimensions is particularly tantalizing. However, since $|\alpha| \sim 1$ in these string models, the ultimate experimental goal is quite challenging, namely to set limits $|\alpha| \lesssim 1$ for $\lambda \lesssim 10^{-3}$ m. To accomplish this, one has to be able to sense and distinguish a force of gravitational strength at these distance scales. Since the current laboratory limits

in this region are orders of magnitude less sensitive than this goal, we discuss briefly some of the difficulties in studying gravity at short distances.

An obvious problem is scaling [9]. Suppose we have two identical spheres of density ρ and radius R. If we wish to test the law of gravity between these spheres at short distances, the dimensions of the spheres have to be made comparable to the small distance scales that we are interested in probing. Since the minimum separation distance between centers is $2R$, this leads to a maximum force

$$F_{max} = \frac{G(4\pi\rho R^3/3)^2}{(2R)^2} \propto R^4. \tag{11}$$

This example illustrates that the gravitational force between macroscopic objects, which is already quite small, decreases rapidly with size and separation of the test masses.

A second problem in searching for new short-ranged forces is that their short range limits the effective mass of a body that can participate in the interaction. Suppose we have a sphere of uniform density and radius R. Since gravity is a long-range force, another identical sphere close by will interact with all of the mass of the first sphere. However, if there exists a new force of range $\lambda \ll R$, then only the layer of material of thickness $\sim \lambda$ on the surface of the sphere will interact with external objects, and hence only a fraction $\sim \lambda/R$ of the total mass participates in interactions. But this problem is actually much worse in general. If we have two spheres nearly touching, it is only the mass within a range λ of the contact point that interacts, which is much less than the fraction λ/R, while the gravitational force is still felt by all the mass. Therefore, even if the new force is intrinsically of gravitational strength ($\alpha = 1$) between point masses when $r \ll \lambda$, this new force between macroscopic bodies will usually be much smaller than the corresponding gravitational force. This situation was not encountered in previous longer range gravity experiments since in those cases, $\lambda \gtrsim L$, where L was the characteristic size of the test bodies used.

A third problem occurs when the separation distance is $\lesssim 10^{-4}$ m, where intermolecular forces become significant. Since these forces have a power-law form $1/r^n$ [30], where n depends on the geometry of the macroscopic bodies, they grow very rapidly as r decreases and overwhelm gravity at very short distances. Distinguishing a force of gravitational strength from this background will be a major challenge.

3.2 Quantitative Example: Parallel Plate Gravity Experiment

Idealized Setup

To better appreciate how these problems might arise in actual experiments, let us now consider a simple experimental setup. Our goal here is to estimate the size of the various effects which might appear, and not to propose the optimal experimental design, and hence we will ignore practical problems which might be encountered when one actually attempts to realize such a design. Since we are searching for forces of very short range, the discussion of the previous section

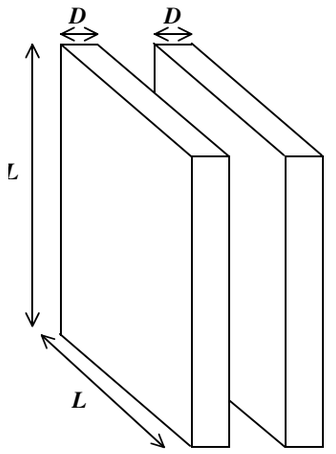


Fig. 2. The idealized parallel plate setup used to quantitatively estimate the relative magnitudes of the gravitational, Casimir, and Yukawa forces. Numerical results were obtained by letting $L = 1$ cm, $D = 1$ mm, $\rho_{\text{copper}} = 8.96 \times 10^3$ m^3, $T = 300$ K, and $d \ll L$.

suggests we should have most, if not all, of the mass of the two test bodies in an experiment contributing in order to realize the largest possible force. This means that we need to have all the mass in one body as close as possible to all the mass of the second body. The simplest way to accomplish this is to use parallel plates as our test bodies [29] which maximizes the "effective mass" for any short-range range force. It then follows that the most appropriate configuration for searching for new forces between macroscopic bodies is a parallel-plate experiment analogous to those used to study the Casimir effect [31].

Let us now consider two identical plates of density ρ, thickness D, area $A = L^2$, separated by a distance d (Figure 2). If we assume $d \ll L$, we can then safely neglect edge effects and calculate the pressures between the plates as if $L = \infty$. In addition, we assume that the plates are perfectly smooth, perfectly conducting, and at temperature $T = 300$ K.

Force Formulas

We begin our investigation of the forces between these plates with the known forces, starting with gravity. For this particular configuration, the gravitational force acting on the plates is given by

$$F_{\text{Gravity}}(d) = -2\pi G\rho^2 L^2 D^2, \tag{12}$$

where the minus sign indicates an attractive force. We see that when the plates are sufficiently close, the gravitational attraction is constant, independent of the separation d.

If there are no stray charges, etc., gravity is the dominant force at large plate separations, but as d decreases, the Casimir force grows rapidly and quickly overwhelms gravity. Calculating the Casimir force for this geometry for real metals can become quite complicated, involving corrections for finite conductivity and surface roughness [32–34]. However, for present purposes we will ignore these difficulties by assuming the plates to be smooth and perfectly conducting over all frequencies, which should be a good approximation as long as the plates are not too close. However, we will include thermal effects which become large when the plate separation is large. The Casimir force between our plates at temperature

T can be written as [35,36]

$$F_{\rm Casimir}(d) = -\frac{\pi^2}{240}\frac{\hbar c L^2}{d^4} - \frac{\pi k T L^2}{d^3}\sum_{n=1}^{\infty} n^2 \ln\left[1 - \exp\left(-\frac{n\pi\hbar c}{kTd}\right)\right] - \frac{\pi^2 (kT)^4 L^2}{45(\hbar c)^3}, \tag{13}$$

where k is Boltzmann's constant and factors of $\hbar$ and c have been included for convenience. Eq. (13) simplifies in the two limiting cases [35,36]:

$$F_{\rm Casimir}(d) = \begin{cases} -1.202\left(\dfrac{kTL^2}{4\pi d^3}\right), & d \gg \dfrac{\pi\hbar c}{kT}, \\ -\dfrac{\pi^2\hbar c L^2}{240 d^4}, & d \ll \dfrac{\pi\hbar c}{kT}. \end{cases} \tag{14}$$

Having obtained expressions for the known forces acting between the plates in this idealized setup, let us now determine the forces arising from possible new interactions. The attractive Yukawa potential between point masses as given by Eq. (10) leads to a force between the plates (with $d \ll L$) given by

$$F_{\rm Yukawa}(d) = -2\pi\alpha\lambda^2 G\rho^2 L^2 \left(1 - e^{-D/\lambda}\right)^2 e^{-d/\lambda}. \tag{15}$$

We then notice that the ratio of this Yukawa force to the gravitational force in Eq. (12) is

$$\frac{F_{\rm Yukawa}(d)}{F_{\rm Gravity}(d)} = \alpha\left(\frac{\lambda}{D}\right)^2 \left(1 - e^{-D/\lambda}\right)^2 e^{-d/\lambda}. \tag{16}$$

If $\lambda \ll D$, then

$$\frac{F_{\rm Yukawa}(d)}{F_{\rm Gravity}(d)} \simeq \alpha\left(\frac{\lambda}{D}\right)^2 e^{-d/\lambda}. \tag{17}$$

Thus, even if the Yukawa coupling is intrinsically of gravitational strength ($\alpha = 1$), the actual Yukawa force is suppressed relative to gravity not only by the usual exponential factor $e^{-d/\lambda}$, but also by $(\lambda/D)^2$ which arises because only a fraction λ/D of the total mass of each plate contributes to the Yukawa force. This effect was discussed earlier and clearly illustrates how a short-ranged force intrinsically of gravitational strength is strongly suppressed in an experiment using macroscopic bodies.

Numerical Results

Having found the general formulas for all the forces that we will be considering, let us now obtain numerical values for the following setup. We assume that the plates have dimensions $L \times L \times D$, where $L = 1$ cm and $D = 1$ mm, which are roughly comparable to the values used in some of the current short distance

Table 1. The magnitudes of the gravitational, Casimir, and Yukawa (using $\alpha = 1$ and $\lambda = 10^{-5}$ m) forces arising in the idealized parallel plate experiment discussed in the text. Here $F_{\rm Back} = F_{\rm Gravity} + F_{\rm Casimir}$ is the total background force against which the signal of $F_{\rm Yukawa}$ must be seen.

d (m)	$F_{\rm gravity}$ (N)	$F_{\rm Casimir}$ (N)	$F_{\rm Yukawa}$ (N)	$F_{\rm Yukawa}/F_{\rm Back}$
10^{-3}	3.4×10^{-12}	4.2×10^{-17}	1.3×10^{-59}	4×10^{-48}
10^{-4}	3.4×10^{-12}	4.0×10^{-14}	1.5×10^{-20}	4×10^{-9}
10^{-5}	3.4×10^{-12}	4.0×10^{-11}	1.2×10^{-16}	3×10^{-6}
10^{-6}	3.4×10^{-12}	1.3×10^{-7}	3.0×10^{-16}	2×10^{-9}
10^{-7}	3.4×10^{-12}	1.3×10^{-3}	3.3×10^{-16}	3×10^{-13}
10^{-8}	3.4×10^{-12}	1.3×10^{1}	3.4×10^{-16}	3×10^{-17}

gravity experiments [9,13]. Since our previous calculations assumed $d \ll L$, we focus our attention on the region 10^{-8} m $\lesssim d \lesssim 10^{-3}$ m. Next we will assume that the plates are made of pure copper which has a density $\rho = 8.96 \times 10^3$ kg/m^3. Except for the new force parameters α and λ, our problem is now completely specified.

Using these numbers, we first calculate the known forces, gravity and Casimir, for the plates. As discussed earlier, the gravitational force under the conditions assumed here is constant and given by Eq. (12). Substituting the parameters given above yields

$$F_{\rm Gravity} = 3.37 \times 10^{-12} \text{ Newton.} \tag{18}$$

To determine the Casimir force for this configuration, we use Eq. (13). The cross-over distance d_c, where temperature-dependent effects become important at $T = 300$ K, is

$$d_c = \frac{\pi \hbar c}{kT} = 2.4 \times 10^{-5} \text{ m} = 24 \ \mu\text{m}. \tag{19}$$

Graphs of the Casimir force using Eq.(14), and the gravitational force between the plates, are shown together in Figure 3, and numerical values of these forces at various distances can be found in Table 1. The gravitational and Casimir forces are equal to each other when $T = 300$ K at $d = 2.3 \times 10^{-5}$ m $= 23$ μm, which just happens to coincide with d_c here. Thus, for $d \lesssim 23$ μm, the Casimir force will dominate gravity in this setup.

Now let us turn to new Yukawa forces, which are characterized by two free parameters, the relative strength α and the range λ. If for illustrative purposes we consider a force of gravitational strength ($\alpha = 1$) and set $\lambda = 10^{-5}$ m, then Eq. (15) yields $F_{\rm Yukawa}(d)$ exhibited in Figure 3 and Table 1. We see that when $d \ll \lambda$, the force becomes constant:

$$F_{\rm Yukawa}(d \ll \lambda) = 3.37 \times 10^{-16} \text{ Newton}, \tag{20}$$

which is $(\lambda/D)^2 = 10^{-4}$ times smaller than the corresponding gravitational force given by Eq. (18). Thus, as explained earlier, even though the Yukawa force

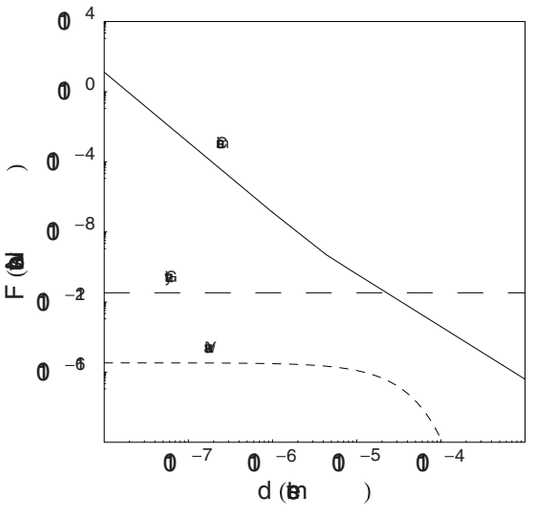


Fig. 3. The Casimir, gravitational, and Yukawa ($\alpha = 1$, $\lambda = 10^{-5}$ m) forces between the parallel plates shown in Fig.2.

between *point* particles is of gravitational strength at short distances, $F_{\rm Yukawa}$ is much smaller than gravity for these macroscopic plates. We also clearly see that $F_{\rm Yukawa}/F_{\rm Back}$ is maximized when $d \sim \lambda$ and falls off rapidly from this plate separation (Fig. 4). This is because $F_{\rm Yukawa}$ levels off when $d \lesssim \lambda$ while $F_{\rm Casimir}$ continues to increase via a power-law ($1/d^4$ if $d \lesssim d_c$).

Constraining New Short–Ranged Yukawa Forces

This analysis using an obviously idealized setup reveals the two critical problems that will be encountered in devising experiments using macroscopic bodies to search for very short–ranged Yukawa interactions of gravitational strength ($\alpha \sim 1$). The first is that the absolute magnitude of such a force will be very small, possibly even smaller than the gravitational force if α is small. Thus, an experiment must be sensitive to the smallest possible forces. Second, since the Casimir background force grows rapidly as the separation decreases, one must be able to extract the signal of a very weak force from a background of very strong intermolecular forces. A direct attack on this problem would be to attempt to calculate as accurately as possible the background forces in a gravity experiment, and to then subtract these from the observed force to set limits using what remains [9,26,28,29]. However, recent experiments studying the Casimir force reveal the difficulty of accurately calculating the background

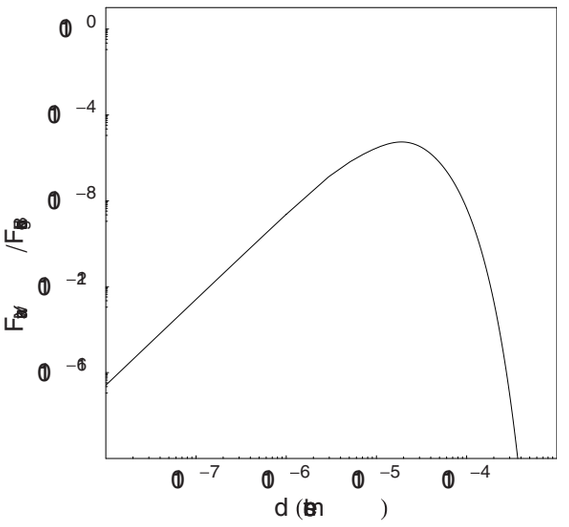


Fig. 4. The ratio of the Yukawa ($\alpha = 1$, $\lambda = 10^{-5}$ m) force to the background force $F_{\text{Back}} = F_{\text{Gravity}} + F_{\text{Casimir}}$ for the parallel plate setup shown in Fig. 2.

forces to better than 1% [25,27,32–34,37,38]. While this approach is still possible, we will describe in the next section a new method of performing a null short-distance gravity experiment specifically designed to directly subtract out the unwanted background effects. Calculating intermolecular forces precisely then becomes much less important.

4 Very Short Distance Null Gravity Experiments

Some of the best constraints on Yukawa interactions come from tests of the WEP [1]. In these experiments, one compares the accelerations of compositionally-different test bodies toward a common source body. Any differences in these accelerations can then be attributed to non-gravitational forces. We wish to utilize the same principle in a short-distance gravity experiment, and thus avail ourselves to the extreme sensitivity of such experiments.

4.1 Null Experiment #1

Inspired by two ongoing short distance experiments [9,13], a possible design for one such experiment is shown in Figure 5. It consists of two parallel plates, a source plate 1 and a detector plate made of two smaller plates 2 and 2′.

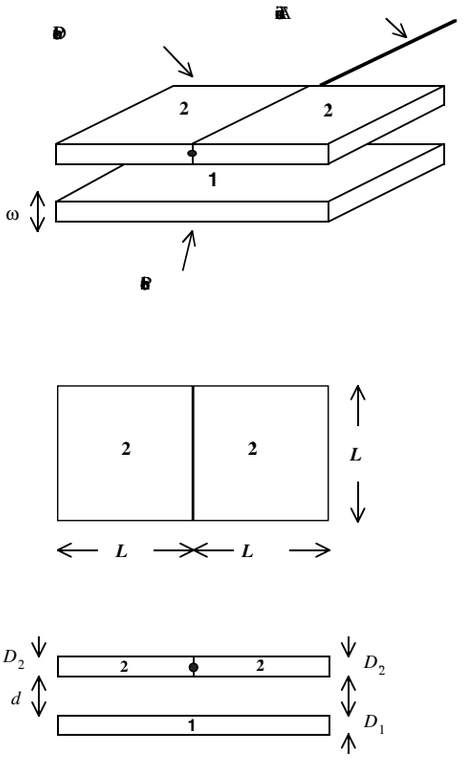


Fig. 5. Schematic design for Very Short Distance Null Experiment #1. See text for details.

The source plate is driven sinusoidally with angular frequency ω such that the separation distance d is given by

$$d(t) = d_0 + d_1 \cos \omega t. \tag{21}$$

Instead of detecting a force, this experiment would be sensitive to a torque, modulated by the frequency ω, about an axis passing along the boundary where plates 2 and $2'$ are joined, as shown in Fig. 5. If we assume that the plates are conducting, the net torque $\tau_{\rm net}$ on the detector plate will arise from contributions

from the gravitational, Casimir, and possibly, Yukawa forces:

$$\begin{aligned}\tau_{\rm net} &= \tau_{\rm Gravity} + \tau_{\rm Casimir} + \tau_{\rm Yukawa} \\ &= \frac{L}{2}\left[(F_2^{\rm Gravity} - F_{2'}^{\rm Gravity}) + (F_2^{\rm Casimir} - F_{2'}^{\rm Casimir})\right. \\ &\quad \left. +(F_2^{\rm Yukawa} - F_{2'}^{\rm Yukawa})\right], \end{aligned} \tag{22}$$

where $F_i^{\rm Gravity}$, $F_i^{\rm Casimir}$, and $F_i^{\rm Yukawa}$ are the gravitational, Casimir, and Yukawa forces on plates $i = 2$ and $2'$ respectively. One then selects plates 2 and $2'$ such that the torque $\tau_{\rm Gravity} + \tau_{\rm Casimir}$ due to background forces vanishes while $\tau_{\rm Yukawa} \neq 0$ if $\alpha_{1i} \neq 0$. [Here we allow for the possibility that α_{12} and $\alpha_{12'}$, the Yukawa couplings between the materials comprising plates 1 and 2, and 1 and $2'$ respectively, are different.]

At very small separations, $F_i^{\rm Gravity}$ will be negligible (and independent of ω to first approximation) , but it is still easy to make $\tau_{\rm Gravity}$ vanish anyway. Using Eq. (12), we see that

$$|F_2^{\rm Gravity} - F_{2'}^{\rm Gravity}| = 2\pi G L^2 \rho_1 D_1 \left|\rho_2 D_2 - \rho_{2'} D_{2'}\right|, \tag{23}$$

where ρ_i and D_i are the density and plate thickness of the ith plate. Then

$$(\rho_2 D_2 = \rho_{2'} D_{2'}) \Rightarrow (M_2 = M_{2'}) \Rightarrow \tau_{\rm Gravity} = 0, \tag{24}$$

so if the detector plates 2 and $2'$ have the same mass, the gravitational torque will vanish.

Of course, the much bigger challenge is choose materials for plates 2 and $2'$ such that the Casimir torque $\tau_{\rm Casimir}$ also vanishes. If the plates were perfectly conducting, this would be the case since Eq. (13) would be identical for all such plates with the same surface area. However, the finite conductivity of real metallic plates becomes very important when the plate separation $d \sim \lambda_P$, where $\lambda_P = 2\pi c/\omega_P$, and ω_P is the plasma frequency of the metal. Still, it was shown recently [32] that the Casimir force between pairs of copper and gold plates are equal to a good approximation for separations $d \gtrsim 10^{-6}$ m at $T = 0$. Such calculations are difficult for real materials, but this certainly raises the hope that it is possible to choose appropriate plates 2 and $2'$ such that

$$F_2^{\rm Casimir} - F_{2'}^{\rm Casimir} \simeq 0 \;\;\Rightarrow\;\; \tau_{\rm Casimir} \simeq 0. \tag{25}$$

At the very least, one should be able to fabricate the plates using two different isotopes of the same element (e.g., ^{24}Mg and ^{26}Mg) such that Eq. (25) is satisfied. The underlying premise of the "iso-electronic" effect (IE) is that to a good approximation the Casimir effect depends on the *electronic* properties of the materials, and hence is largely independent of their *nuclear* properties. By contrast, the gravitational interaction, and virtually all proposed new Yukawa interactions, involve couplings to both electrons and nucleons. Hence, subtracting out the electronic contributions by choosing two isotopes of some material, or by choosing materials with similar electronic properties (such as Cu and Au), we

can enhance the signal from a new Yukawa force while simultaneously reducing the Casimir background.

The remaining torque after the gravitational and Casimir torques have been suppressed might be due to a new Yukawa force. Using Eq. (15), the net torque due to a putative Yukawa force would be

$$\tau_{\rm Yukawa} = 2\pi G\lambda^2 (L/2) L^2 \rho_1 \left(1 - e^{-D_1/\lambda}\right) e^{-d/\lambda} \times \left[\alpha_{12}\rho_2 \left(1 - e^{-D_2/\lambda}\right) - \alpha_{12'}\rho_{2'} \left(1 - e^{-D_{2'}/\lambda}\right)\right]. \tag{26}$$

If $\lambda \ll D_i$, then Eq. (26) simplifies to

$$\tau_{\rm Yukawa} \simeq 2\pi G\lambda^2 (L/2) L^2 \rho_1 e^{-d/\lambda} \left[\alpha_{12}\rho_2 - \alpha_{12'}\rho_{2'}\right]. \tag{27}$$

If the Yukawa force arises from an extra-dimensional modification of Newtonian gravity such that $\alpha_{12} = \alpha_{12'} = \alpha_n$, Eq. (27) reduces to

$$\tau_{\rm Yukawa} \simeq 2\pi \alpha_n G\lambda^2 (L/2) L^2 \rho_1 \rho_2 e^{-d/\lambda} \left(1 - \frac{\rho_{2'}}{\rho_2}\right). \tag{28}$$

This result depends on the difference in the mass densities $\rho_2 - \rho_{2'}$ for a simple reason: When $\lambda \ll D_i$, the force only sees the mass within a distance λ of the surface. Thus, if plates 2 and $2'$ have different densities, the effective mass seen by the Yukawa force will be different and a net torque arises due to the "finite-size" effect discussed earlier. It is then clear that one should choose materials 2 and $2'$ such that ρ_2 and $\rho_{2'}$ differ as much as possible while still ensuring that the Casimir torque vanishes. For the gold/copper and ^{24}Mg/^{26}Mg combinations suggested above,

$$1 - \frac{\rho_{2'}}{\rho_2} \simeq \begin{cases} 0.32 & 2 = \text{Au},\ 2' = \text{Cu}, \\ 0.077 & 2 = {}^{26}\text{Mg},\ 2' = {}^{24}\text{Mg}. \end{cases} \tag{29}$$

The hope is that the suppression factor Eq. (29) is more than compensated by the reduction of the unwanted background torques.

4.2 Null Experiment #2

We conclude this section by briefly describing another possible design for a short distance null experiment motivated by another set of gravity experiments [14,39]. As shown schematically in Figure 6, this experiment consists of two disks, one serving as the source mass, while the other (detector mass) is the pendant of a torsion pendulum. Each disk is divided into alternating wedges made of two different materials 1 and 2. The wedges and materials 1 and 2 are designed such that no Casimir torque would arise when the source disk rotates below the pendulum. For example, as indicated above, 1 and 2 might be gold/copper or ^{24}Mg/^{26}Mg which would significantly reduce the Casimir torque. Then, if the separation between the disks is small, the remaining torque on the pendulum would arise from a putative Yukawa force because the effective mass within a distance of λ will be different for alternating wedges.

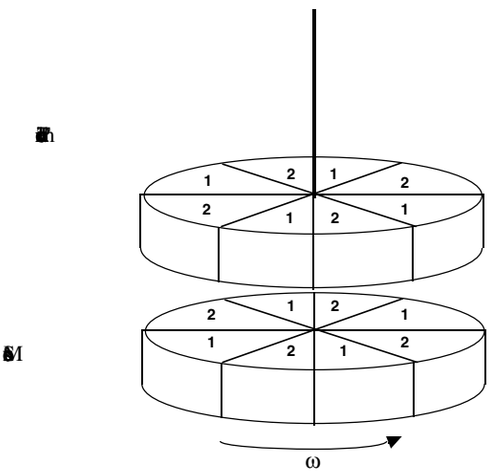


Fig. 6. Schematic design for Very Short Distance Null Experiment #2. See text for details.

5 Discussion

It is clear that there is significant motivation for testing Newtonian gravity at very short distances. However, as we have seen, new problems will face experimentalists who wish to extend current constraints on the Yukawa coupling constant α down to ranges $\lambda \lesssim 10^{-4}$ m. Our aim in this paper has been to point out some of the most obvious difficulties, but there may be others that have passed unnoticed. While we have not addressed the significant issue of improving the sensitivity of experiments to very small forces, we have taken the first steps towards dealing with the problem of detecting small forces against a large intermolecular force background. Our schematic designs for null experiments are meant as illustrations of the principles involved in canceling background forces (the "iso-electronic" and "finite-size" effects), and hence are not intended to suggest optimal designs. It is hoped that experimentalists, who face the realities of imperfect materials and incomplete theories, can extract some useful ideas from these schematic designs or, perhaps, will be able to point to flaws which preclude them from working as actual experiments. Finally, we conclude with the encouraging note that since so little is known about gravity at separations $\lesssim 10^{-4}$ m, virtually *any* good experiment in this region will tell us something new.

Acknowledgments

We wish to thank P. Boynton, G. Carugno, C. Deufel, D. Koltick, A. Lambrecht, J. Mullen, R. Newman, R. Reifenberger, S. Reynaud, and C. Talmadge for very useful discussions. D. Krause also acknowledges the support of Wabash College and Purdue University, and this work was supported in part by the U.S. Department of Energy under Contract No. DE-AC 02-76ER01428.

References

1. E. Fischbach and C. Talmadge: *The Search for Non-Newtonian Gravity* (AIP Press/Springer-Verlag, New York, 1999).
2. I. Ciufolini and J. A. Wheeler: *Gravitation and Inertia* (Princeton University Press, Princeton, 1995).
3. C. M. Will: *Theory and Experiment in Gravitational Physics*, revised edition (Cambridge University Press, New York, 1993).
4. Y. T. Chen and A. Cook: *Gravitational Experiments in the Laboratory* (Cambridge University Press, New York, 1993).
5. E. Fischbach, G.T. Gillies, D.E. Krause, J.G. Schwan, and C. Talmadge: *Metrologia* **29**, 213 (1992).
6. E.G. Adelberger, B.R. Heckel, C.W. Stubbs, and W.F. Rogers: *Annu. Rev. Nucl. Part. Sci.* **41**, 269 (1991).
7. Y. Fujii: *Int. J. Mod. Phys.* **A 6**, 3505 (1991).
8. J.C. Price: in *International Symposium on Experimental Gravitational Physics*, edited by P. Michelson, H. Enke, and G. Pizzella (World Scientific, Singapore, 1988), pp. 436–439.
9. J.C. Long, H.W. Chan, and J.C. Price: *Nucl. Phys.* **B 539**, 23 (1999).
10. R. Onofrio: *Mod. Phys. Lett.* **A 13**, 1401 (1998).
11. S.R. Beane: *Gen. Rel. Grav.* **29**, 945 (1997).
12. S. Dimopoulos and G. F. Giudice: *Phys. Lett.* **B 379**, 105 (1996).
13. G. Carugno, Z. Fontana, R. Onofrio, and C. Rizzo: *Phys. Rev.* **D 55**, 6591 (1997).
14. E.G. Adelberger, et al.: (C. Deufel, personal communication.)
15. J. Polchinski: *String Theory*, Volumes 1 and 2 (Cambridge University Press, New York, 1999).
16. M. Kaku: *Introduction to Superstrings and M-Theory* (Springer-Verlag, New York, 1999).
17. I. Antoniadis, S. Dimopoulos, and G. Dvali: *Nucl. Phys.* **B 516**, 70 (1998); N. Arkani-Hamed, S. Dimopoulus, and G. Dvali: *Phys. Lett.* **B 429**, 263 (1998); I. Antoniadis, N. Arkani–Hamed, S. Dimopoulos, and G. Dvali: *Phys. Lett.* **B 436**, 257 (1998).
18. N. Arkani–Hamed, S. Dimopoulos, and G. Dvali: *Phys. Rev.* **D 59**, 086004 (1999).
19. L. Randall and R. Sundrum: *Phys. Rev. Lett.* **83**, 4690 (1999).
20. G. Shiu and S.H.H. Tye: *Phys. Rev.* **D 58**, 106007 (1998).
21. S. Nussinov and R. Shrock: *Phys. Rev.* **D 59**, 105002 (1999).
22. A. Kehagias and K. Sfetsos: hep-ph/9905417.
23. E.G. Floratos and G.K. Leontaris: *Phys. Lett.* **B 465**, 95 (1999).
24. V.P. Mitrofanov and O. I. Ponomareva: *Sov. Phys. JETP* **67**, 1963 (1988).
25. S.K. Lamoreaux: *Phys. Rev. Lett.* **78**, 5 (1997); **81**, 5475(E) (1998).

26. M. Bordag, B. Geyer, G. L. Klimchitskaya, and V. M. Mostepanenko: *Phys. Rev.* **D 60**, 055004 (1999).
27. U. Mohideen and A. Roy: *Phys. Rev. Lett.* **81**, 4549 (1998); A. Roy, C. -Y. Lin, and U. Mohideen: *Phys. Rev.* **D 60**, 111101 (1999).
28. M. Bordag, B. Geyer, G. L. Klimchitskaya, and V. M. Mostepanenko: in *The Casimir Effect 50 Years Later*, edited by M. Bordag (World Scientific, Singapore, 1999), pp. 39–49; M. Bordag, B. Geyer, G. L. Klimchitskaya, and V. M. Mostepanenko: *Phys. Rev.* **D 58**, 075003 (1999); V. M. Mostepanenko and I. Yu. Sokolov: *Phys. Rev.* **D 47**, 2882 (1993).
29. V. M. Mostepanenko and I. Yu. Sokolov: in *Quantum Gravity*, edited by. M. A. Markov, V. A. Berezin, and V. P. Frolov (World Scientific, Singapore, 1990), pp. 213—232; V. M. Mostepanenko and I.Yu. Sokolov: in *Modern Problems of Theoretical Physics*, edited by P.I. Pronin and Yu.N. Obukhov (World Scientific, Singapore 1991), pp. 175–196.
30. J. Israelachvili: *Intermolecular & Surface Forces*, 2nd edition (Academic Press, New York 1992).
31. H.B.G. Casimir: *Proc. Kon. Ned. Akad. Wet.* **51**, 793 (1948).
32. A. Lambrecht and S. Reynaud: *Eur. Phys. J.* **D 8**, 309 (2000).
33. G.L. Klimchitskaya, A. Roy, U. Mohideen, and V. M. Mostepanenko: *Phys. Rev.* **A 60**, 3487 (1999).
34. S.K. Lamoreaux: *Phys. Rev.* **A 59**, R3149 (1999).
35. J. Mehra: *Physica* **37**, 145 (1967).
36. L.S. Brown and G.J. Maclay: *Phys. Rev.* **184**, 1272 (1969).
37. S.K. Lamoreuax: quant-ph/9907076.
38. S.K. Lamoreuax: *Phys. Rev. Lett.* **83**, 3340; U. Mohideen and A. Roy: *Phys. Rev. Lett.* **83**, 3341 (1999).
39. M. W. Moore, A. Boudreaux, M. DePue, J. Guthrie, R. Legere, A. Yan, and P. E. Boynton: *Class. Quantum Grav.* **11**, A97 (1994).

Relativistic Effects in the Motion of the Moon

Bahram Mashhoon[1] and Dietmar S. Theiss[2]

[1] Department of Physics and Astronomy, University of Missouri-Columbia, Columbia, Missouri 65211, USA
[2] Institute for Theoretical Physics, University of Cologne, 50923 Köln, Germany

Abstract. The main general relativistic effects in the motion of the Moon are briefly reviewed. The possibility of detection of the solar *gravitomagnetic* contributions to the mean motions of the lunar node and perigee is discussed.

1 Introduction

In a recent paper, Gutzwiller has provided an admirable review of the oldest three-body problem, namely, the Sun-Earth-Moon system [1]. Some work on the relativistic theory is mentioned in his paper; however, in view of the recent advances in relativistic celestial mechanics this subject deserves a more complete discussion. Here we provide a brief description of the main relativistic effects.

The lunar laser ranging experiment has opened up the possibility of measuring relativistic effects in the motion of the Moon; indeed, the agreement between the standard general relativistic model that contains over a hundred model parameters and the ranging data accumulated over the past three decades is excellent [2,3]. For instance, the post-fit residuals in the Earth-Moon distance are at the centimeter level [2,3]. Simple theoretical estimates lead to the conclusion that the main relativistic effects in the lunar theory are due to the spin-orbit coupling of the Earth-Moon system in the gravitational field of the Sun. The post-Newtonian influence of the solar field on the lunar motion consists of terms that can be classified as either harmonic (i.e. periodic) or secular (i.e. cumulative) in time. It turns out to be very difficult in practice to separate the harmonic terms from the corresponding Newtonian terms with the same periodicities. In effect, the existence of the post-Newtonian harmonic terms leads to small relativistic corrections in the numerical values of certain model parameters that are thereby adjusted by a fit to the ranging data. To give an example of such harmonic effects, we mention our prediction of a 6 cm relativistic tidal variation in the Earth-Moon distance with a period of 1/2 synodic month [4].

The main secular terms turn out to be essentially due to the precessional motion of the Earth-Moon orbital angular momentum in the field of the Sun. The Earth-Moon system can be thought of as an extended gyroscope in orbit about the Sun; we are interested in the description of the motion of the spin axis of this gyroscope with respect to the "fixed" stars (i.e. the sidereal frame). An ideal pointlike test gyroscope carried along a geodesic orbit would exhibit, in the post- Newtonian approximation, geodetic precession due to the orbital motion around the mass of the source as well as gravitomagnetic precession due

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 310–316, 2001.

to the intrinsic rotation of the source; however, the finite size of the gyroscope in this case (i.e. the orbit of the Moon about the Earth) leads to additional tidal effects. The main post-Newtonian gravitoelectric effect, i.e. geodetic precession, results in the advance of the Moon's node and perigee by about 2 arcseconds per century as first predicted by de Sitter already in 1916 [5]. This motion has been measured by Shapiro *et al.* with an accuracy of about one percent [6]. It is a relativistic three-body effect; therefore, we consider in the next section the restricted three-body problem in general relativity and briefly indicate, in particular, the more subtle post-Newtonian gravitomagnetic contributions to the motions of the Moon's node and perigee that are caused by the rotation of the Sun; indeed, solar rotation induces *cumulative* relativistic tidal effects in the Earth-Moon system [4].

2 Restricted Three-Body Problem in General Relativity

In our previous work [4], we developed a new scheme for the approximate treatment of the restricted three-body problem in general relativity. This coordinate-invariant approach is particularly useful for a reliable theoretical description of relativistic (solar) tidal effects in the motion of the Earth-Moon system. We assume that the Moon follows a geodesic in the gravitational field generated by the Earth and the Sun. This field may be calculated as follows: we first imagine that the Earth follows a geodesic in the solar field. Along this geodesic, we set up a geocentric Fermi coordinate system. This system, which involves the tidal field of the Sun, is then enhanced by taking due account of the field of the Earth in the linear approximation. Tidal effects in general relativity involve the projection of the Riemann tensor onto the tetrad frame of the measuring device. Consider the tidal matrix for a test system ("Earth") in free fall in the gravitational field of a rotating mass ("Sun"). In the standard first-order post-Newtonian treatment, the spatial axes of the local tetrad frame along the orbit are obtained by boosting the background Minkowski axes and adjusting scales to maintain orthonormality; the resulting tidal matrix for an approximately circular geodesic orbit turns out to be sinusoidal in time [7]. In this case, the tetrad frame is not parallel-transported, but its motion involves the Lense-Thirring orbital precession as well as the geodetic (i.e. de Sitter-Fokker) precession of the spatial axes. Once the parallel transport of the spatial axes along the orbit is imposed, the gravitomagnetic (i.e. Schiff) precession of the spatial axes would also appear in the first post-Newtonian order. In this order, the tidal matrix for the parallel-transported axes contains a secular term as well that must therefore be a direct consequence of the Schiff precession of the spatial axes [8], in agreement with our previous work [9-11]. The linear growth of this gravitomagnetic contribution to the tidal field poses a problem for the first post-Newtonian approximation: the non-Newtonian "off-diagonal" part of the tidal matrix can diverge in time [9-11]. To avoid this limitation, we have developed a post-Schwarzschild treatment of gravitomagnetic tidal effects; indeed, the concept of relativistic nutation

provides a natural resolution of this difficulty by limiting the temporal extent of validity of the post-Newtonian approximation [9-11].

Imagine, for instance, a set of three orthogonal test gyroscopes falling freely along an inclined circular geodesic orbit with constant radius r ("astronomical unit") about a slowly rotating central body ("Sun") with mass M and proper angular momentum J. The motion of the spin axes of these torque-free gyroscopes, which constitute a local inertial frame (i.e. the geocentric Fermi frame), is essentially governed by the equations of parallel transport along the geodesic orbit. By solving these equations using the post-Schwarzschild approximation scheme that takes M into account to all orders, it can be shown that the average motion of the gyroscope axes with respect to an effective Newtonian (i.e. sidereal) frame consists of a gravitoelectric precessional motion—i.e. geodetic precession that was first completely analyzed by Fokker— together with a complex gravitomagnetic motion that can be loosely described as a combination of precessional movement and a harmonic nodding movement. The latter motion is a new relativistic effect of a rotating mass and has been referred to as *relativistic nutation* [11]. In the post-Newtonian approximation, the nutational terms over a limited time combine with the other gravitomagnetic precessional terms to give the Schiff precession. To see how this comes about, let us denote by τ the proper time of the geodesic orbit and consider a vector normal to the orbital plane (ecliptic) at the beginning of measurement ($\tau = 0$). Relativisticnutation is a periodic variation of the angle between this vector and a gyroscope axis that is Fermi propagated along the orbit. The leading contribution of relativistic nutation to this angle can be written as

$$\Theta_n \approx \xi \left[\sin(\eta_0 + \omega_F \tau) - \sin \eta_0\right] \sin \alpha, \tag{1}$$

where η_0 is the azimuthal position of the Earth in the ecliptic at $\tau = 0$ measured from the line of the ascending nodes and $\xi = J/Mr^2\omega$. Here $\omega, \omega^2 = GM/r^3$, approximately describes the orbital frequency in the absence of rotation and α denotes the inclination of the orbit with respect to the equatorial plane of the Sun [12]. The frequency of this nutational oscillation is the Fokker frequency $\omega_F \approx \frac{3}{2}\epsilon\omega$, where $\epsilon = GM/c^2 r$. The nutation amplitude, $\xi \sin \alpha$, does not depend on the speed of light c. This remarkable fact can be traced back to the occurrence of a small divisor [9-11] involving the Fokker frequency. In the post-Newtonian limit of the post-Schwarzschild approximation, Eq. (1) reduces to $\Theta_n \sim \omega_n \tau$, which represents a *precessional* motion with frequency $\omega_n = \xi \omega_F \sin \alpha \cos \eta_0$ about a direction opposite to that of orbital velocity at $\tau = 0$. Thus, relativistic nutation reduces to a *part* of the Schiff precession in the first post-Newtonian approximation. It follows from this analysis that the first post-Newtonian approximation breaks down over timescales of the order of Fokker period $\tau_F = 2\pi/\omega_F$; however, this fact does not diminish the usefulness of the first post-Newtonian approximation for the description of observations in the solar system since in this case the Fokker period is almost immeasurably long (e.g. $\tau_F \simeq 67$ million years for the motion of the Earth about the Sun).

Let us consider the influence of the gravitomagnetic field of the central body ("Sun") on the relative acceleration of two nearby test particles ("Earth" and

"Moon") moving along the circular geodesic orbit. The dominant contributions of the gravitomagnetic field of the central body to the tidal matrix, first calculated by the authors [9-11], are proportional to

$$\omega^2 \xi \sin\alpha \sin\left(\frac{1}{2}\omega_F \tau\right), \tag{2}$$

which is directly proportional to the amplitude of relativistic nutation ($\xi \sin\alpha$) and exhibits a maximum (at $\tau = \tau_F/2$) that is independent of the speed of light c. It follows from Eq. (2) that to first order in $\omega_F \tau \ll 1$, the dominant gravitomagnetic amplitude varies linearly with τ. This secular amplitude originates from a coupling of the *nutation part* of Schiff precession with the amplitude ($\sim \omega^2$) of the Newtonian contribution to the gravity gradient [13]. It should be mentioned in passing that the relativistic quadrupole contributions to the tidal matrix have properties quite similar to the gravitomagnetic tidal effect described here [10].

Let us now turn to the potentially observable effects of the solar gravitomagnetic field on the lunar motion. The lunar path is determined by the Newton-Jacobi equation

$$\frac{d^2 x^i}{d\tau^2} + \frac{Gm}{R^3} x^i = -K_{ij}(\tau) x^j, \tag{3}$$

where $x^i, i = 1, 2, 3$, represent the geocentric Fermi coordinates of the Moon, m is the total mass of the Earth-Moon system and $R(\tau)$ denotes the Earth-Moon distance depending on the proper time τ measured along the geocentric path around the Sun. Here K is the tidal matrix. Equation (3) describes the motion of the Moon with respect to a geocentric local inertial frame [14]. Using the equation of relative motion (3), we have calculated—among other things—the influence of the tidal field of the Sun on the orbital angular momentum of the Moon with respect to the Earth. To express the result with respect to the sidereal frame, we choose as our sidereal reference framethe geocentric Fermi frame at $\tau = 0$. This frame is related to the Fermi frame at time τ by a rotation matrix that incorporates the relativistic precession and nutation of the Fermi frame with respect to the sidereal frame. In the first post-Newtonian approximation, this motion reduces to a (Fokker plus Schiff) precession. Let D denote this rotation matrix, then $\mathcal{L}_i = D_{ij} L_j$, where the sidereal components ($\mathcal{L}_i$) of the orbital angular momentum are obtained from a transformation of the geocentric components (L_j) with

$$D_{ij} = \delta_{ij} - \epsilon_{ijk} \Phi_k \quad , \quad \Phi = \int_0^\tau \omega_{FS}(\tau') d\tau'. \tag{4}$$

Here the analysis is limited to the first post-Newtonian approximation and ω_{FS} represents the frequency of (Fokker plus Schiff) precession. The direction of Schiff precession is not fixed along the Earth's orbit; therefore, Φ contains (cumulative) secular terms (which represent simple precession) together with (harmonic) nutational terms of frequency 2ω and amplitude of order $\alpha\epsilon\xi$. Averaging over the

latter terms, the dominant secular terms in Φ are given by

$$\Phi_1 \sim \frac{1}{2}\alpha\epsilon\xi\omega\tau\sin\eta_0 \quad , \quad \Phi_2 \sim \left(-\frac{3}{2}+\xi\right)\epsilon\omega\tau \quad , \quad \Phi_3 \sim \frac{1}{2}\alpha\epsilon\xi\omega\tau\cos\eta_0 \tag{5}$$

with respect to the geocentric Fermi frame at $\tau = 0$, which has its first axis essentially along the radial position of the Earth, the third axis approximately along the direction of motion of the Earth and the second axis normal to the ecliptic (in a direction opposite to the Earth's orbital angular momentum about the Sun). We note that for $\xi = 0$, Eq. (5) expresses the de Sitter-Fokker effect that has been observed by Shapiro *et al.* [6]. To illustrate our approach, let us use Eq. (3) to determine the value of $\mathbf{L}$, which is the angular momentum of the Moon in a circular orbit about the Earth with respect to the Fermi frame, averaged over orbital motions of the Earth about the Sun (with frequency ω) and the Moon about the Earth (with frequency Ω). Then

$$\frac{d\langle\mathbf{L}\rangle}{d\tau} = \tilde{\omega}\times\mathbf{L_0} \quad , \tag{6}$$

where $\mathbf{L_0}$ is the unperturbed orbital angular momentum with respect to the geocentric Fermi frame and $\tilde{\omega}$ is given by

$$\tilde{\omega}^1 \approx -\tilde{\omega}_0\alpha\epsilon\xi\left(2\sin\eta_0+\frac{3}{2}\omega\tau\cos\eta_0\right), \tag{7}$$

$$\tilde{\omega}^2 \approx \tilde{\omega}_0(1-6\epsilon\xi), \tag{8}$$

$$\tilde{\omega}^3 \approx -\tilde{\omega}_0\alpha\epsilon\xi\left(2\cos\eta_0-\frac{3}{2}\omega\tau\sin\eta_0\right) \tag{9}$$

to first order in the tidal perturbation characterized by the Newtonian regression frequency $\tilde{\omega}_0 = 3\omega^2/4\Omega$, which corresponds to a period of nearly 18 years [15,16]. It is clear from Eqs. (4)-(9) that the motion of $\langle\mathcal{L}\rangle$ can be expressed as a *Newtonian* regression modulated by long-term (secular) *relativistic* perturbations characterized by the de Sitter-Fokker, Schiff and gravitomagnetic tidal effects. To illustrate this point, let us assume for the sake of simplicity that in the absence of relativistic effects the lunar orbital angular momentum undergoes a steady regression of frequency $\tilde{\omega}_0$ and that once relativistic effects are included the average motion in the Fermi frame is one of precession with the frequency given by Eqs. (7)-(9). It then follows that the expression for $\langle\mathcal{L}_2\rangle$, i.e. the average of the second sidereal component of the lunar orbital angular momentum, contains a dominant gravitomagnetic contribution of the form

$$\langle\mathcal{L}_2\rangle_{\rm secular} \approx 2\alpha\beta\epsilon\xi(\mu R_0^2\Omega)\omega\tau\sin(\eta_0+\zeta_0-\tilde{\omega}_0\tau), \tag{10}$$

where β, μ, R_0 and ζ_0 denote, respectively, the inclination of the lunar orbit with respect to the ecliptic ($\approx 5^\circ$), the mass of the Moon, the mean Earth-Moon separation and the longitude of the ascending node of the orbit of the Moon measured from the first axis of the sidereal frame. These simple considerations that are based on an initial circular orbit only indicate the nature of the secular terms involved; clearly, extensive calculations are necessary for a complete treatment.

3 Discussion

The results of our theoretical work are of particular interest for the description of dominant relativistic gravitational effects in the motion of the Moon, especially the gravitomagnetic tidal component of the motion of the orbital angular momentum of the Moon. It is important to point out that the eccentricities of the orbit of the Earth around the Sun and the orbit of the Moon about the Earth should be taken into account; we have ignored them in our preliminary analysis [16]. As lunar laser ranging data further accumulate, it may become possible in the future to deduce the angular momentum of the Sun from the measurement of the solar gravitomagnetic contributions to the mean motions of the lunar node and perigee.

It is interesting to compare our secular gravitomagnetic tidal terms with hypothetical terms that might indicate a temporal variation of the gravitational "constant" G. Our results have thus far been based on a secular term proportional to τ in the tidal matrix K in Eq. (3); however, as can be seen from the middle term in Eq. (3), similar effects could be produced if such a term appears in G instead. We have shown that our predictions are similar to a variation of G in Eq. (3) at the level of 10^{-16} yr^{-1}; moreover, there are significant differences between the two effects that can be used to separate them [4,9]. The present upper limit on $|\dot{G}/G|$ is at the level of 10^{-12} yr^{-1}; therefore, it may be a long while before the gravitomagnetic effects in the motion of the Moon become detectable.

Acknowledgments

We would like to thank Friedrich Hehl for fruitful discussions. We are also grateful to Sergei Kopeikin for helpful comments. This work was supported in part by the Deutsche Forschungsgemeinschaft, Bonn.

References

1. M.C. Gutzwiller: *Rev. Mod. Phys.* **70**, 589 (1998).
2. J.O. Dickey *et al.*: *Science* **265**, 482 (1994).
3. K. Nordtvedt, these proceedings; *Phys. Today* **49**, no. 5, 26 (1996).
4. B. Mashhoon and D.S. Theiss: *Nuovo Cimento* **B 106**, 545 (1991). A misprint in Eq. (64) of this paper should be corrected: The overall sign of the right-hand side of Eq. (64) should be negative. This correction would make that equation compatible with Eq. (8) of the present paper.
5. W. de Sitter: *Mon. Not. Roy. Astron. Soc.* **77**, 155 (1916).
6. I.I. Shapiro, R.D. Reasenberg, J.F. Chandler and R.W. Babcock: *Phys. Rev. Lett.* **61**, 2643 (1988). See also B. Bertotti, I. Ciufolini and P.L. Bender: *Phys. Rev. Lett.* **58**, 1062 (1987).
7. B. Mashhoon, H.J. Paik and C.M. Will: *Phys. Rev.* **D 39**, 2825 (1989).
8. C.A. Blockley and G.E. Stedman: *Phys. Lett.* **A 147**, 161 (1990).
9. B. Mashhoon and D.S. Theiss: *Phys. Rev. Lett.* **49**, 1542 (1982); *Phys. Lett.* **A 115**, 333 (1986); B. Mashhoon: *Gen. Rel. Grav.* **16**, 311 (1984).

10. D.S. Theiss, Ph.D. thesis, University of Cologne (Köln, 1984); *Phys. Lett.* **A 109**, 19, 23 (1985); in *Relativistic Gravity Research*, edited by J. Ehlers and G. Schäfer (Springer, Berlin, 1992), p. 131; in *Proc. First William Fairbank Meeting on Relativistic Gravitational Experiments in Space*, edited by M. Demianski and C.W.F. Everitt (World Scientific, Singapore, 1993), p. 227; in *Proc. Seventh Marcel Grossmann Meeting on General Relativity*, edited by R.T. Jantzen and G. Mac Keiser (World Scientific, Singapore, 1996), p. 1555, p. 1558.
11. B. Mashhoon: *Found. Phys.* **15** (Bergmann Festschrift), 497 (1985). In this paper a misprint in Eq. (22) must be corrected: r_0 in the denominator of the last term must be replaced by c. Furthermore, in Eqs. (11) and (13) the temporal components must be divided by c.
12. For the motion of the Earth about the Sun, $\alpha \approx 7^\circ$, $\epsilon \approx 10^{-8}$ and $\xi \approx 2\times 10^{-5}$ based on the standard value of solar angular momentum [cf. C.W. Allen, *Astrophysical Quantities*, 3rd ed. (Athlone, London, 1973)]. The Fermi frame adopted in this paper is such that the nutation vanishes at $\tau = 0$. This frame can be obtained from the results given in Ref. [11]; see, especially, p. 506.
13. In this connection, it is also important to note that ω_n vanishes for an equatorial orbit ($\alpha = 0$) in contrast to the frequency of the *full* Schiff precession.
14. The precise definition of the notion of a local geocentric frame was formulated in [11] and further developed in S.M. Kopeikin: *Celestial Mech.* **44**, 87 (1988) and V.A. Brumberg and S.M. Kopeikin: *Nuovo Cimento* **B 103**, 63 (1989).
15. It follows from a more complete treatment of the Newtonian problem that the mean motion of the lunar node can be characterized by a backward movement of frequency $\tilde{\omega}_0 N$, which corresponds to a period of about 18.61 years. Similarly, the mean motion of the perigee can be characterized by a forward movement of frequency $\tilde{\omega}_0 P$, which corresponds to a period of about 8.85 years. The theoretical expressions for N and P are rather complicated and depend on ω/Ω as well as the orbital eccentricities, etc. The first two terms of N and P in terms of $\nu = \omega/\Omega$ are given by $N = 1 - 3\nu/8 - \cdots$ and $P = 1 + 75\nu/8 + \cdots$. A detailed discussion of this subtle problem is given by D. Brouwer and G.M. Clemence, *Celestial Mechanics* (Academic Press, New York, 1961), Ch.12, especially pp. 320-323.
16. If our analysis is extended by including a slight eccentricity for the lunar orbit, then the motion of the Runge-Lenz vector indicates an average forward precession of perigee with frequency $\tilde{\omega}_0\,(1 - 6\epsilon\xi)$. The relativistic gravitoelectric correction to this result is of order ϵ^2 in the post-Newtonian scheme; however, if the standard Schwarzschild coordinates are used in the analysis (cf. Ref. [4]), $\tilde{\omega}_0 \to \tilde{\omega}_0(1 + 3\epsilon)$, so that a slight increase in the forward (backward) precession rate of the perigee (node) would occur to first order in the tidal perturbation. This supersedes (and partly corrects) an assertion in Ref. (21) of our Nuovo Cimento paper (cf. Ref. [4]) regarding the comparison of our results with Robertson's "solar effect." However, a more complete analysis of the relativistic three-body problem is necessary for a reliable calculation of the motion of the lunar perigee.

Lunar Laser Ranging – A Comprehensive Probe of the Post–Newtonian Long Range Interaction

Kenneth Nordtvedt

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

Abstract. 30 years of lunar laser ranging (LLR) data has been modeled and fit with several millimeters precision using the general relativistic equations of motion for solar system dynamics. This produces several key tests of that tensor theory of gravity and strongly constrains presence of any supplementary interactions. Earth and Moon fall toward the Sun at rates equal to a couple parts in 10^{13}, confirming both the universal coupling of gravity to matter's *stress–energy tensor*, and gravity's specific non–linear coupling to itself. The expected *deSitter* precession (with respect to the distant 'fixed' stars) of the local inertial frame moving with the Earth–Moon system is confirmed to 3.5 parts in 10^3 precision, and *Newton's constant* indeed shows no cosmological time variation at the few parts in 10^{12} *per year* level. All the types of post–Newtonian terms in the N–body equation of motion — motional, *gravito–magnetic*, non–linear, inductive, etc. — contribute to the measured details of the lunar orbit, so LLR achieves 'near–completeness' as a gravity experiment and probe. The precision of these measurements, especially those connected with lunar orbit frequencies and their rates of change, should further improve as LLR observations continue into the future.

1 Introduction

In the late 1960s I was seeking new ways to test general relativity theory using the new possibilities of space–based experimentts. This led me to calculate the gravitational to inertial mass ratio for celestial bodies which contained appreciable internal gravitational binding energy. Most alternative theories of metric gravity yielded an 'anomalous' result — that this ratio for a body depends on its fractional gravitational binding energy [1]

$$\frac{M(G)}{M(I)} = 1 - \eta \, \frac{G}{2Mc^2} \int \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{y})}{|\boldsymbol{x}-\boldsymbol{y}|} d^3x \, d^3y \tag{1}$$

resulting in a body–dependent acceleration rate in external gravitational fields

$$\boldsymbol{a}_i \;=\; \left(\frac{M(G)}{M(I)}\right)_i \boldsymbol{g}_{\rm ex} \, . \tag{2}$$

Here η is a theory–dependent dimensionless coefficient sensitive to just about every post–Newtonian feature of theory and which vanishes in general relativity.

Shortly after this theoretical work I learned of the forthcoming lunar laser ranging (LLR) capability which would result from the Apollo landings on the Moon. LLR would be sufficiently sensitive to measure the difference between

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 317–329, 2001.

the acceleration rates of Earth and Moon toward the Sun to a scientifically interesting precision; there was a fortunate near–resonance amplification of a synodic month perturbation in the lunar orbit in proportion to any acceleration difference

$$\delta r_{\rm EM} \;\cong\; \left(1+2\frac{\omega}{\omega-\Omega}\right)\left(\left(\frac{M(G)}{M(I)}\right)_{\rm E}-\left(\frac{M(G)}{M(I)}\right)_{\rm M}\right)\times \tag{3}$$

$$\times\frac{g_{\rm Sun}}{\omega^2-(\omega-\Omega)^2}\;\cos(\omega-\Omega)t$$

$$\cong\; 10\,\eta\;\cos(\omega-\Omega)t \qquad \text{meter} \tag{4}$$

in which ω and $\omega-\Omega$ are the lunar orbit's sidereal and synodic frequencies, respectively [2].

The passive laser reflectors placed on the Moon by Apollo astronauts were developed in Robert Dicke's research group at Princeton University. One of their mission goals was to measure the changes in the lunar orbit which would result from any evolution of Newton's G in proportion to the Hubble expansion of the universe

$$\frac{1}{G}\frac{dG}{dt}\;\sim\;H\;\sim\;10^{-10}\quad\text{per year}\,.$$

A third LLR measurement of general and historical interest in gravitation eventually emerged from the data — 'geodetic precession' of the lunar orbit. At the dawn of the general relativistic age, within months of Einstein's publication of his gravitational theory in 1916, Wilhelm deSitter applied the theory to the calculation of post–Newtonian corrections in lunar motion [3]. The most promising outcome of his analysis was a relativistic contribution to the secular precession rate of lunar perigee due to the non–radial motion of the Earth–Moon system through the Sun's gravitational field

$$\Omega_{\rm dS}=\frac{3}{2}\frac{GM_{\rm Sun}}{c^2R^3}\;|\boldsymbol{R}\times\boldsymbol{V}|\,. \tag{5}$$

Angular astronomical observations never became precise enough to see this relativistic precession, but finally in the late 1980s the accumulated LLR data (since 1969) permitted its measurement [6,5], and with ever–increasing precision in subsequent years [7,13]. This precession, which can simply be thought of as the tensor field analog of the electromagnetic field's spin–orbit force, has also acquired the space–time geometrical interpretation of being the precession of the local inertial frame which at each instant accompanies the Earth–Moon system through the Sun's gravity.

Today, 30 years of LLR data allows not only the three measurements described above, but also a comprehensive collection of other measurements which map out the long range, post–Newtonian interaction between bodies in all its details. The ranging continues, and the measurement precision should further improve into the future.

2 Dynamical Equations for Bodies, Light, and Clocks

General relativity theory makes a number of predictions about details of the Moon's orbit which differ from those calculated from either Newtonian gravity or from alternative relativistic theories. That theoretical situation, coupled with an accumulated thirty years of laser ranging measurements to the Moon now reaching sub–centimeter levels of precision, and some fortuitous amplifications of the theory–differentiating perturbations found in the Sun–Earth–Moon dynamical system [8], results in *LLR today providing several of the most precise measurements supporting the hypothesis that general relativity's tensor gravity is a complete description of the interaction among bodies at macroscopic ranges.*

The most scientifically interesting features of post–Newtonian gravity which are measured with high precision in LLR are a consequence of the fully general N–body equations of motion. Analysis groups implement these equations to fit the LLR data in the solar system's barycentric frame. The Sun–Earth–Moon system dynamics is symbolically illustrated in Figure 1, with the rest of the solar system bodies sufficiently considered at the Newtonian level of perturbation. The Earth moves with velocity $\boldsymbol{V}$ and acceleration $\boldsymbol{A}$ with respect to the Sun, while the Moon is moving at velocity $\boldsymbol{V} + \boldsymbol{u}$ and acceleration $\boldsymbol{A} + \boldsymbol{a}$, and for purposes of discussing *preferred frame* effects if gravity is not locally Lorentz–invariant [4], the Sun (solar system) moves with cosmic velocity $\boldsymbol{W}$. The post–Newtonian forces on Earth and Moon from the Sun, each other, and themselves are dependent on these general motions, and there are *non–linear* gravitational forces for which each mass element of the Earth and Moon experiences forces due to the interactive effect of the Sun's gravity with the other mass elements of the same body, or of the other body.

The N–body equation of motion in metric gravity has been formulated in the literature for the completely general case [10]. *Preferred frame* effects (i.e., absence of local Lorentz–invariance in the dynamical equations) then generally will occur in solar system phenomena due to the high speed with which our system moves through the cosmos [4,12]. LLR data has been tested for such effects, and none are found to a part in 10^4 of their expected magnitude, so I will specialize consideration to locally Lorentz–invariant gravitational physics. Similarly, the conservation laws for total energy, momentum, and angular momentum, which we usually expect to hold in the physics of an isolated N–body system, are also not assumed in the general case. But analysis of the LLR data shows no evidence for any conservation law violations, so *a–posteriori*, I specialize to consideration of fully conservative, Lagrangian–based equations of motion.

Decades ago Eddington generalized the discussion of the gravitational physics in the static, spherically symmetric field of the Sun, in order to manifest some of the novel experimental possibilities from alternative gravitational theories. He concluded that two free parameters γ and β were sufficient to encompass the plausible variations in theory. It is remarkable that when the empirically confirmed foundations of local Lorentz–invariance and energy–momentum–angular momentum conservation laws are accepted, then only the two *Eddington* parameters γ and β are needed to parameterize the N–body equations of motion

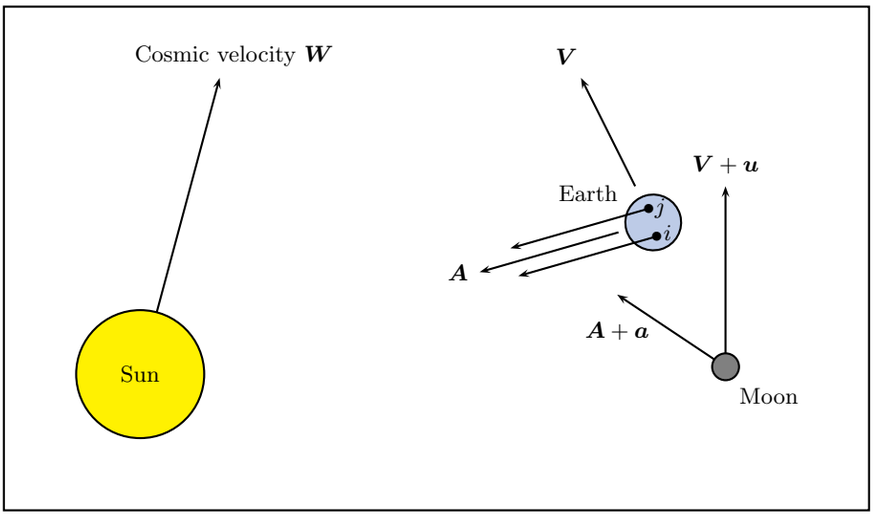


Fig. 1. When formulating the Earth–Moon dynamics in the solar system barycentric frame, there are post–Newtonian force terms acting between Sun, Earth, and Moon which depend on either the velocity or acceleration vectors of both the Earth and Moon. Body *self–accelerations* also result from the inductive inertial forces acting between the mutually accelerating mass elements $(i,\ j)$ within each of these bodies. The intrinsic non–linearity of gravity also produces net external forces on these bodies proportional to not only the presence of other bodies, but also to their internal gravitational binding energies. The motional, accelerative, and non–linear contributions to the three body system's dynamics, taken collectively, make LLR a comprehensive probe of the post–Newtonian dynamics of metric gravity in the general case. If the dynamics is not locally Lorentz invariant, then the velocity $\boldsymbol{W}$ of the solar system through the cosmos leads to novel forces and resulting observable effe cts in LLR proportional to $\boldsymbol{W}$ (or its square); but such effects have not been seen.

for the general dynamical situation. (It is possible that a very small part of the gravitational interaction could be of a finite range *Yukawa* nature, rather than inverse square; this possibility is separately tested with the LLR data and discussed later.)

For N bodies in general motion and configuration, and valid for a broad class of empirically plausible metric theories of gravity, the order $1/c^2$ equations of motion for these bodies have the form

$$A \qquad \boldsymbol{a}_i = \left(1+\frac{\dot{G}}{G}(t-t_0)\right)\left(\frac{M(G)}{M(I)}\right)_i \sum_j \boldsymbol{g}_{ij} \tag{6}$$

$$B \qquad -(2\beta-1)\sum_{j,k}\left(\frac{\mu_k}{r_{ik}}+\frac{\mu_k}{r_{jk}}\right)\boldsymbol{g}_{ij}$$

$$
\begin{aligned}
C \qquad & + \sum_j \left((2\gamma+2)\, \boldsymbol{v}_i \times (\boldsymbol{g}_{ij} \times \boldsymbol{v}_j) + (2\gamma+1)\, \boldsymbol{g}_{ij} \cdot \boldsymbol{v}_j\, \boldsymbol{v}_i \right) \\
D \qquad & + \sum_j \Big(\frac{1}{2} \left((2\gamma+1) v_i^2 + (2\gamma+2) v_j^2 - 3(\boldsymbol{v}_j \cdot \hat{\boldsymbol{r}}_{ij})^2 \right) \boldsymbol{g}_{ij} \\
& \qquad\qquad - (2\gamma+1) \left((\boldsymbol{g}_{ij} \cdot \boldsymbol{v}_j) \boldsymbol{v}_j + (\boldsymbol{g}_{ij} \cdot \boldsymbol{v}_i) \boldsymbol{v}_i \right) \Big) \\
E \qquad & + \frac{1}{2} \sum_j \frac{\mu_j}{r_{ij}} \left((4\gamma+3) \boldsymbol{a}_j + (\boldsymbol{a}_j \cdot \hat{\boldsymbol{r}}_{ij}) \hat{\boldsymbol{r}}_{ij} \right) \\
F \qquad & - \frac{1}{2} v_i^2\, \boldsymbol{a}_i - (\boldsymbol{a}_i \cdot \boldsymbol{v}_i) \boldsymbol{v}_i - (2\gamma+1) \sum_j \frac{\mu_j}{r_{ij}} \boldsymbol{a}_i \,,
\end{aligned}
$$

see page 7 of this Volume for an explanation of the symbols. The body gravitational mass strengths $\mu_i = GM_i$ are indicated along with the Newtonian acceleration vectors

$$\boldsymbol{g}_{ij} = \frac{\mu_j}{r_{ij}^3} \boldsymbol{r}_{ji} \,. \tag{7}$$

Some descriptions of the several lines of this total equation of motion are worthwhile.

Line *A*: Whenever the metric theory *Eddington* parameters γ and β differ from their general relativistic values $\gamma_{\rm GR} = \beta_{\rm GR} = 1$, theoretical consistency requires that the gravitational to inertial mass ratio of celestial bodies depend on the bodies' gravitational self–energy content

$$\frac{M(G)}{M(I)} \;=\; 1 \;-\; (4\beta - 3 - \gamma) \frac{G}{2Mc^2} \int \frac{\rho(\boldsymbol{x}) \rho(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|}\, d^3x\, d^3y + \mathcal{O}(1/c^4) \tag{8}$$

and Newton's coupling parameter G will generally vary in time in proportion to the Hubble expansion rate of the universe

$$\frac{1}{G} \frac{dG}{dt} \sim\; (4\beta - 3 - \gamma)\, H \,. \tag{9}$$

In LLR, it turns out that the most precise way to measure any deviation of β from its general relativistic value is through measurement of the $M(G)/M(I)$ ratio of Earth as given by Eq.(1).

Line *B*: Gravity couples to itself, thereby producing non–linear gravitational forces between bodies.

Line *C*: Just as pairs of moving charges generate magnetic forces between themselves in proportion to the velocities of both charges, pairs of moving masses generate *gravito–magnetic* forces between themselves.

Line *D*: Masses in motion both produce and couple to gravitational fields differently than masses at rest.

Line *E*: Accelerating masses generate inductive gravitational forces on other proximite masses.

Line *F*: The inertia of a mass is altered by its motion and by its proximity to other masses.

LLR involves precisely measuring the round trip time of propagation of light between two separate trajectories, and using a clock moving on one of those trajectories. So in the solar system barycentric and isotropic coordinates employed to express the body equations of motion, there are also model requirements for the post–Newtonian modifications to the light coordinate speed function

$$c(\boldsymbol{r}) = c_\infty \left(1 - (1+\gamma)\, U(\boldsymbol{r})/c^2\right) \tag{10}$$

and for clock rates

$$d\tau = dt \left(1 - U(\boldsymbol{r})/c^2\right) , \tag{11}$$

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

$$U(\boldsymbol{r}) = \sum_j \int \frac{G\rho(\boldsymbol{r}')_j}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3r' . \tag{12}$$

It is necessary to use these equations in the modeling of the LLR data in order to achieve parameter fits consistent with those obtained from geocentrically modeled satellite laser ranging measurements, and to eliminate relatively small spurious range effects due to motion of the Earth and Moon into and out of the Sun's gravitational potential; otherwise these light and clock equations play only a supportive role to the body equations of motion in LLR. These equations for clock rates and light speeds also establish the relationships between the convenient and global solar system barycentric coordinates for space and time locations, and the so–called *proper* local coordinates measured by rulers and clocks.

3 New Long Range Force?

An additional long–range interaction in physical law would generate a force between bodies i and j which is likely to have the static limit structure

$$\boldsymbol{f}_i \;=\; K_i\, \boldsymbol{\nabla}_i\, \frac{K_j}{r_{ij}}\, e^{-\mu\, r_{ij}} . \tag{13}$$

The coupling strengths K_i and K_j, except in special cases, will be body attributes different than total mass–energy (*non–metric* coupling), and the dependence on distance of this force can either be inverse square or Yukawa–like if the underlying field transmitting this force between bodies is not massless. Such a new force will typically produce a difference in the Sun's acceleration of Earth and Moon, because these two bodies are of different compositions — The Earth has a substantial iron core while the Moon is composed of silicate mantle–like materials. This Equivalence Principle violating acceleration difference amounts to

$$|\delta \boldsymbol{a}_{\rm em}/\boldsymbol{g}_s| = \frac{K_s}{GM_s} \left(\frac{K_{\rm m}}{M_{\rm m}} - \frac{K_{\rm e}}{M_{\rm e}}\right) (1+\mu R)\, e^{-\mu R} \tag{14}$$

and it will supplement, even in the $\mu \to 0$ limit, any difference of accelerations resulting from the previously discussed gravitational to inertial mass ratio anomalies of the bodies. LLR has become a sufficiently precise tool for measuring $|\delta \boldsymbol{a}_{\rm em}|$, it now competes favorably with ground–based measurements of the universality of free fall.

4 LLR's Science–Related Range Signals

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

Violation of Universality of Free–Fall

If Earth and Moon fall toward the Sun at different rates due to either of the mechanisms given by Eqs.(8) or (14), then the lunar orbit is polarized along the solar direction. Detailed calculation of this polarization reveals an interesting interactive feedback mechanism acting between this $\cos D$ polarization signal and the $\cos(2D)$ Newtonian solar tide perturbation of the lunar orbit called the *variation*). The result is a range signal enhanced from the simpled estimate given in Eq.(3)

$$\delta r(t)_{\rm me} = \delta_{\rm me}\, \frac{3}{2}\frac{\Omega}{\omega} R\, F(\Omega/\omega)\ \cos D\,, \tag{15}$$

$$\cong 2.9 \times 10^{12}\ \delta_{\rm me}\ \cos D \qquad {\rm cm}\,, \tag{16}$$

with $\delta_{\rm em} = |(\boldsymbol{a}_{\rm e} - \boldsymbol{a}_{\rm m})/\boldsymbol{g}_s|$, R is distance to the Sun, Ω and ω are the sidereal frequencies of solar and lunar motion, and D is the lunar phase measured from new moon. The feedback amplification factor for the lunar orbit is already $F(\Omega/\omega) \cong 1.75$; it grows further with larger orbits and approaches an interesting singularity for an orbit less than twice as large as that of the Moon [8,9]. The most recent fits of the LLR data find no anomalies in the $\cos D$ amplitude to precision of a few millimeters, so from Eq.(16) this constrains $\delta_{\rm me}$ to be less than about 2×10^{-13}! This provides a constraint on a combination of the two *Eddington* parameters

$$|4\beta - 3 - \gamma|\ \le\ 5 \times 10^{-4}\,. \tag{17}$$

Computer integration of the complete Eq.(6) for the Sun–Earth–Moon system dynamics confirms these analytically estimated polarization sensitivities.

From LLR's measured constraint $\delta_{\rm em}\ \le\ 2 \times 10^{-13}$, the curves on the right in Fig.2, labeled *Equivalence Principle* are derived and show the constraints on the strength (relative to Newtonian gravity) as function of force range of any additional *metric* or *non–metric* Yukawa interaction which might supplement the tensor field interaction of general relativity. These curves continue horizontally to infinite range (inverse square force law) in which domain this LLR measurement is presently one of the premier probes of the gravitational interaction.

The Newtonian solar tidal perturbation of the Moon's orbit at second order in the orbit's radius also produces a range oscillation at the synodic frequency; and it is 110 km in amplitude! But this perturbation, discovered by the generation after Newton and called the *parallactic inequality*, has amplitude proportional to $(\Omega/\omega)\ r^2/R$. The least precisely known factor here is the Moon's mean orbital size, but it need only be known to half a part in 10^8 (200 cm) in order that the *parallactic inequality* be determined to the sub–millimeter level and thereby enabling further perturbations of a few millimeters to have significance.

Geodetic Precession of the Local Inertial Frame

If one collects the *difference* between the Sun's acceleration of the Moon and of the Earth which emerge from lines D and F of Eq.(6) because of the different velocities of these bodies, then a particularly interesting part of this acceleration difference is proportional to Vu times the Sun's acceleration, with, as shown in Figure 1, $\boldsymbol{V}$ being the velocity of the Earth relative to the Sun, and $\boldsymbol{u}$ the velocity of the Moon relative to Earth. These terms form deSitter's Coriollis–like acceleration

$$\delta \boldsymbol{a}_{\rm m} = 2\ \boldsymbol{\Omega}_{\rm dS} \times \boldsymbol{u}\,, \tag{18}$$

with

$$\boldsymbol{\Omega}_{\rm dS} = \frac{2\gamma+1}{2}\ \frac{GM_s}{c^2R^3}\ \boldsymbol{R}\times\boldsymbol{V}\,. \tag{19}$$

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

$$\delta r_{\rm me} = re\ \cos A + r_{\rm var}\ \cos(2D) + r_{\rm evc}\ \cos(2D-A) + \ldots\,, \tag{20}$$

with r and e being orbital radius and eccentricity, $r_{\rm var}$ being the amplitude of solar tidal perturbation called *variation*, and $r_{\rm evc}$ being the amplitude of the hybrid *evection* perturbation due to both the solar tidal force and the eccentric motion of the Moon. So the least–squares–fit of the LLR data, which yields best estimates for the two key lunar frequencies, will involve the parameter 'partials'

$$\frac{\partial \delta r_{\rm me}}{\partial \dot{A}} = -t\ \left(re\,\sin A - r_{\rm evc}\,\sin(2D-A)\right)\,, \tag{21}$$

$$\frac{\partial \delta r_{\rm me}}{\partial \dot{D}} = -2t\ \left(r_{\rm var}\,\sin(2D) + r_{\rm evc}\,\sin(2D-A)\right)\,. \tag{22}$$

Additional model details bring measurement of the Moon's out–of–plane motion frequency into the picture as well, though this does not change the key point here. Because these rate signals grow in time, measurement precision of the deSitter precession grows with total time of the LLR experiment, not only because of the growing quantity and quality of the accumulated range measurements, but also because of the linear growth in signal sensitivity. A most recent fit of the LLR data confirms presence of the geodetic precession with precision of 3.5×10^{-3}. The continued rapid improvement of this measurement into the future is expected.

The mean precession rate of the Moon's orbit is measured to better than 0.1 mas/year, which is less than a part in 10^{11} of the orbital rate. Any new Yukawa–like force of the type given by Eq.(13) will produce an additional precession $\delta\omega$ of lunar perigee in amount

$$\frac{\delta\omega}{\omega} \;\cong\; \frac{1}{2}\frac{K_{\rm e}K_{\rm m}}{GM_{\rm e}M_{\rm m}}(\mu R)^2\; e^{-\mu R}\,. \tag{23}$$

For such a force with the *matched* range $1/\mu \;\cong\; 2\times10^{10}$ cm, LLR observations limit the strength of that force to be less than 10^{-11} that of the Newtonian gravity force! This constraint is shown as the left dotted curve in Fig.2 and applies whether the new force is either *metric* or *non–metric*. The curve's secondary hump is due to the solar tidal force's participation in shaping the lunar orbit. The perigee measurement is an excellent supplement (at finite range) of LLR's other key experimental constraint on any new forces, shown by the right curves in this figure, which results from measuring the equality of the Sun's acceleration of Earth and Moon.

Time Evolution of Gravity's Coupling Strength *G*

Time evolution of Newton's coupling parameter G results in proportional evolutions for both the radial size and frequencies of the lunar motion. Similar, but not identical in detail, effects result when a torque (indicated by $\dot{L}$) acts on the orbit

$$\frac{\dot r}{r} = -\frac{\dot G}{G} + 2\frac{\dot L}{L}\,, \tag{24}$$

$$\frac{\dot\omega_n}{\omega_n} = 2\,\frac{\dot G}{G} - 3\frac{\dot L}{L}\,. \tag{25}$$

During the initial years of the LLR experiment it has been the mean orbital radius signal

$$\delta r(t)_{\rm me} = \left(2\,\frac{\dot L}{L} - \frac{\dot G}{G}\right)\; r\;(t-t_0)\,, \tag{26}$$

which has been used to measure $\dot G$; this involved estimating and subtracting the part which results from the orbital torque exerted on the Moon by the ocean tidal bulges on Earth which, because of friction, lag in angle from the direction toward the Moon. The inclination and 18.6 year precession of the lunar orbit's plane

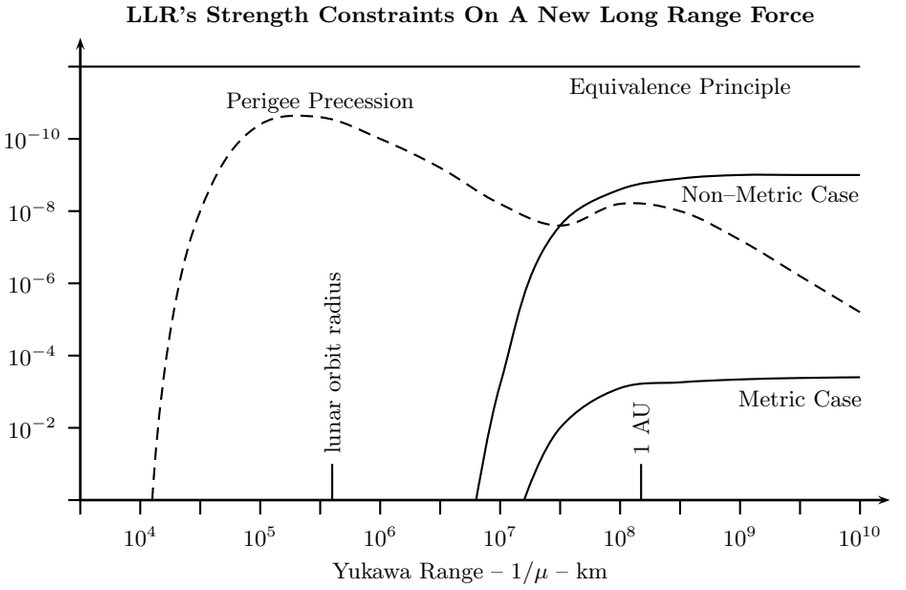


Fig. 2. The constraints provided by LLR on the strength of any additional long range force, as a fraction of that of gravity, are shown as a function of the additional force's Yukawa range. The dashed curve results from measurement of the Moon's perigee precession rate; the two peaks reflecting contributions from both the Earth's gravitational field and the Sun's tidal field. The curves on the right (which horizontally continue to 'infinite range' or inverse–square) result from the measured equality of the Sun's acceleration of Earth and Moon. These latter *Equivalence Principle* results are modeled two ways — by a supplementary *metric* force or by a *non–metric* force.

result in a modulation of the tidal contribution to $\dot{r}$ which helps the separation of the two perturbations after accumulation of sufficient years of data. But the data set produced by LLR has in recent years become sufficiently extended in time so that the range signals associated with frequency shifts, which grow quadratically in time, are becoming dominant in the fit for $\dot{G}$.

$$\delta r(t)_{\rm me} \sim \left(\frac{3}{2}\frac{\dot{L}}{L} - \frac{\dot{G}}{G}\right) r_n\, \omega_n (t-t_0)^2\, sin(\omega_n t - \theta_n)\,. \tag{27}$$

This new era also shows enhanced ability to separate the tidal torque $\dot{L}$ perturbation from any $\dot{G}$ perturbation, since the ratio of orbital size and frequency shifts produced by the two differ somewhat, as seen from Eqs.(24) and (25). And those parts of the lunar frequencies which result from the annual motion around the Sun, though fully responsive to an evolving G, are not disturbed by tidal torque; this also contributes to the separate measurability of $\dot{G}$. Recent fits of almost 30 years of LLR data yield excellent measurement constraint on this

parameter

$$\frac{\dot{G}}{G} \cong (1 \pm 2)\, 10^{-12}\, \mathrm{y}^{-1} \qquad (2\,\sigma) \tag{28}$$

This is about 1/35 the observed Hubble expansion rate of the universe. The size of the perturbation signals associated with $\dot{G}$, which are now growing quadratic in time, should permit LLR to continue into the future being at the cutting edge of precision in supplying measurements of, or constraints on, $\dot{G}$.

5 The Gravitomagnetic Interaction

Line C of the complete N–body gravitational equation of motion given by Eq.(6) represents a post–Newtonian gravitational force proportional to the velocities of both the bodies in the interaction. In analogy with electromagnetic theory, part of this interaction has been called *gravitomagnetism* and represents a force between two 'mass currents'. From line C of Eq.(6), this acceleration is

$$\begin{aligned} \delta \boldsymbol{a}_i \;&=\; (2+2\gamma) \sum_{j \neq i} \frac{G m_j}{c^2 r_{ij}^3} \boldsymbol{r}_{ij}\; \boldsymbol{v}_i \cdot \boldsymbol{v}_j \\ &-\; \sum_{j \neq i} \frac{G m_j}{c^2 r_{ij}^3} \left[(2+2\gamma) \boldsymbol{r}_{ij} \cdot \boldsymbol{v}_i\; \boldsymbol{v}_j \;+\; (1+2\gamma) \boldsymbol{r}_{ij} \cdot \boldsymbol{v}_j\; \boldsymbol{v}_i \right] . \end{aligned} \tag{29}$$

It has been alleged that the presence of gravitomagnetism within the total gravitational interaction has not been experimentally confirmed and measured. Different experiments have been under development to see explicitly the effects of this historically interesting prediction of general relativity. But this gravitomagnetic acceleration already plays a large role in producing the final shape of the lunar orbit, albeit in conjunction with the rest of the total equation of motion, so its presence and strength in the equation of motion can hardly be in doubt. Because both the Earth and Moon are moving in the solar system barycentric frame — the frame in which the dynamical equations are formulated and then integrated into orbits — a gravitomagnetic interaction exists between these two bodies which have velocities $\boldsymbol{V}(t)$ and $\boldsymbol{V}(t) + \boldsymbol{u}(t)$, respectively, as seen in Fig.1. Perturbations to the Earth–Moon distance from the gravitomagnetic acceleration result proportional to both V^2 and Vu, and they are surprisingly large

$$\begin{aligned} \delta r(t)_{\mathrm{me}} \;&\cong\; -\frac{7}{3}\, \frac{V^2}{c^2}\; r\; \cos(2D) \;-\; 5\, \frac{G M_{\mathrm{s}}}{c^2 r} F(\Omega/\omega)\; R \cos D \\ &\cong -930\, \cos(2D) + 1300\, \cos D \quad \mathrm{cm}\,. \end{aligned} \tag{30}$$

As previously discussed, the amplitudes of the lunar motion at both these periods (monthly and semi–monthly) are both determined to better than a centimeter precision in the total orbital fit to the LLR data. It would be impossible to understand this fit of the LLR data without the participation of the gravitomagnetic interaction in the underlying model, and with strength very close to

that provided by general relativity ($\gamma_{\rm GR} = 1$). It has been argued that the lunar orbit, like near–Earth satellite orbits, could be calculated in the geocentric frame in which the velocity $\boldsymbol{V}$ essentially vanishes and gravitomagnetic forces are negligible; therefore LLR 'measurement' of gravitomagnetism is somewhat illusory? But this point of view fails to take local Lorentz invariance (LLI) into consideration. The very ability to reformulate the equations of motion in the geocentric frame without introducing new frame–dependent terms, depends on the LLI of the equations. But it is the entire package of velocity–dependent, post–Newtonian terms which includes the gravitomagnetic terms, lines C plus D of Eq.(6), that produce the LLI; the *Eddington* parameter γ represents the only freedom remaining in the structure of this LLI package. Our confidence in the exhibited structure of this total LLI collection of velocity–dependent terms is proportional to the precision with which the various *preferred frame* effects in the solar system proportional to W^2, WV, and Wu have *not* been seen. LLR, itself, has contributed to establishment of LLI through its null measurement of the third type of effects proportional to Wu [12].

6 Inductive (Inertial) Forces

Inductive forces, shown on line E of Eq.(6) and closely related to *gravitomagnetism*, also play an important role in determining the final lunar orbit. In such forces the acceleration of one mass element induces an acceleration of another proximite mass element (e.g., i and j in Fig.1). From line E of Eq.(6) we have

$$\delta \boldsymbol{a}_i \;=\; \sum_{j\neq i} \frac{Gm_j}{2c^2 r_{ij}} \left[(2\gamma+3)\boldsymbol{a}_j \;+\; \boldsymbol{a}_j \cdot \hat{r}_{ij}\hat{r}_{ij}\right] . \tag{31}$$

These accelerations play a key part in adjusting the inertial masses of the Earth and Moon because of their internal gravitational binding energies; the absence or an anomalous strength of these inductive forces would translate directly into differences between the acceleration rates of these bodies toward the Sun, and a resultant polarization of the Moon's orbit in the solar direction. The forces, Eq.(31), acting between the mass elements of Earth, for example, by themselves lead to an anomalous polarization of the lunar orbit of very large magnitude

$$\delta r(t); \cong \; (4\gamma + 10/3)\; 13\; \cos D \,\text{meter}\,. \tag{32}$$

If the inductive forces are combined with the other post–Newtonian *inertial* forces shown on line F of Eq.(6), only then does the total inertial self force of a body become

$$\begin{aligned} \delta \boldsymbol{f} \;=\; -\; \frac{1}{c^2}\left(\frac{1}{2}\sum_i m_i v_i^2 \;-\; \frac{G}{2}\sum_{i,j}\frac{m_i m_j}{r_i j}\right)\boldsymbol{a} \\ -\; \frac{1}{c^2}\left[\sum_i m_i \boldsymbol{v}_i\, \boldsymbol{v}_i \;-\; \frac{G}{2}\sum_{i,j}\frac{m_i m_j}{r_{ij}^3}\boldsymbol{r}_{ij}\,\boldsymbol{r}_{ij}\right]\cdot \boldsymbol{a}\,. \end{aligned} \tag{33}$$

The first line of this self force is simply the inertial force due to the kinetic energy and gravitational binding energy within the body. The second line represents contributions to the body's internal *virial* which, when totaled over all internal fields, vanishes for a body in internal equilibrium and experiencing negligible external tidal forces. These self forces of a body are actually an integral part of the calculation of the total gravitational to inertial mass ratio of bodies, discussed previously. They were explicitly discussed here in order to show the large size of the inductive force contributions.

The inductive forces also act externally between Earth and Moon, producing a further polarizing perturbation of the lunar orbit. This perturbation, though now only of order the Earth's gravitational potential at lunar distance rather than Earth internal potential, is still orders of magnitude larger than the measurement precision of this polarization achieved in LLR

$$\delta r(t)_{\rm em} \;\cong\; \frac{11}{2}\frac{GM_{\rm e}}{c^2 r}\frac{\Omega}{\omega}\, R\,\cos D \;\cong\; 100\,\cos D \quad {\rm cm}\,. \tag{34}$$

It is actually almost totally canceled by the non–linear force on the Moon proportional to the product of the masses of Sun and Earth. Like the gravitomagnetic interactions, the inductive interactions have been confirmed in LLR to have their general relativistic strength to better than a part in a thousand precision.

Acknowledgement

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

References

1. K. Nordtvedt: *Phys. Rev.* **169**, 1017 (1968).
2. K. Nordtvedt: *Phys. Rev.* **170**, 1186 (1968).
3. W. deSitter: *Mon. Not. R. Astron. Soc.* **77**, 155 (1916).
4. K. Nordtvedt: *Phys. Rev.* **D 7**, 2347 (1973).
5. I.I. Shapiro *et al.*: *Phys. Rev. Lett.* **61**, 2643 (1988).
6. B. Bertotti, I. Ciufolini, and P.L. Bender: *Phys. Rev. Lett.* **58**, 1062 (1987).
7. J.G. Williams, X.X. Newhall, and J.O. Dickey: *Phys. Rev.* **D 53**, 6730 (1996).
8. K. Nordtvedt: *Icarus* **114**. 51 (1995).
9. T. Damour and D. Vokrouhlicky: *Phys. Rev.* **D 53**, 4177 (1996).
10. C. Will and K. Nordtvedt: *Astrophys. J.* **177**, 757 (1972).
11. K. Nordtvedt: *Class. Quant. Grav.* **13**, 1317 (1996).
12. J. Müller, K. Nordtvedt, and D. Vokrouhlicky: *Phys. Rev.* **D 54**, 5927 (1996).
13. J. Müller and K. Nordtvedt: *Phys. Rev.* **D 58**, 2001 (1998).

Testing Relativistic Gravity and Measuring Solar System Parameters via Optical Space Missions

Wei–Tou Ni

Center for Gravitation and Cosmology, Department of Physics, National Tsing Hua University,
Hsinchu, Taiwan, 30055 Republic of China

Abstract. For last thirty years, great advances in the testing of relativistic gravity have come from interplanetary radio ranging and lunar laser ranging. With optical mission concepts in the interplanetary space, testing relativistic gravity can be improved by 3 – 6 orders of magnitude and many solar–system parameters can be measured either for the first time or more precisely. After reviewing briefly dedicated optical mission concepts — SORT, IPLR and ASTROD together with other optical mission or mission concepts which have important implications on testing relativity and astrodynamics — HIPPARCOS, GAIA and LISA, we concentrate on a specific mission concept — ASTROD to discuss various mission goals and capabilities in detail. ASTROD is an optical interferometry mission concept. Optical interferometry missions hold great promises for the testing of relativistic gravity and for the measuring of solar–system parameters. We discuss the determination of relativistic parameters γ , β and the solar quadrupole moment parameter J_2, the measurements of solar Lense–Thirring effect together with the application of laser astrodynamics to solar system studies — solar angular momentum, solar g–modes, asteroid masses, etc.

1 Introduction

We have seen great advances in the testing of relativistic gravity for last thirty years. This is largely due to interplanetary radio ranging and lunar laser ranging. Interplanetary radio ranging and tracking provided more stimuli and progresses at first. However with improved accuracy of 2 – 3 cm from 20 – 30 cm and long–accumulation of observation data, lunar laser ranging reaches similar accuracy in determining relativistic parameters as compared to interplanetary radio ranging. In broad lines, both solar–system radio ranging and lunar laser ranging have tested relativistic–gravity effects to 10^{-3}. If laser ranging can be extended to the whole solar system, the precision of testing relativistic gravity will be dramatically improved. If this is by laser pulse ranging, the precision will be improved by 3 orders of magnitude as in SORT (Solar Orbit Relativity Test) [1] and IPLR (InterPlanetary Laser Ranging) [2]. If this is by laser interferometric ranging, the precision could be enhanced by another 3 orders of magnitude [3]. That is, the improvement from the present basis would be 3 – 6 orders of magnitude.

The SORT mission concept is to use laser pulses and a drag–free spacecraft with a precision clock orbiting around the Sun to measure γ and J_2 (solar quadrupole moment parameter) precisely [1]. IPLR mission concept also uses laser pulses [2].

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 330–343, 2001.

The ASTROD (Astrodynamical Space Test of Relativity using Optical Devices) mission concept is to use drag–free spacecraft in solar orbits employing laser interferometric ranging techniques together with a constellation of Earth orbiting satellites to provide high–precision measurement of relativistic effects, better determination of the orbits of major asteroids, improvement in the measurement of γ, measurement of solar angular momentum via Lense–Thirring effect and the detection of low–frequency gravitational waves and solar oscillations in a single mission [3–10].

SORT, IPLR and ASTROD mission concepts are dedicated relativity and astrodynamical mission concepts using optical devices. There are mission concepts using optical devices which are mainly dedicated to different purposes, but also have important implication on testing relativity and astrodynamics. The HIPPARCOS (HIgh Precision PARallax COllecting Satellite) mission gives the best determination of the PPN parameter γ using light; the precision is only a factor three short of radio observations [11]. The GAIA (Global Astrometric Interferometer for Astrophysics) mission concept [12] proposes to measure γ to 1 ppm; this accuracy is similar to a dedicated relativity mission. LISA [13] is an optical interferometry mission dedicated to the gravitational–wave detection. Testing the equivalence principle [14] and microscopic physics in gravity [15] using optical methods in space is also promising. Astrodynamical space missions using optical interferometry hold great promises for testing relativistic gravity and measuring solar–system parameters. In the following, we concentrate on the analysis and discussion of such a mission concept — ASTROD to illustrate various goals and capabilities in detail.

The objectives of the ASTROD Mission are threefold. The first objective is to discover and explore fundamental physical laws governing matter, space and time via testing relativistic gravity with 3 - 6 orders of magnitude improvement. Relativistic gravity is an important cornerstone of physics, astronomy and cosmology. Its improved test is crucial to cosmology and modern theories of gravitation including superstring theories. Included in this objective is the precise determination of the relativistic parameters β and γ, the improved measurement of $\dot{G}$ and a precise determination of an anomalous, constant acceleration directed towards the Sun.

The second objective of the ASTROD mission is the high–precision measurement of the solar–system parameters. This includes: (i) a measurement of solar angular momentum via Lense–Thirring effect and the detection of solar g–mode oscillations via their changing gravity field, thus, providing a new eye to see inside the Sun; (ii) precise determination of the planetary orbit elements and masses; (iii) better determination of the orbits and masses of major asteroids. These measurements give better solar dynamics and probe the origin of our solar system.

The third objective is to detect and observe gravitational waves from massive black holes and galactic binary stars in the frequency range 50 μHz to 5 mHz. Background gravitational–waves will also be explored.

A desirable implementation is to have two spacecraft in separate solar orbit carrying a payload of a proof mass, two telescopes, two 1 - 2 W lasers, a clock and a drag–free system, together with an Earth reference system. The Earth reference system could be ground stations, Earth satellites and/or spacecraft near Earth–Sun Lagrange points. A proposal is submitted to ESA in response to call for mission proposals for two–flexi–missions F2/F3 with international collaboration [4].

In 1993, we have proposed to use laser astrodynamics to study relativistic gravity and to explore the solar system [5]. With a multi–purpose astrodynamical mission proposed in 1994 [6,7], we reached the ASTROD mission concept. Our ASTROD mission concept is presented in the 1996 COSPAR Assembly [7]. ASTROD in its relation to gravitational–wave detection is presented in the TAMA Gravitational–Wave Workshop [8]. ASTROD's sensitivity in measuring $\dot{G}$ is presented in the Pacific Conference on Gravitation and Cosmology [9]. Preliminary calculation of asteroid perturbations and investigation in the possible determination of asteroid masses through range observations of the ASTROD mission is published in *Planetary and Space Science* [10].

In Section 2, we discuss the ASTROD payload concept and its technological development requirements. In Section 3, we review the orbit simulation with uncertainty distributions for the ASTROD to investigate the accuracy for the determination of relativistic parameters γ, β and solar quadrupole parameter J_2, together with other standard solar–system parameters. In Section 4, we discuss the method to measure the solar angular momentum and low–l solar oscillations. In Section 5, we discuss the strategy and sensitivity of detecting gravitational waves. In Section 6 we discuss alternate mission concepts. In Section 7, we present an outlook.

2 ASTROD Payload Concept and Technological Development Requirements

For the realization of the ASTROD mission concept, it is desirable to have a fleet of drag–free spacecraft in solar orbits together with an Earth reference system. The Earth reference system could be ground stations, Earth satellites and/or spacecraft near Earth–Sun Lagrange points. Each spacecraft in solar orbit communicates and ranges with the Earth reference system. When two of them are close to each other, they can communicate and range with each other. For the fleet of spacecraft in solar orbits, a two spacecraft implementation is to have each spacecraft in separate solar orbit carrying a payload of a proof mass, two telescopes, two 1 - 2 W lasers, a clock, and a drag–free system [6,7]. We call this the basic ASTROD.

With the precision requirement of this mission, drag–free is a must. Field Emission Electric Propulsion (FEEP) system is under development in Centrospazio (Italy) and Bradford Engineering (Holland). Accelerometer is under development in ONERA.

Many of the developments for LISA [13] can be applied to ASTROD. Laser and telescope are two important items. Due to the absolute requirement, ASTROD needs a demanding laser metrology system and the proof mass needs to be specially designed.

In JPL, space flyable Hg^+ frequency standards with 5 kg mass estimate are under development. The demonstrated frequency stability to below 10^{-15} of existing Hg^+ standards is hoped to be maintained or improved. An absolute stabilized laser can also serve as a clock. More thorough study of the payload concept needs to be done in the near future.

There are several technological developments required for this mission:

(i) Weak light phase locking to 100 fW incoming light,
(ii) Heterodyne interferometry schemes and data–analysis technique,
(iii) Light–weight precision space clock and/or absolute–stabilized laser to 10^{-15},
(iv) Accelerometer with noise level 10^{-13} - 3×10^{-15} m/s$^2\sqrt{\text{Hz}}$ from 50 μHz - 5 mHz and absolute stability of 10^{-13} - 10^{-15} m/s^2,
(v) A laser metrology system to monitor the positions of various parts of the spacecraft to facilitate gravitational modelling.

With 100 fW, there still are 5×10^5 photons/s. This is enough for 100 kHz frequency tuning. At present, University of Glasgow has achieved 85 pW weak–light phase–locking, and National Tsing Hua University has achieved 2 pW weak–light phase–locking with 0.2 mW local oscilator. Further improvement is needed in this aspect. Heterodyne interferometry schemes has been proposed by JPL people [16] for unequal arms. These schemes can be adapted to the ASTROD mission. Data–analysis for detecting gravitational waves and solar oscillations are demanding and needs to worked out. Laser stabilization is a subject under active study in University of Konstanz and University of Tokyo. Accelerometer requirement is very severe and needs to be developed in steps. This is in the expertise of ONERA. The absolute requirement may need to be implemented using absolute laser metrology.

3 Orbit Simulation

From the mission orbit design process, we have reached the following orbits for the two spacecraft. Two spacecraft are to be launched with one going into an inner orbit and the other one going into an outer orbit. After two and half years, both will be on the other side of the Sun opposite the Earth and stay near the Sun in apparent position with certain time on one side of the Sun and certain time on both side of the Sun for a total period of 120 days. Second–order relativistic light retardation and solar angular momentum can be measured during this period of time. During working time of the two spacecraft, gravitational waves can be monitored.

In order to specifically calculate the orbit, we assume an initial spacecraft location in geostationary orbit, one located at the midday longitude and one at midnight, and we assume they are launched from these locations to their

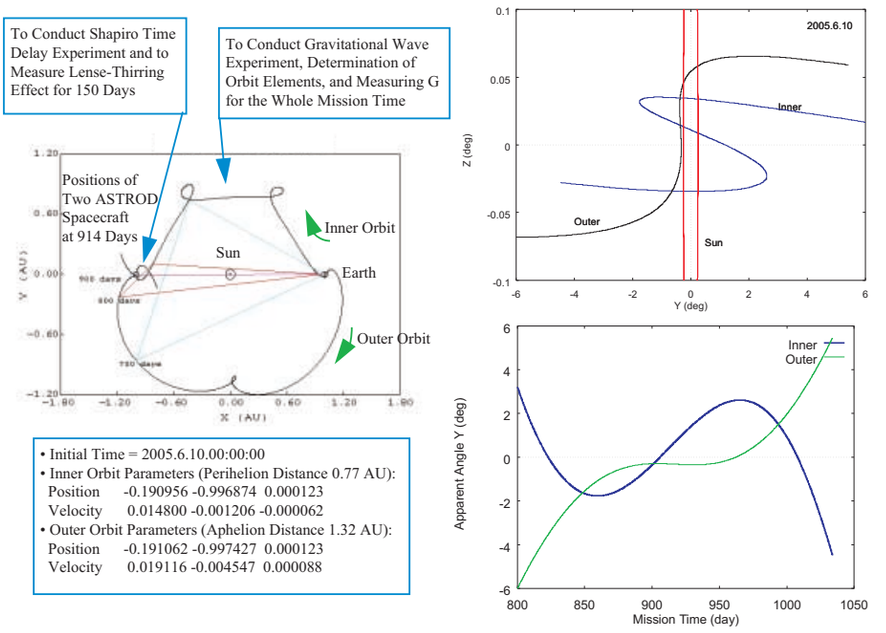


Fig. 1. Left: The inner and outer orbits of two ASTROD spacecraft in the Sun–Earth fixed frame; and the triangles formed by the positions of the two spacecraft and Earth at 700, 800, and 900 days of mission time. Top right: The apparent angles of the two ASTROD spacecraft located respectively on the inner and outer orbits near two and half years after launch. Bottom right: apparant angle Y vs. mission time.

interplanetary trajectories on June 10, 2005. Each spacecraft orbit is propagated using a sixth–order Runge–Kutta method with the force term calculated using the JPL DE403 Ephemeris data. The computed results are shown in Fig. 1. On the left of Fig. 1 the orbits are drawn in the Sun–Earth fixed frame. On the right, the apparent angles of the two spacecraft from 800 to 1034 days after launch as viewed from the Earth are shown. During this period, two–way ranging between each spacecraft and Earth reference system, and between two spacecraft will be implemented. More details of the orbit design process including fine tuning to have a second close–up around 7.5 years can be found in reference [17].

In reference [18], we have implemented an orbit simulation for ASTROD. In this section, we review this simulation. First we establish a post–Newtonian ephemeris and extend it to a stochastic one. With an appropriate modelling of local acceleration noises and instrumental noises, we then determine the accuracy of parameter determination for β (nonlinearity parameter), γ and J_2 (solar quadrupole moment parameter).

3.1 Post–Newtonian Ephemerides and ASTROD Orbits

The barycentric metric with solar quadrupole moment is given as follows

$$
\begin{aligned}
ds^2 = & \left[1 - 2\sum_i \frac{m_i}{r_i} + 2\beta\left(\sum_i \frac{m_i}{r_i}\right)^2 + (4\beta - 2)\sum_i \frac{m_i}{r_i}\sum_{j\neq i}\frac{m_i}{r_{ij}}\right. \\
& -\frac{2\gamma+1}{c^2}\sum_i \frac{m_i}{r_i}\dot{\mathbf{x}}_i^2 - \frac{1}{c^2}\sum_i \frac{m_i}{r_i}(\dot{\mathbf{x}}_i^2 - \mathbf{r}_i\ddot{\mathbf{x}}_i - \frac{1}{r_i^2}(\mathbf{r}_i\dot{\mathbf{x}}_i)^2) \\
& \left.+\frac{m_1 R_1^2}{r_1^3}J_2\left(3\left(\frac{\mathbf{r}_1\cdot\hat{\mathbf{z}}}{r_1}\right)^2 - 1\right)\right]c^2dt^2 + \frac{2}{c}\sum_i \frac{m_i}{r_i}((2\gamma+2)\dot{\mathbf{x}}_i)d\mathbf{x}cdt \\
& -\left[1 + 2\gamma\sum_i \frac{m_i}{r_i}\right](d\mathbf{x})^2\,, \qquad (1)
\end{aligned}
$$

with $\mathbf{r}_i = \mathbf{x} - \mathbf{x}_i, \mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j, m_i = GM_i/c^2, M_i$'s the masses of the bodies with M_1 the solar mass [19]. J_2 is the quadrupole moment parameter of the Sun. $\hat{\mathbf{z}}$ is the unit vector normal to the elliptic plane. The solar angular momentum has a 7° tilt with respect to $\hat{z}$. This is marginally detectable in our simulation. To be simple, we assume that they are in the same direction.

The equations of motion of N–mass problem associated with this metric can be derived from geodetic variational principle and used to build our computer–integrated ephemeris (with $\gamma = \beta = 1$, and $J_2 = 2 \times 10^{-7}$) for nine–planets, the moon and the sun. The positions and velocities at the initial time, 2005.6.10 0:00, are given in advance by the DE403 ephemeris. The evolution is solved by using the 4th–order Runge–Kutta method with the stepsize, h = 0.01 day. The stepsize, h = 0.01 day, is chosen such that the computer efficiency and the numerical accuracy can both approach the requirement. With this tiny stepsize, the numerical discrepancy is within about 10^{-13} AU for all positions and 10^{-13} AU/day for velocities. This ephemeris is then used to calculate the orbits of inner and outer spacecraft, and the light travel times and Shapiro time delays.

3.2 Simulation of ASTROD Ranging Data

Now, we use a stochastic model to generate simulated ASTROD ranging data. The ranging observation Z_k at time instant t_k is expressed as an explicit function $h(\mathbf{u}, t_k)$ of uncertain parameters plus a superposing noise term V_k as follows,

$$Z_k = h(\mathbf{u}, t_k) + V_k\,, \qquad (2)$$

where $\mathbf{u}$=(u_1, u_2, ...) are uncertain parameters to be estimated, e.g., $u_1 = \gamma, u_2 = \beta, u_3 = J_2$ etc. $\mathbf{u}$ normally includes masses and initial states of planets. However, for simulating the uncertainties of relativistic parameter determination, we only need to include a few such parameters.

In the noise term V_k, we include two types of errors:

(1) the uncertainty due to the imprecision of the ranging devices,

(2) unknown accelerations due to the imperfections of the spacecraft *drag–free* system.

We treat the first type of error as white Gaussian random noise superimposed into the ranging measurement. The present achievable timing error is 50 ps [20]. If there are interruptions of interferometric signals, this timing error would constitute a limit for error reductions. Thermal expansion errors and all other errors due to optical devices are required to be below this level. Hence, conservatively, we take 50 ps to be the standard deviation of this type of errors. We assume the mean of this error to be zero. If the interferometric signals are continuously monitored, 1 fs (subwavelength) resolution is there; when corresponding magnitudes of other errors can be minimized, great improvement can be envisaged.

As to the second type of errors, there are unknown accelerations fixed or confined in directional angles with respect to the spacecraft orientation or spacecraft–Sun orientation and unknown accelerations truly random in directions. In this paper, we only consider the unknown accelerations truly random in directions. For modelling, we assume the distribution of these accelerations is Gaussian in magnitude with $\sigma = 10^{-15}$ m/s^2 and change directions every 4 hours (equivalent to 10^{-13} m/s$^2/\sqrt{\text{Hz}}$ for $f \sim 10^{-4}$ Hz). We will consider the others in a separate paper. To generate a sequence of simulated observations, the simulation program numerically integrates the equations of motion adopted for the stochastic model. The equations of motion are unchanged for all solar objects except the spacecraft. For the spacecraft, the equations of motion are modified by adding the noisy accelerations. After the integration and the total light travel time is solved, the first type of noise (50 ps–half–width Gaussian) is added to obtain the ranging simulation. The deviations of two simulated rangings from the fiducial ranging for the inner spacecraft are shown in Fig. 2.

3.3 Estimation of Parameters

We use sequential Kalman filter data processing [21] to obtain the estimated uncertainty of the relativistic parameters to be determined from the mission as

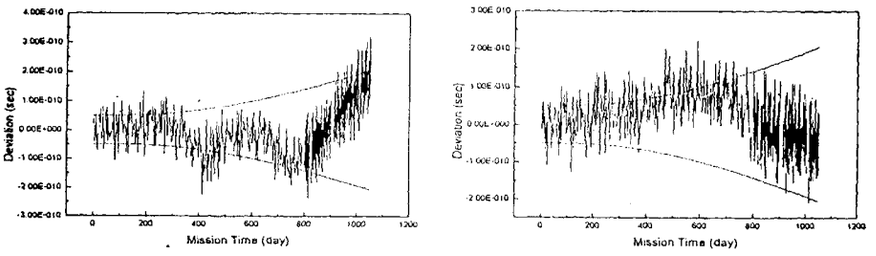


Fig. 2. Deviations of two simulated rangings from the fiducial ranging for the inner spacecraft. The smooth curves are the ± standard deviation of simulations vs. epoch. For 800–1050 days of the mission, the time density is 10 times higher.

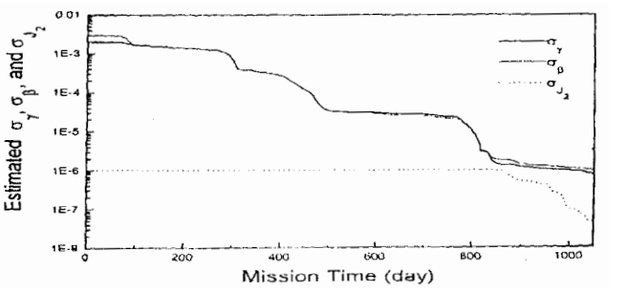


Fig. 3. σ_γ, σ_β, σ_{J-2} as functions of mission epoch.

a function of epoch. The relativistic parameters β and γ, the solar quadrupole moment parameter J_2 together with initial state of the spacecraft and mass parameters of Sun, Mercury, Venus, Earth system, Mars and Jupiter are estimated. The uncertainties of β, γ and J_2 versus mission epoch estimated for inner spacecraft are shown in Fig. 3. With the 50 ps ranging uncertainty, γ can be determined to 7.7×10^{-7}. When ranging uncertainty is decreased, better resolution of γ in the range $10^{-7}-10^{-9}$ is possible. In addition, β and J_2 can be determined to 10^{-6} and 4.5×10^{-8} respectively, with 50 ps ranging uncertainty and 10^{-15} m/s^2 (time scale: 4 hours) accelerometer random noise uncertainty.

Extension of simulation [22] to the determination of the masses of three big asteroids (Ceres, Vesta and Pallas) [10], the rate of change of the gravitational constant, and an anomalous constant acceleration directed towards the Sun [23] has been completed with good results.

4 Solar Angular Momentum and Solar Oscillations

To measure the solar Lense–Thirring effect, we need the difference, $t_1 - t_2$, in the two round trip propagations, ERS (Earth Reference System; Earth–system basis) $\rightarrow$ S/C 1 (Spacecraft 1) $\rightarrow$ S/C 2 (Spacecraft 2) $\rightarrow$ ERS and ERS $\rightarrow$ S/C 2 $\rightarrow$ S/C 1 $\rightarrow$ ERS. Newtonian calculation of $t_1 - t_2$ for 800 - 1034 days after launching gives about 10 ms. The Lense–Thirring effect variation for this period of time is about 100 ps and has a totally different signiture. For a laser stability of 10^{-15}-10^{-13}, the accuracy for measuring the round–trip time difference $t_1 - t_2$ (about 10 ms) is 10 as - 1 fs and the sensitivity of measurement of the Lense–Thirring effect is 10^{-7} – 10^{-5} by means of this interferometric measurement. From a preliminary theoretical modelling, we advocate that the Lense–Thirring effect can be determined to 0.1% or better.

Since the solar Lense–Thirring effect is proportional to the solar angular momentum, this measurement of the solar Lense–Thirring effect will give an accurate value of the solar angular momentum. At present, there is about 10% uncertainty [24] in the solar angular momentum in various solar models [25–28].

The strain sensitivity of ASTROD spacecraft will reach 10^{-20} - $10^{-21}/\sqrt{\text{Hz}}$ in the 50 μHz - 5 mHz frequency range. The low l p–modes solar oscillations will be

readily detectable through their changing gravity field if there is no background "noise" masks these signals. The g–mode sensitivity of ASTROD is about 2 orders of magnitude better than LISA and SOHO. ASTROD will have a good chance to detect and observe these g–mode oscillations. This will create a new eye to see inside the Sun. However, all will depend on whether the background noise can be resolved. Hils, Bender and Webbink [29] estimated that there are about 3×10^6 short–period white–dwarf binaries producing a large gravitational wave background in the relevant frequency band. The recent discovery of two white–dwarf binaries [30,31] (periods: 3.47 h and 4 h) give input for a better estimate of the population. We are analyzing in the following how this background noise can be lifted for the gravitational–wave detection of the solar g–mode, p–model and f–mode oscillations. From more precise orbit determinations, the solar quadrupole moment and higher moments can be measured. This together with solar angular momentum measurement and solar oscillation measurement constitute a gravity eye to see inside the Sun.

Efforts to see inside the Sun have stimulated important works in solar neutrino experiments and helioseismological observations. Recently, Schutz [32], Gough [33], and Cutler and Lindblom [34] suggested the possibility that solar oscillations might be observable by measuring their associated gravitational perturbations. With a detailed analysis, Cutler and Lindblom [34] concluded that LISA may be confusion limited at the relevant frequencies due to the galactic background from short–period white dwarf binaries and present estimates of the number of these binaries would require the solar modes to have energies above about 10^{33} erg to be observable by LISA.

ASTROD has better sensitivity in the frequency–band considered. In the first TAMA Workshop on Gravitational Wave Detection, we reported that the confusion limit might be lifted for the gravitational detection of the solar oscillations for ASTROD due to higher strain sensitivity and its orbit configuration [8]. In the third Amaldi Conference on Gravitational Waves, we present a scheme for the sparation of the gravitational–wave signals and the solar oscillation signals [35]. We review the scheme in the following.

The change of external gravitational potential, $\delta U^{(s)}_{n\ell m}$, due to the $({}^{s}_{n\ell m})$ mode oscillation of the sun can be written in the following simple form:

$$\delta U^{(s)}_{n\ell m} = -\frac{4\pi}{2\ell+1}\frac{\bar{\xi}^{(s)}_{n\ell m}}{R_\odot}\frac{GM_\odot}{R_\odot}\left(\frac{R_\odot}{r}\right)^{\ell+1} Y_{\ell m}e^{i\omega t}. \tag{3}$$

The surface helioseismology favors p–mode detection while gravitational helioseismology favors g–mode detection. Table 1 compiles some important parameters for selected solar oscillations.

For the ASTROD, the strain sensitivity for one year intergation for Signal–to–Noise–Ratio $S/N = 5$ threshold is 10^{-23} in the frequency range 100 μHz - 1 mHz. With this strain sensitivity, the $\ell = 2$ mode detection threshold is about 1 mm for $\bar{\xi}$. This is about 2 orders of magnitude more sensitive than LISA. The time constants for solar oscillations are long — over 10^6 yr for low l g–mode oscillations [36] and over 2 – 3 months for low l p–mode oscillations [37]. The

distance to the Sun of the inner ASTROD spacecraft varies from 0.77 AU to 1 AU and that of the outer ASTROD spacecraft varies from 1 AU to 1.32 AU. When the spacecraft move in solar orbits, the amplitude and direction of the solar oscillation signals receive deep modulations in addition to the modulations due to spacecraft motion and orientation. The time constant for the gravitational radiation (or orbit evolution) of the Close White Dwarf Binaries (CWDB) is more than 10^6 yr, and hence the CWDB confusion background is steady in the inertial space. This background is modulated only by the orientations and motions of spacecraft, not by the distances and orientations of the spacecraft relative to the Sun. With this extra modulation — deep in magnitude and direction, the detectability of the solar oscillation signals reaches at least 5 orders lower than the confusion limit, i.e., to the instrumental noise floor.

Table 1. Frequencies, energies, and surface velocity amplitudes of various l=2 modes of solar oscillations with $\bar{\xi}$=1mm [35]

Mode	Frequency (μHz)	Energy (erg)	Surface radial velocity (mm/s)
g_3	220.4	2.82×10^{27}	0.183
g_2	254.0	3.93×10^{27}	0.308
g_4	192.2	4.08×10^{27}	0.138
p_1	381.6	5.13×10^{27}	1.15
p_2	514.4	3.56×10^{28}	8.20
p_3	663.6	1.96×10^{29}	35.5

5 G–Wave Detection

For gravitational–wave detection, we compare the following two paths through the whole missions except for the period 850 - 1000 days: (i) ERS → S/C 1 → ERS → S/C 2 → ERS; ERS → S/C 2 → ERS → S/C 1 → ERS. The time difference of the two paths varies up to 0.34 s. A laser linewidth of 1 Hz would keep the interference coherent. The sensitivity 10^{-20}–$10^{-21}/\sqrt{\text{Hz}}$ would be comparable to that of LISA, but more sensitive to lower frequency due to long path. The round–trip interference for measurement of the Lense–Thirring effect also has some sensitivity on the gravitational wave [8]. Various methods of Doppler tracking and noise substraction schemes [16] can also be applied here for detection of gravitational waves.

Now we discuss the ASTROD sensitivity for gravitational waves following our presentation in the LISA Symposium (ref. [21] in [8]). Let the distance between the spacecraft (S/C 1) in the inner orbit and the ERS be l_1, and that for the spacecraft (S/C 2) in the outer orbit and the ERS be l_2. To make a Michelson

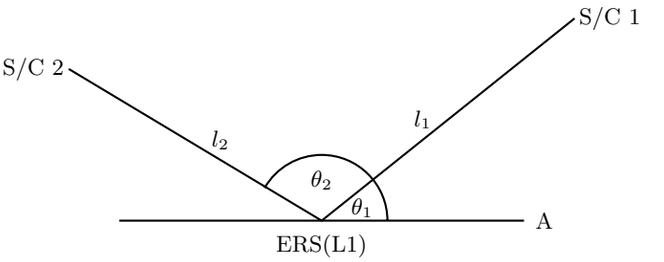


Fig. 4. Gravitational–wave detection topology. Path I: ERS → S/C 1 → ERS → S/C 2 → ERS. Path II: ERS → S/C 2 → ERS → S/C 1 → ERS.

interferometer from these two arms would require a linewidth of 1 mHz. To minimize the armlength difference, we propose to compare the following two paths through the whole mission except for the period 850 - 1000 days: (I) ERS→S/C 1→ERS→S/C 2→ERS; (II) ERS→S/C 2→ERS→S/C 1→ERS. The time difference of the two paths varies only up to 0.34 s. A laser with linewidth of 1 Hz would keep the interference coherent. The phase sonsitivity of this topology to gravitational wave is calculated as follows. In the plane containing S/C 1, S/C 2 and ERS (for simplicity, we use L1 as ERS), draw a reference line L1→A as in Fig. 4. Let the angle between ERS→S/C 1 and ERS→A be θ_1, and that between ERS→S/C 2 and ERS→A be θ_2. First consider a monochromatic gravitational wave with + polarization and strain amplitude h_+ coming in perpendicular to the orbital plane. The change of length scale in the L1→A direction at time t is $h_+ \sin(2\pi f_G t)$. For laser light travelling through Path I and Path II to return at t simultaneously, the optical path difference is

$$\Delta l = 4h_+(c/f_G)(\cos(2\theta_1) + \cos(2\theta_2))\times \\ [\cos(2\pi f_G(\tau_1 - \tau_2)) + \cos(2\pi f_G(\tau_1 + \tau_2))]\cos(2\pi f(t + \phi_0))\,, \quad (4)$$

with $\tau_1 \equiv 2l_1/c$, and $\tau_2 \equiv 2l_2/c$. Hence phase difference of the laser light is

$$\Delta\phi = 4h_+(f/f_G)(\cos(2\theta_1) + \cos(2\theta_2))\times \\ [\cos(2\pi f_G(\tau_1 - \tau_2)) + \cos(2\pi f_G(\tau_1 + \tau_2))]\cos(2\pi f(t + \phi_0))\,. \quad (5)$$

This is the laser phase senstivity for our gravitational detection topology. For $\times$ polarization, simply rotate by 45°. For non–perpendicular incidence, insert a cosine factor.

The theoretical sensitivity limit (shot noise limit) in the strain for gravitational-wave detection is inversely proportional to $P^{1/2}l$ with P the received power and l the distance. Since P is inversely proportional to l^2 and $P^{1/2}l$ is constant, this sensitivity limit is independent of the distance. For 1 - 2 W emitting power, the limit is $10^{-21}/\sqrt{\text{Hz}}$. As noted in the LISA study, making the arms longer shifts the whole time–integrated sensitivity curve to lower frequencies while leaving the bottom of the curve at the same level. With the same laser power, the ASTROD sensitivity would be shifted to lower frequency by a factor up to 60 (30 in average) if the technological requirements in Section 2 are met.

For one–spacecraft ranging with one leg on earth, 10 ps accuracy in the transit time variation is achievable. With 10^5 - 10^6 Hz repetition rate and a distance of 2 AU, the sensitivity can reach $10^{-17}/\sqrt{\text{Hz}}$ when an ultrastable clock of comparable stability is available in the future. In the T2L2 project study of satellite time transfer by laser light, precision and accuracy are aimed at the order of 10 ps. For gravitational–wave detection, this implies a sensitivity of 10^{-16}-$10^{-17}/\sqrt{\text{Hz}}$.

6 Mini–ASTROD and Super–ASTROD

ASTROD is a relativity mission concept encompassing multipurposes. With its technological requirements achieved, it would give a gravitational–wave sensitivity similar to LISA, but shifted to lower frequencies. ASTROD would complement LISA in probing the early Universe and study strong–field black hole physics.

A down–scaled version, i.e., Mini–ASTROD with one spacecraft carrying a payload of a telescope, two lasers, and a clock will test the optical scheme and yet give important scientific results. These scientific results include a better measurement of γ to 1 ppm [38], a better sensitivity (1 - 2 orders better) in using optical Doppler tracking method for detecting gravitational wave [39,40], and a potential of measuring the solar angular momentum via Lense–Thirring effect [41]. It is important to do things in appropriate steps and we are now studying this mini–ASTROD mission concept in more details.

With the advance of laser technology and the development of space interferometry, one can envisage 15 W (or more) compact laser power and 2 - 3 fold increase in pointing ability. With these developments, one can increase the distance from 2 AU for ASTROD to 10 AU (2×5 AU) and the spacecraft would be in orbits similar to Jupiter's. Four spacecraft would be ideal for a dedicated gravitational–wave mission (super–ASTROD).

7 Outlook

Optical space missions will be important in testing relativistic gravity and measuring solar–system parameters. Laser Astrodynamics in the solar system envisages ultraprecision tests of relativistic gravity, provision of a new eye to see the solar interior, precise measurement of $\dot{G}$, monitoring of the solar–system mass loss, and detection of low–frequency gravitational waves to probe the early Universe and study strong–field black hole physics together with astrophysics of binaries. One spacecraft and multi–spacecraft mission concepts — SORT, Mini–ASTROD, ASTROD and Super–ASTROD are in line for more thorough mission studies. In view of their importance both in fundamentals and in technology developments, mission concepts along this line will be realized and fruitful.

References

1. T. Melliti, F. Fridelance, and E. Samain: "Study of gravitational theories and of the solar quadrupole moment with the SORT experiment: Solar Orbit Relativity Test," *in preparation for Astronomy and Astrophysics; and references therein.*
2. K. Nordtvedt and D. Smith: private communication.
3. W.-T. Ni, J.-T. Shy, S.-M. Tseng, X. Xu, H.-C. Yeh, W.-Y. Hsu, W.-L. Liu, S.-D. Tzeng, P. Fridelance, E. Samain, A. M. Wu: "Progress in Mission Concept Study and Laboratory Development for the ASTROD—Astrodynamical Space Test of Relativity using Optical Devices," in *the Proceedings of SPIE*, Vol. 3316: Small Spacecraft, Space Environments, and Instrumentation Technologies, pp. 105-116, ed. F. A. Allahdadi, E. K. Casani, and T. D. Maclay, 1997; and references therein.
4. A. Bec-Borsenberger, J. Christensen-Dalsgaard, M. Cruise, A. Di Virgilio, D. Gough, M. Keiser, A. Kosovichev, C. Lämmerzahl, J. Luo, W.-T. Ni, A. Peters, E. Samain, P. H. Scherrer, J.-T. Shy, P. Touboul, K. Tsubono, A.-M. Wu, H.-C. Yeh: "Astrodynamical Space Test of Relativity using Optical Devices ASTROD—*A Proposal Submitted to ESA in Response to Call for Mission Proposals for Two Flexi-Mission F2/F3,*" January 31, 2000.
5. W.-T. Ni: Laser Astrodynamical Mission: Concepts, Proposals and Possibilities, Plenary talk given in *the Second William Fairbank Conference on Relativistic Gravitational Experiments in Space, 13-16 December, 1993, Hong Kong.*
6. W.-T. Ni, A.-M. Wu, and J.-T. Shy: A mission concept to measure second-order relativistic effect, solar angular momentum and low-frequency gravitational waves, in *Proceedings of the Seventh Marcel Grossmann Meeting on General Relativity (Stanford, California, July 23-30, 1994)*, (World Scientific, Singapore 1996), p. 1519.
7. W.-T. Ni, M.C.W. Sandford, C. Veillet, A.-M. Wu, P. Fridelance, E. Samain, G. Spalding, and X. Xu: Astrodynamical space test of relativity using optical devices, *presented to 31st COSPAR Scientific Assembly (Birmingham, 14-21, July, 1996)(National Tsing Hua University Preprint GP-074, July, 1996).*
8. W.-T. Ni: ASTROD and gravitational waves, in *Gravitational Wave Detection*, edited. by K. Tsubono, M.-K. Fujimoto and K. Kuroda, Universal Academy Press, Tokyo, Japan, p. 117 (1997).
9. W.-T. Ni: ASTROD mission concept and measurement of the temporal variation of the gravitation constant, pp. 309-320 in *Proceedings of the Pacific Conference on Gravitation and Cosmology, February 1-6, 1996, Seoul, Korea, ed. Y. M. Cho, C. H. Lee and S.-W. Kim* (World Scientific, Singapore 1996).
10. Z.-Y. Su, et al.: Asteroid perturbations and possible determination of asteroid masses through the ASTROD space mission, *Planetary and Space Science* **47**, 339 (1999).
11. M. Frœschlé, F. Mignard and F. Arenou: in *Proceedings of the ESA Symposium Hipparcos-Venice'97*, 13-16 May, Venice, Italy, ESA SP-402 (July 1997), p. 49.
12. L. Lindegren and M. A. C. Perryman: The GAIA Concept, in Proceedings of a Joint RGO-ESA Workshop on Future Possibilities for Astrometry in Space, Cambridge, UK, 19-21 June 1995 (ESA SP-379, September 1995), p. 23.
13. LISA Pre-Phase A Reprot, 2^{nd} edition, July, 1998 (MPQ233, Garching).
14. W.-T. Ni: Equivalence Principle, Lorentz Invariance and the Microscopic Origin of Gravity, invited talk given at the Inauguration Conference of the Asia Pacific Center for Theoretical Physics, June 4-10, 1996, Seoul, Korea.
15. ACES (Atomic Clock Ensemble onboard of the International Space Station), Mission Duration: beginning 2002-2005, Neuchatel Workshop, 12 June 1998.

16. J.W. Armstrong, F.B. Estabrook, and M. Tinto: "Time-delay interferometry for space-based gravitational wave searches", *Phys. Rev. D*, in press; and references therein.
17. A.–M. Wu, X. Xu and W.–T Ni: *Int. J. Mod. Phys.* **D 9**, 201 (2000).
18. D.–W. Chiou, and W.–T. Ni: "ASTROD Orbit Simulation and Accuracy of Relativistic Parameter Determination", *Advances in Space Research* **25**, 1259 (2000).
19. V.A. Brumberg: *Essential Relativistic Celestial Mechanics* (Adam Hilger, Bristol 1991); and references therein.
20. E. Samain, J. F. Mangin, C. Veillet, J. M. Torre, P. Fridelance, J. E. Chabaudie, D. Féraudy, M. Glentzlin, J. Pham Van, M. Furia, A. Journet, G. Vigouroux: "Millimetric Lunar Laser Ranging at OCA (Observatoire de la Côte d'Azur)", *Astron. Astrophys. Suppl.* **130**, 235 (1998).
21. G. J. Biermann: *Factorization Methods for Discrete Sequential Estimation*, (Academic Press, New York 1997); and references therein.
22. D.-W. Chiou, and W.-T. Ni: "Orbit simulation for the determination of relativistic and solar-system parameters for the ASTROD space mission," to be presented in COSPAR 2000 (Warsaw).
23. J. D. Anderson *et al.*: *Phys. Rev. Lett.* **81**, 2825 (1998).
24. X. Xu and W.-T. Ni: "Solar angular momentum calculation," *National Science Council report* (1997).
25. L. Paternó, S. Sofia, and M.P. Di Mauro: *Astronomy and Astrophysics* **314**, 940 (1996).
26. Y. Elsworth *et al.*: "Slow rotation of the sun's interior," *Nature* **376**, 669 (1995).
27. M.J. Thompson, J. Toomre, and E.R. Anderson *et al.*: *Science* **727**, 1300 (1996).
28. J. Schou, S. Tomczyk, and M.J. Thompson: "A measurement of the rotation rate in the deep solar interior", in *Proceedings of Fourth SOHO Workshop: Helioseismology, Pacific Grove, California, 2-6 April 1995 (ESA SP-376, June 1995).*
29. D. Hills, P.L. Bender, and R.F. Webbink: *Astrophysics Journal* **360**, 75 (1990).
30. T.R. Marsh: *Mon. Not. R. Astron. Soc.* **275**, L1 (1995).
31. T.R. Marsh, V.S. Dhilon, and S.R. Duck: *Mon. Not. R. Astron. Sec.* **275**, 828 (1995).
32. B.F. Schutz: in G. Pizzella and E. Coccia (eds.): *Processdings of the First Amaldi Conference on Gravitational Wave Experiments* (World Scientific, Singapore 1995).
33. D. Gough: *Nature* **376**, 120 (1996).
34. C. Cutler, and L. Lindblom: *Phys. Rev. D* **54**, 1287 (1996).
35. W.-T. Ni and X. Xu: Separation of the gravitational-wave signals and the solar oscillation signals, in S. Meshkov (ed.): *Proceedings of Third Amaldi Conference on Gravitational Waves* (AIP 2000), in press .
36. P. Kumar, and E.J. Quataert: *The Astrophys. J.* **458**, L83 (1996).
37. J.W. Harvey *et al.*: *Science* **272**, 1284 (1996).
38. C. Veillet: "TROLL, a proposal for the ESA M3 mission," 1993.
39. G. Giamepieri, R.W. Hellings, M. Tinto, and J.E. Faller: *Opt. Commun.* **123**, 669 (1996).
40. R. Hellings, G. Giamepieri, L. Maleki, M. Tinto, K. Danzmann, J. Hough, and D. Robertson: *Opi. Commun.* **124**, 313 (1996).
41. B. Bertotti and D. Giampieri: *Class. Quantum Grav.* **9**, 777 (1992).

Clocks for Length and Time Measurement

Fritz Riehle

Physikalisch–Technische Bundesanstalt (PTB), 38116 Braunschweig, Germany

1 Introduction

The evolution of various fields of science, technology, trade or legal metrology is intimately connected with the ability to relate measurements with each other that were performed at different places and different instants of time. For this purpose a practical system of units of measurement i.e. the International System of Units (SI) has been established by international cooperation [1]. In this SI, the metre and the second represent the base units of length and time, respectively. From all units these two can be realized with by far the highest accuracy since they are based on frequency measurements and most accurate clocks. In contrast to clocks based on mechanical properties of macroscopic bodies, e.g., pendulum clocks, quartz clocks or pulsars, the frequency reference for a suitable oscillator in atomic clocks is mainly determined by the intrinsic properties of suitable absorbers like atoms, molecules or ions. These atomic properties are determined by fundamental constants resulting from the basic interactions between the elementary constituents of matter. Following the generally accepted idea that the properties of each atomic absorber of a selected species are the same, identical clocks can be set up in any desired number and at any desired place.

Consequently, the definition of the second is based on an atomic frequency: *The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the hyperfine levels of the ground state of the caesium-133 atom [2].* In the SI, this definition is also the basis for the length unit which is related to the second by a fundamental constant i.e. the speed of light. *The metre is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second [3].* This definition can be applied directly to measure large distances, the distance to the moon, e.g., as has been done with unprecedented accuracy (see contributions of K. Nordtvedt, this volume on p.317). The large value of the speed of light, however, prevents one from measuring distances on a laboratory scale with the required accuracy. The definition corresponds to a fixed speed of light of c = 299 792 458 m/s and at the same time fixes the ratio between the vacuum wavelength λ and the frequency $\nu = c/\lambda$ of a light wave. Hence, the vacuum wavelength of a light wave used for accurate interferometric distance measurements is known, in principle, with the same accuracy as the frequency of the light wave. Consequently, time measurements as well as distance measurements rely on frequency standards or clocks, i.e. oscillators with known frequencies. At present, clocks that are used for time measurements predominantly operate in the radio frequency (rf) regime whereas

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 347–368, 2001.

for length measurements optical clocks are used. In the context of general relativity, the definitions of the second and the meter have to be considered as units of proper time and proper length, respectively. Depending on the required accuracy, it is often necessary to apply general relativity to metrology [4].

In this lecture, Sect. 2 concentrates on the basics of atomic clocks and reviews the status of rf clocks (Sect. 3) and optical clocks (Sect. 4). In Sect. 5 an example of the required frequency links between both regimes is given that now allows to use rf clocks and optical clocks in the same way for length and time measurements. Sect. 6 concludes with optical frequency standards for the realization of the meter.

2 Basics of Atomic Clocks

2.1 Definitions

The definition of the second [2] does not include a definite instruction how to build a clock. However, the statement "the Caesium-133 atom" makes some implicite assumptions. First, the transition frequency is assumed to be the same in all Caesium-133 atoms used in the clock. This requires that the clock is operated in such a way that the ideal conditions of atoms isolated and at rest are fulfilled. In reality, these ideal conditions can be realized only approximately and there will be frequency offsets due to spurious effects. The accuracy of a clock can be *measured* only if a more accurate clock is available. Otherwise, as it is the case for the best primary standards the magnitude of the uncertainty of the knowledge about these offsets is a measure of the accuracy (or uncertainty) one has to attribute to a real clock [5]. In this sense, the uncertainty represents the *estimated* magnitude of a possible deviation of its frequency from the unperturbed atomic frequency. Within a limited measurement time τ the actual mean value of the frequency of a clock may differ by a larger magnitude as compared to the one given by the uncertainty resulting from various noise processes associated with the realization of the clock e.g. by the shot–noise of the atomic sample or noise in the electronics. A generally agreed measure of the frequency instability of a clock is given by the relative two–sample standard deviation, or Allan standard deviation $\sigma_y(2,\tau)$ [6]

$$\sigma_y(2,\tau) = \sqrt{\left\langle \frac{(\overline{y}_{k+1} - \overline{y}_k)^2}{2} \right\rangle}. \tag{1}$$

Here, $\overline{y}_k$ is the average value of the normalized frequency difference $y(t)$ in an arbitrary time interval t_k to $t_k + \tau$ between the clock under investigation and a suitable reference. The normalized frequency is used rather than the frequency to allow the comparison of the performance of clocks operating at very different frequencies. For short times τ the frequency stability is often limited by the counting statistics

$$\sigma_y(2,\tau) = const \times \left(Q \times \frac{S}{N} \times \sqrt{\frac{\tau}{\mathrm{s}}} \right)^{-1} \tag{2}$$

where $Q \equiv \nu/\Delta\nu$ is the quality factor of the resonance depending on the linewidth $\Delta\nu$ and the center frequency ν, S/N is the signal–to–noise ratio in a bandwidth of 1 Hz. The constant depends on the lineshape of the resonance line and is of the order unity.

2.2 Perturbing Effects and Remedies

The ideal reference in an atomic clock or frequency standard would therefore comprise of an absorber at rest in an environment free of perturbations by other particles or fields with a high line quality factor Q and a strong signal when probed by an external oscillator. The art of making a perfect clock depends on the ability to realize this ideal situation as close as possible.

Limited Interaction Time The basic limitation to the quality factor Q is given by the frequency uncertainty $\Delta\nu \cong 1/T$ resulting from the time T the absorbers are interrogated. Consequently, one is interested first to use absorbers with narrow line–width transitions and second to allow for the required long interaction time of the external field with the quantum absorber. Such long–lived states are provided e.g. by the ground states of atoms and ions splitted due to the hyperfine interaction leading to frequency separations in the rf domain up to several 10 GHz (see table 1). Furthermore, atoms, ions or molecules sometimes have long–lived electronic states with optical transitions corresponding to several hundred terahertz (see e.g. Table 2).

To utilize the potential of these lines one has to allow for an interaction time of milliseconds and above. Long interaction times can be achieved by several methods that find application in various frequency standards. There are long beam machines based on atomic or molecular beams of low velocities together with separated field excitation [7] (Ramsey excitation) as used in the the conventional Cs atomic clocks (Sect. 3.1). In the hydrogen maser the atoms are kept inside the interaction volume by containing them in a bulb. Laser cooled ballistical atoms allow easily for interaction times of a millisecond and this time can be increased to about a second in a so–called atomic fountain (Sect. 3.3). The longest interaction time can be achieved with ionized particles in a so–called ion trap in the harmonic potential well (see Sect. 3.4).

Velocity of the Absorbers Even if an absorber with a very small natural line width has been chosen the center of the resonance line and its width can be seriously affected by the velocity of the absorber. The shift of the absorbed frequency resulting from the first–order Doppler effect $\Delta\nu(1^{\text{st}}\text{ order}) = \boldsymbol{k} \cdot \boldsymbol{v}/2\pi$ can be as high a 10^{-6} for thermal absorbers. This shift depends on the direction of the wave vector $\boldsymbol{k}$ and on the direction of the velocity $\boldsymbol{v}$ of the absorbers. Consequently, its influence can be largely avoided by a perpendicular alignment of both vectors as it is done in a beam apparatus. Other methods determine the influence of the first–order Doppler shift and subsequently correct the corresponding frequency offset e.g. by re–directing one of the vectors

by a so–called beam reversal technique. These methods cannot be applied to the second–order Doppler shift resulting from the time dilation of the absorber which is $\Delta\nu(2^{\text{nd}}\text{ order}) = -\nu \cdot v^2/(2c^2)$ when observed in the laboratory frame. For the aforementioned thermal velocities it can be as high as $\Delta\nu/\nu \cong 10^{-12}$. This shift scales with the square of the velocity of the absorbers and can only be reduced by orders of magnitude if absorbers of low velocities are utilized. The use of slow absorbers also increases the interrogation time of the field with the atoms and therefore leads to a reduced transit–time broadening.

With neutral atoms two basically different approaches are used to reduce the influence of the first–order Doppler effect. The first one selects a group of slow absorbers out of a broad velocity distribution e.g. by using transversal excitation of absorbers in collimated molecular and atomic beams or by selecting only the slow molecules out of the thermal distribution [8] leading to a reduction of the relevant velocity spread by several orders of magnitude. The second approach reduces the width of the velocity distribution by cooling the ensemble. The positive result of this approach is that more absorbers are compressed into the desired velocity range and the signal used for frequency stabilization of the oscillator to be locked is increased. Laser cooling of atoms now represents a well established and acknowledged technique [9–11] to reach low velocities. Two–level atoms with a fast cycling transition can be cooled to the so–called Dopppler limit $k_{\text{B}}T = \hbar\gamma/2$, where γ is the line width of the cooling transition. As an example, the root–mean–square velocity corresponding to this temperature is 9 cm/s for Cs. Often, laser cooling is performed in a magneto–optical trap (MOT) [12]. In a MOT, photons from six orthogonal laser beams with frequencies below the atomic resonance are absorbed by the atoms with velocity components counterpropagating to the laser beams. The subsequent emission of photons from the spontaneous decay of these excited atoms is isotropically whereas the absorbed photons always have the same momentum $\boldsymbol{k}$ of the laser beam. Hence, the atomic velocity after a large number of absorption/emission cycles is reduced. In the MOT, a magnetic field increasing into all spatial directions with the zero of the field located at the center of the six crossing beams shifts the energy levels of the atoms by the Zeeman effect. Atoms outside the center of the MOT experience a stronger repelling force due to the bigger Zeeman shift than the ones close to the center. Consequently, the combined influence of the light field and the magnetic field leads to cooling and trapping of the atoms in a MOT. The achievable velocity can be significantly lower as given by the Doppler limit in the case of multi–level atoms, due to the mechanism of the so–called polarization gradient cooling [10] leading to an rms velocity of 1.1 cm/s in Cs. Raman cooling methods can achieve even lower velocities [13]. Laser cooling works also very efficient for trapped ions [14], where specific cooling methods for the bound ions as e.g. resolved–optical–sideband cooling allow to cool a single ion close to the zero point energy [15].

Interaction with Fields and Particles The interaction of the absorbers with external fields perturbs the quantum states of the absorbers and consequently

Table 1. Ground–state splitting due to the hyperfine structure

H	1 420 405 751.770 Hz ± 0.003 Hz	H–maser	[5]
^{87}Rb	6 834 682 610.904 29 Hz ± 0.000 09 Hz	Rb–clock	[19]
^{133}Cs	9 192 631 770 Hz	Cs–atomic clock	[5]
$^{171}Yb^{+}$	12 642 812 118.466 ± 0.002 Hz	Yb^{+}–ion trap	[20]
$^{199}Hg^{+}$	40 507 347 996.841 59 ± 0.000 44 Hz	Hg^{+}–ion clock	[21]

leads to frequency splitting, shifts or broadenings that, even small, often cannot be tolerated when the highest accuracy is aimed at. Examples include Stark and Zeeman shifts resulting from static magnetic and electric fields but also from ac–Stark shifts as a result of trapping fields, stray light or thermal radiation [16]. The interaction with other particles of the same species [17] or of other species [18] may lead to collisional shifts. The degree to which these influences can be suppressed and corrected for is reflected in the achieved accuracy of a clock.

3 Clocks Based on rf Transitions

A number of excellent frequency standards are based on the hyperfine separation of the ground state of atoms and ions (see table 1). This splitting is due to the coupling between the magnetic moment associated with the total spin J of the electron shell and the nuclear spin I. In ^{133}Cs e.g. the coupling of $I = 7/2\ \hbar$ and J=1/2 $\hbar$ leads to the $F = I + J = 4$ and $F = I - J = 3$ states which split in the magnetic field (Fig.1). The Cs clock uses the transition with the smallest sensitivity to magnetic fields, i.e. between the $F = 4, m_F = 0$ and $F = 3, m_F = 0$ states with a quadratic dependence of $\Delta\nu \cong 0.0427 \times B^2$ Hz/$(\mu T)^2$.

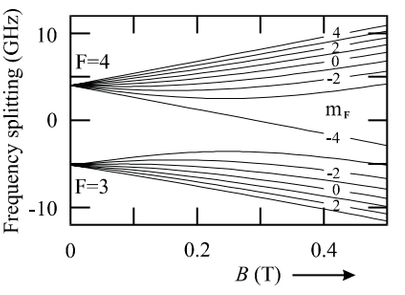


Fig. 1. Hyperfine structure of Caesium in a magnetic field

3.1 Cs Beam Machine

Most Cs atomic clocks operated today use Cs atoms in an atomic beam inside a high vacuum chamber. Their principle (Fig.2) does not differ much from the ancestor of the Cs atomic clocks developed at the National Physical Laboratory, England [22]. The Cs atoms effusing from an oven kept at a temperature of about 450 K have a most probable velocity of about 250 m/s. Due to the small energy separation between the $F = 3$ and $F = 4$ states both levels are in fact almost equally populated in the thermal beam. Detection of the transitions

between these levels induced by an external oscillator can be performed with an optimal signal–to–noise ratio in a large sample of atoms only if most atoms are initially in the same state. It is therefore necessary either to prepare all atoms in the same state or to select atoms from one state. The latter procedure is applied in conventional Cs atomic beam machines in which a magnetic polarizer with an inhomogeneous magnetic field realized by a two–, four–, or six–pole magnet is used to deflect the atoms according to their magnetic moment. Only atoms in the state $F = 4$ enter the first interaction zone of the Ramsey cavity fed by the rf derived from a voltage controlled quartz crystal oscillator (VCXO). The interaction with the rf field transfers the atom into a coherent superposition between the $F = 4$ and the $F = 3$ state. The temporal evolution of the quantum mechanical state of the Cs atom can be thought of as an oscillation with a frequency corresponding to the energy difference of these states. After the interaction of the atom with the field in the second zone of the Ramsey cavity the probability to find the atom in the $F = 4$ state or in the F=3 state depends on the external rf field being in phase or out of phase with the atomic oscillator at the second interaction zone. Hence, the number of atoms in either the $F = 4$ or in the $F = 3$ state oscillates as a function of the frequency of the external oscillator thereby leading to the Ramsey interference structure [23] similar to the ones shown later (Fig.4 and Fig.5). In Fig.2 the magnetic analyzer deflects the atoms in the F=3 state into the detector. The detector signal is used to stabilize the frequency of the VCXO to the atomic transition frequency. The largest contributions to the frequency uncertainties in these clocks results from cavity phase shifts due to the imperfect field distribution in the cavity and from inhomogeneities of the magnetic C–field. As a representative of the state–of–the art primary beam–machine clocks, PTB's CS1 has a stability of $5 \times 10^{-12}\sqrt{\tau/\mathrm{s}}$ and an accuracy of 7×10^{-15} [24].

3.2 Optically Pumped Cs Clocks

The short–term stability of the Cs beam machines is limited by the shot noise of the number n of atoms in the beam since the signal–to–noise ratio (2) increases with $\sqrt{n}$. The stability can be increased if all atoms can be prepared in the desired state rather than selecting one particular state with only 1/16 of the total number of atoms leaving the oven (see Fig.1). Optical pumping techniques can be utilized to accumulate atoms in the desired state. Similarly, the detection of the excited atoms can be performed by laser induced optical fluorescence. Several Cs beam machines based on these principles have been designed and operated (see e.g. [25–28]). A stability of about $1 \times 10^{-12}\sqrt{\tau/\mathrm{s}}$ [29,28] and an accuracy of 1×10^{-14} has been reported [28] for these types of clocks.

3.3 Atomic Fountain Clocks

Laser cooling of Cs atoms can lead to very low temperatures of about 2 μK corresponding to velocities of a few centimeters per second that might allow interaction times of seconds and more. The utilization of such a long interaction

time meets with difficulties, because gravitation accelerates the atoms and, in contrast to ions to be discussed in more detail in Sect. 3.4 and Sect. 4.1, up to now no means have been found to keep neutral atoms at rest in the gravitational field without seriously perturbing the resonance frequency. Applying Ramsey's separated–field technique to ballistically free flying atoms leads to the development of a of a fountain clock in which interaction times of about a second can be realized. In an atomic fountain a cloud of cold atoms is launched vertically through an interaction region with a velocity of a few meters per second. The gravitational acceleration g forces the atoms to slow down and fall back thereby passing the same interaction zone a second time. Similarly, as in the Cs atomic clock with a thermal beam, the resulting resonance feature displays a Ramsey interference structure with a resolution that is determined by the time T between the two interactions. This time is calculated from the time T' necessary for the atoms to climb up to the apogee and the same time to fall down

$$T = 2T' = 2\sqrt{\frac{2H}{g}}. \tag{3}$$

For a realistic apparatus with a height between the resonator and the apogee of H= 1 m this time is $T = 0.9$ s and the necessary starting velocity of the atoms is $v = \sqrt{2gH} = 4.5$ m/s. The low velocities of the atoms in the atomic cloud are prerequisites for an efficient operation of such an atomic fountain inasmuch as the number of atoms returning into the microwave cavity after the ballistic

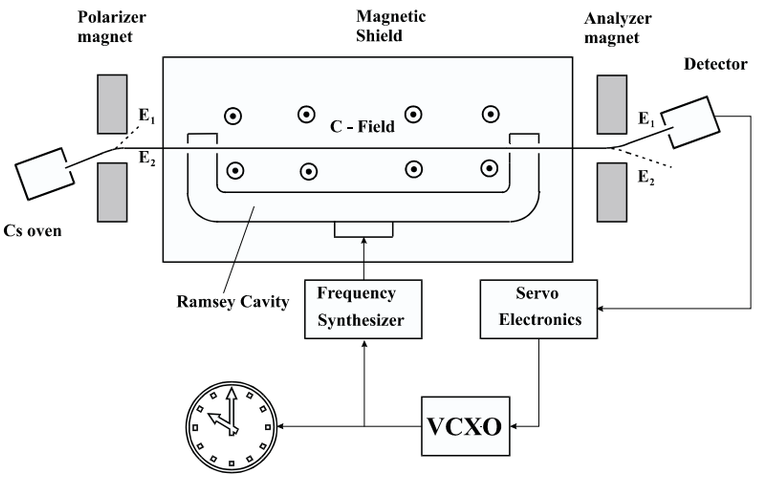


Fig. 2. Schematics of a Cs atomic clock. The field lines of the homogeneous magnetic C–field are perpendicular to the paper plane. The states E_1 and E_2 correspond to all $F = 3$ states and the $F = 4, m_F = -4$ state and to the $F = 4, -3 \leq m_F \leq +4$ states, respectively (see Fig. 1).

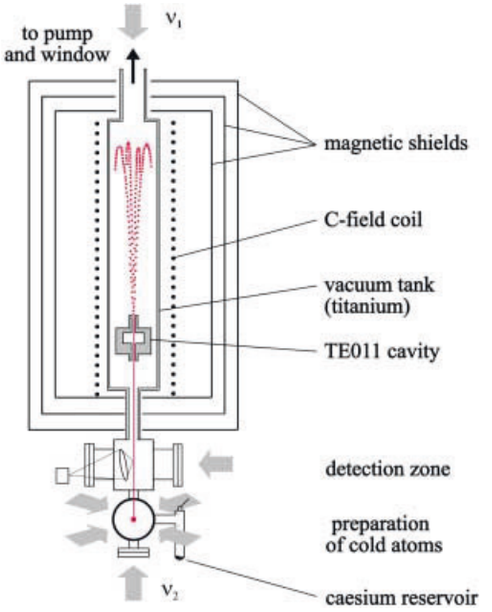


Fig. 3. Cs atomic fountain clock.

flight depends on the transverse velocity. From a cloud of caesium atoms with a temperature $T = 2\ \mu$K about 10% of the atoms find their way through the 1 cm hole of the resonator in contrast to about 0.2% that would enter the same region if the sample was cooled only to the Doppler limit of 125 μK [30]. The first successful predecessor of an fountain experiment using laser cooled sodium atoms in a pulsed rf cavity demonstrated a line width of 2 Hz [31]. Following the realization of the first fountain clocks [32,33] it has been realized soon that collisions between the cold atoms may limit the achievable accuracy in a Cs fountain [17]. It is now known that the associated shifts are much less pronounced in rubidium [34,35]. Now a variety of atomic fountains using Cs or Rb atoms are operated or under investigation for atomic clocks in several institutions [36–41,34,35,42]. Even though these realizations differ somewhat in their design, each apparatus uses three sections (see Fig.3), a trapping and cooling section, an excitation section containing the microwave cavity and the zone for the ballistical flight and a section for the detection. The cloud of cold atoms is often prepared from a thermal vapour fed by a Cs reservoir at a base pressure of about typically 10^{-6} Pa. In a MOT [12] about 10^7 Cs atoms are collected and further cooled in an optical molasses to about 2 μK and sometimes further cooled by means of Raman cooling techniques [43].

In the next step the atoms have to be launched upward without increasing their temperature. This is conveniently done in a moving molassis where the frequency ν_1 of the downward pointing vertical laser beam is red detuned by $\delta\nu$ and the frequency ν_2 of the upward pointing beam is blue detuned by $\delta\nu$ with respect to the original frequency ν. The corresponding phase front e.g. the ones for $\delta\omega t + kz = 2\pi\delta\nu t + 2\pi z/\lambda = 0$ move with the velocity $v = z/t$ upwards. The atoms are now exposed to this "walking wave" with velocity

$$v = \lambda\delta\nu \tag{4}$$

and are instantly laser cooled into equilibrium with the walking wave and hence are moving up with the velocity given in (4). When the vertical beams are switched off the atoms follow the trajectories of a free ballistic flight. The long flight time between the two interactions leads to Ramsey spectra with high resolution (Figs. 4 and 5).

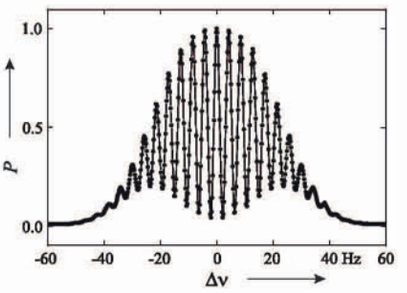


Fig. 4. Measured probability P to excite Cs atoms in the fountain clock of PTB with low resolution.

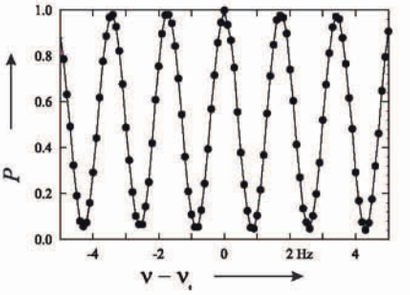


Fig. 5. Excerpt of the high resolution Ramsey spectrum similar to Fig.4 with a height of the atomic fountain of about 0.6 m.

Recently, a stability of 5 $\times 10^{-14}(\tau/\mathrm{s})^{-1/2}$ [44] and accuracy of 1-2 $\times 10^{-15}$ [45,36] has been reported for Cs fountain clocks with smaller uncertainties anticipated.

3.4 Clocks Based on Ion Traps

To keep microscopic particles at rest at a fixed position in space a strong binding force pointing to this point is required. Due to the weak interaction of neutral atoms or molecules with electric and magnetic fields, strong fields or field gradients are required which may seriously perturb the atomic energy levels. For ionized particles, however, much smaller fields can be used for confinement in so–called ion traps. There are several advantages of ion traps when used for frequency standards which will be briefly addressed here. First, storage times of days and more allow to probe ultra–narrow lines with virtually no broadening due to limited interaction times. Second, the concentration of the absorbers into a small volume allows effective application of the methods of laser cooling and detection of the induced signals originating from a single spot. The reduction of the velocity and the confinement to regions smaller than the wavelength of the probing radiation (Dicke regime) allows the reduction of the Doppler effect in all orders. According to Dicke [46] the first–order Doppler effect of a single–photon transition of wavelength λ can be suppressed if the absorber is kept in a range smaller than $\lambda/2$. Third, the use of ultra–high vacuum reduces perturbations from the coupling to the environment. Fourth, the strong interaction with other ions can be avoided by the use of a "mono–ion oscillator as potential ultimate laser frequency standard" as has been proposed by Dehmelt already in 1982 [47]. The use of ion traps in frequency standards has been the subject of a number of recent reviews (see e. g. [48–51]).

In an ion trap an electric field $\boldsymbol{E}(\boldsymbol{r})$ can be used that leads to a net force on an ion with charge e and mass m pointing to the center of the trap at any space point inside the trapping volume. The corresponding potential $\Phi(x, y, z)$ has a parabolic shape and may be represented by $\Phi \propto \left(ax^2 + by^2 + cz^2\right)$. For the

constants determining this potential the condition $a + b + c = 0$ can be derived from the Laplace equation. This condition is e.g. fulfilled by the two particular solutions $a = 1, b = -1, c = 0$ corresponding to a linear quadrupole trap and $a = b = 1, c = -2$ representing the three–dimensional Paul trap [52].

The two–dimensional quadrupole potential can be generated by a set of four hyperbolic electrodes where the potentials of adjacent electrodes change polarities. This case can be approximated by four rods (see Fig.6 b). In the case of a static field in the radial direction a particle of electric charge $+e$ will be repelled from positively charged electrodes experiencing a repulsive force towards the center at $r = 0$. According to the linear dependence of the field strength E on the coordinate similar to Hooke's law we expect the ion to perform a harmonic oscillation in this direction. Along the orthogonal direction, however, the ion is accelerated towards the nearest negative electrode. To confine the ion for both directions, in the rf or Paul trap [52] the potential of both pairs of electrodes is alternated periodically

$$\Phi(x, y, t) = \frac{U_{\mathrm{ac}}}{2r_0^2}\,(x^2 - y^2)\cos(\omega t)\,. \tag{5}$$

The alternating potential leads under certain conditions [53] to a time–averaged net force that attracts the ions towards the nodal line. In Fig.6 b) two additional ring electrodes confine the ions in the axial direction.

In a three dimensional Paul trap the potential has the form

$$\Phi(x, y, z, t) = \frac{U_{\mathrm{dc}} + V_{\mathrm{ac}}\cos(\omega t)}{r_0^2 + 2z_0^2}\,(x^2 + y^2 - z^2)\,. \tag{6}$$

A *dc* voltage U_{dc} is added in general to the alternating rf voltage V_{ac} with a driving frequency $\nu = \omega/2\pi$ to vary the ratio between the restoring forces in axial and radial direction. The positive sign in (6) leads to a hyperbolic surface with rotational symmetry around the $z-$ axis generated by a ring electrode of an inner radius r_0 (Fig.6a). The negative sign results in two branches of the hyperbola of revolution separated by the distance $2z_0 = \sqrt{2}r_0$ also exhibiting rotational symmetry with respect to the $z-$ axis.

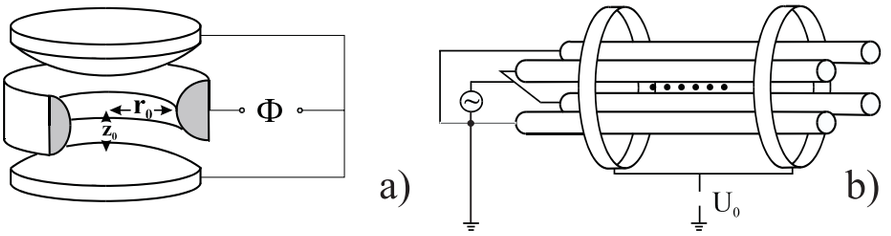


Fig. 6. Radio frequency ion traps a) Paul trap with hyperbolic end caps and a ring electrode b) linear trap.

Inside the trap the trajectory of the ion is a superposition of a driven motion at the frequency ω referred to as the micromotion and a low–frequency motion called the secular motion.

Ion traps of both kinds shown in Fig.6 have been used with many species for frequency standards (for an overview see e.g. [50]). Using rf transitions, most work has probably been done using mercury and ytterbium ions. The 12.6 GHz clock transition in a cloud of about 50 000 $^{171}Yb^{+}$ ions in a hyperbolic Paul trap has been utilized [18,54]. The ions were cooled with a buffer gas and the frequency has been determined by extrapolation to a vanishing relativistic, collisional and Stark effect shift with a relative uncertainty of 7 $\times 10^{-13}$. The largest contribution to the systematic uncertainty resulted from the temperature of the ions of about 1800 K. Fisk et al [20] used two linear Paul traps and demonstrated a relative uncertainty of 2 $\times 10^{-13}$. The Allan standard deviation as a measure of the frequency instability was $\sigma_y(2,\tau) = 4.7 \times 10^{-14}(\tau/\mathrm{s})^{-1/2}$.

The 40.5 GHz clock transition in ^{199}Hg has been investigated and utilized as a frequency standard by several groups (see e.g. [21,55–57]). Devices with about 2 $\times 10^{6}$ ions cooled with He buffer gas have been designed by Cutler et al [56] and the frequency was measured with a relative uncertainty of 2.5 $\times 10^{-13}$. A mercury linear ion trap standard operated at the JPL gave a stability of $\sigma_y(2,\tau) = 6.5 \times 10^{-14}(\tau/\mathrm{s})^{-1/2}$ [57]. Slightly lower stability of $\sigma_y(2,\tau) = 3.3 \times 10^{-13}(\tau/\mathrm{s})^{-1/2}$ but excellent low relative uncertainty from systematic effects of only 3.4 parts in 10^{15} was reported for a few ions in a cryogenic linear trap of the National Institute of Standards and Technology [21].

4 Optical Clocks

The major advantage of frequency standards operating in the optical rather than in the rf range is based on their high frequencies. With the same interaction time i.e. with the same resolved line width the quality factor Q can be higher by several orders of magnitude. The short wavelength of the radiation in the optical domain, however, makes it difficult to keep the absorbers in the Dicke regime.

4.1 Clocks Based on Ion Traps

Ion traps allow to confine a single atom in the Dicke regime. For highest accuracy the number of trapped ions has to be restricted to a single ion in a hyperbolic Paul trap or to a few ions in a linear Paul trap since the ions have to be confined at the field free center or nodal line of the traps. The sensitive detection of the transition of an absorber consisting of a single ion is achieved by the method of electron shelving [47] where a strong transition shares the ground state with the weak clock transition. Suppose, that the fluorescence radiation from the strong transition is monitored. The transition of the ion into the long lived state immediately is identified by the disappearance of the fluorescence radiation from the strong transition which persists as long as the electron is shelved in the long–lived state. This scheme allows signal detection with unity detection efficiency.

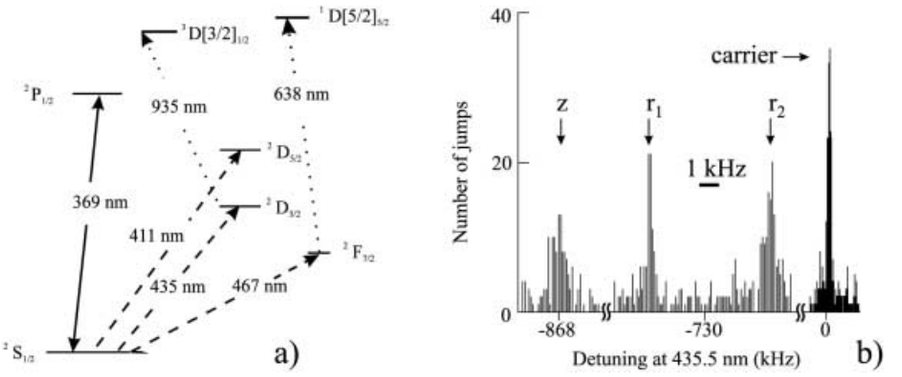


Fig. 7. a) Partial energy diagram of the Yb ion b) Spectrum of the transition of a single Yb ion in a Paul trap as measured by quantum jumps [66]

There is a number of suitable candidate ions for optical frequency standards [58] including $^{88}Sr^+$ [59,60] at 674 nm, $^{115}In^+$ [61–63] at 236.5 nm, $^{199}Hg^+$ [64] at 282 nm or $^{171}Yb^+$. Ytterbium has several transitions at 411 nm [65], 435 nm [66], or 467 nm [67] (see Fig.7 a). The strong transition at 369 nm is used for laser cooling and detection. The line at 467 nm has a natural line width of 5 $\times 10^{-10}$ Hz corresponding to $Q \cong 10^{24}$ [67], but due to its low transition rate it has been detected but not utilized yet. The 435 nm transition has been recorded with a line width of 80 Hz [66] by detecting the quantum jumps (see Fig.7 b) where the carrier together with the sidebands due to the oscillation frequencies of the ion in the trap in radial (r_1, r_2) and vertical (z) direction can be identified. Similar line widths have been observed with the mercury ion and it is expected that optical frequency standards based on single ions soon will reach relative uncertainties at 10^{-15}.

4.2 Optical Frequency Standards Based on Neutral Atoms

From the proposed optical frequency standards based on neutral atoms (see e.g [68,69]) hydrogen [70], xenon [71,72] and the alkaline earth metals have received the most attention. The narrow intercombination line of the latter ones was utilized e.g. in Mg [73], Ca [74,75], or Sr [76]. Here the recent work of an optical frequency standard based on ^{40}Ca is reported.

4.3 Expanding Cloud of Cold Ballistic Ca Atoms

Ca atoms from the low velocity tail of the Boltzmann distribution ($v_{prob} \cong 600$ m/s, $T_{oven} \cong 900$ K) are cooled and trapped in a MOT [12] (see Fig.8). The Ca intercombination line at $\lambda = 657$ nm is interrogated by the radiation from a high–resolution dye laser spectrometer [77] or a diode laser spectrometer [78]. The laser pulses for the excitation are generated by acousto–optic modulators

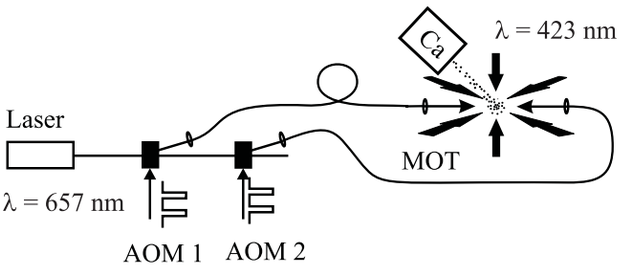


Fig. 8. Experimental setup for an optical Ca standard.

(AOMs) (Fig.8) used as "light switches". The radiation at $\lambda \cong 423$ nm used to cool, deflect, and trap the Ca atoms in the MOT are produced either by a dye laser or a frequency doubled diode laser system. Up to 10^7 atoms are loaded during about twenty milliseconds. To avoid systematic frequency shifts due to the Zeeman effect of the magnetic field of the MOT and light shifts due to the 423 nm radiation, both the magnetic field and the light of the trapping laser have to be shut off before interrogating the Ca atoms. The atomic cloud expands according to the root–mean–square (rms) velocity of the ballistic atoms of $v_{rms} \cong 1$ m/s. In order to achieve high spectral resolution combined with a good signal–to–noise ratio (S/N), the method of separated field excitation (optical Ramsey excitation) is applied. Short pulses of 1 μs duration are used to excite a significant part of the cold ensemble of atoms. The corresponding time sequence consisted of atom trapping ($t_1 \cong 15$ ms), turning off the trapping fields ($t_2 \cong 0.5$ ms), separated–field excitation, and detection ($t_4 \cong 0.5$ ms). If the laser frequency is tuned close to the resonance, the fluorescence intensity contains a contribution which varies with the cosine of the laser detuning ($\nu - \nu_{Ca}$) times $2\pi t$ (Fig.9). In contrast to rf Ramsey excitation, in the optical regime the atoms are not in the Dicke

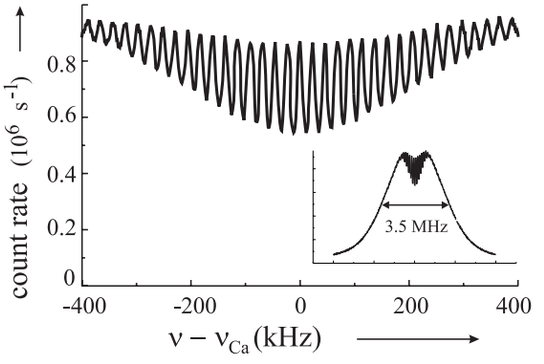


Fig. 9. Optical Ramsey resonances excited in an expanding cloud of Ca atoms released from a MOT .

regime. Hence, three [79] or four [80] interaction zones have to be used to excite Doppler–free optical Ramsey resonances. Bordé has shown that the resonance structures can be interpreted as atom interferences generated by the excitation with separated fields [81]. The narrowest fringe width obtained with the Ca standard was below 400 Hz. PTB operates two different systems [74,82] allowing one to compare the transition frequencies of the intercombination line measured at two independent ensembles of Ca atoms and to check the reproducibility of the frequency delivered by such a Ca standard. Arelative uncertainty below 10^{-13} was derived and it was shown that for an optimized system a relative uncertainty of about 10^{-15} [82] can be expected. A stability of $1.3 \times 10^{-14}(\tau/\mathrm{s})^{-1/2}$ was obtained for a similar system at NIST [75].

5 Measurement of Optical Frequencies

A frequency stabilized laser only represents an optical frequency standard if its frequency is compared directly or indirectly with the frequency of the primary standard of time and frequency i.e. with the Cs atomic clock. This comparison meets with some difficulties due to the fact that both frequencies differ by almost five orders of magnitudes. Therefore, up to now not many frequencies of optical frequency standards were measured. Some recent measurements include the optical Ca frequency standard [83,74], optical transitions in atomic hydrogen [84,85], the Sr^+ ion standard [60], or the In^+ ion standard [63]. As an example of a so–called frequency chain, PTB's phase–coherent frequency chain from the Cs atomic clock to the optical regime is shown in Fig.10. This chain is described in more detail elsewhere [83] and it suffices to recall that it basically applies the method of harmonic mixing thereby connecting the Ca standard via various intermediate oscillators (diode laser, colour center laser, CO_2 lasers, methanol laser, backward wave oscillator, Gunn oscillator, and H–maser) to the Cs clock. As an example, consider the uppermost stage of Fig.10. The frequency of a diode laser is tuned near half of the frequency of the Ca stabilized laser. The radiation of the diode laser is frequency doubled in a non–linear crystal and the beat note between this radiation and the radiation of the Ca stabilized laser is monitored with a photo diode. If the signal of the photo diode is phase locked using a Phase Locked Loop (PLL) the frequency ratio of both lasers is kept constant and can be used to determine the frequency of the Ca stabilized laser, provided that the frequency of the diode laser and the frequency of the PLL are known. In the same way, the frequency ratios of all other stages can be related. Such a frequency chain allows a phase–coherent optical frequency measurement which, in principle, does not contribute to the uncertainty of the measured frequency value provided, cycle slips can be excluded in each multiplication step.

Several frequency measurements during more than two years have been performed at PTB using two different MOTs, different resolutions, and different stabilization schemes. The weighted mean of all frequency measurements up to now is $\nu_{Ca} = 455\ 986\ 240\ 494.13\ (12)$ kHz. This uncertainty makes the Ca stabilized laser one of the most accurate optical frequency standards in the visible,

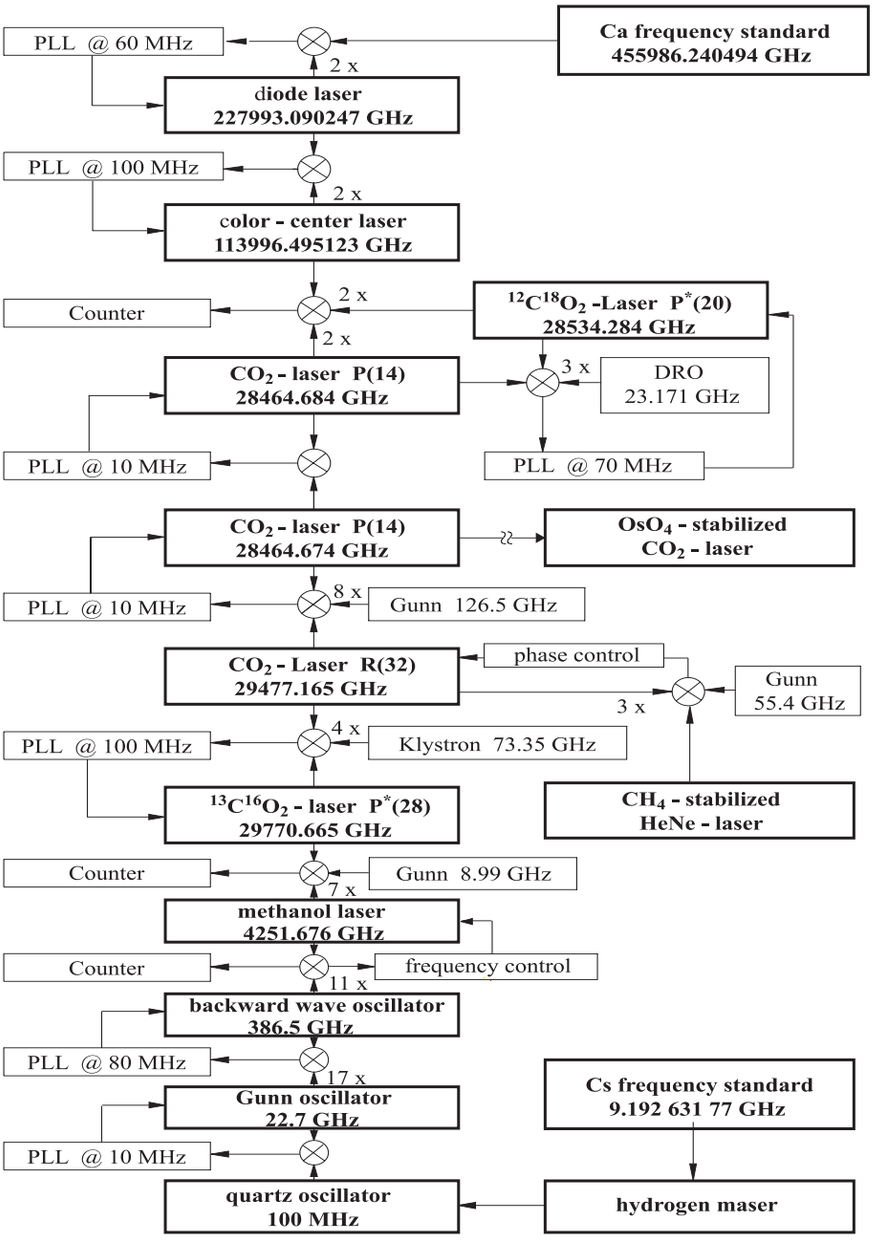


Fig. 10. Frequency measurement chain of PTB.

Table 2. Radiations recommended by the CIPM for the realization of the meter [86]

Atom or Molecule	Transition	Wavelength (μm)	Rel. Standard Uncertainty
^{1}H	$1S - 2S$	0.243 134 624 6260	8.5×10^{-13}
$^{127}I_2$	32 – 0 $R(56)$, a_{10}	0.532 245 03614	7.0×10^{-11}
$^{127}I_2$	11 – 5, $R(127)$, a_{13} or (i)	0.632 991 398 22	2.5×10^{-11}
^{40}Ca	$^3P_1 - {}^1S_0$; $\Delta m_J = 0$	0.657 459 439 291 7	6.0×10^{-13}
$^{88}Sr^+$	$5^2S_{1/2} - 4^2D_{5/2}$	0.674 025 590 95	1.3×10^{-10}
CH_4	7 – 6, ν_3, P(7), $F_2^{(2)}$	3.392 231 397 327	3.0×10^{-12}

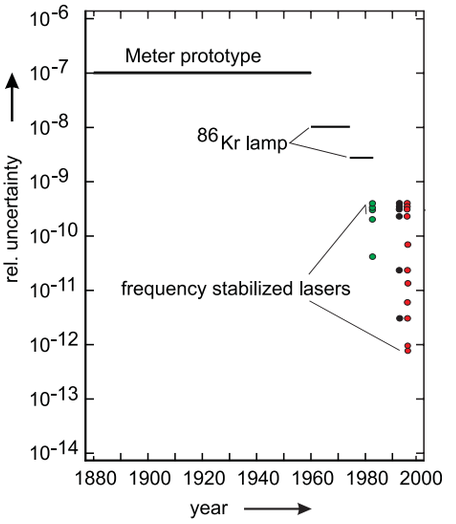


Fig. 11. Progress in the accuracy of the realization of the length unit during the past in the International System of Units (SI) according to the definitions of the CGPM [89,90] and the recommendations of the CIPM [91,92,86].

today. Consequently, this standard has been recommended recently by the CIPM [86] for the realization of the meter with the lowest uncertainty. The instability of a frequency measurement including the stabilities of the optical Ca standard and of the hydrogen maser was shown to decrease with $1/\sqrt{\tau}$ for integration times between $\tau = 1$ s and $\tau = 1000$ s. It reaches its flicker floor at approximately 10^{-14}. Frequency measurement chains used so far like the one of Fig. 10 are expected to be replaced by chains that use less oscillators. The progress of new chains based on optical frequency interval divider chains [84], optical frequency comb generators [87] or mode–locked lasers [88] is expected to allow the measurement of optical frequencies with less effort and may eventually lead to optical clocks competitive or superior to the best rf clocks.

6 Optical Frequency Standards for the Realization of the Meter

Optical frequency standards generate optical reference frequencies which are required in precision laser spectroscopy, optical communication and in length and frequency metrology. From the optical frequency standards developed at various frequencies in many laboratories the International Committee of Weights

and Measures (CIPM) has evaluated and selected several radiations of such stabilized lasers and recommends them as references for the realization of the meter and for precision spectroscopy (see Table 2, [91,92,86]). Their wavelengths presently range from the near Ultra-Violet (UV) (243 nm) to the Infra-Red (IR) (10.3 μm). The advantage of frequency standards for the realization of the meter is visualized from Fig.11. The pace of the development of optical frequency standards and clocks with reduced uncertainties does not seem to slow down as can be concluded from the fact that the uncertainties of the optical Sr [60] and the hydrogen 1S-2S frequency or of the Ca stabilized laser [74] already are considerably lower than given in Table 2.

Acknowledgements

The author is very grateful to his colleagues Drs. J. Helmcke, U. Sterr, H. Schnatz, A. Bauch, S. Weyers and Chr. Tamm for many fruitful discussions and the permission to reproduce their figures. The hospitality of the Heraeus Foundation is gratefully acknowledged.

References

1. Bureau International des Poids et Mesures (1998) Le système international d'unités (SI). Pavillon de Breteuil, F-92310 Sèvres, France, http://www.bipm.fr/pdf/si-brochure.pdf
2. Bureau International des Poids et Mesures (Ed.) (1967/1968) Comptes Rendus des séances de la 13^e CGPM. Pavillon de Breteuil, F-92310 Sévres, France. BIPM
3. Bureau International des Poids et Mesures (Ed.) (1983) Comptes Rendus des séances de la 17^e CGPM. Pavillon de Breteuil, F-92310 Sévres, France. BIPM
4. Guinot B. (1997) Application of general relativity to metrology. Metrologia. **34**, 261–290
5. Sydnor R.L., Allan D.W. (Eds.) (1997) Handbook Selection and Use of Precise Frequency and Time Systems. Radiocommunication Bureau of the International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland
6. Allan D.W. (1966) Statistics of atomic frequency standards. Proceedings of the IEEE. **54**, 221–230
7. Ramsey N.F. (1950) A molecular beam resonance method with separated oscillating fields. Phys. Rev. **78**, 695–699
8. Bagayev S.N., Chebotayev V.P., Dmitriyev A.K., Om A.E., Nekrasov Y.V., Skvortsov B.N. (1991) Second–order Doppler–free spectroscopy. Appl. Phys. **B 52**, 63–66
9. Phillips W.D. (1998) Laser cooling and trapping of neutral atoms. Rev. Mod. Phys. **70**, 721–741
10. Cohen–Tannoudji C.N. (1998) Manipulating atoms with photons. Rev. Mod. Phys. **70**, 707–719
11. Chu S. (1998) The manipulation of neutral particles. Rev. Mod. Phys. **70**, 685–706
12. Raab E.L., Prentiss M., Cable A., Chu S., Pritchard D.E. (1987) Trapping of neutral sodium atoms with radiation pressure. Phys. Rev. Lett. **59**, 2631–2634

13. Davidson N., Lee H., Kasevich M., Chu S. (1994) Raman cooling of atoms in two and three dimensions. Phys. Rev. Lett. **72**, 3158–3161
14. Wineland D.J., Drullinger R.E., Walls F.L. (1978) Radiation–pressure cooling of bound resonant absorbers. Phys. Rev. Lett. **40**, 1639–1642
15. Diedrich F., Bergquist J.C., Itano W.M., Wineland D.J. (1989) Laser cooling to the zero–point energy of motion. Phys. Rev. Lett. **62**, 403–406
16. Bauch A., Schröder R. (1997) Experimental verification of the shift of the Cesium hyperfine transition frequency due to blackbody radiation. Phys. Rev. Lett. **78**, 622–625
17. Gibble K., Chu S. (1993) Laser-cooled Cs frequency standard and a measurement of the frequency shift due to ultracold collisions. Phys. Rev. Lett. **70**, 1771–1774
18. Bauch A., Schnier D., Tamm C. (1996) Microwave spectroscopy of ^{171} Yb^{+} stored in a Paul trap. In: Bergquist J.C. (Ed.) *Proceedings of the Fifth Symposium on Frequency Standards and Metrology*, Singapore, New Jersey, London, HongKong. World Scientific, 387–388
19. Bize S., Sortais Y., Santos M.S., Mandache C., Clairon A., Salomon C. (1999) High–accuracy measurement of the ^{87}Rb ground–state hyperfine splitting in an atomic fountain. Europhys. Lett. **45**, 558–564
20. Fisk P.T.H., Sellars M.J., Lawn M.A., Coles C. (1997) Accurate measurement of the 12.6 GHz "clock" transition in trapped $^{171}Yb^{+}$ ions. IEEE Trans. Ultrason. Ferroelec. Freq. Contr. **44**, 344–354
21. Berkeland D.J., Miller J.D., Bergquist J.C., Itano W.M., Wineland D.J. (1998) Laser-cooled mercury ion frequency standard. Phys. Rev. Lett. **80**, 2089–2092
22. Essen L., Parry J.V.L. (1957) The Caesium resonator as a standard of frequency and time. Phil Trans Roy. Soc. **A 250**, 45–69
23. Ramsey N.F. (1990) Experiments with separated oscillatory fields and hydrogen masers. Rev. Mod. Phys. **66**, 541–552
24. Bauch A., Fischer B., Heindorff T., Schröder R. (1998) Performance of the PTB reconstructed primary clock CS1 and an estimate of its current uncertainty. Metrologia. **35**, 829–845
25. Ohshima S.I., Nakadan Y., Ikegami T., Koga Y., Drullinger R., Hollberg L. (1989) Characteristics of an optically pumped Cs frequency standard at the NRLM. IEEE Trans. Instrum. Meas. **IM 38**, 533 – 536
26. Rovera G.D., de Clercq E., Clairon A. (1994) An analysis of major frequency shifts in the LPTF optically pumped primary frequency standard. IEEE Trans. Ultrason. Ferroelec. Frequ. Contr. **41**, 457 – 461
27. Lee W.D., Shirley J.H., Lowe J.P., Drullinger R.E. (1995) The accuracy evaluation of NIST-7. IEEE Trans. Instrum. Meas. **IM 44**, 120 – 123
28. Lee W.D., Drullinger R.E., Shirley J.H., Nelson C., Jennings D.A., Mullen L.O., Walls F.L., Parker T.E., Hasegawa A., Fukuda K., Kotake N., Kajita M., Morikawa T. (1999) Accuracy evaluations and frequency comparisons of NIST-7 and CRL-01. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 62–65
29. de Clercq E., Makdissi A. (1996) Current status of the LPTF optically pumped Cs beam standard. In: Bergquist J.C. (Ed.) *Proceedings of the Fifth Symposium on Frequency Standards and Metrology*, Singapore, New Jersey, London, HongKong. World Scientific, 409–410
30. Ghezali S., Laurent P., Lea S., Clairon A. (1996) An experimental study of the spin–exchange frequency shift in a laser–cooled cesium fountain frequency standard. Europhys. Lett. **36**, 25–30

31. Kasevich M.A., Riis E., Chu S., DeVoe R.G. (1989) rf spectroscopy in an atomic fountain. Phys. Rev. Lett. **63**, 612–615
32. Clairon A., Salomon C., Guellati S., Phillips W.D. (1991) Ramsey resonance in a Zacharias fountain. Europhys. Lett. **16**, 165–170
33. Clairon A., Ghezali S., Santarelli G., Laurent P., Lea S.N., Bahoura M., Simon E., Weyers S., Szymaniec K. (1996) Preliminary accuracy evaluation of a cesium fountain frequency standard. In: Bergquist J. (Ed.) *Proceedings of the 5th Symposium on Frequency Standards and Metrology*, Singapore. World Scientific, 49–59
34. Sortais Y., Bize S., Nicolas C., Santos M., Mandache C., Santarelli G., Salomon C., Clairon A. (1999) An evaluation of the collisional frequency shift in a ^{87}Rb cold atom fountain. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 34–38
35. Fertig C., Legere R., Süptitz W., Gibble K. (1999) Laser–cooled Rb fountain clocks. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 39–42
36. Jefferts S.R., Mekkhof D.M., Shirley J.H., Parker T.E., Levi F. (1999) Preliminary accuracy evaluation of a cesium fountain primary frequency standard at NIST. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 12–15
37. Weyers S., Bauch A., Griebsch D., Hübner U., Schröder R., Tamm C. (1999) First results of PTB's atomic caesium fountain. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 16–19
38. Burt E., Swanson T., Ekstrom C. (1999) Cesium fountain development at USNO. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 20–23
39. Whibberley P.B., Henderson D., Lea S.N. (1999) Development of a caesium fountain primary frequency standard at the NPL. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 24–26
40. Huang M.S., Yao A., Peng J.L., Chen C.C., Hsu S., Hsiao J.M., Kou C., Liao C.S. (1999) Compact cesium atomic fountain clock. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 27–29
41. Liji W., Changhua W., Bingying H., Mingshou L., Jin Q., Wangxi J. (1999) Design & preliminary results of NIM cesium fountain primary frequency standard. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 30–33
42. Dudle G., Joyet A., Fretel E., Berthoud P., Thomann P. (1999) An alternative cold cesium frequency standard: The continuous fountain. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 77–80
43. Kasevich M., Chu S. (1992) Laser cooling below a photon recoil with three–level atoms. Phys. Rev. Lett. **69**, 1741–1744
44. Santarelli G., Laurent P., Lemonde P., Clairon A., Mann A.G., Chang S., Luiten A.N., Salomon C. (1999) Quantum projection noise in an atomic fountain: A high stability cesium frequency standard. Phys. Rev. Lett. **82**, 4619–4622
45. Simon E., Laurent P., Clairon A. (1998) Measurement of the Stark shift of the Cs hyperfine splitting in an atomic fountain. Phys. Rev. A. **57**, 436–439

46. Dicke R.H. (1953) The effect of collisions upon the Doppler width of spectral lines. Phys. Rev. **89**, 472–473
47. Dehmelt H.G. (1982) Mono–ion oscillator as potential ultimate laser frequency standard. IEEE Trans. Instrum. Meas. **IM-31**, 83–87
48. Drullinger R.E., Rolston S.L., Itano W.M. (1993) Primary atomic–frequency standards: New developments. In: Stone W.R. (Ed.) *Review of Radio Science 1993 - 1996*, Oxford, New York. Oxford University Press, 11–41
49. Blatt R., Gill P., Thompson R.C. (1992) Current perspectives on the physics of trapped ions. J. Mod. Opt. **39**, 193–220
50. Fisk P.T.H. (1997) Trapped–ion and trapped–atom microwave frequency standards. Rep. Prog. Phys. **60**, 761–817
51. Major F.G. (1998) The Quantum Beat. Springer-Verlag, New York, Berlin, Heidelberg
52. Paul W., Raether M. (1955) Das elektrische Massenfilter. Z. Phys. **140**, 262–273
53. Paul W. (1990) Electromagnetic traps for charged and neutral particles. Rev. Mod. Phys. **62**, 531–540
54. Tamm C., Schnier D., Bauch A. (1995) Radio–frequency laser double–resonance spectroscopy of ^{171}Yb ions and determination of line shifts of the ground–state hyperfine resonance. Appl. Phys. B. **60**, 19–29
55. Major F.G., Werth G. (1973) High-resolution magnetic hyperfine resonance in harmonically bound ground-state ^{199}Hg ions. Phys. Rev. Lett. **30**, 1155–1158
56. Cutler L.S., Giffard R.P., McGuire M.D. (1985) Thermalization of ^{199}Hg ion macromotion by a light background gas in an rf quadrupole trap. Appl. Phys. B. **36**, 137–142
57. Tjoelker R.L., Prestage J.D., Maleki L. (1996) Record frequency stability with mercury in a linear ion trap. In: Bergquist J.C. (Ed.) *Proceedings of the Fifth Symposium on Frequency Standards and Metrology*, volume 31, Singapore, New Jersey, London, HongKong. World Scientific, 33–38
58. Helmcke J., Morinaga A., Ishikawa J., Riehle F. (1989) Optical frequency standards. IEEE Trans. Instrum. Meas. **IM 38**, 524–532
59. Marmet L., Madej A.A., Siemsen K.J., Bernard J.E., Bradley G., Whitford B.G. (1997) Precision frequency measurement of the $^2S_{1/2}-^2D_{5/2}$ transition of Sr^+ with a 674-nm diode laser locked to an ultrastable cavity. IEEE Trans. Instrum. Meas. **IM 46**, 169–173
60. Bernard J.E., Madej A.A., Marmet L., Whitford B.G., Siemsen K.J., Cundy S. (1999) Cs– based frequency measurement of a single, trapped ion transition in the visible region of the spectrum. Phys. Rev. Lett. **82**, 3228–3231
61. Peik E., Hollemann G., Walther H. (1994) Laser cooling and quantum jumps of a single indium ion. Phys. Rev. A. **49**, 402–408
62. Nagourney W., Torgerson J., Dehmelt H. (1999) Optical frequency standard based upon single laser–cooled Indium ion. In: Dubin D.H.E., Schneider D. (Eds.) *Trapped charged particles and fundamental physics*, volume 457 of *AIP Conference Proceedings*, Woodbury, New York. American Institute of Physics, 343–347
63. von Zanthier J., Abel J., Becker T., Fries M., Peik E., Walther H., Holzwarth R., Reichert J., Udem T., Hänsch T., Nevsky A., Skvortsov M., Bagayev S. (1999) Absolute frequency measurement of the $^{115}In^+$ $5s^2\ ^1s_0$–$5s5p\ ^3p_0$ transition. Opt. Commun. **166**, 57–63
64. Rafac R.J., Young B.C., Cruz F.C., Beall J.A., Bergquist J.C., Itano W.M., Wineland D.J. (1999) $^{199}Hg^+$ optical frequency standard: Progress report. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 676–681

65. Taylor P., Roberts M., Gateva-Kostova S.V., Clarke R.B.M., Barwood G.P., Rowley W.R.C., Gill P. (1997) Investigation of the $^2S_{1/2}$–$^2D_{5/2}$ clock transition in a single ytterbium ion. Phys. Rev. A. **56**, 2699–2704
66. Tamm C., Engelke D., Bühner V. (2000) Spectroscopy of the electric-quadrupole transition $^2S_{1/2}$(F=0) - $^2D_{3/2}$(F=2) in trapped $^{171}Yb^+$. Phys. Rev. A. accepted for publication
67. Roberts M., Taylor P., Barwood G.P., Gill P., Klein H.A., Rowley W.R.C. (1997) Observation of an electric octupole transition in a single ion. Phys. Rev. Lett. **78**, 1876–1879
68. Ertmer W., Blatt R., Hall J.L. (1983) Some candidate atoms and ions for frequency standards research using laser radiative cooling techniques. In: Phillips W.D. (Ed.) *Laser Cooled and Trapped Atoms*. U.S. National Bureau of Standards special publication Vol. 653, Reading, Massachusetts, 154 – 161
69. Hall J.L., Zhu M., Buch P. (1989) Prospects for using laser–prepared atomic fountains for optical frequency standards applications. J. Opt. Soc. Am. B. **6**, 2194–2205
70. Schmidt-Kaler F., Leibfried D., Seel S., Zimmermann C., König W., Weitz M., Hänsch T.W. (1995) High–resolution spectroscopy of the $1S-2S$ transition of atomic hydrogen and deuterium. Phys. Rev. A. **51**, 2789–2800
71. Rolston S.L., Phillips W.D. (1991) Laser-cooled neutral atom frequency standards. Proceedings IEEE. **79**, 943–951
72. Walhout M., Sterr U., Witte A., Rolston S.L. (1995) Lifetime of the metastable 6s $[1/2]_0$ clock state in xenon. Opt. Lett. **20**, 1192–1194
73. Ruschewitz F., Peng J.L., Hinderthür H., Schaffrath N., Sengstock K., Ertmer W. (1998) Sub–kilohertz optical spectroscopy with a time domain atom interferometer. Phys. Rev. Lett. **80**, 3173–3176
74. Riehle F., Schnatz H., Lipphardt B., Zinner G., Trebst T., Helmcke J. (1999) The optical Calcium frequency standard. IEEE Trans. Instrum. Meas. **IM 48**, 613–617
75. Oates C.W., Bondu F., Fox R.W., Hollberg L. (1999) A diode–laser optical frequency standard based on laser–cooled Ca atoms: Sub–kilohertz spectroscopy by optical shelving detection. Eur. Phys. J. D. **7**, 449–460
76. Dinneen T.P., Vogel K.R., Arimondo E., Hall J.L., Gallagher A. (1999) Cold collisions of Sr^*-Sr in a magneto–optical trap. Phys. Rev. A. **59**, 1216–1222
77. Helmcke J., Snyder J.J., Morinaga A., Mensing F., Gläser M. (1987) New ultra–high resolution dye laser spectrometer utilizing a non–tunable reference resonator. Appl. Phys. B. **43**, 85–91
78. Vassiliev V., Velichansky V., Kersten P., Trebst T., Riehle F. (1998) Subkilohertz enhanced–power diode–laser spectrometer in the visible. Opt. Lett. **23**, 1229–1231
79. Baklanov Y.V., Dubetsky B.Y., Chebotayev V.P. (1976) Non–linear Ramsey resonances in the optical region. Appl. Phys. **9**, 171–173
80. Bordé C.J., Salomon C., Avrillier S., Van Lerberghe A., Bréant C., Bassi D., Scoles G. (1984) Optical Ramsey fringes with travelling waves. Phys. Rev. A. **30**, 1836–1848
81. Bordé C.J. (1989) Atomic interferometry with internal state labelling. Phys. Lett. A. **140**, 10–12
82. Riehle F., Schnatz H., Lipphardt B., Zinner G., Trebst T., Binnewies T., Wilpers G., Helmcke J. (1999) The optical Ca frequency standard. In: *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and The IEEE International Frequency Control Symposium*, 700–705
83. Schnatz H., Lipphardt B., Helmcke J., Riehle F., Zinner G. (1996) First phase–coherent frequency measurement of visible radiation. Phys. Rev. Lett. **76**, 18–21

84. Udem T., Huber A., Gross B., Reichert J., Prevedelli M., Weitz M., Hänsch T.W. (1997) Phase-coherent measurement of the hydrogen 1S-2S transition frequency with an optical frequency interval divider chain. Phys. Rev. Lett. **79**, 2646–2649
85. de Beauvoir B., Nez F., Julien L., Cagnac B., Biraben F., Touahri D., Hilico L., Acef O., Clairon A., Zondy J.J. (1997) Absolute frequency measurement of the $2S - 8S/D$ transitions in hydrogen and deuterium: New determination of the Rydberg constant. Phys. Rev. Lett. **78**, 440–443
86. Quinn T.J. (1999) Practical realization of the definition of the metre (1997). Metrologia. **36**, 211–244
87. Udem T., Reichert J., Hänsch T.W., Kourogi M. (1998) Accuracy of optical frequency comb generators and optical frequency interval divider chains. Opt. Lett. **23**, 1387–1389
88. Telle H.R., Steinmeyer G., Dunlop A.E., Stenger J., Sutter D.H., Keller U. (1999) Carrier-envelope offset phase control: A novel concept for absolute frequency measurement and ultra-short pulse generation. Appl. Phys. B. **69**, 327–332
89. L'Ecole Polytechnique, du Bureau des Longitudes (Ed.) (1890) Comptes Rendus des séances de la $1^{\text{è}re}$ CGPM 1889. Quai des Grands-Augustins, 55, France. Gauthier-Villars et Fils
90. Bureau International des Poids et Mesures (Ed.) (1960) Comptes Rendus des séances de la 11^{e} CGPM. Quai des Grands-Augustins, 55, France. Gauthier-Villars & C^{ie}
91. Editor's note (1984) Documents concerning the new definition of the metre. Metrologia. **19**, 163–177
92. Quinn T.J. (1993/94) Mise en pratique of the definition of the Metre (1992). Metrologia. **30**, 523–541

SpaceTime Mission: Clock Test of Relativity at Four Solar Radii

Lute Maleki and John Prestage

Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Drive, Pasadena, California 91109, USA

Abstract. SpaceTime is a mission concept developed to test the Equivalence Principle. The mission is based on a clock experiment that will search for a violation of the Equivalence Principle through the observation of a variation of the fine structure constant, α. A spatio–temporal variation of α is expected in some string theories aimed at unifying gravity with other forces in nature. SpaceTime uses a special tri–clock instrument on a spacecraft which approaches the sun to within four solar radii. The instrument consists of three trapped ion clocks based on mercury, cadmium, and ytterbium ions, in the same environment. This configuration allows for a differential measurement of the frequency of the clocks and the cancellation of perturbations common to the three. The observation of any frequency drift between each of the clocks, as the tri–clock instrument approaches the sun, signals the existence of a scalar partner to the tensor gravity. Some relevant details of the mission design are discussed in the paper.

1 Introduction

The unification of gravity with other forces of nature is arguably the most urgent problem in theoretical physics. Yet the unification program, initiated by Albert Einstein shortly after his introduction of general relativity, has proved to be a difficult challenge and remains an open problem today. The lack of a clear path to unification is even more frustrating since the two underlying theories, the quantum field theory and general relativity, have been separately immensely successful. Quantum field theory is widely regarded as the most successful theory in physics, capable of reproducing details of interaction for all matter. The predictions of this theory have been upheld by the experimental scrutiny in numerous tests. General relativity, on the other hand, has widened our picture of reality, and helped us consider the birth of the Universe, the cosmological evolution, and exotic objects such as the black hole. In the past eighty years since its introduction, this theory has also withstood the most exacting scrutiny that experimental physics has been able to devise [1]. Experimental tests and observations performed in vastly different conditions, ranging from the weak gravity of solar system, to the strong fields of binary pulsars, have all failed to reveal any violation of general relativity. In fact the failure of these very elaborate experiments with their impressive sensitivity in finding any violation has branded them with the label "Null Experiments", prompting the cynics to declare these tests as measuring zero with ever higher accuracy!

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 369–380, 2001.

Yet the impetus behind devising other, more sensitive, tests of general relativity is the unification, itself. Amongst promising candidates for the unification program, string theories stand out since they evidently have the ingredients necessary to accomplish the task. All viable string theories, nevertheless, require a massless scalar field partner to the tensor gravity, in contradiction to Einstein's Equivalence Principle (EEP), the pillar on which general relativity and other metric theories of gravity stand. This reality has led to a widespread belief among theorists that EEP must be violated, and experiments with requisite sensitivity can detect such violation. Thus the experimental tests of EEP now hold the promise of uncovering the range of validity of general relativity. And since some string theories make specific predictions regarding the nature and strength of the EEP violating scalar fields [2], sensitive tests of the EEP also stand to identify viable string theories.

That a violation of EEP will signal evidence for new physics extends beyond the promise of the unified fields; it directly confronts some of the most pressing questions in cosmology. There have been a number of recent suggestions that the so called "cosmological constant problem" may be linked to the presence of a scalar (quintessence) field which could be manifested by a violation of the Equivalence Principle [3].

Among various approaches to test general relativity, atomic clock tests are particularly significant since they represent a direct test of the coupling of gravity to the electromagnetic field. In a recent paper [4] we have extended the ability of atomic clocks in testing general relativity to a local test of the variation of the fine structure constant, α. A variation of the fine structure constant will signal a violation of the EEP [1]. On the other hand, a spatiotemporal variation of α, the coupling constant of the electromagnetic field, on the Hubble time scale is also implied by the dilaton and other scalar fields (moduli), which are the necessary partners of the tensor gravity field in string theories [5]. So a clock comparison test can fulfill the promise of aiding the unification program by pointing to the viability, or not, of string theories predicting a variation of fundamental coupling constants.

A clock test near the sun is particularly suitable to test the variation of α, since it can do so with a sensitivity far beyond that achievable on or near Earth. Finally, clock tests are also useful as a needed complement to "free fall" tests of general relativity. In the latter type of tests, the specifics of the coupling of gravity to any, or various, matter fields must be inferred from apportioning any observed difference in free fall of two, or more, test masses to various types of (mass) energy. Since the coupling of the dilaton field to the Coulomb energy dominates, when combined with clock tests (which are strictly based on atomic transitions driven only by the electromagnetic field) free fall tests will offer a vastly clearer picture of the details of the coupling of gravity to matter fields.

In the following section we will motivate a mission concept to fly a special tri–clock to within four solar radii of the sun in search for the observation of a varying α. This mission is referred to as SpaceTime. We will describe the tri–

clock instrument in section 3, and the mission design in section 4. The conclusions will be given in section 5.

2 Motivation

Atomic clocks are based on electronic transitions in neutral or ionized atoms which result in the emission or absorption of a photon. The frequency of the photon is fixed by the difference in the energy levels that the photon connects, and the inverse of the frequency determines a time interval. This simple relation provides the means for a sensitive test of any variability in the fine structure constant, owing to the electromagnetic nature of the energy in the electronic structure of the atom.

A change in α with time can be detected by comparing the frequency of two clocks based on dissimilar atoms, and through measuring the frequency drift between them. In the case of clocks based on hyperfine transitions, the α dependence of the hyperfine energy in each atom differs, depending on the atomic number, Z. This arises from the presence of a relativistic term in the hyperfine energy splitting, the so called Casimir correction factor $F(Z\alpha)$, which accounts for the interaction between the magnetic field of the nucleus and that of the valence electron in the finite nuclear volume. It has been shown [4] that a time variation of α may be related to the variation in the transition frequency of atoms 1 and 2, on which two clocks are based, via,

$$\frac{d}{dt}\ln\frac{A_1}{A_2} = [L_d F(\alpha Z_1) - L_d F(\alpha Z_2)]\frac{1}{\alpha}\frac{d\alpha}{dt}\,. \tag{1}$$

Here A_i $(i = 1, 2)$ is related to the hyperfine transition frequency, and $L_d F(Z_i)$ represents the sensitivity of the atom with a nuclear charge Z_i to the time variation of α.

The difference in the sensitivity of various hyperfine clocks to the time variation of the fine structure constant naturally points to experiments where clocks based on different atoms are compared for a time, and their relative frequency drift is measured to obtain $\dot{\alpha}$. Such a test was previously performed by comparing a mercury ion clock and a hydrogen maser for a period of about six months; it provided an upper limit for $\dot{\alpha}/\alpha$ to be 3.7×10^{-14}/yr.

There are two ways to improve the current limit on α variation set with clocks: Improved clock stability or longer measurement intervals for clock comparison. The first approach is possible, but based on the best available clock technologies, is limited to about one order of magnitude improvement. The second approach is difficult to implement, since the results improve with the observation time only linearly. Furthermore, the longer the observation time with separate clocks, the harder it is to exclude any environmentally caused frequency drift from spoiling the data.

An equivalent approach to testing for a time varying fine structure "constant" is a search for a spatial variation of α in a strong gravitational potential. The presence of any new long range force based on a scalar field will result in an

additional frequency shift in the clocks, the value of which will depend on the nuclear charge of the atoms of each clock. Since 98 % of the mass of the solar system resides in the sun, comparison of the frequency of clocks near the sun provides for the highest achievable sensitivity for such a test. At four solar radii, the sun's gravitational potential (divided by c^2, where c is the speed of light) is $5{\times}10^{-7}$, so clocks accurate to about 10^{-16} will enable measurement of differential frequency shifts at the level of 10^{-10} of general relativity's universal prediction.

This notion is the basis for the proposed SpaceTime mission, which attempts to combine the opportunity of the increased sensitivity at the proximity of Sun with a unique "tri–clock", to provide for an unambiguous test of α variation. The sensitivity of such an experiment may be compared with any test of time variation in α in the following manner. The variation of α with the variation of the gravitational potential may be written as:

$$\frac{\delta\alpha}{\alpha} = S\,\frac{\delta U}{c^2}\,, \tag{2}$$

where S is the sensitivity of the experiment (10^{-10} in the scenario suggested above) and U is the Newtonian potential. Taking the time derivative of the above equation leads to:

$$\frac{\dot{\alpha}}{\alpha} = S\,\frac{\dot{U}}{c^2}\,. \tag{3}$$

The time variation of U, for a closed universe, is related to the Hubble constant H by $\dot{U}/c^2 \simeq H$. Thus the equivalent sensitivity of SpaceTime to the detection of any time variation in α is:

$$\frac{\dot{\alpha}}{\alpha} = S \times H \simeq 10^{-20}/\text{year} \tag{4}$$

for H less than 10^{-10}.

This sensitivity should be contrasted to the limit of $\dot{\alpha}/\alpha \leq 10^{-16}$ placed by comparing the ratio of isotopes 147 and 149 of Sumarium in the natural uranium fission reactor in the Oklo mine in Gabon [6]. It also exceeds the recent results of the observational astronomy with the tantalizing hint of a possible α variation, [7] by four orders of magnitude.

3 The Tri–clock Instrument

The tri–clock instrument design is key to the sensitivity of SpaceTime to an α variation, and is based on the trapped ion frequency standards of the JPL. The interest in ion traps as the basis for atomic clocks stems from the general requirements for achieving high performance. The achievable stability of atomic clocks depends on the energy level structure of the active atom, as well as the environment in which the atom undergoes the clock transition. The first condition determines the achievable stability inherent to the atom via the sensitivity of the atomic transition to the environmental perturbations. The second determines the degree to which the environmental perturbations can be reduced with

respect to the atomic energy level sensitivity. The trapped ion frequency standard fulfills both these conditions, since ions with inherently high insensitivity to environmental changes are readily confined in traps which provide an essentially benign environment. The relatively large hyperfine level separation of ions, compared to neutral atoms, make them inherently less sensitive to environmental perturbations, such as the ambient magnetic fields

Research in the past two decades in the field of trapped ion clocks has already produced impressive results with the development of a standard with stability corresponding to $7 \times 10^{-14}/\sqrt{\tau}$ [8], and has demonstrated the highest measured stability for averaging intervals longer than 10^5 s. A brief description of the operation of the trapped ion frequency standard is given below, for completeness. The interested reader is referred to the literature [9–11] for more information.

In trapped ion clocks, ions of the atomic species of interest are confined in an electrodynamic trap. The trap may be constructed of electrodes of various geometry. In the most widely used trap type, four rods arranged in a cylindrical geometry form the electrodes [9]. A radio frequency (rf) voltage applied to the trap keeps the neighboring electrodes at opposite potentials at any instant of time. Ions are created in the trap by the collision of electrons generated from a hot filament on a background vapor of the atom of interest. Once generated within the trap, ions experience the quadrupolar rf potential created by the electrodes. This quadrupolar potential is zero along the axis of the trap, and increases as ions get closer to the electrodes. Ions experience a force resulting from the interaction with the field which varies in direction as the field oscillates. If the trap rf frequency is high enough, it results in a net force directed towards the trap axis (the zero field position) at all times, and ions are thus trapped.

Once trapped, the population of the two hyperfine levels of the ion are nearly the same at room temperature, as required by the Boltzman distribution. This implies that ions must be prepared in the desired state, from which the clock transition takes place. The state preparation is accomplished by optical pumping, whereby ground state ions are first pumped from one of the ground state hyperfine sublevels (typically the highest lying in energy) to the first electronic excited state, with light from a resonance lamp or a laser. Upon their decay, excited atoms make a transition to all hyperfine sublevels of the ground state. Since the lifetime of the excited state is short (only a few nanoseconds), the applied light depletes the upper hyperfine sublevel, while increasing the population of the lower sublevels. When the optical pumping is complete, and the upper hyperfine sublevel is completely empty, the atom is prepared for the clock transition which can be induced from one of the lower sublevels with population, to an empty sublevel.

The next step in the operation of the clock entails inducing an interaction between the atoms, and the microwave radiation from a local oscillator (LO) source at the clock frequency. It is this cycle in the clock operation which requires that the ions be shielded from any interaction with the environment. Since hyperfine transitions are due to magnetic dipole interactions between the electron and the applied field, shielding from external magnetic fields is of paramount importance.

For the same reason, it is practically more advantageous to apply a small, yet fixed, magnetic field to the trap, while shielding it from the outside magnetic field variations.

If the frequency of the LO exactly matches the clock transition, the depleted upper hyperfine sublevel becomes populated again, and can be used to scatter photons from the resonance lamp again. Otherwise, the clock transition does not populate the upper sublevel, and the atoms remain transparent to the light from the resonance lamp. This information (scattered photons from the ion) is used in a feedback loop that controls the frequency of the LO to ensure its coincidence with the clock transition in the ion. The output of the LO, disciplined to the atomic transition, constitutes the clock output.

There is one more important point about the LO which should be considered here. While the output of the LO is disciplined to the atomic transition, the LO is free running (i.e. without feedback control) for the time intervals corresponding to ion generation, optical pumping, clock transition, and clock interrogation cycles. Most of these cycles are short, on the order of a second long, but the microwave transition must be long, since the inverse of its length determines the width of the atomic line. In high performance clocks, the atomic line is as narrow as practical since it directly determines the achievable stability. This length of the microwave transition cycle typically ranges from about ten seconds to as long as a minute. A major obstacle in the realization of ultra–high performance atomic clocks, as required in precision measurements, is the availability of a LO whose noise during the free running interval is low. It has been shown [12] that the noise of the LO during the free running cycle can limit the long term stability of the atomic clocks based on cyclic feedback.

For the SpaceTime mission we have modified the scheme outlined above to achieve the best performance. Here by "best performance" we mean one that can satisfy the following requirements: 1) a reliable, power efficient, compact, and low cost instrument, compatible with the space environment; 2) the requisite stability of a part in 10^{-16} in about 70000 s; and 3) the capability to produce unambiguous results.

At the core of our design is a modified architecture of the ion trap, to include two segments (Fig.1) [13].

In this configuration, the generation of ions, their optical pumping, and the final interrogation for the clock transition takes place in the first trap segment. The interaction of ions with the applied microwave signal from the LO occurs in the second segment. Ions, being charged particles, can be gently "shuttled" back an forth between the two segments by applying a small dc potential for to the "end cap" separating the two segments. Thus only the second trap segment, where ions interact with the microwave signal, needs to be isolated from the environmental perturbations. The major advantage of this architecture is that the requirement of the environmental shielding is significantly reduced, leading to considerable reduction in size and mass. This ability is crucial to the realization of any spaceborne clock.

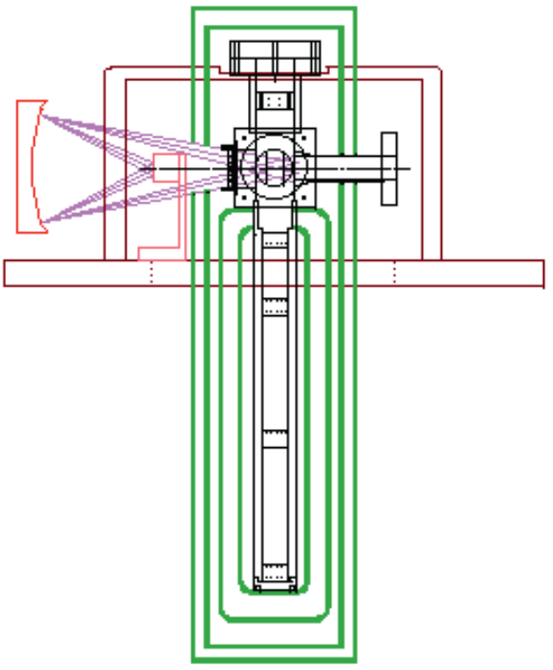


Fig. 1. Line drawing of the Shuttle Trap.

The second important modification is a geometry whereby three of these traps share the same vacuum chamber, applied potentials, and magnetic and thermal environments, as shown in Fig.2.

Ions of three different atomic species are confined in different traps, forming the basis for three separate clocks. The cycles of ion generation, optical pumping, clock transition, and interrogation are common to all ions in the three traps. Each trap confining potential, though, is set to trap only one of the three species. The three potentials are nevertheless derived from the same source, and reduced to the required value. The hyperfine transition in each of the ion species is driven at the same cycle. While the transition frequency for each of the three species is different, they are derived from a single local oscillator. Three different resonant lamps optically pump, and later interrogate each of the ion species. This arrangement implies that most environmental perturbations will be common to ions confined in any of the three traps. Thus environmentally induced drifts can be caused to be cancelled. This is another crucial aspect of the SpaceTime clock experiment. Comparison of three independent clocks would leave the task of unraveling the influence of the environment's perturbations on each clock, from a "true" signal to modeling with associated uncertainty.

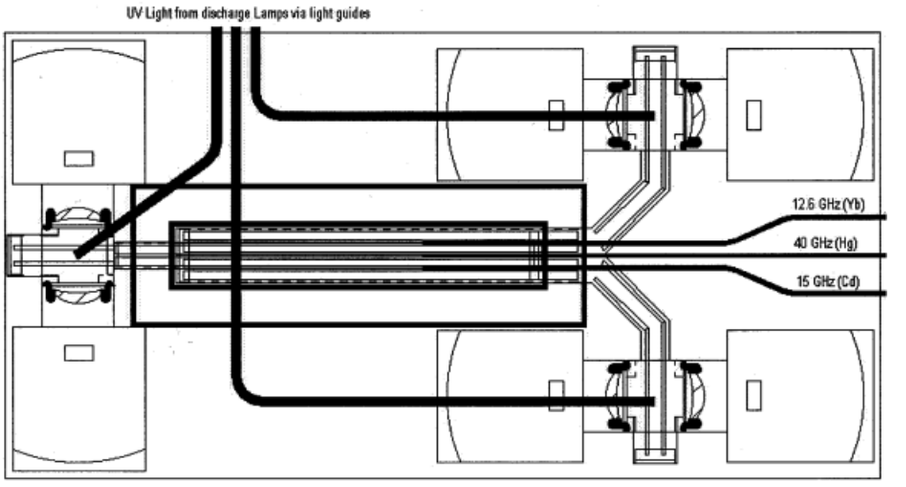


Fig. 2. Layout of the Tri-clock instrument.

For the SpaceTime mission, the specific design of the tri–clock calls for using mercury, cadmium, and ytterbium ions in the three traps. The hyperfine energy levels of each ions, at 40.5, 14, and 12.5 GHz are shown in Fig.2. Each of these ions has been previously used in the realization of an atomic clock [11]. In the tri–clock instrument the optical pumping and the clock interrogation cycles use resonant lamps. By manufacturing lamps which include all three species of appropriate isotope, we plan to realize redundancy, another important consideration in designing space instruments.

We plan to use the redundancy in the three clocks to measure, and further eliminate, any environmental perturbations with different signature on different ion species. In particular, the value of the any ambient magnetic field may be obtained by using one of the three ions as the "probe" to measure the field, which can then be used to cancel out its effect on the remaining two clocks.

A major advantage realized with the tri–clock architecture is the use of a single local oscillator, from which all three hyperfine signals can be derived. This approach will ensure that the noise of the local oscillator during its free running interval is a common noise to all three clocks, and thus can be cancelled out to a great extent. The extent of the cancellation is in the ratio of the frequencies, with the largest noise on the 40.5 GHz (mercury transition), and the smallest for the 12.5 GHz (ytterbium transition.) This cancellation is large enough that a high stability quartz oscillator would suffice as an LO to achieve 10^{-16} stability at 7000 s, as required for SpaceTime.

The tri–clock architecture will thus allow the major sources of noise and perturbation to become common to all three clocks. This enables the cancellation of all the common noise when the three clocks are inter–compared, to a level that will allow the requisite sensitivity for finding a varying α. This feature, together

with the specific dependence of a varying α on the nuclear charge (Z) of each ion species, will produce an unambiguous signal as the three clocks are inter–compared. Finally, the SpaceTime experiment based on a differential clock shift measurement with the tri–clock instrument eliminates the need for comparing a spaceborne clock with a ground based clock. This latter approach would greatly complicate the mission by need for a stable, high data rate communication link, and associated need for the modeling of the media through which it propagates. The task of modeling alone, owing to the proximity to Sun, is difficult and subject to uncertainties. In short, the design of the tri–clock instrument provides for simplification of the mission, and reduction of weight, all of which significantly help the cost of the experiment. These benefits of the tri–clock complement the expected high performance and ensure that its signal are unambiguously interpreted.

4 Mission Design

The SpaceTime is designed to fit the NASA Mid–sized Explorer mission type (MIDEX). This mission type has specific limits set for the total cost, and the schedule. The goal of the SpaceTime mission design is to deliver the clock payload to the close vicinity of the Sun while minimizing complexity and risk. SpaceTime launches from Cape Canaveral Air Station (CCAS) on a Delta 7925 launch vehicle. This proven vehicle simplifies all aspects of the mission and allows us to reduce risk and cost, and return breakthrough science in the shortest time possible. After the first two stages burn, the third stage spins up to 60 rpm and injects the spacecraft towards Jupiter. The target at Jupiter is to within 8.68 Jupiter radii, which will lead to a gravity assist to reduce the spacecraft energy, allowing it to fall toward the Sun with a 4 solar radii pass. The SpaceTime mission plans for two trajectory correction maneuvers (TCMs) to achieve its trajectory. The first maneuver will occur approximately 10 days after launch, and the second about one year after launch. Since SpaceTime differential redshift measurements allow a 10% error in the solar flyby radius, we target to $4.2\,R_s \pm 0.2\,R_s$. This allows tolerable large navigation uncertainties at the Jupiter flyby.

After the Jupiter encounter, the spacecraft falls rapidly toward the Sun, and will take only about 9 hours to travel from $10\,R_s$ to $4\,R_s$. This trajectory greatly simplified navigation, and the tracking scheme allows for a Doppler track once every 1-3 months. This in turn simplifies mission operations and reduces cost at no increase to mission risk.

4.1 Spacecraft

The SpaceTime spacecraft design is simple and reliable. This design leads to reduced cost and risk in mission operations and allows the science data return at its critical point near the Sun. The SpaceTime spacecraft utilizes extensive heritage from past spacecraft, including Mars Pathfinder, Cassini, DS-1, Lewis, MSP '98, and Defense Systems programs. Comprehensive analysis has produced

the best small spacecraft design for the SpaceTime mission within the budget constraints. The primary purpose of the spacecraft is to take the SpaceTime clocks as close as possible to the Sun and to allow the acquisition and transmission of data from the clocks as the spacecraft nears the Sun. Figs. 3 and 4 show a spacecraft close–up and a near–solar pass configuration. The spacecraft integrates a blowdown mono–propellant propulsion system and the necessary spacecraft electronics to support the mission. The spacecraft electronics and the clock are all within a single thermal enclosure.

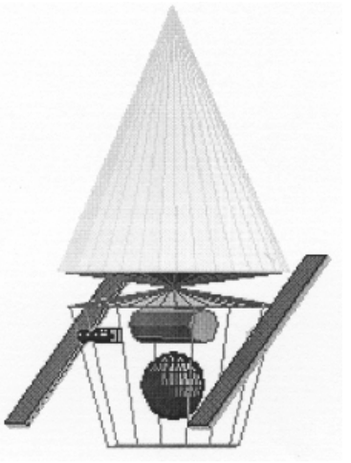


Fig. 3. The spacecraft.

The large, conical solar shield protects the entire spacecraft during the 4 solar radii encounter. The thermal control system of the SpaceTime spacecraft will utilize this thermal shield system that will isolate its flight elements from the solar flux of a $4\,R_s$ (from the solar center) close passage. This shield system is based on a new technology and consists of a solar blocking element (primary shield), IR shields, High Temperature MLI, and support structure. The system is a simpler version of the Solar Probe shield system. The first element of this shield will operate at about 2000 K during close solar approach, but the shield system will provide thermal isolation that will allow the spacecraft elements behind it to operate at nominal temperatures (0° to +40° C).

The spacecraft thermal control system components behind the shield system utilizes flight–proven elements such as thermal conduction isolation, thermal surfaces (paints, films), MLI, louvers, and electric heaters/thermostats. This design will control spacecraft element temperatures for all flight regimes from launch to Jupiter flyby, and close solar approach.

The SpaceTime spacecraft current best estimate dry mass is 114 kg, including the 20-kg science payload, with contingencies of 30% in both mass and power. The fuel load is about 30 kg for a launch mass of around 185 kg, the maximum allowable mass that can be launched on a Delta 7925 with Star 30c kick stage to a C3 of $120\,\mathrm{km}^2/\mathrm{sec}^2$. The attitude control subsystem (ACS) subsystem provides attitude knowledge and control, and solar array articulation and control. The Medium Gain Antenna (MGA) is body–mounted, perpendicular to the spacecraft centerline. Twin star trackers and an Inertial Measurement Unit (IMU) provide primary attitude determination, with four analog Sun sensors serving in a functional redundancy role as backup pointing devices. During the Jovian gravity assist, the ACS system will operate in an all–stellar mode so that power can be conserved by turning off the IMU.

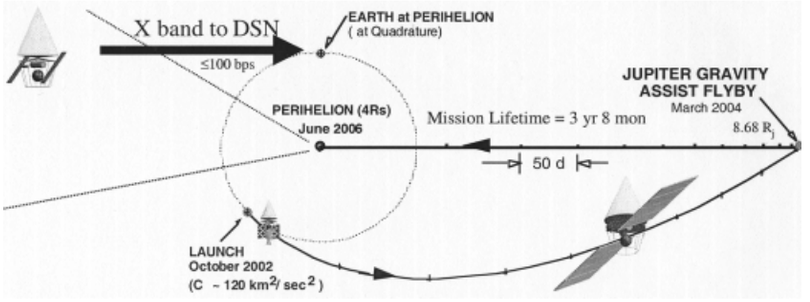


Fig. 4. picture trajectory

The Command and Data Handling (C&DH) subsystem provides instrument science data storage, spacecraft commanding, interfaces to the X–Band transponder, ACS sensors and actuators, and spacecraft timing.

The communications subsystem supports X–band command uplink and data downlink using the Small Deep Space Transponder, a 3-watt Solid State Power Amplifier, one 0.6-meter diameter MGA, and dual low–gain patch antenna (LGA). The MGA provides 100 bps data transmission near Jupiter and 100 bps during the solar flyby. The LGAs provide emergency command capability.

The power subsystem provides power collection, storage, control, and distribution. The Electrical Power Subsystem (EPS) is comprised of the solar array, battery, charge control unit (CCU), power distribution and drive unit (PDDU), and pyro initiator unit (PIU). Because of the profile of this mission, the SpaceTime spacecraft carries two sets of featherable solar arrays, a small secondary battery and a high–energy primary battery for the critical final phase of the mission when the solar arrays can no longer operate. The low temperature solar panels are populated with $6.7\,\mathrm{m}^2$ of silicon solar cells, providing 1500 W average power near Earth. The low temperature solar arrays are sized for stay–alive power near Jupiter. The low temperature array is jettisoned after crossing Mercury's orbit at 0.3 AU. The high temperature portion of the solar array consists of 0.15 m2 of Gallium Indium Phosphide cells at 6% efficiency and provides power after crossing 0.3 AU. These arrays will provide up to 200W of power as close as 0.1 to 0.15 AU when feathered to a 70-degree angle of incidence. This will allow the SpaceTime clocks to be calibrated prior to the final phase of the experiment. The high temperature arrays will also be jettisoned to prevent an imbalance in control forces near the Sun. The High–Energy Density (HED) primary battery, a 300 A-hour Lithium Carbon Fluoride (LiCF) package with a long shelf life, provides the 3730 W-hr necessary during the 33-hour solar flyby after the solar arrays are jettisoned at about 0.13 AU.

5 Conclusions

In this article we have described a new mission design for a sensitive clock test of Einstein's Equivalence Principle. The mission includes a payload consisting of

a unique tri–clock instrument, based on cadmium, ytterbium, and mercury ions. The instrument will fly to within four solar radii of the Sun, where a differential drift between the three clocks will signal a violation of the Equivalence Principle. The instrument is designed to eliminate or minimize all ambient noise sources, and with the designed stability of 1×10^{16}, enables an unambiguous test at a part in 10^{13} level. This same measurement, interpreted as a search for a variation of the fine structure constant *alpha*, will test the validity of string based theories that specify a scalar field partner, such as the dilaton, to the tensor gravity. An important feature of this mission is that it is based on proven, existing technology, and can be flown within the cost and schedule constraints of NASA's MIDEX (Medium-sized Explorer) program.

Acknowledgements

The mission concept SpaceTime is the result of a concept study performed at the Jet Propulsion Laboratory by a team. Members of the science team in the SpaceTime proposal are John Anderson, John Armstrong, Thibault Damour, Lute Maleki (PI), Ken Nordvedt, John Prestage, Michael Soffel, and Robert Vessot. The mission design team included Freda Cheung, Elaine Hansen, Rob Maddock, Steve Matousek (Proposal Manager), Bob Miyake, Bill Moore, George Sprague, Ellie Trevarthen, Jay Wyatt and numerous others.

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. C.M. Will: *Theory and Experiment in Gravitational Physics* (Cambridge University press, Cambridge 1981; Revised edition 1993).
2. T. Damour: in L. Maleki (Ed.) *Proceedings of the Workshop on the Scientific Applications of Clocks in Space* (JPL Publication 97-15, 1996), p. 13.
3. Y. Fujii: *Prog. Theor. Phys.* **99**, 599 (1988).
4. J.D. Prestage, R.L. Tjoelker, and L. Maleki: *Phys. Rev. Lett.* **74**, 3511 (1995).
5. T. Damour and A. M. Polyakov: *Nuc. Phys.* **B 423**, 542 (1994).
6. T. Damour and F. Dyson: *Nucl. Phys.* **B 480**, 37 (1996).
7. J.K. Webb, V.V. Flambaum, C.W. Churchill, M.J. Drinkwater, and J.D. Barrow: *Phys. Rev. Lett.* **82**, 888 (1999).
8. R.L. Tjoelker *et al.*: in J. Vig (ed.): *Proceedings of 1996 IEEE International Frequency Control Symposium, 1996*, p. 1073.
9. J.D. Prestage, R.L. Tjoelker, G.J. Dick, and L. Maleki: *J. Mod. Optics* **39**, 232 (1992).
10. L. Maleki: *Proc. Europ. Freq. and Time Forum*, Beconson, France, March 1995.
11. P.T.H. Fisk: *Rep. Prog. Phys.* **60**, 761 (1997).
12. G.J. Dick and C.A. Greenhall: *Proc. 1998 IEEE International Freq. Control Symp.*, 99 (1998).
13. J.D. Prestage, R.L. Tjoelker, G.J. Dick, and L. Maleki: *Proc. 1993 IEEE International Frequ. Control Symp.* (Salt Lake City, USA 1993), p. 144.

Pulsar Timing – Strong Gravity Clock Experiments

Norbert Wex

Max–Planck–Institut für Radioastronomie
Auf dem Hügel 69
53121 Bonn, Germany

Abstract. 25 years ago, in summer 1974, Joseph Taylor and Russell Hulse discovered the first binary pulsar, a pulsar in orbit with a compact companion which itself is most likely a neutron star. This pulsar, denoted PSR B1913+16, turned out to be the most exciting laboratory for testing relativistic gravity theories. Before the discovery of PSR B1913+16 all gravity experiments were confined to our solar system with its very weak gravitational fields. Hence, it has been possible to test gravity theories only in the first post–Newtonian approximation. Binary pulsars take us beyond the weak–field context because of their high orbital velocity and/or the strong self–gravitational fields of neutron stars.

To date, more than 70 binary pulsars have been discovered, most of them in orbit with a neutron star or a white dwarf. Many binary pulsars belong to a group of so–called millisecond pulsars which have very short rotational periods (< 20 ms) and slowdown rates of typically 10^{-20}, proving to be extremely accurate clocks. This gave rise to a variety of new gravity experiments, like tests for the strong equivalence principle.

After a brief introduction to pulsars, the technical and theoretical aspects of binary–pulsar gravity experiments are reviewed. The latest results are presented and an outlook is given to future improvements of these experiments.

1 The Pulsar Population

Pulsars were discovered in 1967 by Antony Hewish and Jocelyn Bell at Cambridge University. During a study of rapid fluctuations of extragalactic radio sources caused by the interstellar medium they came across point like sources which were emitting radio signals with extraordinary regularity [18]. Very soon after their discovery, pulsars were understood to be highly magnetized rapidly rotating neutron stars in our Galaxy, becoming apparent as 'cosmic lighthouses' with excellent rotational stability. Over the past 30 years, systematic surveys with the world's largest radio telescopes have revealed more than 1200 pulsars. About 500 of them were found only recently in the on–going Parkes–multibeam Galactic–plane survey [9] which proved to be extraordinarily successful and lead to the discovery of a number of exciting single and binary pulsars. To date, rotational periods of pulsars, P, span nearly 4 orders of magnitude ranging from 1.56 ms to 8.51 s. The observed period changes, $\dot{P}$, explained by the loss of rotational energy due to magnetic dipole radiation and/or relativistic particle winds, range from 10^{-21} to a few times 10^{-11}, corresponding to magnetic surface fields of about 10^8 to 10^{14} Gauss, respectively. The majority of pulsars are clustered

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 381–399, 2001.

around a P of one second and a $\dot{P}$ of 10^{-15} (see Fig. 1). It is believed that these pulsars were born in supernova explosions with rotational periods of a few tens of milliseconds and then spun down into the region of 'normal' pulsars within 10^6 to 10^7 years.

There is a group of pulsars whose characteristic age ($\equiv P/2\dot{P}$) typically exceeds 10^9 years and whose magnetic surface fields lie below 10^{10} Gauss. Most of them are found to be members of binary systems and are located to the bottom left of the large group of 'normal' pulsars in Fig. 1. It is understood that these pulsars were born as 'normal' pulsars in binary systems. Then, after 10^8 to 10^9 years the rotation became too slow and the pulsar–emission mechanism ceased to work. In the meantime the companion star evolved, started to fill its Roche lobe and transfer matter onto the neutron–star. The infall of matter spun up the neutron star and suppressed the magnetic field by several orders of magnitude, leaving a fast rotating 'recycled' pulsar with a weak magnetic field behind. Depending on the initial mass, the companion star will either form a second neutron star in a supernova explosion and create, in cases where the supernova explosion does not unbind the binary, an eccentric double–neutron–star system (high mass progenitor), or end up as a white dwarf in a very circular orbit about a pulsar with rotational periods of, typically, a few milliseconds (low mass progenitor).

About 5% of all pulsars known to date have been identified as *binary pulsars*, i.e. pulsars which are members of binary systems. There are two binary pulsars which are in orbit with a main–sequence B–star [21,24] and there is a third pulsar suspected to orbit a K–supergiant star [32]. The rest of the binary–pulsar population is believed to have a compact star as a companion (white dwarf or neutron star). Since in most cases the orbital separation is much larger than the size of the compact companion, neither mass–transfer between the two components of the binary system is present, nor tidal effects are having any significant influence on their orbital dynamics ('clean' binary system). Therefore, the orbital dynamics of most binary–pulsar systems should follow a simple 'point–particle' model. Presently six binary pulsars are thought to have a neutron–star as their companion. The range of their orbital periods is from four hours to more than two weeks and in all six cases the orbital eccentricity is high (see Table 1 for details).

Table 1. Binary pulsars with neutron–star companions.

PSR	P (ms)	P_b (days)	e	discovery	comments	Ref.
J1141−6545	394	0.20	0.17	1999	maybe white–dwarf companion	[32]
J1518+4904	40.9	8.63	0.25	1994		[36]
B1534+12	37.9	0.42	0.27	1990		[59]
J1811−1736	104	18.8	0.83	1997		[30]
B1913+16	59.0	0.32	0.62	1974	Hulse–Taylor pulsar	[19]
B2127+11C	30.5	0.33	0.68	1988	in globular cluster M15	[1]

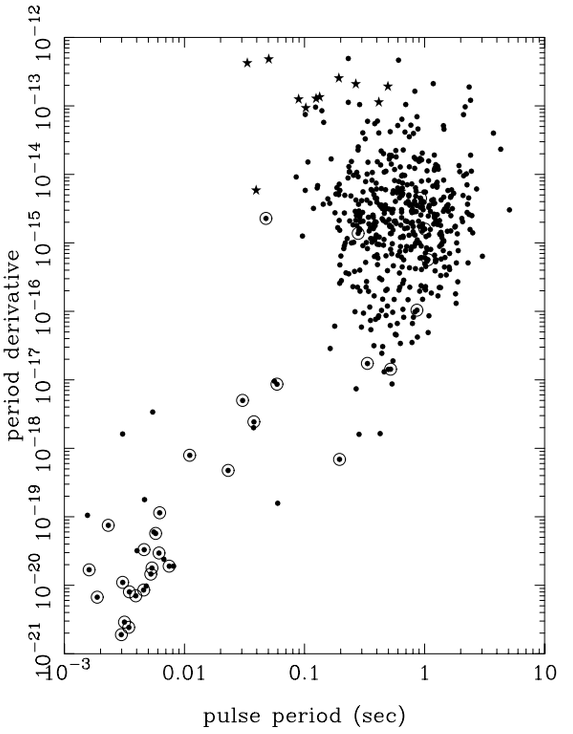


Fig. 1. Plot of period derivative, $\dot{P}$, versus pulse period, P, for pulsars taken from [47,48]. Pulsars that are members of a binary systems are marked with a circle and pulsars associated with a supernova remnant are marked with a star. For an updated version of this P-$\dot{P}$-digram see [32].

While the binary pulsar PSR B2303+49 was suspected to be a member of a double–neutron–star system, more recent optical observations using the 10–m Keck telescope indicate a massive white dwarf as the companion to PSR B2303+49 [52]. Also PSR J1141−6545 could, instead of a neutron star, have a massive white dwarf as a companion as pointed out by Tauris and Sennels [44]. In fact, most binary pulsars are in orbit with a white dwarf companion. While the orbital periods range from 1.6 hours to 3.3 years the orbital eccentricities for these systems are generally very small, in some cases even less than 10^{-5}. All of these systems suffered a phase of mass transfer in the past which was very effective in circularizing the binary orbit.

For a general review on pulsars see [31,29].

2 Pulsar Timing

Pulsar timing has become an important subfield of observational astronomy and proved to provide a powerful tool for many areas of physical and astrophysical

research. In particular, the high–precision timing of millisecond pulsar ($P < 20$ ms) has a wide range of applications:

- Time–keeping metrology
- Planetary ephemerides
- High–precision astrometry (position and proper motion of pulsars)
- Probing the cosmological gravitational–wave background
- Tests of relativistic theories of gravity
- Properties of the interstellar medium
- Evolution of binary–star systems

See [23,5] for a review on pulsar timing applications.

A complete discussion of the observing procedures and instrumentation used in pulsar–timing experiments is beyond the scope of this article. We therefore only summarize the basic ideas. For a more detailed description the reader is referred to review articles, like [3]. Due to the distribution of free electrons in the interstellar medium, pulses at higher radio frequencies propagate at a higher group velocity and, therefore, arrive earlier at the telescope than pulses at lower frequencies. To correct for this propagation effect across the bandwidth of the receiver one uses either filterbanks to divide the observing bandpass into small channels or a 'coherent dedispersion' system, which samples the raw telescope voltages and then removes the effects of dispersion using a software filter. In practice the second method yields a clearly better timing precision.

Most of the pulsars are weak radio sources and therefore an addition of many thousands of pulses is required in order to produce a pulse profile with a good signal–to–noise ratio. Apart from this, a high number of individual pulses has to be added, even for strong pulsars, to produce a stable integrated profile, since individual pulses vary quite dramatically (see Fig. 2). Therefore the pulse signal is folded over a time span of typically several minutes at the predicted topocentric pulse period in order to produce a total–intensity profile. Cross–correlation with a standard profile with high signal–to–noise ratio, obtained during previous observations, determines the observed rotational phase of the pulsar ϕ as a function of the telescope time τ or, equivalently, the time of arrival (TOA) of a chosen point of reference in the pulse profile. For a pulse profile with good signal–to–noise ratio the location of the point of reference with respect to the time stamp can be determined with high precision, typically $10^{-4} \times P$.

The TOA, which is measured using the time τ of the atomic clock stationed at the radio–telescope site, has to be connected with the time T of the pulse emission as measured in the co–moving frame of the pulsar. Thereby, one has to correct for a number of propagation and time–dilation effects within the solar system and, in case of binary pulsars, for propagation and time–dilation effects in the binary system. This transformation can be written as

$$\begin{aligned} T = \tau - \tau_0 + \Delta_C - D/f^2 + \Delta_{\mathrm{Roemer},\odot} + \Delta_{\mathrm{Shapiro},\odot} + \Delta_{\mathrm{Einstein},\odot} \\ + \Delta_{\mathrm{Roemer},b} + \Delta_{\mathrm{Shapiro},b} + \Delta_{\mathrm{Einstein},b} \ , \end{aligned} \tag{1}$$

where τ_0 denotes a reference epoch and Δ_C an offset between the observatory–master clock and the reference standard of terrestrial time which is determined

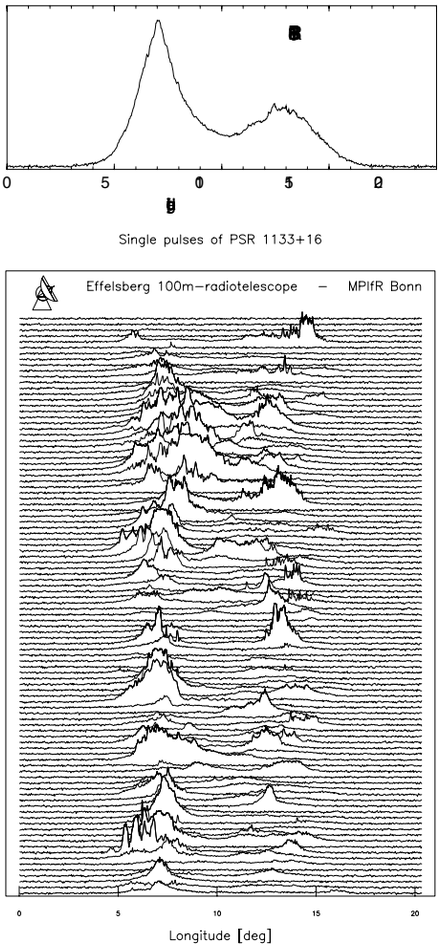


Fig. 2. A sequence of 100 individual pulses of the strong pulsar PSR B1133+16. The sum of all the individual pulses forms the characteristic integrated pulse profile as shown in the box of the top.

using the Global Positioning System (GPS). The dispersive delay equals to D/f^2, where D is proportional to the column density of free electrons between pulsar and observer and f is the observing radio frequency. The parameter D and possible long–term variations of D can be determined by using observations at different radio–frequencies. In Effelsberg, for instance, we use 1.4 GHz as our prime timing frequency and in addition we make use of 0.8 and 2.7 GHz observations [27]. $\Delta_{\mathrm{Roemer},\odot}$, $\Delta_{\mathrm{Shapiro},\odot}$, and $\Delta_{\mathrm{Einstein},\odot}$ account for propagation delays and time dilation effects within the solar system. If the pulsar is a member of a binary system similar terms accounting for these effects within the binary system are needed (index b in equation (1)). Full details on all of the terms in

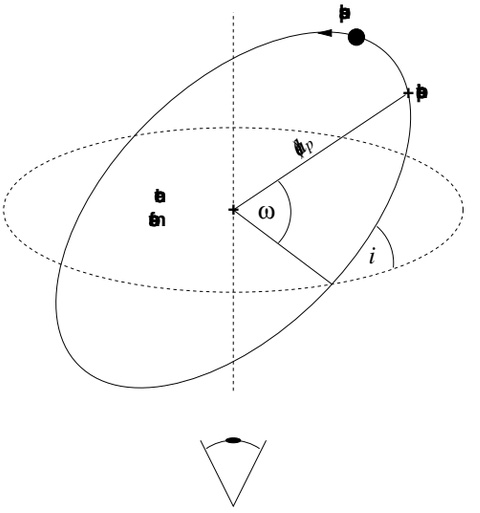


Fig. 3. Definition of the five Keplerian parameters for a pulsar in a binary orbit: The location of periastron (point of closest approach to the center–of–mass of the binary system) is given by the longitude of periastron, ω. The orbital period, P_b, is the time elapsed between two consecutive periastron passages and T_0 is the date of a chosen periastron passage. The periastron distance is given by the eccentricity of the binary orbit, e, and the semi–major axis of the pulsar orbit, a_p. From timing observations, however, only the combination $a_p \sin i$ (projected semi–major axis) can be extracted, where i is the orbital inclination.

equation (1) can be found in [11,49,17]. Each of these effects is described by a set of parameters which form three major groups:

Spin parameters: Rotational frequency ν of the pulsar and its time derivatives.
Astrometric parameters: Position of the pulsar on the sky, proper motion, and parallax.
Binary parameters: 5 Keplerian parameters (see Fig. 3) and a set of post–Keplerian (PK) parameters, which account for relativistic corrections to a Newtonian timing model.

In particular for millisecond pulsars, where in many cases TOA measurements with sub–microsecond precision are achievable, timing parameters can be determined with excellent accuracy. For instance, the positional coordinates for some millisecond pulsars are known within a few micro–arc–seconds and orbital periods are typically determined with a fractional precision of 10^{-10}. In addition, while 'normal' pulsars show large random irregularities in their periods on time–scales of months and years, known as *timing noise* (see [31] and references therein), recycled pulsars are found to have a high rotational stability. Only a few recycled pulsars show deviations from a simple spin–down model on time–scales of a few years ('red noise') like PSR B1937+21 — the pulsar with the shortest rotational period (1.56 ms) — as shown in Fig. 4. The combination of rotational

stability and high timing–accuracy provide perfect conditions for high precision 'clock comparison' experiments which have the power to test strong–field gravity.

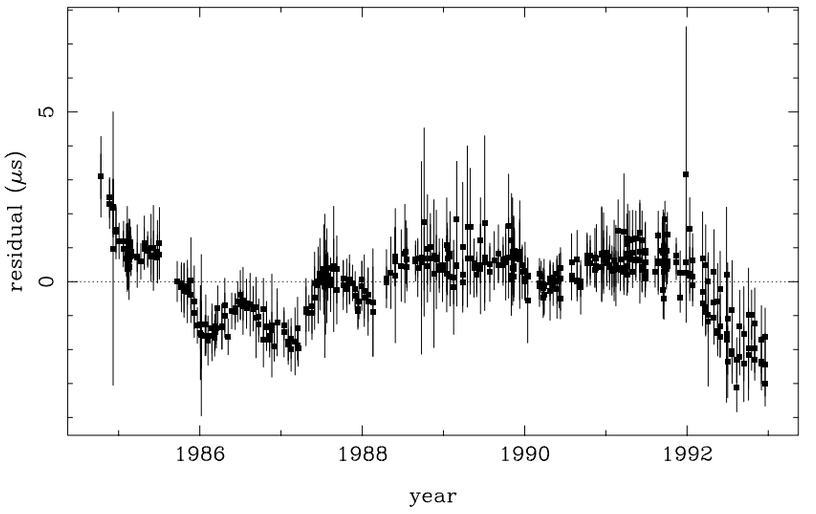


Fig. 4. Post–fit residuals for the millisecond pulsar PSR B1937+21 after fitting for astrometric parameters, and the period and period derivative [25]. The unmodelled long–term variations in the residuals are understood to be caused by timing noise which strongly compromises the use of this pulsar as a long–term time standard.

The important point for testing relativistic effects in binary pulsar systems is that the Keplerian and PK parameters can be measured in a phenomenological manner, independently of the choice of a specific theory of gravity. One expects that in a particular theory of gravity, the PK parameters can be written as functions of the pulsar and companion masses, m_p and m_c, and the well–determined Keplerian parameters. In *general relativity*, for instance, the five most important PK parameters are given in a first approximation by [11]:

$$\dot{\omega} = 3\, T_\odot^{2/3} \left(\frac{P_b}{2\pi}\right)^{-5/3} \frac{1}{1-e^2}\, (m_p + m_c)^{2/3} \,, \tag{2}$$

$$\gamma = T_\odot^{2/3} \left(\frac{P_b}{2\pi}\right)^{1/3} e\, \frac{m_c(m_p + 2m_c)}{(m_p + m_c)^{4/3}} \,, \tag{3}$$

$$\dot{P}_b = -\frac{192\pi}{5}\, T_\odot^{5/3} \left(\frac{P_b}{2\pi}\right)^{-5/3} \frac{(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4)}{(1-e^2)^{7/2}}\, \frac{m_p\, m_c}{(m_p + m_c)^{1/3}} \,, \tag{4}$$

$$r = T_\odot\, m_c \,, \tag{5}$$

$$s = T_\odot^{-1/3} \left(\frac{P_b}{2\pi}\right)^{-2/3} x\, \frac{(m_p + m_c)^{2/3}}{m_c} \,, \tag{6}$$

where $\dot{\omega}$ denotes the relativistic advance of periastron, γ is the amplitude of the combined effect of relativistic time dilation and gravitational redshift, $\dot{P}_b$ denotes the time derivative of the orbital period due to gravitational wave damping, and r and s are two parameters related to the Shapiro propagation delay caused by the gravitational field of the companion. The masses m_p and m_c are expressed in units of solar masses ($M_\odot$), and we use the notations $s \equiv \sin i$ and $T_\odot \equiv GM_\odot/c^3 = 4.925490947\,\mu$s. G denotes the Newtonian constant of gravity and c the speed of light. The projected semi–major axis x is defined as $x \equiv a_p \sin i/c$.

The measurement of the five Keplerian parameters and two post Keplerian parameters will allow to determine the masses of pulsar and companion within a given theory of gravity. Once the masses are known with sufficient precision all the other PK parameters can be determined for this system. Hence, the measurement of more than two PK parameters allows to test the theory of gravity being used. More precisely, the measurement of n PK parameters determines n curves in the two dimensional m_p–m_c plane whose shape and position depend on the theory of gravity being applied. If the theory of gravity and the simple theoretical model give an accurate description for the binary system, the n curves meet at one point [17].

3 Binary Pulsars and Gravity Experiment I. Double–Neutron–Star Binaries

Out of the 6 known double–neutron–star systems presently only two of them allow the determination of more than two PK parameters, which is necessary to conduct tests as outlined in the previous section. In the following we summarize the results for these two binary–pulsar systems.

3.1 PSR B1913+16

PSR B1913+16 was the first binary pulsar to be discovered. It was 25 years ago when Joseph Taylor and Russell Hulse found this double–neutron–star system during a systematic search for new pulsars using the 305–m Arecibo radiotelescope [19]. At present three PK are measured for this system with high precision: the relativistic advance of periastron $\dot{\omega}$, the Einstein delay γ, and a change in the orbital period $\dot{P}_b$. Figure 5 shows the corresponding three curves in the m_p–m_c plane based on equations (2) to (4) and the *observed* values for these three parameters.

Before the observed $\dot{P}_b$ can be compared to the theoretical value, we must apply a correction which accounts for the acceleration of the binary–pulsar system with respect to the solar system caused by the gravitational potential of the Galaxy and an apparent acceleration due to the proper motion of the binary system (Shklovskii effect). For PSR B1913+16 these corrections can be estimated with necessary precision and as a result, $\dot{\omega}$, γ, and $\dot{P}_b^{\text{corrected}}$ are in excellent agreement (0.25%) with the predictions of general relativity [49,16,45,46]. In particular, the observed change in the orbital period confirms that the binary

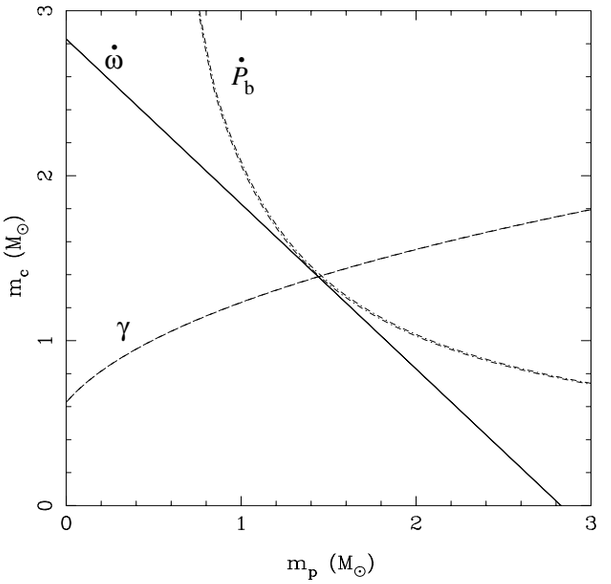


Fig. 5. Mass–mass diagram for the PSR B1913+16 system. Labelled curves illustrate 68% confidence ranges for indicated PK parameters.

system is losing orbital energy at the rate which is predicted by the quadrupole formula of general relativity. Apart from this test, PSR B1913+16 is providing the most precise neutron–star masses ever measured: $m_p = 1.4411 \pm 0.0007$ and $m_p = 1.3879 \pm 0.0007$. [1]

Unfortunately, there is little hope that the precision of this $\dot{\omega}$–γ–$\dot{P}_b$ test can improve much further, since this would require a high precision in the determination of the pulsar distance and a good knowledge of the gravitational field of our Galaxy in order to correct for acceleration effects present in the observed $\dot{P}_b$. Presently the only way to determine the distance of PSR B1913+16 is based on models of the free electron distribution in our Galaxy, which rarely allow a precision better than 20%.

3.2 PSR B1534+12

With the discovery of PSR B1534+12 in 1990 by Wolszczan [59] a second double–neutron–star binary became available for testing relativity in strong field regimes. PSR B1534+12 is significantly brighter than PSR B1913+16, and its pulse has

[1] Note, these are the observed masses of pulsar and companion, which are related to the intrinsic (true) masses by $m_{\rm intrinsic} = D m_{\rm obs}$. The parameter D is a Doppler factor characterizing the radial velocity, V_r, of the binary–pulsar system. $D \simeq 1 - V_r/c$ cannot be determined from timing observations and is set equal to one during the fitting process.

a narrow peak, leading to more precise timing measurements. In addition, the orbit is seen nearly edge–on and, as a consequence, the Shapiro propagation delay caused by the gravitational field of the companion star can easily be extracted from the timing data. Thus, in addition to $\dot{\omega}$, γ, and $\dot{P}_b$ two PK parameters characterizing the Shapiro propagation delay are known for this double–neutron–star system, i.e. the "range" r and the "shape" s of the Shapiro delay [41]. The resulting five curves in the mass–mass digram are shown in Fig. 6.

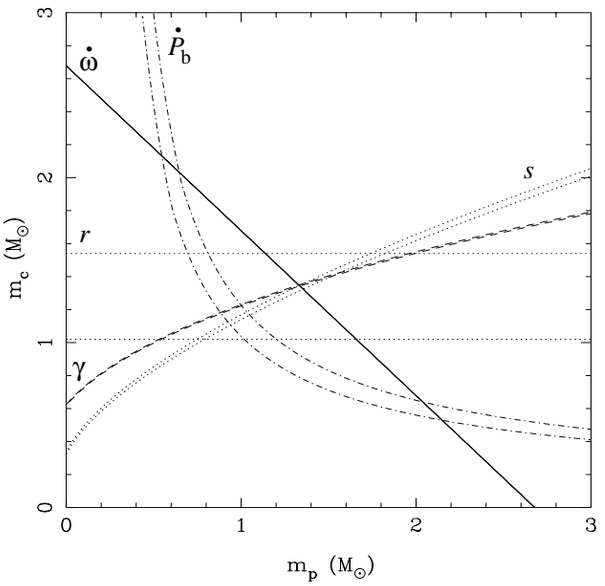


Fig. 6. Mass–mass diagram for the PSR B1534+12 system. Labelled curves illustrate 68% confidence ranges for indicated PK parameters.

The four curves for $\dot{\omega}$, γ, r, and s have a common solution in the m_p–m_c parameter space ($m_p = 1.334 \pm 0.002$, $m_p = 1.344 \pm 0.002$) leading to non–radiative test of relativistic gravity, complementing the (radiative) $\dot{\omega}$–γ–$\dot{P}_b$ test for PSR B1913+16 [50].

The fifth curve corresponding to the observed $\dot{P}_b$ does not agree with the solution for the $\dot{\omega}$–γ–r–s test. As in the case of PSR B1913+16 the observed $\dot{P}_b$ is the sum of an apparent and an intrinsic change of the orbital period. However, for PSR B1534+12 the non–intrinsic part of $\dot{P}_b$ ammounts to more than 30% of the observed $\dot{P}_b$. The $\dot{P}_b$ curve does agree with the other four curves if one assumes a pulsar distance of 1.08±0.15 kpc [42]. In order to perform a $\dot{\omega}$–γ–$\dot{P}_b$ test for this system a precise distance estimation is needed. A reliable value for the distance could either come from the measurement of a timing parallax or the measurement of a parallax distance using VLBI observations.

4 Binary Pulsars and Gravity Experiment II. Small–Eccentricity Binary Pulsars

The majority of binary pulsars is found to be in orbit with a white–dwarf companion. Due to the mass transfer in the past, these systems have very small orbital eccentricities. Therefore, the location of periastron is not at all a prominent feature in pulsar timing observations, and neither $\dot{\omega}$ nor γ were measured for any of these binary pulsars. In fact, the only PK parameters measured with reasonable accuracy for a small–eccentricity binary pulsar are the two Shapiro parameters, r and s, in case of PSR B1855+09 [25]. However, since the orbital period of this system is 12.3 days, the expected $\dot{P}_b$ is by far too small to be of any importance for timing observations and, consequently, there is no third PK parameter which would allow the kind of test conducted with the two double–neutron–star systems above.

On the other hand, many alternative theories of gravity, tensor–scalar theories for instance, predict effects that depend strongly on the difference between the gravitational self–energy per unit mass ($\epsilon \equiv E^{\rm grav}/mc^2$) of the two masses of a binary system. While this difference in binding energies is comparably small for double–neutron–star systems, it is large in neutron star–white dwarf systems since for a white dwarf $\epsilon \sim 10^{-4}$ while for a $1.4 M_\odot$ neutron star $\epsilon \approx 0.15$.

Damour and Eposito–Farèse developed a field–theory based framework for discussing and interpreting experimental tests of relativistic gravity at a second post–Newtonian (2PN) level [12]. This framework is based on a class of tensor–multiscalar theories in which gravity is mediated by a tensor field together with one or several scalar fields. Within this tensor–multiscalar framework, 2PN deviations from general relativity can be fully described by two parameters, ε and ζ. Damour and Eposito–Farèse come to the conclusion that it is extremely difficult to measure these parameters in the solar system. In a second paper [13] they go even further and demonstrate that certain classes of tensor–scalar theories develop nonperturbative strong field effects which lead to significant deviations from general relativity in conditions involving strong gravitational fields. Since such strong field effects only occur in neutron stars above a certain critical mass, they are completely inaccessible by solar–system experiments and, at present, could only be detected in binary pulsar experiments. It was pointed out by Damour and Eposito–Farèse in [14], that future LIGO–VIRGO observations of inspiralling compact binaries are not competitive with present binary–pulsar tests in their discriminating probing power of strong–field gravity effects. Thus, binary pulsars are and will continue to be superb tools for testing strong–field aspects of gravity.

While double–neutron–star systems, like PSR 1534+12 and PSR B1913+16, already tightly constrain tensor–multiscalar theories of gravity, there are theoretical predictions for which neutron star–white dwarf systems provide better testing grounds, like the existence of *gravitational dipole radiation* and the *violation of the strong equivalence principle*.

4.1 Gravitational Dipole Radiation

Unlike general relativity, many alternative theories of gravity predict the presence of all radiative multipoles — monopole and dipole, as well as quadrupole and higher multipoles [56]. For binary systems scalar–tensor theories, for instance, predict a loss of orbital energy which at highest order is dominated by scalar dipole radiation. As a result, the orbital period, P_b, should change according to

$$\dot{P}_b^{(\mathrm{dipole})} = -\frac{4\pi^2 G_* M_\odot}{c^3 P_b} \, \frac{1+e^2/2}{(1-e^2)^{5/2}} \, \frac{m_p m_c}{m_p+m_c} \, (\alpha_p - \alpha_c)^2 \, + \mathcal{O}\left(\frac{v^5}{c^5}\right) , \quad (7)$$

where $-\alpha_p m_p$ and $-\alpha_c m_c$ are the total scalar charge of pulsar and companion, respectively. G_* is the 'bare' gravitational constant and the masses m_p and m_c are measured in units of solar masses $M_\odot$. For a white dwarf companion $|\alpha_c| \ll 1$ and thus the expression $(\alpha_p - \alpha_c)^2$ in equation (7) can be of the order one if the pulsar develops a significant amount of scalar charge. In this case the gravitational wave damping of the orbit is completely dominated by the emission of gravitational dipole radiation. Note, in double–neutron–star systems where the two components have similar mass the term $(\alpha_p - \alpha_c)^2$ is very small and, therefore, gravitational dipole radiation would be of much less importance than in neutron star–white dwarf systems.

PSR J1012+5307 is a 5.3 ms pulsar in a 14.5 h circular orbit with a low mass white–dwarf companion. Since its discovery in 1993 [35] this pulsar has been timed on a regular basis using the Jodrell Bank 76–m and the Effelsberg 100–m radiotelescope, sometimes achieving a timing accuracy of 500 ns after just 10 min of integration [27,28]. In addition, the white–dwarf companion appears to be relatively bright ($V = 19.6$) and shows strong Balmer absorption lines. Based on white dwarf model calculations, a companion mass of $m_c = 0.16 \pm 0.02$ and a distance of 840 ± 90 pc was derived [51,8]. Further, a reliable radial velocity curve for the white dwarf has been extracted, which then, in combination with the pulsar timing information, gave a mass for the pulsar of $m_p = 1.64 \pm 0.22$. Since $e \simeq 0$ for this binary system, we find from equation (7)

$$\dot{P}_b^{(\mathrm{dipole})} \approx -5 \times 10^{-10} \, \alpha_p^2 \, . \quad (8)$$

On the other hand, the change of the orbital period as predicted by general relativity amounts to -10^{-14}, which is more than four orders of magnitude less if $|\alpha_p|$ is of the order one. A comparison with the limits on $\dot{P}_b$, as extracted from pulsar–timing observations, i.e. $\dot{P}_b = (0.1 \pm 1.8) \times 10^{-13}$, sets an upper limit of

$$|\alpha_p| < 0.02 \, . \quad (9)$$

Using the 2PN parameter ζ, α_p^2 in equation (8) can be replaced by $c_p^2 \zeta$ where c_p is the compactness of the neutron star. For a medium equation of state $c_p \approx 0.21 \, m_p/M_\odot$ holds and we find

$$-3 \times 10^{-3} < \zeta < 3 \times 10^{-3} \, , \quad (10)$$

which is the so far tightest limit on the parameter ζ [28].

Provided regular timing observations are conducted on a monthly basis, limits on the change of the orbital period for this system should improve by a factor of five for this system within the next three years.

4.2 Violation of the Strong Equivalence Principle

The *strong equivalence principle (SEP)* requires the universality of free fall of all objects in an external gravitational field regardless of their mass, composition and fraction of gravitational self–energy. While all metric theories of gravity share the property of universality of free fall of test particles (weak equivalence principle), many of them, which are considered as realistic alternatives to general relativity, predict a violation of the strong equivalence principle. It was pointed out by Nordtvedt [37] that laboratory–size bodies possess a negligible fraction of gravitational self–energy and therefore laboratory–free–fall experiments indicate nothing about a violation of the SEP. Nordtvedt [38] further pointed out that lunar–laser–ranging experiments have the potential to test a violation of the SEP due to the difference in fraction of gravitational self–energy between the Earth ($\epsilon \sim -4.6 \times 10^{-10}$) and the Moon ($\epsilon \sim -0.2 \times 10^{-10}$) which are both exposed to the (external) gravitational field of the Sun. A violation of the SEP can be understood as an inequality between the gravitational mass, m_g, and the inertial mass, m_i, which can be written as function of ϵ:

$$\frac{m_g}{m_i} \equiv 1 + \delta(\epsilon) = 1 + \eta\epsilon + \mathcal{O}\left(\epsilon^2\right) \ . \tag{11}$$

While the analysis of lunar–laser–ranging data tightly constrains the 'Nordtvedt parameter' η [58,33] it indicates nothing about a violation of the SEP in strong–field regimes, i.e. terms of higher order in ϵ, due to the smallness of ϵ for solar–system bodies. For neutron stars, however, $\epsilon \sim 0.15$ and thus binary–pulsars with white–dwarf companions ($\epsilon \sim 10^{-4}$) provide ideal laboratories for testing a violation of the SEP due to nonlinear properties of the gravitational interaction [15].

In case of a violation of the SEP the eccentricity vector of a small–eccentricity binary–pulsar system with a white–dwarf companion exposed to the external gravitational field of the Galaxy evolves due to Fig. 7. For the length of the 'induced' eccentricity vector $\mathbf{e}_F$ one finds

$$e_F \propto (\delta_p - \delta_c)\ P_b^2\ a_\perp \ . \tag{12}$$

Therefore, small eccentricity binaries with long orbital periods are most sensitive to a violation of the SEP. Since neither θ nor e_R are known quantities in our analysis we have to treat them in a statistical manner in order to get limits for e_F. The angle θ, for instance, is assumed to be uniformly distributed between 0 and 2π. On the other hand, given a certain $(\delta_p - \delta_c) \simeq \delta_p$, i.e. a certain e_F for a given binary pulsar, the observed eccentricity, e, sets an upper limit on θ which is independent of e_R. If we now compare the limit on θ which we get for every

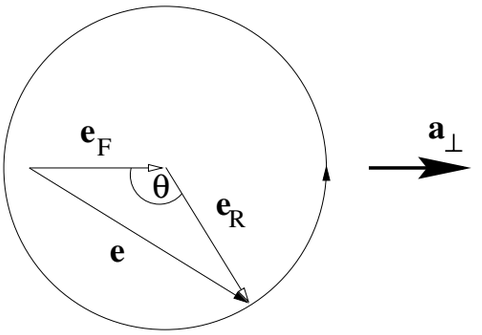


Fig. 7. Evolution of the eccentricity vector $\mathbf{e}$ (vector of length e pointing to periastron) for a small eccentricity binary system in case of a violation of the SEP [15]. $\mathbf{e}$ is a superposition of the constant vector $\mathbf{e}_\delta$ and the vector $\mathbf{e}_R$ which is turning in the orbital plane with the rate of the relativistic advance of periastron, i.e. $\theta = \theta_0 + \dot{\omega}t$. The vector $\mathbf{a}_\perp$ denotes the projection of the Galactic acceleration onto the orbital plane.

observed small–eccentricity binary–pulsar system with a uniform distribution using Monte–Carlo simulations we obtain an safe upper limit for δ_p which is

$$|\delta_p| < 0.009 \quad (95\% \text{ C.L.}) \tag{13}$$

(For details see [55]).

In the framework of tensor–multiscalar theories δ_p can be expressed using the 2PN parameters ε and ζ. For a medium equation of state one finds

$$\delta_p \approx 0.045 \left(\frac{m_p}{M_\odot}\right)^2 \left(\frac{\varepsilon}{2} + \zeta\right) . \tag{14}$$

Consequently, assuming a typical neutron–star mass of $1.4 M_\odot$ one obtains

$$\left|\frac{\varepsilon}{2} + \zeta\right| < 0.1 \quad (95\% \text{ C.L.}) . \tag{15}$$

Since ζ is already tightly constrained by equation (10) we find a safe upper limit for $|\varepsilon|$ of 0.2. Further restrictions on ε can be derived from the PSR B1534+12 and PSR B1913+16 tests (see [12] for details).

4.3 Violation of Local Lorentz Invariance and Conservation Laws

If gravity is mediated in part by a long–range vector field or by a second tensor field one expects the global matter distribution in the Universe to select a preferred frame for the gravitational interaction [57]. At the post–Newtonian level, gravitational effects associated with such a *violation of the local Lorentz invariance* of gravity are characterized by two theory dependent parameters α_1 and α_2. The close alignment of the Sun's spin axis with the total angular momentum of

the solar system yields a tight bound on the second parameter, $\alpha_2 < 2.4 \times 10^{-7}$ [39]. If α_1 were different from zero, a binary system which moves with respect to the global matter distribution in the Universe would suffer a secular change of the orbital eccentricity e. In fact, for small eccentricity binary systems the evolution of the eccentricity vector $\mathbf{e}$ is the same as illustrated in Fig. 7 for a violation of the SEP. This time, however, $\mathbf{a}_\perp$ represents the direction of motion with respect to the preferred frame projected onto the orbital plane and the 'induced' eccentricity vector $\mathbf{e}_F$ has the property

$$|e_F| \propto \alpha_1 \, |m_p - m_c| \, P_b^{1/3} \, w_\perp \, , \tag{16}$$

where $w_\perp$ is the projected velocity of the binary system with respect to the global matter distribution, i.e. the cosmic microwave background. Again we can perform a Mote–Carlo analysis as outlined in the previous section and derive an upper limit for the parameter α_1 from this [55]:

$$|\alpha_1| < 1.2 \times 10^{-4} \quad (95\% \text{ C.L.}) \tag{17}$$

This limit is slightly better than the limit obtained from lunar–laser–ranging data [34] and, more importantly, also holds for strong gravitational–field effects which could occur in the strong–field regions of neutron stars. Due to its small eccentricity, $e < 1.7 \times 10^{-6}$ (95% C.L.), and high velocity with respect to the cosmic microwave background ($w \approx 500$ km/s), PSR J1012+5307 turns out to be the most important binary system for this kind of analysis [28].

In theories of gravity which violate the local Lorentz invariance and the momentum conservation law, a rotating self–gravitating body will suffer a self–acceleration which is given by [40]

$$\mathbf{a}_{\text{self}} = -\frac{\alpha_3}{3} \, \epsilon \, \mathbf{w} \times \boldsymbol{\Omega} \tag{18}$$

where α_3 is a theory dependent parameter and ϵ is the fraction of gravitational self–energy of the body moving with velocity $\mathbf{w}$ with respect to the preferred frame. $\boldsymbol{\Omega}$ denotes the rotational velocity of the body. Again, binary pulsars are ideal probes for this kind of self–acceleration effects [6]. A careful analysis analogous to the previous analyses (SEP, local Lorentz invariance), which appropriately takes care of selection effects, gives

$$|\alpha_3| < 1.5 \times 10^{-19} \quad (95\% \text{ C.L.}) \tag{19}$$

as a safe upper limit for a combined violation of local Lorentz invariance and momentum conservation [55].

5 Geodetic Precession

In general relativity the proper reference frame of a freely falling object suffers a precession with respect to a distant observer (geodetic precession). This was

tested with high precision in the static gravitational field of the sun using the Earth–Moon reference system [58,34]. In a binary–pulsar system this geodetic precession leads to a relativistic spin–orbit coupling, analogous to spin–orbit coupling in atomic physics. As a consequence, the spin of the pulsar precesses about the orbital angular momentum with an angular frequency [4,7]

$$\Omega^{\rm prec} = \left(\frac{2\pi}{P_b}\right)^{5/3} T_\odot^{2/3} \, \frac{m_c(4m_p+3m_c)}{2(m_p+m_c)^{4/3}} \, \frac{1}{1-e^2} \, . \tag{20}$$

Since the angular momentum of the orbital motion is much larger than the spin of the pulsar it practically represents a fixed direction in space, defined by the orbital plane of the binary system. For PSR B1913+16, for instance, $\Omega^{\rm prec}$ amounts to 1.21° per year, i.e. it takes 300 years for a full 360° precession. If the spin of the pulsar is sufficiently inclined with respect to the orbital angular momentum, the geodetic precession of the pulsar will notably change the spin–direction of the pulsar. As the angle between the pulsar–spin axis and the line–of–sight to the pulsar changes, our view of the pulsar–emission region changes and, consequently, a secular evolution of the observed pulse profile is expected.

For PSR B1913+16 long–term changes in the 1.4 GHz pulse profile were first reported ten years ago by Weisberg, Romani & Taylor [53]; (see Fig. 8 for a plot of the pulse profile). They observed a change in the relative amplitude of the leading and trailing sub–pulse of the PSR 1913+16 pulse–profile which was understood as the first detection of the effects of geodetic precession in a binary system. Based on a simple model for the intensity of pulsar–radio emission, Istomin [20] derived a value of 25° for the angle between the spin axis of the pulsar and the orbital angular momentum and made the prediction that the pulsar will become unobservable near the year 2020, when geodetic precession moves the line–of–sight outside the emission region of PSR B1913+16. Cordes, Wasserman, and Blaskiewicz [10] used polarization data to constrain the orientation of the pulsar and argued that a change in the separation between the two sub–pulses, as expected from a simple hollow–cone model for the pulsar–radio emission, has not been observed so far due to an unfavourable phase in the precession. In 1998, eventually, Kramer [26] reported the detection of changes in the sub–pulse separation for PSR B1913+16 based on five years of observations at the 100–m Effelsberg radiotelescope and showed that these changes are in agreement with a hollow–cone model. New Arecibo observations [46] and Effelsberg polarization data [22] lead to improved restrictions on the orientation and emission–geometry parameters of PSR B1913+16 (based on pure geometrical arguments). It is worth noting, that these restrictions still allow for a geometry where the line of sight will never move outside the emission cone and the pulsar therefore is always visible.

Only recently Stairs *et al.* [43] reported a secular evolution of the PSR B1534+12 ($\Omega^{\rm prec} = 0.51°/{\rm yr}$) pulse profile. In both cases, PSR B1913+16 and PSR B1534+12, the changes in the observed pulse profiles are generally accepted as a qualitative test for the presence of geodetic precession in these systems. To date, however, in non of these two cases one can derive useful limits to the rate of

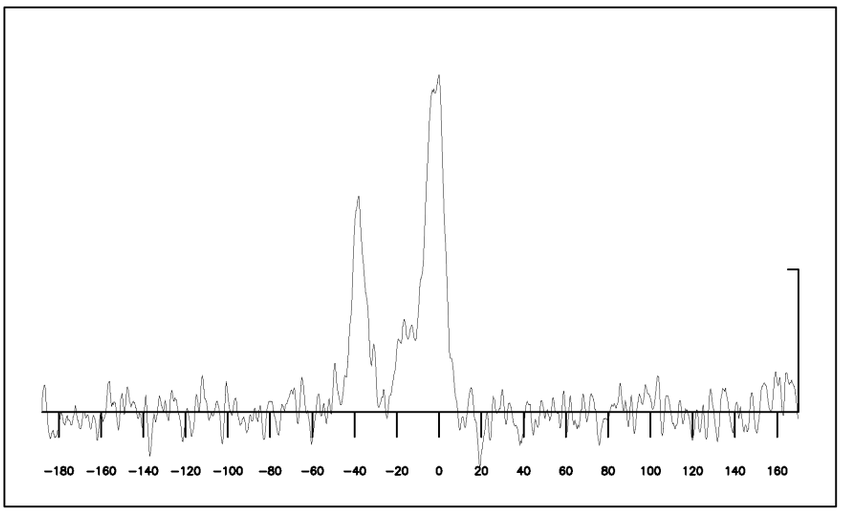


Fig. 8. Pulse profile of PSR B1913+16 at 1.41 GHz.

precession, $\Omega^{\rm prec}$. Due to the uncertainties in pulsar emission it remains an open question whether one day these experiments will allow for a quantitative test of geodetic precession. PSR B1534+12 seems to be the more promising candidate since its magnetic geometry is well constrained from polarization observations [2]. For pulsar astronomy geodetic precession provides a unique opportunity to obtain two–dimensional information on the structure of a pulsar emission zone [53,54].

Two more pulsars from Table 1 have a significant rate of geodetic precession, both PSR B2127+11C and PSR J1141−−6545, exceeding PSR B1913+16 in its precession rate with $\Omega^{\rm prec} = 1.3^\circ/{\rm yr}$ and $\Omega^{\rm prec} = 1.7^\circ/{\rm yr}$, respectively. While PSR B2127+11C is a weak pulsar where it is difficult to get a pulse profile with sufficient signal–to–noise ratio, the newly discovered pulsar PSR J1141–6545 is likely to show interesting pulse–profile changes within the next few years.

References

1. S. B. Anderson, P. W. Gorham, S. R. Kulkarni, T. A. Prince, and A. Wolszczan: *Nature* **346**, 42 (1990).
2. Z. Arzoumanian, J. A. Phillips, J. H. Taylor, A. Wolszczan: *Astrophys. J.* **470**, 1111 (1996).
3. D. C. Backer: 'High Accuracy Timing and Positional Astronomy'. In: *Galactic High–Energy Astrophysics, High–Accuracy Timing and Positional Astronomy*, ed. by J. van Paradijs, H. M. Maitzen, Lecture Notes in Physics, **418**, p. 193 (1993).
4. B. M. Barker and R. F. O'Connell: *Phys. Rev.* **D 12**, 329 (1975).
5. J. F. Bell: 'Radio Pulsar Timing'. In: *Satellite and Ground Based Studies of Radio Pulsars, proceedings of the 31st Scientific Assembly of COSPAR, 1996*, astro-ph/9610145.

6. J. F. Bell and T. Damour: *Class. Quantum Grav.* **13**, 3121 (1996).
7. G. Börner, J. Ehlers, and E. Rudolph: *Astron. Astrophys.* **44**, 417 (1975).
8. P. J. Callanan, P. M. Garnavich, D. Koester: *Month. Not. R. Astron. Soc.* **298**, 207 (1998).
9. F. Camilo, A. G. Lyne, R. N. Manchester, *et al.*: 'The Parkes Multibeam Pulsar Survey'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series, astro-ph/9911185.
10. J. M. Cordes, I. Wasserman, and M. Blaskiewicz: *Astrophys. J.* **349**, 546 (1990).
11. T. Damour and N. Deruelle: *Ann. Inst. H. Poincaré* **44**, 263 (1986).
12. T. Damour and G. Esposito-Farése: *Phys. Rev.* **D 53**, 5541 (1996).
13. T. Damour and G. Esposito-Farése: *Phys. Rev.* **D 54**, 1474 (1996).
14. T. Damour and G. Esposito-Farése: *Phys. Rev.* **D 58**, 042001 (1998).
15. T. Damour and G. Schäfer: *Phys. Rev. Lett.* **66**, 2550 (1991).
16. T. Damour and J. H. Taylor: *Astrophys. J.* **366**, 501 (1991).
17. T. Damour and J. H. Taylor: *Phys. Rev.* **D 45**, 1840 (1992).
18. A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins: *Nature* **217**, 709 (1968).
19. R. A. Hulse and J. H. Taylor: *Astrophys. J. Lett.* **195**, 51 (1975).
20. Ya. N. Istomin: *Sov. Astron. Lett.* **17**, 301 (1991).
21. S. Johnston, R. N. Manchester, A. G. Lyne, M. Bailes, V. M. Kaspi, G. Qiao, N. D'Amico: *Astrophys. J. Lett.* **387**, 37 (1992).
22. A. Karastergiou, M. Kramer, N. Wex, and A. von Hoensbroech: 'Geodetic Precession and the Binary Pulsar B1913+16'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series.
23. V. M. Kaspi: 'Millisecond Pulsar Timing: Recent Advances'. In: *Millisecond Pulsars: A Decade of Surprise*, ed. by A. S. Fruchter, M. Tavani, and D. C. Backer, ASP Conf. Series, Vol. 72, pp. 345–356 (1995).
24. V. M. Kaspi, S. Johnston, J. F. Bell, *et al.*: *Astrophys. J. Lett.* **423**, 43 (1994).
25. V. M. Kaspi, J. H. Taylor, and M. F. Ryba: *Astrophys. J.* **428**, 713 (1994).
26. M. Kramer: *Astrophys. J.*, **509**, 856 (1998).
27. Ch. Lange, N. Wex, O. Doroshenko, and M. Kramer: 'Pulsar Timing at the Effelsberg Radiotelescope'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series.
28. Ch. Lange, F. Camilo, N. Wex, *et al.*: in prep.
29. D. R. Lorimer: 'Radio Pulsars — An Observer's Perspective'. To appear in: *Proceedings of the NATO ASI "The Neutron Star - Black Hole Connection"*, astro-ph/9911324.
30. A. G. Lyne, F. Camilo, R. N. Manchester, *et al.*: Month. Not. R. Astron. Soc., in press, astro-ph/9911313.
31. A. G. Lyne and F. Graham-Smith: *Pulsar Astronomy* (Cambridge University Press, Cambridge 1998).
32. R. N. Manchester, A. G. Lyne, and F. Camilo, *et al.*: 'Timing the Parkes Multibeam Pulsars'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series, astro-ph/9911319.
33. J. Müller, M. Schneider, K. Nordtvedt, and D. Vokrouhlicky: In: *Proceedings of the 8th Marcel Grossman Meeting, Jerusalem 1997.*

34. J. Müller, K. Nordtvedt, and D. Vokrouhlický: *Phys. Rev.* **D 54**, R5927 (1996).
35. L. Nicastro, A. G. Lyne, D. R. Lorimer, P. A. Harrison, M. Bailes, B. D. Skidmore: *Month. Not. R. Astron. Soc. Lett.* **273**, 68 (1995).
36. D. J. Nice, R. W. Sayer, and J. H. Taylor: *Astrophys. J. Lett.* **466**, 87 (1996).
37. K. Nordtvedt: *Phys. Rev.* **169**, 1014 (1968).
38. K. Nordtvedt: *Phys. Rev.* **170**, 1186 (1968).
39. K. Nordtvedt: *Astrophys. J.* **320**, 871 (1987).
40. K. Nordtvedt and C. M. Will: *Astrophys. J.* **177**, 775 (1972).
41. I. H. Stairs, Z. Arzoumanian, F. Camilo, A. G. Lyne, D. J. Nice, J. H. Taylor, S. E. Thorsett, A. Wolszczan: *Astrophys. J.* **505**, 352 (1998).
42. I. H. Stairs, D. J. Nice, S. E. Thorsett, J. H. Taylor: 'Recent Arecibo Timing of the Relativistic Binary PSR B1534+12'. *Proceedings of XXXIV Rencontres de Moriond, "Gravitational Waves and Experimental Gravity" January 23–30, 1999*, astro-ph/9903289.
43. I. H. Stairs, S. E. Thorsett, J. H. Taylor, and Z. Arzoumanian: 'Geodetic Precession in PSR B1534+12'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series, astro-ph/9911198.
44. T. M. Tauris and T. Sennels: *Astron. & Astrophys.*, submitted, astro-ph/9909149 (1999).
45. J. H. Taylor: *Rev. Mod. Phys.* **66**, 711 (1994).
46. J. H. Taylor: 'A quarter century of pulsars and relativity'. *Proceedings of XXXIV Rencontres de Moriond, "Gravitational Waves and Experimental Gravity" January 23–30, 1999.*
47. J. H. Taylor, R. N. Manchester, and A. G. Lyne: *Astrophys. J. Suppl.* **88**, 529 (1993).
48. J. H. Taylor, R. N. Manchester, A. G. Lyne, and F. Camilo: Catalog of 706 pulsars, `ftp://pulsar.princeton.edu/pub/catalog/`.
49. J. H. Taylor and J. M. Weisberg: *Astrophys. J.* **345**, 434 (1989).
50. J. H. Taylor, A. Wolszczan, T. Damour, and J. M. Weisberg: *Nature* **355**, 132 (1992).
51. M. H. van Kerkwijk, P. Bergeron, and S. R. Kulkarni: *Astrophys. J. Lett.* **467**, 89 (1996).
52. M. H. van Kerkwijk and S. R. Kulkarni: *Astrophys. J. Lett.* **516**, 25 (1999).
53. J. M. Weisberg, R. W. Romani, and J. H. Taylor: *Astrophysical J.* **347**, 1030 (1989).
54. J. M. Weisberg and J. H. Taylor: 'General Relativistic Precession of the Spin Axis of Binary Pulsar B1913+16: First Two Dimensional Maps of the Emission Beam'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series.
55. N. Wex: 'Small-eccentricity binary pulsars as laboratories for strong field gravity'. To appear in: *Proceedings of IAU Colloquium 177, Pulsar Astronomy — 2000 and Beyond*, ed. by M. Kramer, N. Wex, R. Wielebinski, ASP Conf. Series.
56. C. M. Will: *Theory and experiment in gravitational physics*, (Cambridge University Press, Cambridge 1993).
57. C. M. Will and K. Nordtvedt: *Astrophys. J.* **177**, 757 (1972).
58. J. G. Williams, X. X. Newhall, and J. O. Dickey: *Phys. Rev.* **D 53**, 6370 (1996).
59. A. Wolszczan: *Nature* **350**, 688 (1991).

Relativistic Phase Shifts for Dirac Particles Interacting with Weak Gravitational Fields in Matter–Wave Interferometers

Christian J. Bordé[1,2,3], Jean–Claude Houard[1],, and Alain Karasiewicz[1]

[1] Laboratoire de Gravitation et Cosmologie Relativistes, Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Cedex 05, Paris, France
[2] Laboratoire de Physique des Lasers, UMR 7538 CNRS, Université Paris–Nord, Avenue J.–B. Clément, 93430 Villetaneuse, France
[3] Institut für Quantenoptik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany

Abstract. We present a second–quantized field theory of massive spin one–half particles or antiparticles in the presence of a weak gravitational field treated as a spin two external field in a flat Minkowski background. We solve the difficulties which arise from the derivative coupling and we are able to introduce an interaction picture. We derive expressions for the scattering amplitude and for the outgoing spinor to first–order. In several appendices, the link with the canonical approach in General Relativity is established and a generalized stationary phase method is used to calculate the outgoing spinor. We show how our expressions can be used to calculate and discuss phase shifts in the context of matter–wave interferometry (especially atom or antiatom interferometry). In this way, many effects are introduced in a unified relativistic framework, including spin–gravitation terms: gravitational red shift, Thomas precession, Sagnac effect, spin–rotation effect, orbital and spin Lense–Thirring effects, de Sitter geodetic precession and finally the effect of gravitational waves. A new analogy with the electromagnetic interaction is pointed out.

1 Introduction

The development of high accuracy atom interferometers, used as clocks in the microwave or in the optical domain, as inertial sensors (gyros, gravimeters, gradiometers...) or for the determination of atomic masses and of the fine structure constant [1–11], requires now a framework to describe the interference of atom waves in a rigorous way. On one hand, one needs to investigate general relativistic effects including those involving the spin of the atoms and, on the other hand, it is necessary to take into account the statistical properties (bosonic or fermionic) of the interfering particles, given the development of coherent atom wave sources and also for a proper treatment of the detection noise. One would also like to be able to discuss the propagation of antimatter in interferometers in the presence of gravitation and as suggested in reference [22] the properties of coherent antimatter waves (generated by an antiatom laser such as an antihydrogen Bose–Einstein condensate). This is possible only within relativistic quantum field theory. Atoms (or antiatoms), in a given eigenstate of the internal atomic

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 403–438, 2001.

Hamiltonian, are considered as elementary particles having a rest mass fixed by the energy of the atomic level and a spin equal to the total angular momentum of the atom in that level. In this paper, we shall consider only Dirac particles for illustration. This is the simplest example of particles with spin which still contains most interesting effects related to spin and applies to neutron or electron interferometry as a special case. The generalization to others spin values is possible along similar lines with Dirac–type equations (Bargmann–Wigner [27], de Broglie fusion method [28], Durand [29]). Our overall goal is to introduce gravitation and general relativistic effects at the quantum level of modern atomic physics and quantum optics experiments. For this, we propose an extension of our first paper on atom interferometry in General Relativity [12], which includes now a second–quantization scheme for the atom waves in the presence of gravitational and electromagnetic fields. The point of view adopted in this paper is the extrinsic point of view using purely quantum field theory in a flat Minkowski background. The connection with the canonical intrinsic approach, using Dirac equation in curved space–time, is made in Appendices A and B. The reader who wishes to start with this canonical approach is thus invited to read first these appendices.

In the main text, we begin right away with the minimal coupling Lagrangian in flat space–time, which is identical to the one derived in curved space–time for a standard choice of tetrads. Then, we proceed with the quantization of the Dirac field and we emphasize the difficulties which arise because of the derivative coupling. These difficulties are solved in a consistent scheme which allows also to define an interaction picture. The evolution operator and the S–matrix are constructed and we demonstrate explicitly a conjecture of Gupta. These results are used to derive formulas for the transition amplitude and for the outgoing spinor in the weak–field approximation, first in configuration space and second in the momentum representation. An expansion in the perturbation wave vector ($\hbar k/mc$) is used to retrieve various physical effects, some of which are well–known. A new analogy with the electromagnetic interaction is presented which includes all components of the field $h^{\mu\nu}$ and generalizes gravitoelectric and gravitomagnetic interactions. In Appendices C and D, calculations of the outgoing spinor are sketched, first with a generalized stationary phase method in configuration space and second in the momentum representation.

2 Lagrangian Theory

Considered as a field theory in flat spacetime, the theory describing the interaction of matter with a given gravitational field will be defined by a Lagrangian density of the following form [15,19]:

$$\mathcal{L} = \mathcal{L}_0 - \tfrac{1}{2} h^{\mu\nu} T_{\mu\nu} \,, \tag{1}$$

where $h^{\mu\nu}$ is the given external field, and where $\mathcal{L}_0$ is the free Lagrangian density of the matter field and $T_{\mu\nu}$ the corresponding stress–energy tensor. For a Dirac

field, one has respectively[1], in symmetrical form,

$$\mathcal{L}_0 = \frac{\hbar c}{2}\left[\overline{\Psi}\left(i\gamma^\mu\overrightarrow{\partial}_\mu - \frac{mc}{\hbar}\right)\Psi + \overline{\Psi}\left(-i\gamma^\mu\overleftarrow{\partial}_\mu - \frac{mc}{\hbar}\right)\Psi\right], \tag{2}$$

$$T_{\mu\nu} = -\eta_{\mu\nu}\mathcal{L}_0 + \frac{\hbar c}{4}\left[\overline{\Psi}\left(i\gamma_\mu\overrightarrow{\partial_\nu} - i\overleftarrow{\partial_\nu}\gamma_\mu\right)\Psi + \overline{\Psi}\left(i\gamma_\nu\overrightarrow{\partial_\mu} - i\overleftarrow{\partial_\mu}\gamma_\nu\right)\Psi\right]. \tag{3}$$

The total Lagrangian density (1) then becomes

$$\mathcal{L} = \left(1+\tfrac{1}{2}h\right)\mathcal{L}_0 - \tfrac{1}{4}i\hbar c h^{\mu\nu}\,\overline{\Psi}\left(\gamma_\mu\,\overrightarrow{\partial}_\nu - \overleftarrow{\partial}_\nu\gamma_\mu\right)\Psi, \tag{4}$$

where[2] $h = h^\mu{}_\mu = \eta_{\mu\nu}h^{\mu\nu}$. This Lagrangian density can also be considered as obtained from the Lagrangian density valid in General Relativity for the interaction of the Dirac field with a prescribed gravitational field in the linear approximation (see the Appendices).

The equations for Ψ and $\overline{\Psi}$ derived from (4) are the following[3]:

$$[\left(1+\tfrac{1}{2}h\right)(i\gamma^\mu\overrightarrow{\partial}_\mu - mc/\hbar) - \frac{i}{2}h^{\mu\nu}\gamma_\mu\overrightarrow{\partial}_\nu - \frac{i}{4}\partial_\nu h^{\mu\nu}\gamma_\mu + \frac{i}{4}\partial_\mu h\gamma^\mu]\Psi = 0, \tag{5}$$

$$\overline{\Psi}[(-i\gamma^\mu\overleftarrow{\partial}_\mu - mc/\hbar)\left(1+\tfrac{1}{2}h\right) + \frac{i}{2}\overleftarrow{\partial}_\nu\gamma_\mu h^{\mu\nu} + \frac{i}{4}\partial_\nu h^{\mu\nu}\gamma_\mu - \frac{i}{4}\partial_\mu h\gamma^\mu] = 0. \tag{6}$$

They admit the conserved current

$$j^\mu = c\overline{\Psi}[\gamma^\mu + \tfrac{1}{2}h\gamma^\mu - \tfrac{1}{2}h^{\mu\nu}\gamma_\nu]\Psi, \tag{7}$$

which coincides with the usual Dirac current when the gravitational field vanishes.

In order to stress some other differences and analogies with electromagnetism, we may write the equation for Ψ with a covariant derivative in the usual sense of non–Abelian gauge field theories in flat space–time

$$i\gamma^\nu(\overrightarrow{\partial}_\nu + \frac{i}{4}\sigma^{\lambda\mu}\partial_\lambda h_{\mu\nu} - \frac{1}{2}h_\nu{}^\alpha\overrightarrow{\partial}_\alpha)\Psi - \frac{mc}{\hbar}\Psi = 0 \tag{8}$$

where

$$\sigma^{\mu\nu} = \frac{i}{2}\left(\gamma^\mu\gamma^\nu - \gamma^\nu\gamma^\mu\right) \tag{9}$$

---

[1] The conventions used here for the metric and for the Dirac equation and matrices are generally those of [17]. In particular, the signature of the metric is taken as $(+,-,-,-)$. Greek indices $\mu,\nu,...$ run from 0 to 3 and latin indices run from 1 to 3. The space–time 4–vector is written $x = (x^0, \boldsymbol{x}) = (ct, \boldsymbol{x})$. The partial derivatives with the right arrow act on the right and those with the left arrow on the left.

[2] The term $(h/2)\mathcal{L}_0$ in (4) comes from the term containing $\mathcal{L}_0$ in (3), and was omitted in [16]. However, $\mathcal{L}_0$ which vanishes when the free Dirac equation is satisfied, does not vanish here.

[3] If one considers these equations as first–order equations with respect to the $h^{\mu\nu}$'s, the factor $(1 + h/2)$ in the first term can be replaced by 1. But it can be shown that the equations so obtained cannot be derived from a Lagrangian, if this latter is restricted to depend linearly on Ψ and $\overline{\Psi}$, and to admit first–order derivatives only.

and where the factor $\left(1+\frac{1}{2}h\right)$ has been removed. The Poincaré generators are associated with gauge fields and a local gauge invariance[4]

$$\Psi'(x) = \left[1+\frac{i}{8}\sigma^{\lambda\mu}\left(\partial_\lambda\xi_\mu-\partial_\mu\xi_\lambda\right)-\xi^\lambda\partial_\lambda\right]\Psi(x) \tag{10}$$

$$h'_{\mu\nu}(x) = h_{\mu\nu}(x)-\partial_\mu\xi_\nu-\partial_\nu\xi_\mu \quad . \tag{11}$$

3 Quantization

We want now to proceed with the quantization of the field Ψ submitted to the interaction defined by (4). When they are applied directly, the standard methods lead to some difficulties coming from the presence of a derivative coupling in the Lagrangian. These problems are, first, briefly discussed, then, a solution is presented allowing the quantization together with the definition of the interaction picture.

3.1 Difficulties with the Derivative Coupling

Two methods can be used, *a priori*, according to whether the Lagrangian is taken under a symmetrical or a asymmetrical form.

Symmetrical Lagrangian. Starting from the symmetrical Lagrangian (4), the usual anticommutation relations will be obtained from the following expression of the conjugate momentum

$$\Pi = 2\,\frac{\partial\mathcal{L}}{\partial\dot{\Psi}}\,, \qquad \text{with} \qquad \dot{\Psi}=\partial_t\Psi\,, \tag{12}$$

which gives

$$\Pi = i\hbar\overline{\Psi}[(1+\tfrac{1}{2}h)\gamma^0-\tfrac{1}{2}h^{0\mu}\gamma_\mu], \tag{13}$$

and

$$\{\Psi_\alpha(x),\Pi_\beta(y)\}_{x^0=y^0} = i\hbar\delta_{\alpha\beta}\,\delta(\boldsymbol{x}-\boldsymbol{y}) \tag{14}$$

where α and β are spinorial indices.

By introducing the matrix (depending on the coordinates)

$$\gamma^{-1} = (1+\tfrac{1}{2}h)\gamma^0-\tfrac{1}{2}h^{0\mu}\gamma_\mu\,, \tag{15}$$

these formulas can be rewritten

$$\Pi = i\hbar\overline{\Psi}\gamma^{-1}, \tag{16}$$

[4] Differences with usual Yang–Mills theories come from the non–commutation of the Lorentz generators with the Dirac matrices and from the fact that the translation generators act on space–time itself.

and

$$\{\Psi_\alpha(x), \overline{\Psi}_\beta(y)\}_{x^0=y^0} = (\gamma(x))_{\alpha\beta}\,\delta(\boldsymbol{x}-\boldsymbol{y}). \tag{17}$$

For the free field the matrix $\gamma(x)$ reduces to γ^0. However, it appears that the Heisenberg equations, which are generally equivalent to the equations of motion, are not satisfied here under their most usual forms. To write the formulas in a condensed form, let us introduce the two operators

$$\overrightarrow{\mathcal{D}} = \hbar c\left[(1+\tfrac{1}{2}h)(-i\gamma^k\overrightarrow{\partial_k}+mc/\hbar)+\frac{i}{2}h^{\mu k}\gamma_\mu\overrightarrow{\partial_k}+\frac{i}{4}(\partial_\nu h^{\mu\nu}-\partial^\mu h)\gamma_\mu\right], \tag{18}$$

$$\overleftarrow{\mathcal{D}} = \hbar c\left[(i\gamma^k\overleftarrow{\partial_k}+mc/\hbar)(1+\tfrac{1}{2}h)-\frac{i}{2}\overleftarrow{\partial_k}\gamma_\mu h^{\mu k}-\frac{i}{4}(\partial_\nu h^{\mu\nu}-\partial^\mu h)\gamma_\mu\right]. \tag{19}$$

One then finds that the Hamiltonian density $\Pi\dot{\Psi}-\mathcal{L}$ can be written as

$$\mathcal{H} = \frac{1}{2}[\overline{\Psi}\overrightarrow{\mathcal{D}}\Psi+\overline{\Psi}\overleftarrow{\mathcal{D}}\Psi], \tag{20}$$

while the field equation reads

$$i\hbar c\gamma^{-1}\partial_0\Psi = \overrightarrow{\mathcal{D}}\Psi. \tag{21}$$

The relation $\overline{\overrightarrow{\mathcal{D}}\Psi} = \overline{\Psi}\overleftarrow{\mathcal{D}}$ and the formula $(\gamma^{-1})^\dagger = \gamma^0\gamma^{-1}\gamma^0$ then give the conjugate equation

$$-i\hbar c\partial_0\overline{\Psi}\,\gamma^{-1} = \overline{\Psi}\overleftarrow{\mathcal{D}}. \tag{22}$$

Using an integration by parts, the total Hamiltonian H can be transformed into

$$H = \int(d^3x)\overline{\Psi}\overrightarrow{\mathcal{D}}\Psi+\frac{i\hbar c}{2}\int(d^3x)\overline{\Psi}(\partial_0\gamma^{-1})\Psi. \tag{23}$$

The anticommutation relation (17) then gives the commutator

$$[H,\Psi] = -\gamma\overrightarrow{\mathcal{D}}\Psi-\frac{i}{2}\hbar c(\gamma\partial_0\gamma^{-1})\Psi\,, \tag{24}$$

or, given the field equation,

$$[H,\Psi] = -i\hbar c\partial_0\Psi-\frac{i}{2}\hbar c(\gamma\partial_0\gamma^{-1})\Psi. \tag{25}$$

Similarly, one has

$$\left[H,\overline{\Psi}\right] = -i\hbar c\partial_0\overline{\Psi}-\frac{i}{2}\hbar c\overline{\Psi}(\partial_0\gamma^{-1})\gamma, \tag{26}$$

then

$$[H,\Pi] = -i\hbar c\partial_0\Pi+\frac{i}{2}\hbar c\Pi\gamma\partial_0\gamma^{-1}. \tag{27}$$

The second terms in the right members of (25), (26) and (27) are unusual, since the fields Ψ, $\overline{\Psi}$ and Π behave as if, considered as functions of some fundamental dynamical variables, they were also explicitly dependent on the time. From the expression of Π, it is in fact obvious that, among $\overline{\Psi}$ and Π one of them, at least, explicitly depends on time. But this is not obvious for Ψ.

Asymmetrical Lagrangian. The preceding difficulty takes another form if, as it is more usual [18], the Lagrangian (4) is replaced by the asymmetrical one

$$\begin{aligned}\mathcal{L}' &= \mathcal{L} + \frac{i\hbar}{2}\partial_\mu j^\mu \\ &= \hbar c\overline{\Psi}\left[(1+\tfrac{1}{2}h)(i\gamma^\mu\overrightarrow{\partial_\mu} - mc/\hbar)\Psi - \frac{i}{2}h^{\mu\nu}\gamma_\mu\overrightarrow{\partial_\nu}\Psi - \frac{i}{4}\partial_\nu h^{\mu\nu}\gamma_\mu\Psi + \frac{i}{4}\partial_\mu h\gamma^\mu\Psi\right].\end{aligned} \tag{28}$$

The field equations are left unchanged, and the conjugate field is again:

$$\Pi' = \frac{\partial\mathcal{L}'}{\partial\dot{\Psi}} = i\hbar\overline{\Psi}\gamma^{-1}. \tag{29}$$

Since $\Pi' = \Pi$, the anticommutation relation of Ψ and $\overline{\Psi}$ is identical to (17), but the Hamiltonian is now

$$H' = \int (d^3x)\overline{\Psi}\,\overrightarrow{\mathcal{D}}\Psi = \int (d^3x)\overline{\Psi}\,\overleftarrow{\mathcal{D}}\Psi - i\hbar c\int (d^3x)\overline{\Psi}(\partial_0\gamma^{-1})\Psi. \tag{30}$$

It follows that the usual Heisenberg equations are satisfied. In fact, one has

$$[H', \Psi] = -\gamma\,\overrightarrow{\mathcal{D}}\Psi = -i\hbar c\partial_0\Psi\,, \tag{31}$$

$$\left[H', \overline{\Psi}\right] = \overline{\Psi}\,\overleftarrow{\mathcal{D}}\,\gamma - i\hbar c\overline{\Psi}(\partial_0\gamma^{-1})\gamma = -i\hbar c\partial_0\overline{\Psi} - i\hbar c\overline{\Psi}(\partial_0\gamma^{-1})\gamma, \tag{32}$$

then

$$[H', \Pi'] = -i\hbar c\partial_0\Pi'. \tag{33}$$

The equations satisfied by Ψ and Π' are those of variables having no explicit time dependence, which is the rule for canonical variables. On the contrary, expression (29) shows that $\overline{\Psi}$ depends explicitly on time since γ^{-1} does, and equation (32) is in agreement with this dependence.

However, the Hamiltonian H' is not Hermitian. In fact, since one has $(\overline{\Psi}\,\overrightarrow{\mathcal{D}}\Psi)^\dagger = \overline{\Psi}\,\overleftarrow{\mathcal{D}}\Psi$, it follows that

$$H'^\dagger = \int (d^3x)\overline{\Psi}\,\overleftarrow{\mathcal{D}}\Psi = H' + i\hbar c\int (d^3x)\overline{\Psi}(\partial_0\gamma^{-1})\Psi. \tag{34}$$

The asymmetry of the equations for Ψ and $\overline{\Psi}$ is a consequence of this lack of Hermiticity.

Trouble with the Interaction Picture. A common shortcoming of the two preceding methods of quantization is the absence of a coherent definition of the interaction picture. In fact, if this picture was defined, there would be a unitary operator U, such that the corresponding field variables ψ and ϖ would be given by

$$\psi = U\Psi U^{-1}\,, \qquad \varpi = U\Pi U^{-1}. \tag{35}$$

Moreover, these variables would be free fields so that one would have the relation $\varpi = i\hbar\overline{\psi}\gamma_0$. Such a relation is not compatible with (35) since one has $\Pi = i\hbar\overline{\Psi}\gamma^{-1}$, $\gamma^{-1} \neq \gamma_0$, and since U must be unitary.

In what follows, the quantization is defined in a way such that the preceding difficulties do not appear. In particular, the interaction picture will be defined, allowing the construction of the transition probabilities and that of the S matrix.

3.2 A Coherent Method of Quantization

The afore–mentioned problems will be solved by a change of variables eliminating the derivative coupling. The Lagrangian density (4) can be written under the form

$$\mathcal{L} = \frac{i\hbar c}{2}\left(\overline{\Psi}\gamma^{-1}(\partial_0\Psi) - (\partial_0\overline{\Psi})\gamma^{-1}\Psi\right) - \frac{1}{2}\overline{\Psi}(\overrightarrow{\mathcal{D}} + \overleftarrow{\mathcal{D}})\Psi. \tag{36}$$

Let us introduce the new field Θ by the formula $\Psi = \Lambda\Theta$, where Λ is a matrix to be determined, which depends on the coordinates. One has

$$\overline{\Psi}\gamma^{-1}(\partial_0\Psi) = \overline{\Theta}\gamma^0\Lambda^\dagger\gamma^0\gamma^{-1}\Lambda(\partial_0\Theta) + \overline{\Theta}\gamma^0\Lambda^\dagger\gamma^0\gamma^{-1}(\partial_0\Lambda)\Theta. \tag{37}$$

The terms containing the time derivatives of Θ and $\overline{\Theta}$ in (36) will be those of the free Dirac Lagrangian if one has $\gamma^0\Lambda^\dagger\gamma^0\gamma^{-1}\Lambda = \gamma^0$ or, assuming the inversibility of Λ, if $\gamma^0\gamma^{-1} = \Lambda^{\dagger-1}\Lambda^{-1}$. By writing the matrix Λ under the form $\Lambda = MU$ where M is Hermitian and positive [34] and U unitary, the preceding equation becomes

$$\gamma^0\gamma^{-1} = (M^{-1})^2 = I + \tfrac{1}{2}h - \tfrac{1}{2}h^{0\mu}\gamma^0\gamma_\mu. \tag{38}$$

If $h_{\mu\nu}$ is sufficiently small, this matrix is inversible and defines a unique matrix M positive–definite [34], while U may be arbitrary. However, to ensure that the asymptotic states deduced from Θ or Ψ have the same physical interpretation, it is necessary that U goes to the identity when $t \to \pm\infty$. More precisely, we will impose the condition $U = I$, by which, when $h_{\mu\nu} = 0$, the field Θ is a free field as Ψ.

The Lagrangian density takes the following form as a function of Θ:

$$\mathcal{L} = \hbar c\left[\frac{i}{2}\left(\overline{\Theta}\gamma^0(\partial_0\Theta) - (\partial_0\overline{\Theta})\gamma^0\Theta\right) + \frac{i}{2}\left(\overline{\Theta}\Gamma^k(\partial_k\Theta) - (\partial_k\overline{\Theta})\Gamma^k\Theta\right) + \frac{1}{2}\overline{\Theta}\Gamma\Theta\right], \tag{39}$$

where Γ^k and Γ are defined by

$$\Gamma^k = \gamma^0\Lambda^\dagger\gamma^0\left[(1+\tfrac{1}{2}h)\gamma^k - \tfrac{1}{2}h^{\mu k}\gamma_\mu\right]\Lambda, \tag{40}$$

$$\Gamma = i\left(\gamma^0\Lambda^{-1}\partial_0\Lambda + \Gamma^k\Lambda^{-1}\partial_k\Lambda\right) - i\gamma^0\left((\partial_0\Lambda^\dagger)\Lambda^{\dagger-1}\gamma^0 + (\partial_k\Lambda^\dagger)\Lambda^{\dagger-1}(\gamma^0\Gamma^k\gamma^0)\right)\gamma^0 - 2\frac{mc}{\hbar}(1+\tfrac{1}{2}h)\gamma^0\Lambda^\dagger\gamma^0\Lambda\,. \tag{41}$$

Let us note the Hermiticity relations

$$(\Gamma^k)^\dagger = \gamma^0\Gamma^k\gamma^0\,, \qquad \Gamma^\dagger = \gamma^0\Gamma\gamma^0\,. \tag{42}$$

The field equations, equivalent to those of Ψ and $\overline{\Psi}$, are now

$$i\left(\gamma^0(\partial_0\Theta)+\Gamma^k(\partial_k\Theta)\right)+\frac{1}{2}\left(\Gamma+i(\partial_k\Gamma^k)\right)\Theta=0\,, \tag{43}$$

$$-i\left((\partial_0\overline{\Theta})\gamma^0+(\partial_k\overline{\Theta})\Gamma^k\right)+\frac{1}{2}\overline{\Theta}\left(\Gamma-i(\partial_k\Gamma^k)\right)=0, \tag{44}$$

while the current reads

$$j^\mu=c\,\overline{\Theta}\gamma^0\Lambda^\dagger\gamma^0\left[(1+\tfrac{1}{2}h)\gamma^\mu-\tfrac{1}{2}h^{\mu\nu}\gamma_\nu\right]\Lambda\Theta\,, \tag{45}$$

or, more explicitly[5],

$$j^0=c\overline{\Theta}\gamma^0\Theta, \qquad\qquad j^k=c\overline{\Theta}\Gamma^k\Theta\,. \tag{46}$$

The conjugate momentum Π_Θ of Θ has the same form as in the free–field case

$$\Pi_\Theta=2\,\frac{\partial\mathcal{L}}{\partial\dot{\Theta}}=i\hbar\overline{\Theta}\gamma^0\,. \tag{47}$$

It follows that the anticommutation relation of Θ and $\overline{\Theta}$ is the usual one

$$\{\Theta_\alpha(x),\overline{\Theta}_\beta(y)\}_{x^0=y^0}=\gamma^0_{\alpha\beta}\,\delta(\boldsymbol{x}-\boldsymbol{y}). \tag{48}$$

It is equivalent to the anticommutation relation (17) of Ψ and $\overline{\Psi}$. The Hamiltonian density $\mathcal{H}_\Theta$ now reads

$$\mathcal{H}_\Theta=-\frac{i\hbar c}{2}\left(\overline{\Theta}\Gamma^k(\partial_k\Theta)-(\partial_k\overline{\Theta})\Gamma^k\Theta\right)-\frac{\hbar c}{2}\overline{\Theta}\Gamma\Theta, \tag{49}$$

which gives for the total Hamiltonian

$$H_\Theta=-\frac{\hbar c}{2}\int(d^3x)\left[i\left(\overline{\Theta}\Gamma^k(\partial_k\Theta)-(\partial_k\overline{\Theta})\Gamma^k\Theta\right)+\overline{\Theta}\Gamma\Theta\right], \tag{50}$$

or, equivalently,

$$H_\Theta=-\hbar c\int(d^3x)\overline{\Theta}\left[i\Gamma^k(\partial_k\Theta)+\frac{1}{2}\left(\Gamma+i(\partial_k\Gamma^k)\right)\Theta\right], \tag{51}$$

$$=\hbar c\int(d^3x)\left[i(\partial_k\overline{\Theta})\Gamma^k-\frac{1}{2}\overline{\Theta}\left(\Gamma-i(\partial_k\Gamma^k)\right)\right]\Theta. \tag{52}$$

This operator is Hermitian, and from the preceding expressions and, from the field equations (43) and (44), one checks the Heisenberg equations

$$[H_\Theta,\Theta]=-i\hbar c\partial_0\Theta\,,\qquad\left[H_\Theta,\overline{\Theta}\right]=-i\hbar c\partial_0\overline{\Theta}\,,\qquad[H_\Theta,\Pi_\Theta]=-i\hbar c\partial_0\Pi_\Theta\,. \tag{53}$$

[5] The expression of j^0 is identical to that of the free–field case. In Appendix B this property is taken as a condition allowing the introduction of the field Θ in the framework of the linearized theory of General Relativity.

The difficulties discussed in Section 3.1 have disappeared. In particular, the basic variables Θ, $\overline{\Theta}$ and Π_Θ look like variables having no explicit dependence on time. On the contrary, the variables Ψ, $\overline{\Psi}$ and Π, initially considered, depend explicitly on time, since one has

$$\Psi = \Lambda\Theta\,, \qquad \overline{\Psi} = \overline{\Theta}\gamma^0\Lambda^\dagger\gamma^0\,, \qquad \Pi = \Pi_\Theta\Lambda^{-1}, \tag{54}$$

and since the matrix Λ generally depends on time.

Let us remark, however, that the Hamiltonian H_Θ differs from the Hamiltonian H introduced at the beginning in terms of Ψ. It is convenient to consider H from (23) as the integral of the density

$$\mathcal{H}_\Psi = \overline{\Psi}\left[\overrightarrow{\mathcal{D}} + \frac{i\hbar c}{2}(\partial_0\gamma^{-1})\right]\Psi, \tag{55}$$

and H_Θ from (51) as the integral of the density

$$\mathcal{H}_\Theta = \hbar c\overline{\Theta}\left[-i\Gamma^k\overrightarrow{\partial_k} - \frac{1}{2}\left(\Gamma + i(\partial_k\Gamma^k)\right)\right]\Theta. \tag{56}$$

A rather tedious transformation of the former expression, using the expressions of γ^{-1}, Γ^k, $\overrightarrow{\mathcal{D}}$, and the definition of Θ, leads to the formula

$$\mathcal{H}_\Psi = \mathcal{H}_\Theta + \frac{i\hbar c}{2}\overline{\Theta}\gamma^0\left[\Lambda^{-1}(\partial_0\Lambda) - (\partial_0\Lambda^\dagger)\Lambda^{\dagger-1}\right]\Theta. \tag{57}$$

Conversely, one has

$$\mathcal{H}_\Theta = \mathcal{H}_\Psi + \frac{i\hbar c}{2}\overline{\Psi}\gamma^0\left[\Lambda^{\dagger-1}(\partial_0\Lambda^{-1}) - (\partial_0\Lambda^{\dagger-1})\Lambda^{-1}\right]\Psi\,. \tag{58}$$

With the choice made above of the matrix U, one has $\Lambda = M = M^\dagger$, with M defined by (38). The two preceding formulas can then be written

$$\mathcal{H}_\Psi = \mathcal{H}_\Theta + \frac{i\hbar c}{2}\overline{\Theta}\gamma^0\left[M^{-1}, \partial_0 M\right]\Theta, \tag{59}$$

$$\mathcal{H}_\Theta = \mathcal{H}_\Psi + \frac{i\hbar c}{2}\overline{\Psi}\gamma^0\left[M^{-1}, \partial_0 M^{-1}\right]\Psi. \tag{60}$$

It is noticeable that the equality $\mathcal{H}_\Psi = \mathcal{H}_\Theta$ is valid at first order with respect to the $h_{\mu\nu}$'s. In fact, since $\partial_0 M$ and $\partial_0 M^{-1}$ are first–order quantities, this approximation is obtained by taking, for the first term in the commutators in (59) and (60), the zeroth–order approximation of M^{-1}, that is the matrix unity.

4 Interaction Picture and the *S*–Matrix

With the Lagrangian (39) the interaction picture is easily defined, since the corresponding conjugate momentum Π_Θ has the same form as that of the free theory. This allows us to define the evolution operator in that picture, then the S–matrix.

4.1 Evolution Operator and Transition Amplitudes

The field in the interaction picture will be denoted by θ. Let us recall that this operator is obtained from the Heisenberg operator Θ by a unitary transformation such that the field equation becomes the free one [32]. Accordingly, the evolution equation of the state vector reads

$$i\hbar\frac{d}{dt}\left|\Phi(t)\right\rangle = H_I(t)\left|\Phi(t)\right\rangle, \tag{61}$$

where, in the absence of derivative coupling, which is the case for the Lagrangian (39), the Hamiltonian $H_I(t)$ is equal to the interaction Hamiltonian expressed in terms of θ. From the expression (50) of the total Hamiltonian one gets

$$H_I(t) = \int (d^3x)\mathcal{H}_{\rm int}(x), \tag{62}$$

with

$$\mathcal{H}_{\rm int} = -\frac{i\hbar c}{2}\overline{\theta}(\Gamma^k - \gamma^k)(\partial_k\theta) + \frac{i\hbar c}{2}(\partial_k\overline{\theta})(\Gamma^k - \gamma^k)\theta - \frac{\hbar c}{2}\overline{\theta}(\Gamma + 2\frac{mc}{\hbar})\theta. \tag{63}$$

The evolution operator in the interaction picture is, from (61), the solution of the following equation together with the initial condition [21] :

$$i\hbar\frac{d}{dt}U(t,t_0) = H_I(t)U(t,t_0), \qquad U(t_0,t_0) = I. \tag{64}$$

The perturbation theory is then obtained from the integral equation, equivalent to (64),

$$U(t,t_0) = I - \frac{i}{\hbar}\int_{t_0}^{t} H_I(\tau)U(\tau,t_0)d\tau. \tag{65}$$

In what follows, we are interested in the transition amplitudes to first order with respect to the $h_{\mu\nu}$'s. These amplitudes will be obtained from the first–order approximation with respect to the Hamiltonian H_I of the U–operator, namely

$$U^{(1)}(t,t_0) = -\frac{i}{\hbar}\int_{t_0}^{t} H_I(\tau)d\tau. \tag{66}$$

By introducing the Hamiltonian density (63) in normal form[6], and the initial and final states of the transition, we will consider the amplitudes

$$\left\langle\Phi_f\right| U^{(1)}(t,t_0)\left|\Phi_i\right\rangle = -\frac{i}{\hbar c}\int_{t_0}^{t}(d^4x)\left\langle\Phi_f\right| : \mathcal{H}_{\rm int}(x) : \left|\Phi_i\right\rangle, \tag{67}$$

[6] See [21]. To first order this prescription suppresses an infinite contribution, due to the energy of the vacuum, in the transition amplitudes for the antiparticles only. It can be seen, therefore, as an expression of the symmetry between particles and antiparticles.

where the pair of colons : : denotes as usual the normal product.

The first–order expressions needed for the evaluation of (67) are now derived from (38) which gives

$$\Lambda = M = I - \frac{h}{4} + \frac{1}{4} h^{0\mu}\gamma^0\gamma_\mu, \tag{68}$$

then, from (40) and (41),

$$\Gamma^k - \gamma^k = \frac{1}{2}\left[h^{00}\gamma^k - h^{0k}\gamma^0 - h^{\mu k}\gamma_\mu\right] \tag{69}$$

$$\Gamma + 2\frac{mc}{\hbar} = -\frac{mc}{\hbar}h^{00} + \frac{i}{4}(\partial_k h^{0\mu})(\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k). \tag{70}$$

The corresponding expression for the Hamiltonian density $\mathcal{H}_{\rm int}$ is

$$\begin{aligned}\mathcal{H}_{\rm int} = &-\frac{i\hbar c}{4}\overline{\theta}\left[h^{00}\gamma^k - h^{0k}\gamma^0 - h^{\mu k}\gamma_\mu\right](\partial_k\theta)\\ &+\frac{i\hbar c}{4}(\partial_k\overline{\theta})\left[h^{00}\gamma^k - h^{0k}\gamma^0 - h^{\mu k}\gamma_\mu\right]\theta\\ &+\frac{\hbar c}{2}\overline{\theta}\left[\frac{mc}{\hbar}h^{00} - \frac{i}{4}(\partial_k h^{0\mu})(\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k)\right]\theta\,.\end{aligned} \tag{71}$$

In Section 5, the expression of the Hamiltonian $H_I(t)$ which will be used is the expression obtained from the space integral of (71) by performing the integration by parts of the term containing $\partial_k\overline{\theta}$, which gives

$$H_I(t) = \int (d^3x)\,\overline{\theta(x)}\gamma^0\mathcal{V}_G(x)\theta(x), \tag{72}$$

where the operator $\mathcal{V}_G(x)$, acting on $\theta(x)$, is given by

$$\begin{aligned}\mathcal{V}_G(x) = &\frac{\hbar c}{2}\gamma^0\left[\frac{mc}{\hbar}h^{00} + \frac{i}{4}\partial_k h_{0j}\gamma^0(\gamma^k\gamma^j - \gamma^j\gamma^k)\right.\\ &\left.+\frac{i}{2}(2\partial_k h^{0k}\gamma^0 + \partial_k h^{jk}\gamma_j - \partial_k h^{00}\gamma^k)\right]\\ &+\frac{i\hbar c}{2}\gamma^0\left[2h^{0k}\gamma^0 + h^{jk}\gamma_j - h^{00}\gamma^k\right]\partial_k\,.\end{aligned} \tag{73}$$

This form of the Hamiltonian is closely related to the equation of motion (43) of the Heisenberg field Θ. In fact, with the help of (69) and (70), one checks that this equation can be written:

$$\left(i\hbar c\gamma^\mu\partial_\mu - mc^2\right)\Theta = \gamma^0\,\mathcal{V}_G\Theta. \tag{74}$$

In what follows, the initial and final states, which appear in (67), are some one–particle or antiparticle states. They are defined from positive or negative energy solutions χ_i and χ_f of the free Dirac equation

$$(i\gamma^\mu\overrightarrow{\partial_\mu} - mc/\hbar)\chi_k = 0\,, \qquad \overline{\chi_k}(i\gamma^\mu\overleftarrow{\partial_\mu} + mc/\hbar) = 0\,, \qquad k = i, f. \tag{75}$$

Denoting by $|\chi_k\rangle$ the corresponding states, one has then to calculate the matrix element

$$\langle\chi_f| : \overline{\theta(x)}\gamma^0\,\mathcal{V}_G(x)\theta(x) : |\chi_i\rangle\,. \tag{76}$$

As a free field, the operator θ can be written

$$\theta(x) = \sum_{r=1}^{2}\int (d^3p)\left[b_r(\boldsymbol{p})\chi^{(+)}_{\boldsymbol{p},r}(x) + d^\dagger_r(\boldsymbol{p})\chi^{(-)}_{\boldsymbol{p},r}(x)\right]\,, \tag{77}$$

where $b_r(\boldsymbol{p})$ and $d_r(\boldsymbol{p})$ are the annihilation operators for the particles or antiparticles, respectively, and $\chi^{(\pm)}_{\boldsymbol{p},r}$ the positive or negative energy solutions of the free Dirac equation given by [17,21]

$$\chi^{(+)}_{\boldsymbol{p},r}(x) = \frac{1}{(2\pi\hbar)^{3/2}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r)}(\boldsymbol{p})e^{i(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)/\hbar}, \tag{78}$$

$$\chi^{(-)}_{\boldsymbol{p},r}(x) = \frac{1}{(2\pi\hbar)^{3/2}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}v^{(r)}(\boldsymbol{p})e^{-i(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)/\hbar}, \tag{79}$$

with $E(\boldsymbol{p}) = c\sqrt{p^2+m^2c^2}$, $p = \|\boldsymbol{p}\|$. In terms of these, any solution χ with positive or negative energy can be written

$$\chi(x) = \sum_r \int (d^3p)\chi^{(\pm)}_{\boldsymbol{p},r}(x)(\chi^{(\pm)}_{\boldsymbol{p},r},\chi)\,, \tag{80}$$

the scalar product of two solutions being defined by

$$(\chi_1,\chi_2) = \int (d^3x)\overline{\chi_1}(x)\gamma^0\chi_2(x)\,. \tag{81}$$

From (80) we have the following expression of the one–particle states

$$|\chi\rangle = \sum_r \int (d^3p)(\chi^{(+)}_{\boldsymbol{p},r},\chi)b^\dagger_r(\boldsymbol{p})\,|\phi_0\rangle\,, \tag{82}$$

where $|\phi_0\rangle$ is the vacuum state, the correspondence $\chi \to |\chi\rangle$ preserving the scalar product. Denoting by $\theta^{(+)}$ the positive frequency part of θ, this last formula implies

$$\begin{aligned}\theta^{(+)}(x)\,|\chi\rangle &= \sum_r \int (d^3p)\chi^{(+)}_{\boldsymbol{p},r}(x)b_r(\boldsymbol{p})\,|\chi\rangle \\ &= \sum_r \int (d^3p)\chi^{(+)}_{\boldsymbol{p},r}(x)(\chi^{(+)}_{\boldsymbol{p},r},\chi)\,|\phi_0\rangle \\ &= \chi(x)\,|\phi_0\rangle\,.\end{aligned} \tag{83}$$

The matrix element (76) reduces to

$$\langle\chi_f|\,\overline{\theta^{(+)}(x)}\gamma^0\,\mathcal{V}_G(x)\theta^{(+)}(x)\,|\chi_i\rangle = \overline{\chi_f(x)}\gamma^0\,\mathcal{V}_G(x)\chi_i(x), \tag{84}$$

so that, from (72), the amplitude (67) reads

$$\langle \chi_f |\, U^{(1)}(t,t_0)\, |\chi_i\rangle = -\frac{i}{\hbar c}\int_{t_0}^{t} (d^4x)\overline{\chi_f(x)}\gamma^0\, \mathcal{V}_G(x)\chi_i(x). \tag{85}$$

This formula will be analyzed in more detail in Section 5.

4.2 *S*–Matrix

The S–matrix can be defined when the external field vanishes at the limit of infinite time[7]. It is obtained from the U operator by taking the limits $t_0 \to -\infty$ and $t \to +\infty$, and the corresponding amplitudes can then be given from (85). However, it is convenient here to return to the formula (67) in which the Hamiltonian density is given by (71). In fact, we want to show that this expression can be transformed into a covariant one, the expression already given by Gupta [16].

The initial and final states being the same as in the calculation leading to (85), formula (67) with the expression (71) of $\mathcal{H}_{\rm int}$ gives, for the first–order S–matrix,

$$\begin{aligned}\langle \chi_f |\, S^{(1)}\, |\chi_i\rangle = -i\int (d^4x)\Big\{ &\frac{1}{2}\overline{\chi_f}\left[\frac{mc}{\hbar}h^{00} - \frac{i}{4}(\partial_k h^{0\mu})(\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k)\right]\chi_i \\ &+ \frac{i}{4}\overline{\chi_f}\left[h^{\mu k}\gamma_\mu + h^{0k}\gamma^0 - h^{00}\gamma^k\right](\partial_k\chi_i) \\ &-\frac{i}{4}(\partial_k\overline{\chi_f})\left[h^{\mu k}\gamma_\mu + h^{0k}\gamma^0 - h^{00}\gamma^k\right]\chi_i\Big\}\,. \end{aligned}\tag{86}$$

This expression can be simplified by integrating by parts the term containing the derivative $\partial_k h^{0\mu}$. This leads to the two terms

$$\frac{i}{8}h^{0\mu}\left[(\partial_k\overline{\chi_f})(\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k)\chi_i + \overline{\chi_f}(\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k)(\partial_k\chi_i)\right].$$

In each of these, by introducing either of the formulas

$$\begin{aligned}\gamma^k\gamma^0\gamma_\mu - \gamma_\mu\gamma^0\gamma^k &= 2[\delta^0_\mu\gamma^k - \delta^k_\mu\gamma^0 - \gamma_\mu\gamma^0\gamma^k] \\ &= 2[\delta^k_\mu\gamma^0 - \delta^0_\mu\gamma^k + \gamma^k\gamma^0\gamma_\mu]\,, \end{aligned}\tag{87}$$

one can insert the derivatives in the combinations $\gamma^k\partial_k\chi_i$ or $\partial_k\overline{\chi_f}\gamma^k$, yielding

$$\frac{i}{4}h^{0k}\overline{\chi_f}\gamma^0(\overleftarrow{\partial_k} - \overrightarrow{\partial_k})\chi_i + \frac{i}{4}h^{0\mu}\left[\overline{\chi_f}(\delta^0_\mu - \gamma_\mu\gamma^0)(\gamma^k\partial_k\chi_i) - (\partial_k\overline{\chi_f}\gamma^k)(\delta^0_\mu - \gamma^0\gamma_\mu)\chi_i\right]. \tag{88}$$

Adding this contribution to the remaining terms in (86), one gets

$$\begin{aligned}\langle \chi_f |\, S^{(1)}\, |\chi_i\rangle = -i\int (d^4x)\Big\{ &\frac{mc}{2\hbar}h^{00}\overline{\chi_f}\chi_i + \frac{i}{4}h^{\mu k}\overline{\chi_f}\gamma_\mu(\overrightarrow{\partial_k} - \overleftarrow{\partial_k})\chi_i \\ &-\frac{i}{4}h^{\mu 0}\left[\overline{\chi_f}\gamma_\mu\gamma^0(\gamma^k\partial_k\chi_i) - (\partial_k\overline{\chi_f}\gamma^k)\gamma^0\gamma_\mu\chi_i\right]\Big\}\,. \end{aligned}\tag{89}$$

[7] Here, this condition is realized, for example, in the case of a gravitational wave.

Finally, using the Dirac equation in the last bracket, this expression reduces to

$$\langle\chi_f|\, S^{(1)}\, |\chi_i\rangle = -i \int (d^4x) \frac{i}{4} h^{\mu\nu} \overline{\chi_f} \left(\gamma_\mu \overrightarrow{\partial_\nu} - \overleftarrow{\partial_\nu} \gamma_\mu\right) \chi_i. \tag{90}$$

This formula agrees with the rule given by Gupta [16] : the expression appearing under the integral sign is, up to the sign, obtained from the interaction Lagrangian $\mathcal{L} - \mathcal{L}_0$ in (4) by replacing Ψ by χ_i and $\overline{\Psi}$ by $\overline{\chi_f}$ [8]. This result is identical with the one valid in the case of a nonderivative coupling, where the interaction picture exists directly. More generally, the validity of the rule asserted by Gupta, implying the use of the interaction Lagrangian defined by (4) and the covariant form of the propagators, can be proved at higher orders in the quantized theory defined in Section 3.

5 Calculation of the Relativistic Phase Shifts in the Weak–Field Approximation

In this final section we use the tools and the material derived in the previous sections to make an explicit calculation of the various contributions to a gravitationally induced phase shift in matter–wave interferometry. We restrict ourselves to one–particle or one–antiparticle states. The application of the formalism to many–particle states and coherent beams of massive particles of different spins will be developed in another publication. We assume that the incoming particles or antiparticles are described by the state vector $|\chi(t_0)\rangle = |\chi_i\rangle$ at some time t_0 before interaction (t_0 can be conveniently taken to be $-\infty$). At a later time t, this state evolves into $|\chi(t)\rangle$ which interferes with a reference beam described by $|\chi_{\rm ref}\rangle$ with which it is recombined in the final beam splitter. In practice, $|\chi_{\rm ref}\rangle$ is produced by the other arm of the interferometer and in many cases, one will have $|\chi_{\rm ref}\rangle \equiv |\chi_i\rangle$.

We are thus interested in the spinorial wave function for one–(anti)particle states:

$$\chi(x) = \langle\phi_0|\, \theta(x)\, |\chi(t)\rangle \tag{91}$$

where $\theta(x)$, $|\chi(t)\rangle$, $|\phi_0\rangle$ are respectively the free–field operator, the one–(anti)particle and the vacuum state vectors in the interaction representation. It is easily shown that, when pair creations are neglected, this expression is equivalent to the Heisenberg field amplitude $\langle 0, in|\, \Theta(x)\, |\Phi\rangle$.

The interference signal itself is given by the projection:

$$\begin{aligned} \int d^3x\, \chi^\dagger_{\rm ref}(x)\chi(x) &= \int d^3x\, \langle\chi_{\rm ref}|\, \theta^\dagger(x)\, |\phi_0\rangle\, \langle\phi_0|\, \theta(x)\, |\chi(t)\rangle \\ &= \langle\chi_{\rm ref}|\, U(t,t_0)\, |\chi(t_0)\rangle \end{aligned} \tag{92}$$

[8] The contribution of the term $(h/2)\mathcal{L}_0$ of (4) vanishes in the first order considered here, since χ_i and χ_f are solutions of the free Dirac equation. But this property is limited to the first order.

where the space integral is over some detection volume. More generally one should consider a detection hypersurface $\sigma(x)$ and the projection:

$$\int_\sigma d\sigma^\mu \, \overline{\chi}_{\rm ref}(x)\gamma_\mu\chi(x) = \int_\sigma d\sigma^\mu \; \langle\chi_{\rm ref}|\,\overline{\theta}(x)\,|\phi_0\rangle\,\gamma_\mu\,\langle\phi_0|\,\theta(x)\,|\chi(t)\rangle \; . \tag{93}$$

In this paper, we shall limit ourselves to the calculation of the amplitude:

$$\langle\chi_{\rm ref}|\,\chi(t)\rangle = \langle\chi_{\rm ref}|\,U(t,t_0)\,|\chi(t_0)\rangle \; . \tag{94}$$

and take the phase of this complex amplitude as the phase contribution of the perturbing interaction. The resulting spinor (91) will also be derived using two different methods: first, in configuration space and second, in the momentum representation.

5.1 Calculation in Configuration Space

The evolution equation of the state vector in the interaction picture is

$$i\hbar\frac{d}{dt}\,|\chi(t)\rangle = H_I(t)\,|\chi(t)\rangle\,, \tag{95}$$

where the Hamiltonian $H_I(t)$ is

$$H_I(t) = \int d^3x \;\; \theta^\dagger(x)\mathcal{V}_G(x)\theta(x) \; , \tag{96}$$

and where the operator $\mathcal{V}_G(x)$, acting on the field operator $\theta(x)$, is given to first order by (73)

$$\begin{aligned}\mathcal{V}_G = {} & \frac{1}{2}mc^2\gamma^0h^{00} + \frac{i\hbar c}{8}\partial_k h_{0j}(\gamma^k\gamma^j - \gamma^j\gamma^k) \\ & + \frac{i\hbar c}{4}\gamma^0(2\partial_k h^{0k}\gamma^0 + \partial_k h^{jk}\gamma_j - \partial_k h^{00}\gamma^k) \\ & + \frac{i\hbar c}{2}\gamma^0\left[2h^{0k}\gamma^0 + h^{jk}\gamma_j - h^{00}\gamma^k\right]\partial_k\end{aligned} \tag{97}$$

that we shall write:

$$\mathcal{V}_G(x) = A(x) + \frac{i\hbar}{2}\partial_j B^j(x) + i\hbar B^j(x)\partial_j = A(x) + \frac{1}{2}\left\{i\hbar\partial_j, B^j(x)\right\}_+ \tag{98}$$

with:

$$\begin{aligned}A(x) \;\; &= \frac{1}{2}mc^2\gamma^0h^{00} + \frac{\hbar c}{4}\sigma^{kj}\partial_k h_{0j} \\ B^j(x) &= \frac{c}{2}\gamma^0\left[2h^{0j}\gamma^0 + h^{kj}\gamma_k - h^{00}\gamma^j\right] \; .\end{aligned} \tag{99}$$

From equation (91) we check that the evolution of the one-(anti)particle spinor is governed by the equation:

$$i\hbar\partial_t\chi = -i\hbar c\gamma^0\gamma^j\partial_j\chi + mc^2\gamma^0\chi + \mathcal{V}_G\chi \tag{100}$$

to which we may add terms corresponding to diagonal magnetic dipole and off–diagonal electric dipole interactions [10,12]. This equation has been used in references [10] and [12] to discuss all the terms that lead to a phase shift in an interferometer.

To obtain the corresponding amplitude, we can start directly from the integral form of (95):

$$|\chi(t)\rangle = |\chi(t_0)\rangle - \frac{i}{\hbar c}\int_{t_0}^{t} d^4x'\theta^\dagger(x')\mathcal{V}_G(x')\theta(x')\,|\chi(t')\rangle \tag{101}$$

So that, to first order:

$$\begin{aligned}
\langle\chi_{\rm ref}|\,\chi^{(1)}(t)\rangle &= \langle\chi_{\rm ref}|\,U^{(1)}(t,t_0)\,|\chi_i\rangle \\
&= -\frac{i}{\hbar c}\int_{t_0}^{t} d^4x'\chi^\dagger_{\rm ref}(x')\mathcal{V}_G(x')\chi_i(x') \\
&= -\frac{i}{\hbar c}\int_{t_0}^{t} d^4x\left\{\chi^\dagger_{\rm ref}\left[A(x)+\frac{i\hbar}{2}\partial_j B^j(x)+i\hbar B^j(x)\partial_j\right]\chi_i\right\} \\
&= \int_{t_0}^{t} d^4x\left\{\chi^\dagger_{\rm ref}\left[\frac{h^{00}}{2c}\gamma^0\partial_t+\boldsymbol{h}\cdot\boldsymbol{\nabla}-\frac{1}{2}\boldsymbol{\alpha}\cdot\overset{\Rightarrow}{h}\cdot\boldsymbol{\nabla}\right.\right. \\
&\quad \left.\left.+\frac{i}{4}\boldsymbol{\Sigma}\cdot\boldsymbol{\nabla}\times\boldsymbol{h}-\frac{1}{4}\left(\boldsymbol{\nabla}h^{00}+\boldsymbol{\nabla}\cdot\overset{\Rightarrow}{h}\right)\cdot\boldsymbol{\alpha}+\frac{1}{2}\boldsymbol{\nabla}\cdot\boldsymbol{h}\right]\chi_i\right\}
\end{aligned} \tag{102}$$

which follows also from Gupta's form. We have used the definitions:

$$\boldsymbol{\alpha}=\gamma^0\boldsymbol{\gamma}=\begin{pmatrix}0 & \boldsymbol{\sigma}\\ \boldsymbol{\sigma} & 0\end{pmatrix},\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\sigma} & 0\\ 0 & \boldsymbol{\sigma}\end{pmatrix},\boldsymbol{h}=\{h^{0k}\},\overset{\Rightarrow}{h}=\{h^{ij}\}\,. \tag{103}$$

The calculation of the spinor itself, by a stationary phase method in configuration space, is outlined in Appendix C.

In equation (102), the first three terms lead to the familiar phase shifts of the Linet–Tourrenc formula [23], the fourth term is the spin–rotation interaction and the last two terms ensure hermiticity. But, because of the Dirac matrices, the interpretation of the various terms in configuration space is not so transparent and, in previous works non–relativistic limits have been taken either directly as in our own work [12] or through a Foldy–Wouthuysen transformation as in [13] in the case of inertial fields, to put equations (100) and (102) in a form where the significance of the terms is more obvious.

In this paper, we may as well take advantage of the flat Minkowski space–time and, therefore, we will use rather the momentum representation in the following. As we shall see, the interpretation of the terms is then much easier, even in their relativistic form.

5.2 Calculation in the Momentum Representation

The free–field operator θ is written as before as:

$$\theta(x)=\sum_{r=1}^{2}\int d^3p\left[b_r(\boldsymbol{p})\chi^{(+)}_{\boldsymbol{p},r}(x)+d^\dagger_r(\boldsymbol{p})\chi^{(-)}_{\boldsymbol{p},r}(x)\right], \tag{104}$$

where $b_r(\boldsymbol{p})$ and $d_r(\boldsymbol{p})$ are the annihilation operators for the particles or antiparticles, respectively, and $\chi^{(\pm)}_{\boldsymbol{p},r}$ are the positive or negative energy solutions of the free Dirac equation given by (78) and (79).

Let us introduce the Fourier transforms $\widetilde{A}(\boldsymbol{k},t)$, $\widetilde{B}^j(\boldsymbol{k},t)$, $\widetilde{h}_{\mu\nu}(\boldsymbol{k})$:

$$A(x) = \frac{1}{(2\pi)^{3/2}} \int d^3k\ \widetilde{A}(\boldsymbol{k},t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \tag{105}$$

$$B^j(x) = \frac{1}{(2\pi)^{3/2}} \int d^3k\ \widetilde{B}^j(\boldsymbol{k},t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \tag{106}$$

$$h_{\mu\nu}(x) = \frac{1}{(2\pi)^{3/2}} \int d^3k\ \widetilde{h}_{\mu\nu}(\boldsymbol{k},t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \tag{107}$$

$$\mathcal{V}_G(x) e^{i\boldsymbol{p}\cdot\boldsymbol{x}/\hbar} = \frac{1}{(2\pi)^{3/2}} \int d^3k\ \widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}} e^{i\boldsymbol{p}\cdot\boldsymbol{x}/\hbar} \tag{108}$$

$$\widetilde{\mathcal{V}_G}(\boldsymbol{k},\boldsymbol{p},t) = \widetilde{A}(\boldsymbol{k},t) - \widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\left(\boldsymbol{p} + \tfrac{1}{2}\hbar\boldsymbol{k}\right) . \tag{109}$$

Let us expand $\chi(x)$, $\overline{\chi_{\rm ref}}(x)$ and $\chi_i(x)$ in plane waves using the expansions of $\overline{\theta^{(\pm)}}(x)$ and $\theta^{(\pm)}(x)$. For particles, the output spinor is:

$$\begin{aligned}
\chi(x) &= \langle\phi_0|\,\theta^{(+)}(x)\,|\chi(t)\rangle \\
&= \langle\phi_0|\sum_{r=1}^{2}\int d^3p\ b_r(\boldsymbol{p})\chi^{(+)}_{\boldsymbol{p},r}(x)\,|\chi(t)\rangle \\
&= \frac{1}{(2\pi\hbar)^{3/2}}\sum_{r=1}^{2}\int d^3p \sqrt{\frac{mc^2}{E(\boldsymbol{p})}} u^{(r)}(\boldsymbol{p}) e^{i(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)/\hbar}\langle 1_{\boldsymbol{p},r}\,|\chi(t)\rangle
\end{aligned} \tag{110}$$

with $|1_{\boldsymbol{p},r}\rangle = b^\dagger_r(\boldsymbol{p})|\phi_0\rangle$ and with, using again (101),

$$\begin{aligned}
\langle\chi_{\rm ref}|\,\chi(t)\rangle &= \langle\chi_{\rm ref}|\,U(t,t_0)\,|\chi_i\rangle \\
&= \langle\chi_{\rm ref}\,|\chi_i\rangle - \frac{i}{\hbar c}\int_{t_0}^{t} d^4x'\chi^\dagger_{\rm ref}(x')\mathcal{V}_G(x')\chi_i(x') \qquad (111)\\
&= \langle\chi_{\rm ref}\,|\chi_i\rangle - \frac{i}{\hbar}\frac{1}{(2\pi\hbar)^3}\sum_{r,r'}\int_{t_0}^{t} dt' \int d^3x' \int d^3p\ d^3p' \langle\chi_{\rm ref}\,|1_{\boldsymbol{p},r}\rangle \\
&\quad \sqrt{\frac{mc^2}{E(\boldsymbol{p})}} u^{(r)\dagger}(\boldsymbol{p}) e^{-i(\boldsymbol{p}.\boldsymbol{x}'-E(\boldsymbol{p})t')/\hbar}\frac{1}{(2\pi)^{3/2}}\int d^3k\ \widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p}',t') e^{i\boldsymbol{k}\cdot\boldsymbol{x}'} \\
&\quad \sqrt{\frac{mc^2}{E(\boldsymbol{p}')}} u^{(r')}(\boldsymbol{p}') e^{i(\boldsymbol{p}'\cdot\boldsymbol{x}'-E(\boldsymbol{p}')t')/\hbar}\langle 1_{\boldsymbol{p}',r'}\,|\chi_i\rangle \qquad (112)\\
&= \langle\chi_{\rm ref}\,|\chi_i\rangle - \frac{i}{\hbar}\sum_{r,r'}\int_{t_0}^{t} dt' \int d^3p \int \frac{d^3k}{(2\pi)^{3/2}}\langle\chi_{\rm ref}\,|1_{\boldsymbol{p}+\hbar\boldsymbol{k},r}\rangle \\
&\quad \sqrt{\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}} u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t') u^{(r')}(\boldsymbol{p}) \\
&\quad e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})]t'/\hbar}\langle 1_{\boldsymbol{p},r'}\,|\chi(t')\rangle \qquad (113)
\end{aligned}$$

and similar expressions for antiparticles. In many cases $|\chi_i\rangle$ and $|\chi_{\rm ref}\rangle$ can be conveniently taken as plane waves, but it is usually more interesting to consider wave packets. Replacing $|\chi_{\rm ref}\rangle$ by $|1_{\boldsymbol{p},r}\rangle$, we check that the momentum representation $\langle 1_{\boldsymbol{p},r}\,|\chi(t)\rangle$ of $\chi(x)$ satisfies:

$$i\hbar\partial_t\langle\phi_0|b_r(\boldsymbol{p})|\chi(t)\rangle = \sum_{r'}\int \frac{d^3k}{(2\pi)^{3/2}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p}-\hbar\boldsymbol{k})}} \\ u^{(r)\dagger}(\boldsymbol{p})\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p}-\hbar\boldsymbol{k},t)u^{(r')}(\boldsymbol{p}-\hbar\boldsymbol{k}) \\ e^{i[E(\boldsymbol{p})-E(\boldsymbol{p}-\hbar\boldsymbol{k})]t/\hbar}\langle\phi_0|b_{r'}(\boldsymbol{p}-\hbar\boldsymbol{k})|\chi(t)\rangle. \quad (114)$$

This equation is, in momentum representation, the analogous of equation (100) in configuration space and we shall give its first–order solution later. It leads to a discrete set of coupled equations for a fixed or negligible recoil momentum. We illustrate below, in the case of the scattering amplitude, how the matrix element, which appears in the second member of (113) and (114) can be evaluated.

To first order, the scattering amplitude (113) is:

$$\begin{aligned}&\langle\chi_{\ \rm ref}|\chi^{(1)}(t)\rangle \\ &\quad=\langle\chi_{\rm ref}|U^{(1)}(t,t_0)|\chi_i\rangle \\ &\quad=-\frac{i}{\hbar}\sum_{r,r'}\int_{t_0}^{t}dt'\int d^3p\int\frac{d^3k}{(2\pi)^{3/2}}\langle\chi_{\rm ref}|1_{\boldsymbol{p}+\hbar\boldsymbol{k},r}\rangle\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}} \\ &\qquad\qquad u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t')u^{(r')}(\boldsymbol{p}) \\ &\qquad\qquad e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})]t'/\hbar}\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle\ . \end{aligned} \quad (115)$$

The next step takes benefit from the smallness of $\hbar\boldsymbol{k}/mc$ or of $\hbar\boldsymbol{k}c/E(\boldsymbol{p})$ to expand the various quantities in this expression to first order in these parameters. The energy $E(\boldsymbol{p}+\hbar\boldsymbol{k})$ can be expanded in a Taylor series:

$$E(\boldsymbol{p}+\hbar\boldsymbol{k}) = E(\boldsymbol{p}) + \frac{\hbar\boldsymbol{k}\cdot\boldsymbol{p}c^2}{E(\boldsymbol{p})} + \frac{(\hbar k)^2c^2}{2E(\boldsymbol{p})} + \ldots = E(\boldsymbol{p}) + \hbar\boldsymbol{k}\cdot\boldsymbol{v} + \hbar\delta + \ldots\,, \quad (116)$$

where δ is the recoil shift. From the general transformation law of spinors in Lorentz boosts (see for example [26]):

$$u(\boldsymbol{p}) = \left[\cosh\left(\frac{\varphi}{2}\right) + \widehat{\boldsymbol{n}}\cdot\boldsymbol{\alpha}\sinh\left(\frac{\varphi}{2}\right)\right]u(0)\,, \quad (117)$$

where $\widehat{\boldsymbol{n}}$ is the unit vector along $\boldsymbol{p}$, and $\tanh\varphi = |\boldsymbol{p}|/(\gamma mc)$, we derive the following infinitesimal transformation for spinors:

$$u(\boldsymbol{p}+\delta\boldsymbol{p}) = \left\{1 + \frac{\widehat{\boldsymbol{n}}\cdot\boldsymbol{\alpha}}{2}\tanh\varphi\frac{\widehat{\boldsymbol{n}}\cdot\delta\boldsymbol{p}}{p} + \frac{1}{2}\sinh\varphi\left[\frac{\boldsymbol{\alpha}\cdot\delta\boldsymbol{p}}{p} - (\widehat{\boldsymbol{n}}\cdot\boldsymbol{\alpha})\frac{\widehat{\boldsymbol{n}}\cdot\delta\boldsymbol{p}}{p}\right] \\ + i\sinh^2\left(\frac{\varphi}{2}\right)\frac{\widehat{\boldsymbol{n}}\times\delta\boldsymbol{p}}{p}\cdot\boldsymbol{\Sigma}\right\}u(\boldsymbol{p}) \quad (118)$$

we have to first order in $\hbar\boldsymbol{k}/mc$:

$$u^\dagger(\boldsymbol{p}+\hbar\boldsymbol{k}) = u^\dagger(\boldsymbol{p})\{1+\frac{1}{2\gamma}\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\alpha} - \frac{i}{2(\gamma+1)}\frac{\boldsymbol{p}}{mc}\times\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\Sigma}\}\,, \tag{119}$$

where $\gamma = 1/\sqrt{1-\beta^2}$, $\boldsymbol{k} = \boldsymbol{k}_\parallel + \gamma\boldsymbol{k}_\perp$, the indices $\parallel$ and $\perp$ designate vector parts respectively parallel and perpendicular to $\boldsymbol{p}$. The term proportional to $\boldsymbol{\alpha}$ represents a boost (velocity change) and the term proportional to $\boldsymbol{\Sigma}$ a rotation (Thomas precession).

Rather than to calculate directly

$$\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t)u^{(r')}(\boldsymbol{p}) \tag{120}$$

it is simpler to calculate first the matrix element:

$$\begin{aligned}
&\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\frac{1}{2}\hbar\boldsymbol{k})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p}-\frac{1}{2}\hbar\boldsymbol{k})}}u^{(r)\dagger}(\boldsymbol{p}+\tfrac{1}{2}\hbar\boldsymbol{k})\widetilde{\mathcal{V}_G}(\boldsymbol{k},\boldsymbol{p}-\tfrac{1}{2}\hbar\boldsymbol{k},t)u^{(r')}(\boldsymbol{p}-\tfrac{1}{2}\hbar\boldsymbol{k})\\
&= \frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p}+\tfrac{1}{2}\hbar\boldsymbol{k})\left[\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right]u^{(r')}(\boldsymbol{p}-\tfrac{1}{2}\hbar\boldsymbol{k}) \qquad (121)\\
&= \frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left[\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right]u^{(r')}(\boldsymbol{p})\\
&\quad+\frac{1}{4\gamma}\frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left\{\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\alpha},\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right\}_-u^{(r')}(\boldsymbol{p})\\
&\quad-\frac{i}{4(\gamma+1)}\frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left\{\frac{\boldsymbol{p}}{mc}\times\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\Sigma},\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right\}_+u^{(r')}(\boldsymbol{p})\,, \qquad (122)
\end{aligned}$$

where $\{A,B\}_\pm$ designate (anti)commutators. The first line gives:

$$\begin{aligned}
&\frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left[\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right]u^{(r')}(\boldsymbol{p})\\
&\qquad=\left[\frac{E(\boldsymbol{p})\widetilde{h}_{00}}{2}-c\boldsymbol{p}\cdot\widetilde{\boldsymbol{h}}+\frac{c^2}{2E(\boldsymbol{p})}\boldsymbol{p}\cdot\overset{\Rightarrow}{\widetilde{h}}\cdot\boldsymbol{p}\right]\delta_{rr'}\\
&\qquad\quad-\frac{i\hbar c}{4\gamma}(\boldsymbol{k}\times\widetilde{\boldsymbol{h}})\cdot w^{(r)\dagger}(\boldsymbol{\sigma}_\perp+\gamma\boldsymbol{\sigma}_\parallel)w^{(r')}\\
&\qquad=\frac{c^2}{2E(\boldsymbol{p})}p^\mu\widetilde{h}_{\mu\nu}p^\nu\delta_{rr'}-\frac{i\hbar c}{4\gamma}(\boldsymbol{k}\times\widetilde{\boldsymbol{h}})\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}\,, \qquad (123)
\end{aligned}$$

where $w^{(r)}$ are Pauli two-component spinors corresponding either to helicity eigenvalues or to the two values of the z-component of the spin in the rest frame, and where

$$\boldsymbol{a} = (\boldsymbol{\sigma}_\perp+\gamma\boldsymbol{\sigma}_\parallel) \tag{124}$$

is the spatial part of the Thomas–Pauli–Lubanski 4–vector operator [30,31]. The second line gives the term:

$$\frac{1}{4\gamma}\,\frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left\{\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\alpha},\widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right\}_{-}u^{(r')}(\boldsymbol{p})$$
$$=\frac{i\hbar c^2}{4E(\boldsymbol{p})\gamma}(\boldsymbol{k}\times\overset{\widetilde{\Rightarrow}}{h}\cdot\boldsymbol{p})\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \tag{125}$$

The last line gives the Thomas precession terms:

$$-\frac{i}{4(\gamma+1)}\,\frac{mc^2}{E(\boldsymbol{p})}u^{(r)\dagger}(\boldsymbol{p})\left\{\frac{\boldsymbol{p}}{mc}\times\frac{\hbar\boldsymbol{k}}{mc}\cdot\boldsymbol{\Sigma}\ ,\ \widetilde{A}(\boldsymbol{k},t)-\widetilde{\boldsymbol{B}}(\boldsymbol{k},t)\cdot\boldsymbol{p}\right\}_{+}u^{(r')}(\boldsymbol{p})$$
$$=\frac{i\hbar\widetilde{h}_{00}}{4m(\gamma+1)}(\boldsymbol{k}\times\boldsymbol{p})\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}-\frac{i\hbar c}{2m(\gamma+1)}\frac{\boldsymbol{p}\cdot\widetilde{\boldsymbol{h}}}{E(\boldsymbol{p})}(\boldsymbol{k}\times\boldsymbol{p})\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}$$
$$+\frac{i\hbar}{4E^2(\boldsymbol{p})m(\gamma+1)}\left[(\boldsymbol{k}\times\boldsymbol{p})\cdot\overset{\widetilde{\Rightarrow}}{h}\cdot\boldsymbol{p}\right]\boldsymbol{p}\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \tag{126}$$

The last line can be rewritten to yield:

$$\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\frac{1}{2}\hbar\boldsymbol{k})}}\sqrt{\frac{mc^2}{E(\boldsymbol{p}-\frac{1}{2}\hbar\boldsymbol{k})}}u^{(r)\dagger}(\boldsymbol{p}+\tfrac{1}{2}\hbar\boldsymbol{k})\widetilde{\mathcal{V}_G}(\boldsymbol{k},\boldsymbol{p}-\tfrac{1}{2}\hbar\boldsymbol{k},t)u^{(r')}(\boldsymbol{p}-\tfrac{1}{2}\hbar\boldsymbol{k})$$
$$=\frac{c^2}{2E(\boldsymbol{p})}p^\mu\widetilde{h}_{\mu\nu}p^\nu\delta_{rr'}-\frac{i\hbar c}{4\gamma}\left[\boldsymbol{k}\times\left(\widetilde{\boldsymbol{h}}-\overset{\widetilde{\Rightarrow}}{h}\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}$$
$$+\frac{i\hbar}{2m(\gamma+1)}\left[(\boldsymbol{k}\times\boldsymbol{p})\frac{c^2p^\mu\widetilde{h}_{\mu\nu}p^\nu}{2E^2(\boldsymbol{p})}\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \tag{127}$$

If we replace now $\boldsymbol{p}$ by $\boldsymbol{p}+\hbar\boldsymbol{k}/2$ in order to calculate (115) this introduces the additional terms:

$$\left[\frac{\hbar c^2\boldsymbol{k}\cdot\boldsymbol{p}}{4E(\boldsymbol{p})}\left(\widetilde{h}_{00}-\frac{c^2\boldsymbol{p}\cdot\overset{\widetilde{\Rightarrow}}{h}\cdot\boldsymbol{p}}{E^2(\boldsymbol{p})}\right)-c\frac{\hbar\boldsymbol{k}}{2}\cdot\left(\widetilde{\boldsymbol{h}}-\frac{\overset{\widetilde{\Rightarrow}}{h}\cdot\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\delta_{rr'} \tag{128}$$

If we introduce the 4-vector κ^μ:

$$\kappa^0c=\boldsymbol{k}\cdot\boldsymbol{v} \tag{129}$$
$$\boldsymbol{\kappa}\ \ =\boldsymbol{k} \tag{130}$$

which corresponds to the energy–momentum 4–vector exchanged during the interaction, these terms can be rewritten:

$$\frac{c^2}{2E(\boldsymbol{p})}\left\{-p^\mu\widetilde{h}_{\mu\nu}p^\nu\frac{\hbar\kappa^0c}{E(\boldsymbol{p})}+\frac{\hbar\kappa^\mu}{2}\widetilde{h}_{\mu\nu}p^\nu\right\}\delta_{rr'} \tag{131}$$

Our final result is thus, for the scattering amplitude:

$$\begin{aligned}
&\langle\chi_{\rm ref}|\, U^{(1)}(t,t_0)\,|\chi(t_0)\rangle \\
&= -\frac{i}{\hbar}\sum_{r,r'}\int_{t_0}^{t}dt'\int(d^3p)\int\frac{d^3k}{(2\pi)^{3/2}}\,\langle\chi_{\rm ref}|\,1_{\boldsymbol{p}+\hbar\boldsymbol{k},r}\rangle \\
&\qquad\left\{\frac{c^2}{2E(\boldsymbol{p})}\,(p^\mu+\hbar\kappa^\mu)\,\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{2E(\boldsymbol{p})}\right)\delta_{rr'}\right. \\
&\qquad\quad+\frac{i\hbar}{2m(\gamma+1)}\left[(\boldsymbol{k}\times\boldsymbol{p})\frac{c^2p^\mu\widetilde{h}_{\mu\nu}p^\nu}{2E^2(\boldsymbol{p})}\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \\
&\qquad\quad\left.-\frac{i\hbar c}{4\gamma}\left[\boldsymbol{k}\times\left(\widetilde{\boldsymbol{h}}-\overset{\widetilde{\Rightarrow}}{h}\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}\right\} \\
&\qquad e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})]t'/\hbar}\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle\ . \qquad (132)
\end{aligned}$$

To obtain the outgoing spinor, one can replace $\langle\chi_{\rm ref}|$ by $\langle\phi_0|\,\theta(x)$ in the previous expression, which gives this spinor as a sum of outgoing plane–wave spinors:

$$\begin{aligned}
\chi(x) =&\chi_i(x)-\frac{i}{\hbar}\frac{1}{(2\pi\hbar)^{3/2}}\sum_{r,r'}\int_{t_0}^{t}dt'\int d^3p\int\frac{d^3k}{(2\pi)^{3/2}}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r)}(\boldsymbol{p})e^{i[\boldsymbol{p}.\boldsymbol{x}-E(\boldsymbol{p})t/\hbar]} \\
&\left\{\frac{c^2}{2E(\boldsymbol{p})}p^\mu\widetilde{h}_{\mu\nu}\,(p^\nu-\hbar\kappa^\nu)\left(1+\frac{\hbar\kappa^0 c}{2E(\boldsymbol{p})}\right)\delta_{rr'}\right. \\
&+\frac{i\hbar}{2m(\gamma+1)}\left[(\boldsymbol{k}\times\boldsymbol{p})\frac{c^2p^\mu\widetilde{h}_{\mu\nu}p^\nu}{2E^2(\boldsymbol{p})}\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \\
&\left.-\frac{i\hbar c}{4\gamma}\left[\boldsymbol{k}\times\left(\widetilde{\boldsymbol{h}}-\overset{\widetilde{\Rightarrow}}{h}\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}\right\}e^{i[E(\boldsymbol{p})-E(\boldsymbol{p}-\hbar\boldsymbol{k})]t'/\hbar}\langle 1_{\boldsymbol{p}-\hbar\boldsymbol{k},r'}\,|\chi_i\rangle \\
&\qquad (133)
\end{aligned}$$

in which an explicit phase factor is associated with each outgoing plane wave component and which can also be obtained directly from the first–order solution of equation (114).

One can also perform the calculation presented in Appendix D, which gives this spinor in the form of Dirac matrices multiplying the initial spinor plane wave components

$$\begin{aligned}
\chi(x) = \chi_i(x) &- \frac{i}{\hbar}\frac{1}{(2\pi\hbar)^{3/2}}\sum_{r'}\int_{t_0}^{t}dt'\int(d^3p)\int\frac{d^3k}{(2\pi)^{3/2}}\frac{c^2}{2E(\boldsymbol{p})} \\
&\left\{(p^\mu+\hbar\kappa^\mu)\,\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{E(\boldsymbol{p})}\right)-\frac{i\hbar}{2}\kappa_\rho\sigma^{\rho\nu}\widetilde{h}_{\mu\nu}p^\mu\right\} \qquad (134) \\
&e^{i\boldsymbol{k}.\boldsymbol{x}}e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})](t'-t)/\hbar}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p})e^{i(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)/\hbar}\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle\ .
\end{aligned}$$

In this formula, the integral over $\boldsymbol{p}$ can be calculated by assuming that the initial wave packet has a very narrow width in momentum space around a central value

$\boldsymbol{p}_0$. The initial wave packet $\chi_i(x)$ can then be factorized. If this approximation is not sufficient an expansion of the wave packet around $\boldsymbol{p}_0$ can be used [22]. The $\boldsymbol{k}$ integral can be performed by turning each term linear in $\boldsymbol{k}$ into a spatial derivative and the result of Appendix C is recovered for the spinor in configuration space. Finally the time integral can be worked out in many cases and expresses energy conservation [22].

For the scattering amplitude, the comparison of equation (132) with equation (102) shows new terms directly related to the momentum exchange: a generalized Thomas precession and a generalized spin–gravitation interaction. To illustrate how this phase shift calculation is done from equation (132) we shall rewrite this equation without the terms that obviously do not contribute to the phase and use expression (116) for the energy difference, in which we neglect the recoil shift δ

$$\begin{aligned}\delta\varphi = -\frac{1}{\hbar}\sum_{r,r'}\int_{t_0}^{t} dt' \int (d^3p)\alpha^*_{\mathrm{ref}}(\boldsymbol{p})\alpha_i(\boldsymbol{p})\beta^*_{r,\mathrm{ref}}\beta_{r',i}\Bigg\{ \frac{c^2}{2E(\boldsymbol{p})} p^\mu h_{\mu\nu}(\boldsymbol{x}_0+\boldsymbol{v}t',t')p^\nu\delta_{rr'} \\ +\frac{\hbar}{2m(\gamma+1)}\left[\frac{c^2 p^\mu \boldsymbol{\nabla} h_{\mu\nu}(\boldsymbol{x}_0+\boldsymbol{v}t',t')p^\nu}{2E^2(\boldsymbol{p})}\times\boldsymbol{p}\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')} \\ -\frac{\hbar c}{4\gamma}\left[\boldsymbol{\nabla}\times\left(\boldsymbol{h}(\boldsymbol{x}_0+\boldsymbol{v}t',t')-\overrightarrow{\boldsymbol{h}}(\boldsymbol{x}_0+\boldsymbol{v}t',t')\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}\Bigg\},\end{aligned} \tag{135}$$

where we also made explicit the centers of the wave packets and their polarization:

$$\langle 1_{\boldsymbol{p},r}\,|\chi_i\rangle = e^{-i\boldsymbol{p}.\boldsymbol{x}_0/\hbar}\alpha_i(\boldsymbol{p})\beta_{r,i},\langle\chi_{\mathrm{ref}}|\,1_{\boldsymbol{p}+\hbar\boldsymbol{k},r}\rangle \simeq e^{i(\boldsymbol{p}+\hbar\boldsymbol{k})\cdot\boldsymbol{x}_0/\hbar}\alpha^*_{\mathrm{ref}}(\boldsymbol{p})\beta^*_{r,\mathrm{ref}} \tag{136}$$

(the assumption that the wave packet is broad enough for $\alpha^*_{\mathrm{ref}}(\boldsymbol{p}+\hbar\boldsymbol{k})\simeq\alpha^*_{\mathrm{ref}}(\boldsymbol{p})$ has been made).

If we assume, for simplicity, that the reference wave packet is identical to the unperturbed wave packet: $\alpha_{\mathrm{ref}}(\boldsymbol{p})\beta_{r,\mathrm{ref}}\equiv\alpha_i(\boldsymbol{p})\beta_{r,i}$ and that they are very narrow in momentum space around a central value $\boldsymbol{p}$, the phase simplifies to:

$$\begin{aligned}\delta\varphi = & -\frac{1}{\hbar}\int_{t_0}^{t} dt'\left\{\frac{c^2}{2E(\boldsymbol{p})}p^\mu h_{\mu\nu}(\boldsymbol{x}_0+\boldsymbol{v}t',t')p^\nu\right. \\ & +\frac{\gamma}{m(\gamma+1)}\left[\frac{c^2 p^\mu\boldsymbol{\nabla}h_{\mu\nu}(\boldsymbol{x}_0+\boldsymbol{v}t',t')p^\nu}{2E^2(\boldsymbol{p})}\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}} \\ & \left.-\frac{c}{2}\left[\boldsymbol{\nabla}\times\left(\boldsymbol{h}(\boldsymbol{x}_0+\boldsymbol{v}t',t')-\overrightarrow{\boldsymbol{h}}(\boldsymbol{x}_0+\boldsymbol{v}t',t')\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot\overline{\boldsymbol{s}}\right\},\end{aligned} \tag{137}$$

where $\overline{\boldsymbol{s}}$ is the mean spin vector[9]

$$\overline{\boldsymbol{s}} = \sum_{r,r'} \beta^*_{r,i}\beta_{r',i}\hbar w^{(r)\dagger}\boldsymbol{a}w^{(r')}/2\gamma \tag{139}$$

In fact the phase calculation is usually more involved since the previous formula applies only to the case of straight unperturbed trajectories. In practice however, one cannot always ignore the fact that, when calculating the phase to first–order for a given term of the Hamiltonian, the motion of the particles is affected by other terms. One example is the calculation of the gravitational shift within the atom beam splitters, in which one cannot ignore the important effects of the diffracting electromagnetic field on the trajectories of the particles [10,38–40]. Gravitational phase shifts have to be calculated along these trajectories. Another example is the gravity field itself, which, on earth, gives parabolic trajectories for atoms. The phase shift for the other terms in equation (137) has to be calculated along these parabolas. A convenient way to achieve these calculations is to replace $\boldsymbol{x}_0 + \boldsymbol{v}t'$ and $\boldsymbol{v}$ in equation (137) by the classical trajectory $\{\boldsymbol{x}(t'), \boldsymbol{v}(t')\}$ obtained in the $ABCD$ formalism developed in references [22,37].

Expression (137) displays all the terms which may lead to a gravitational phase shift in a matter–wave interferometer. They are summarized in Table 1 where one finds successively:

- the terms involving h_{00} lead to the gravitational shift ($h_{00} = -2\ \boldsymbol{g}\cdot\boldsymbol{r}/c^2$), to shifts involving higher derivatives of the gravitational potential and to the analog of the Thomas precession (spin–orbit coupling corrected by the Thomas factor).
- the terms which involve $\boldsymbol{h} = \{h^{0k}\}$, give the Sagnac effect in a rotating frame ($\boldsymbol{h} = \boldsymbol{\Omega}\times\boldsymbol{r}/c$), the spin–rotation coupling and a relativistic correction (analogous to the Thomas term for h_{00}). They describe also the Lense–Thirring effects coming from inertial frame–dragging by a massive rotating body, which is a source for $\boldsymbol{h}$.
- the other terms, which involve the tensor $\overset{\Rightarrow}{h} = \{h^{ij}\}$ describe genuine General Relativity effects such as the effect of gravitational waves and de Sitter geodetic precession (which also includes the Thomas term for h_{00}).

Our expressions are valid for spins 0 and 1/2 and may be conjectured to be valid for arbitrary spin if $\boldsymbol{\sigma}/2$ is replaced by the corresponding spin operator $\boldsymbol{S}$.

The reader will find calculations of the phases corresponding to these various terms in references [3,7,10,35,41–44]. In these calculations, one should never

[9] More generally, if we use $a_i(\boldsymbol{p}, r)$ instead of $\alpha_i(\boldsymbol{p})\beta_{r,i}$, the mean spin vector should be written:

$$\overline{\boldsymbol{s}} = \langle\chi_i|\,\frac{\hbar}{2}\int(d^3x\,)\theta^\dagger(x)\boldsymbol{\Sigma}\theta(x)\,|\chi_i\rangle = \sum_{r,r'}\int(d^3p)a^*_i(\boldsymbol{p},r)a_i(\boldsymbol{p},r')\hbar w^{(r)\dagger}\boldsymbol{a}w^{(r')}/2\gamma \tag{138}$$

Table 1. Classification of the various energy terms entering the expression of the phase shift. The factor γ is the time dilation factor and should not be confused with the PPN parameter γ_{PPN} which can also be introduced (for de Sitter precession it gives the familiar factor $(\gamma_{PPN}+1/2)$).

Corresponding energy term V	$h_{\mu\nu}$	Name of the effect
$\frac{1}{2}Eh_{00}$	Newtonian potential: $h_{00}=2U/c^2=-2\boldsymbol{g}\cdot\boldsymbol{x}/c^2$	Gravitational red shift
	or acceleration field $h_{00}=2\boldsymbol{a}\cdot\boldsymbol{x}/c^2$	Acceleration shift
	Gravity gradient $\boldsymbol{g}(z)\cdot\boldsymbol{x}=-\left(g-g'z/2\right)z$	
	or curvature $R_{0i0j}x^ix^j$	
	Fermi gauge: $h_{00}^F=\ddot{h}_+(t-z/c).(x^2-y^2)/2+\ddot{h}_\times(t-z/c)xy$	gravitational waves
$\frac{\gamma}{2m(\gamma+1)}\left(\boldsymbol{\nabla}h_{00}\times\boldsymbol{p}\right)\cdot\overline{\boldsymbol{s}}$	$h_{00}=2U/c^2$ gives $V=\frac{1}{mc^2}\frac{\gamma}{\gamma+1}\left[\boldsymbol{\nabla}U\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}}$	Thomas precession
$-c\boldsymbol{p}\cdot\boldsymbol{h}$	Rotating frame: $\boldsymbol{h}=\boldsymbol{\Omega}\times\boldsymbol{x}/c$ gives $V=-\boldsymbol{\Omega}\cdot\boldsymbol{L}$	Sagnac effect
	h_{0i} given by the Lense–Thirring metric	Lense–Thirring (orbital)
$-(c/2)\left[\boldsymbol{\nabla}\times\boldsymbol{h}\right]\cdot\overline{\boldsymbol{s}}$	$\boldsymbol{h}=\boldsymbol{\Omega}\times\boldsymbol{x}/c$ gives $V=-\boldsymbol{\Omega}\cdot\overline{\boldsymbol{s}}$	Spin-rotation interaction
	h_{0i} given by the Lense-Thirring metric	Lense–Thirring (spin)
$-\frac{\gamma}{m(\gamma+1)}\left[\boldsymbol{\nabla}\left(c\boldsymbol{p}\cdot\boldsymbol{h}/E\right)\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}}$		$\sim$ Thomas for rotation

Table 1. (Continued)

$c^2\boldsymbol{p}.\overrightarrow{\overrightarrow{h}}.\boldsymbol{p}/2E$	Schwarzschild metric in isotropic coordinates: $h_{00}=h_{11}=h_{22}=h_{33}=2U/c^2$ gives $V=p^2U/E$ in addition to EU/c^2 from h_{00}	
	Einstein gauge: $h_{11}=-h_{22}=h_+(t-z/c)$, $h_{12}=h_{21}=h_\times(t-z/c)$	Effect of gravitational waves
$(c/2)\left[\boldsymbol{\nabla}\times\left(\overrightarrow{\overrightarrow{h}}\cdot\boldsymbol{p}c/E\right)\right]\cdot\overline{\boldsymbol{s}}$	Schwarzschild metric: $U=-GM/r$ $h_{00}=h_{11}=h_{22}=h_{33}=2U/c^2$ gives $V=\dfrac{1}{mc^2}\dfrac{1}{\gamma}\left[\boldsymbol{\nabla}U\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}}$ in addition to $V=\dfrac{1}{mc^2}\dfrac{\gamma}{\gamma+1}\left[\boldsymbol{\nabla}U\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}}$ from h_{00}	de Sitter or geodetic precession
	Einstein gauge: h_{ij}	Interaction of the spin with gravitational waves
$\dfrac{\gamma}{2m(\gamma+1)}\left[\boldsymbol{\nabla}\left(c^2\boldsymbol{p}\cdot\overrightarrow{\overrightarrow{h}}\cdot\boldsymbol{p}/E^2\right)\times\boldsymbol{p}\right]\cdot\overline{\boldsymbol{s}}$		$\sim$ Thomas for gravitation

forget that the external field $h_{\mu\nu}$ acts not only on the atoms but also on other components of the experiments, such as mirrors and laser beams and that, depending on the chosen gauge, additional contributions may enter in the final expression of the phase which should, of course, be gauge independent.

In present experiments on the earth gravity measurements, the relative sensitivity is $\delta g/g \simeq 3.10^{-9}$ after 60 seconds and the absolute accuracy 5.10^{-9}[6,35]. For rotations, the best sensitivity achieved up to now is $6.10^{-10}\mathrm{rad.s}^{-1}\mathrm{Hz}^{-1/2}$[36] but these numbers are expected to improve rapidly in the near future, especially in space experiments, in which general relativistic effects should become detectable. An accurate measurement of the effect of gravitation and inertia on antimatter also appears as a possibility already discussed in reference [47] with a transmission–grating interferometer, although we believe that an antiatom interferometer using laser beams for the antihydrogen beam splitters (Ramsey–Bordé interferometers) would be better suited for such an experiment. The formalism introduced in this paper to deal with antiatoms should be useful to discuss such experiments, especially when coherent beams of antihydrogen will be produced either by Bose–Einstein condensation and/or by stimulated bosonic amplification.

5.3 Analogy with the Electromagnetic Interaction

The formulas that we have derived for the transition amplitude and for the ougoing spinor strongly suggest analogies with the electromagnetic field case. To emphasize these analogies, let us introduce the following pseudo–potential 4–vector [10]

$$\widetilde{A}^{\mu} = \frac{1}{2}\widetilde{h}^{\mu\nu}p_{\nu}\,, \tag{143}$$

then the Linet–Tourrenc term, which appears also in the generalized Thomas precession is simply

$$\frac{c^2}{2E(\boldsymbol{p})}p_{\mu}\widetilde{h}^{\mu\nu}p_{\nu} = \frac{1}{\gamma}u_{\mu}\widetilde{A}^{\mu}\,, \tag{144}$$

where u_{μ} is the 4–velocity p_{μ}/m.

The corresponding field

$$\widetilde{\Phi}_{\mu\nu} = -i\left(\kappa_{\mu}\widetilde{A}_{\nu} - \kappa_{\nu}\widetilde{A}_{\mu}\right) = (\widetilde{\boldsymbol{E}}/c, \widetilde{\boldsymbol{B}}) \tag{145}$$

[10] More rigorously, one should introduce [33]

$$\widetilde{A}^{\mu} = \frac{1}{2}\widetilde{h}^{\mu\nu}(p_{\nu} + \hbar\kappa_{\nu}/2)\,, \tag{140}$$

which stems directly from a compact form of the interaction Hamiltonian

$$\mathcal{V}_G = \frac{c}{4}\alpha^{\mu}h_{\mu\nu}p^{\nu} + h.c. = \frac{c}{4}\left\{\alpha^{\mu}h_{\mu\nu}, p^{\nu}\right\}_{+} \tag{141}$$

$$\text{with } p^0 = -\alpha^j p_j + \gamma^0 mc \text{ and } p_j = i\hbar\partial_j \tag{142}$$

appears in the outgoing spinor (135) through

$$-\frac{i\hbar}{2}\kappa_\rho\sigma^{\rho\nu}\widetilde{h}_{\mu\nu}p^\mu=\frac{\hbar}{2}\sigma^{\rho\nu}\widetilde{\Phi}_{\rho\nu}$$

and in the generalized spin-gravitation interaction term

$$-\frac{i\hbar c}{4\gamma}\left[\boldsymbol{k}\times\left(\widetilde{\boldsymbol{h}}-\overset{\widetilde{\Rightarrow}}{h}\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}=-\frac{\hbar c^2}{2\gamma E(\boldsymbol{p})}w^{(r)\dagger}\boldsymbol{a}w^{(r')}\cdot\widetilde{\boldsymbol{B}}\,. \quad (146)$$

This new correspondence between the gravitational interaction and the electromagnetic interaction generalizes the so–called gravitoelectric and gravitomagnetic interactions introduced by de Witt [45] and Papini [46].

Appendix A: Dirac Equation in Curved Space–Time

For a given space–time manifold, together with its metric tensor $g_{\mu\nu}$, the Lagrangian density of the Dirac field reads [14]

$$\mathcal{L}=\frac{\hbar c}{2}\sqrt{-g}\,\overline{\Psi}\left[i\gamma^\alpha\overrightarrow{\mathcal{D}}_\alpha-mc/\hbar\right]\Psi+\frac{\hbar c}{2}\sqrt{-g}\,\overline{\Psi}\left[-i\overleftarrow{\mathcal{D}}_\alpha\gamma^\alpha-mc/\hbar\right]\Psi, \quad (147)$$

with

$$\overrightarrow{\mathcal{D}}_\alpha=e_{\widehat{\alpha}}{}^\mu\left[\overrightarrow{\partial}_\mu-\frac{i}{4}e^\nu_{\widehat{\beta}}\nabla_\mu e_{\widehat{\gamma}\nu}\sigma^{\beta\gamma}\right],\quad \overleftarrow{\mathcal{D}}_\alpha=\left[\overleftarrow{\partial}_\mu+\frac{i}{4}e^\nu_{\widehat{\beta}}\nabla_\mu e_{\widehat{\gamma}\nu}\sigma^{\beta\gamma}\right]e_{\widehat{\alpha}}{}^\mu\,. \quad (148)$$

In these formulas the $e_{\widehat{\alpha}}$'s are four-vector fields constituting an orthonormal tetrad (tetrad indices have a hat), that is satisfying the relations

$$g_{\mu\nu}\,e_{\widehat{\alpha}}{}^\mu e_{\widehat{\beta}}{}^\nu=\eta_{\alpha\beta}, \quad (149)$$

with $(\eta_{\alpha\beta})=diag(+1,-1,-1,-1)$. A change of tetrad is possible according to the formula

$$e'_{\widehat{\alpha}}{}^\mu=\Lambda^\beta{}_\alpha e_{\widehat{\beta}}{}^\mu, \quad (150)$$

where Λ is a matrix of the Lorentz group defined at each point of the spacetime manifold. In such a change the Lagrangian density (147) is left invariant if the Dirac field is correspondingly transformed according to the law

$$\Psi'=S(\Lambda)^{-1}\Psi, \quad (151)$$

the matrix $S(\Lambda)$ being the usual transformation matrix under a Lorentz transformation of a Dirac spinor [17]. For an infinitesimal transformation

$$\Lambda_{\alpha\beta}=\eta_{\alpha\beta}+\varepsilon_{\alpha\beta},\qquad \varepsilon_{\alpha\beta}=-\varepsilon_{\beta\alpha}, \quad (152)$$

one has

$$S(\Lambda)^{-1}=I+\frac{i}{4}\varepsilon_{\alpha\beta}\sigma^{\alpha\beta}. \quad (153)$$

The field equations read

$$\left[i\gamma^\alpha \overrightarrow{\mathcal{D}}_\alpha - mc/\hbar\right]\Psi = 0\,, \qquad \overline{\Psi}\left[-i\overleftarrow{\mathcal{D}}_\alpha\gamma^\alpha - mc/\hbar\right] = 0. \tag{154}$$

With the invariance of $\mathcal{L}$ under the phase transformations $\Psi \to \Psi e^{-i\alpha}, \overline{\Psi} \to \overline{\Psi}e^{i\alpha}$, is associated the current density

$$j^\mu = \sqrt{-g}\, J^\mu\,, \qquad J^\mu = ce^\mu_{\widehat{\alpha}}\,\overline{\Psi}\gamma^\alpha\Psi, \tag{155}$$

where J^μ is a four-vector invariant under a change of tetrad, and one has the conservation relation in either one of the two equivalent forms

$$\partial_\mu j^\mu = 0\,, \qquad \nabla_\mu J^\mu = 0\,. \tag{156}$$

Appendix B: Weak-Field Approximation

It is now assumed that space-time admits a coordinate system (x^μ) in which the metric tensor takes the form

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \qquad\qquad |h_{\mu\nu}| \ll 1. \tag{157}$$

According to that hypothesis, the $h_{\mu\nu}$'s will be considered as first-order quantities, and the subsequent calculations will be valid at this order. To determine the corresponding form of the Lagrangian, it suffices to construct the tetrads associated with (157). By putting $e_{\widehat{\alpha}\mu} = \eta_{\alpha\mu} + f_{\alpha\mu}$, where $f_{\alpha\mu}$ is of the first order, one obtains from (149) and (157),

$$f_{\alpha\beta} + f_{\beta\alpha} = h_{\alpha\beta}. \tag{158}$$

The general solution of these equations is of the form

$$f_{\alpha\beta} = \tfrac{1}{2}h_{\alpha\beta} + \epsilon_{\alpha\beta}\,, \tag{159}$$

in which the $\epsilon_{\alpha\beta}$'s are first-order quantities only restricted by the antisymmetry condition $\epsilon_{\alpha\beta} = -\epsilon_{\beta\alpha}$. One then has

$$e_{\widehat{\alpha}\mu} = \eta_{\alpha\mu} + \tfrac{1}{2}h_{\alpha\mu} + \epsilon_{\alpha\mu}. \tag{160}$$

Introducing the free Dirac Lagrangian (2) and the associated stress-energy tensor (3), the Lagrangian density (147) calculated at the first order and corresponding to (160) reads[11]

$$\mathcal{L} = \mathcal{L}_0 - \tfrac{1}{2}h^{\mu\nu}T_{\mu\nu} + \frac{i\hbar c}{2}\epsilon^{\mu\nu}\overline{\Psi}(\gamma_\mu\overrightarrow{\partial}_\nu - \overleftarrow{\partial}_\nu\gamma_\mu)\Psi - \frac{\hbar c}{8}\partial_\mu\epsilon_{\nu\rho}\overline{\Psi}(\gamma^\mu\sigma^{\nu\rho} + \sigma^{\nu\rho}\gamma^\mu)\Psi. \tag{161}$$

[11] The indices of $h_{\mu\nu}$ and $\epsilon_{\mu\nu}$ are raised with the help of $\eta^{\mu\nu}$.

The corresponding field equations are the following

$$0 = [(1+\tfrac{1}{2}h)(i\gamma^\mu \overrightarrow{\partial}_\mu - mc/\hbar) - \frac{i}{2}h^{\mu\nu}\gamma_\mu \overrightarrow{\partial}_\nu + \frac{i}{4}(\partial_\mu h - \partial_\nu h^\nu{}_\mu)\gamma^\mu + i\epsilon^{\mu\nu}\gamma_\mu \overrightarrow{\partial}_\nu - \frac{1}{4}\partial_\mu \epsilon_{\nu\rho}\gamma^\mu \sigma^{\nu\rho}]\Psi\,, \tag{162}$$

$$0 = \overline{\Psi}[(-i\gamma^\mu \overleftarrow{\partial}_\mu - mc/\hbar)(1+\tfrac{1}{2}h) + \frac{i}{2}\overleftarrow{\partial}_\nu \gamma_\mu h^{\mu\nu} - \frac{i}{4}(\partial_\mu h - \partial_\nu h^\nu{}_\mu)\gamma^\mu - i\overleftarrow{\partial}_\nu \gamma_\mu \epsilon^{\mu\nu} - \frac{1}{4}\partial_\mu \epsilon_{\nu\rho}\sigma^{\nu\rho}\gamma^\mu]\,. \tag{163}$$

The terms depending on $h_{\mu\nu}$ are identical to those appearing in (5) and (6). The same is true for the current density, which is now

$$j^\mu = c\overline{\Psi}[\gamma^\mu + \tfrac{1}{2}h\gamma^\mu - \tfrac{1}{2}h^{\mu\nu}\gamma_\nu - \epsilon^{\mu\nu}\gamma_\nu]\Psi. \tag{164}$$

Considered independently of the context, the equations (162) and (163) are invariant under the transformations of the Poincaré group provided that the $h_{\mu\nu}$'s and the $\epsilon_{\mu\nu}$'s are transformed like the components of a second–rank tensor, and Ψ by the corresponding transformation law of a spinor. However, the weak–field character of $h_{\mu\nu}$ is not conserved by any finite Lorentz transformation except by the rotations. Moreover, the $h_{\mu\nu}$'s being the basic quantities, the condition that the $\epsilon_{\mu\nu}$'s are of first order is naturally interpreted by assuming that these latter are some linear functions of the former, that is one has

$$\epsilon_{\mu\nu} = \alpha_{\mu\nu\rho\sigma}h^{\rho\sigma}. \tag{165}$$

Such a relation is not compatible with the general tensor transformation law, on account of the symmetric or antisymmetric character of $h_{\mu\nu}$ or $\epsilon_{\mu\nu}$, but can be made compatible with the rotations by a suitable choice of the coefficients $\alpha_{\mu\nu\rho\sigma}$.

The rotational invariance of (165) is equivalent to the relations

$$\alpha^{\mu'\nu'\rho'\sigma'} = R_\mu{}^{\mu'}R_\nu{}^{\nu'}R_\rho{}^{\rho'}R_\sigma{}^{\sigma'}\alpha^{\mu\nu\rho\sigma}, \tag{166}$$

where R can be any rotation matrix. The corresponding solutions are given by[12]

$$\alpha^{0i0j} = \tfrac{1}{2}A\delta^{ij}\,, \quad \alpha^{0ijk} = 0\,, \quad \alpha^{jk0i} = \tfrac{1}{2}B\varepsilon^{ijk}\,, \quad \alpha^{ijkl} = 0\,, \tag{167}$$

where A and B are some arbitrary parameters. If, in addition, the invariance under parity is postulated, one has to take $B = 0$, giving

$$\epsilon_{0i} = -\epsilon_{i0} = -Ah_{0i}\,, \quad \epsilon_{ij} = 0\,, \tag{168}$$

then, instead of (160),

$$e_{\widehat{0}0} = 1 + \tfrac{1}{2}h_{00}\,, \quad e_{\widehat{0}i} = (\tfrac{1}{2} - A)h_{0i}\,, \quad e_{\widehat{i}0} = (\tfrac{1}{2} + A)h_{0i}\,, \quad e_{\widehat{ij}} = \eta_{ij} + \tfrac{1}{2}h_{ij}\,. \tag{169}$$

[12] As usual, the latin indices can take the values 1,2 or 3, and ε^{ijk} is the completely antisymmetric symbol.

In particular the choice $A = 0$ corresponds to the tetrad

$$e_{\widehat{\alpha}\mu} = \eta_{\alpha\mu} + \tfrac{1}{2}h_{\alpha\mu}\,, \tag{170}$$

which we call the *standard tetrad* associated with $h_{\mu\nu}$. The corresponding Lagrangian is identical to the Lagrangian (1), therefore the equations (5) and (6) are recovered from (162) and (163) by letting $\epsilon_{\mu\nu} = 0$. We will continue to designate the corresponding field by Ψ.

If we write explicitly the spinorial connection in this weak-field approximation

$$-\frac{i}{4}e^{\nu}_{\widehat{\beta}}\nabla_{\mu}e_{\widehat{\gamma}\nu}\sigma^{\beta\gamma} = \frac{i}{4}\sigma^{\lambda\mu}\partial_{\lambda}h_{\mu\nu} \tag{171}$$

the equation for Ψ can also be written in a simple form, analogous to the electromagnetic case, and which will find an interpretation in terms of gauge fields in the flat space–time approach of the main text

$$i\gamma^{\nu}(\overrightarrow{\partial}_{\nu} + \frac{i}{4}\sigma^{\lambda\mu}\partial_{\lambda}h_{\mu\nu} - \frac{1}{2}h_{\nu}{}^{\alpha}\overrightarrow{\partial}_{\alpha})\Psi - \frac{mc}{\hbar}\Psi = 0. \tag{172}$$

Another choice of tetrad can be made in relation with the expression of the vector current. From (155) written in the weak–field case, one has in general

$$j^{\mu}/c = (1 + \tfrac{1}{2}h)\overline{\Psi}\gamma^{\mu}\Psi - \tfrac{1}{2}h^{\mu\nu}\overline{\Psi}\gamma_{\nu}\Psi - \epsilon^{\mu\nu}\overline{\Psi}\gamma_{\nu}\Psi, \tag{173}$$

and then

$$j^{0}/c = (1 + \tfrac{1}{2}h)\overline{\Psi}\gamma^{0}\Psi - (\tfrac{1}{2}h^{0\nu} + \epsilon^{0\nu})\overline{\Psi}\gamma_{\nu}\Psi. \tag{174}$$

This quantity will be proportional to the usual density $\overline{\Psi}\gamma^{0}\Psi$ if one has $\epsilon^{0i} = -\frac{1}{2}h^{0i}$, which is obtained with the choice $A = \frac{1}{2}$. Denoting by Ψ' the corresponding field, one has[13]

$$j^{0} = c(1 + \tfrac{1}{2}h^{i}{}_{i})\overline{\Psi'}\gamma^{0}\Psi'\,. \tag{175}$$

It is then possible to introduce the new field Θ by [12]

$$\Theta = (1 + \tfrac{1}{2}h^{i}{}_{i})^{\frac{1}{2}}\Psi' \simeq (1 + \tfrac{1}{4}h^{i}{}_{i})\Psi'\,, \tag{176}$$

such that

$$j^{0} = c\overline{\Theta}\gamma^{0}\Theta, \tag{177}$$

as in the free–field case. The change of field $\Psi \to \Psi'$ corresponds to a change of tetrad as defined above, so that, from (160) and (170), one has the formulas (151), (152) and (153) written with an infinitesimal parameter ε equal to $-\epsilon$ that is such that $\varepsilon_{ij} = 0$, $\varepsilon_{0i} = -\varepsilon_{i0} = \frac{1}{2}h_{0i}$. This gives

$$\Psi' = (I - \tfrac{1}{4}h_{0i}\gamma^{0}\gamma^{i})\Psi, \tag{178}$$

then, from (176),

$$\Theta = (I + \tfrac{1}{4}h - \tfrac{1}{4}h_{0\mu}\gamma^{0}\gamma^{\mu})\Psi. \tag{179}$$

The field Θ is identical to the field introduced in Section 3.2 for the purpose of defining the quantization and the interaction picture.

[13] Let us note that, in the expression of the conserved charge corresponding to (175), the volume element $(1 + \frac{1}{2}h^{i}{}_{i})d^{3}x$ is, in the first–order approximation, the 3–volume associated with the spatial metric defined from the coordinate system (x^{μ})[20].

Appendix C: A Stationary Phase Calculation

The outgoing spinor can be calculated directly in configuration space with the help of a stationary phase formula. In the particle case, from the definition

$$\chi(x) = \langle \phi_0 | \theta^{(+)}(x) | \chi(t) \rangle \tag{180}$$

and the integral equation

$$|\chi(t)\rangle = |\chi(t_0)\rangle - \frac{i}{\hbar c} \int_{t_0}^{t} (d^4x') \theta^{(+)\dagger}(x') \mathcal{V}_G(x') \theta^{(+)}(x') |\chi(t')\rangle \tag{181}$$

one can derive, to first order, the expression

$$\chi(x) = \chi_i(x) - \frac{i}{\hbar c} \int_{t_0}^{t} (d^4x') \langle \phi_0 | \theta^{(+)}(x) \overline{\theta^{(+)}(x')} | \phi_0 \rangle \gamma^0 \mathcal{V}_G(x') \langle \phi_0 | \theta^{(+)}(x') | \chi_i \rangle, \tag{182}$$

which, with the help of standard formulas, can be transformed into

$$\chi(x) = \chi_i(x) - \frac{1}{\hbar c} \int_{t_0}^{t} d^4x' S^{(+)}(x - x') \gamma^0 \mathcal{V}_G(x') \chi_i(x') \tag{183}$$

$$= \chi_i(x) - \frac{i}{\hbar c} \sum_r \int d^3p \chi^{(+)}_{\boldsymbol{p},r}(x) \int_{t_0}^{t} (d^4x') \overline{\chi^{(+)}_{\boldsymbol{p},r}(x')} \gamma^0 \mathcal{V}_G(x') \chi_i(x'). \tag{184}$$

By introducing (85), one finds

$$\chi(x) = \chi_i(x) - \frac{i}{\hbar c} \sum_r \int d^3p \chi^{(+)}_{\boldsymbol{p},r}(x)\ i\hbar c \langle \chi^{(+)}_{\boldsymbol{p},r} | U^{(1)}(t, t_0) | \chi_i \rangle. \tag{185}$$

In this last expression the matrix element can be submitted to the same transformations as those leading from (86) to (90), yielding

$$\chi(x) = \chi_i(x) - i \sum_r \int d^3p \chi^{(+)}_{\boldsymbol{p},r}(x) \int_{t_0}^{t} d^4x'\ \frac{i}{4} h^{\mu\nu}(x') \overline{\chi^{(+)}_{\boldsymbol{p},r}(x')} \left(\gamma_\mu \overrightarrow{\partial'_\nu} - \overleftarrow{\partial'_\nu} \gamma_\mu \right) \chi_i(x')$$

$$= \chi_i(x) - \frac{i}{4} \int_{t_0}^{t} d^4x'\, h^{\mu\nu}(x') S^{(+)}(x - x') \left(\gamma_\mu \overrightarrow{\partial'_\nu} - \overleftarrow{\partial'_\nu} \gamma_\mu \right) \chi_i(x'). \tag{186}$$

By introducing the expression

$$S^{(+)}(x - x') = \frac{ic}{2(2\pi\hbar)^3} \int \frac{d^3p}{E(\boldsymbol{p})} (\gamma^\mu p_\mu + mc) e^{-ip(x-x')/\hbar} \Big|_{p_0 = +E(\boldsymbol{p})/c}, \tag{187}$$

one finally gets for a plane wave with the momentum $\boldsymbol{p}_0$

$$\chi(x) = \chi_i(x) - \frac{ic^2}{8\hbar} \int_{t_0}^{t} dt' \int d^3x' h^{\mu\nu}(x') \times$$

$$\times \frac{1}{(2\pi\hbar)^3} \int \frac{d^3p}{E(\boldsymbol{p})} (p + p_0)_\nu (\gamma^\rho p_\rho + mc) e^{-ip(x-x')/\hbar} \gamma_\mu \chi_i(x')\,. \tag{188}$$

The elementary stationary phase formula[14] applied successively to the integral over p, then to the integral over x', yields the Linet–Tourrenc result [23], namely :

$$\chi(x) = \left[1 - \frac{ic^2}{2\hbar}\frac{p_{0\mu}p_{0\nu}}{E(\overrightarrow{p_0})}\int_{t_0}^{t} dt' h^{\mu\nu}(\boldsymbol{x} - \boldsymbol{v}_0(t-t'), t')\right]\chi_i(x)\,, \qquad \boldsymbol{v}_0 = \frac{\boldsymbol{p}_0 c^2}{E(\boldsymbol{p}_0)}\,. \tag{189}$$

This expression is the beginning of an asymptotic expansion in powers of $\hbar$ of the form

$$\chi(x) = \left[1 - \frac{ic^2}{2\hbar}\int_{t_0}^{t} dt'\, F(\boldsymbol{x} - \boldsymbol{v}_0(t-t'), t')\right]\chi_i(x), \tag{190}$$

in which F is defined by an expansion in integer powers of $\hbar$ whose first term corresponds to the Linet–Tourrenc formula. This expansion can be obtained from a generalization of the stationary phase formula given by Hörmander [25]. The application of this general formula to the two integrals appearing in (188) yields, after some complicated calculations, the following expression of the outgoing spinor (to be derived in the next Appendix by a simpler method)

$$F = \left\{\frac{p_\mu p_\nu h^{\mu\nu}}{E(\boldsymbol{p})} + \frac{i\hbar}{2E(\boldsymbol{p})}\left[p_\nu \partial_i h^{\mu\nu}\gamma^i\gamma_\mu + p_\mu \partial_i h^{\mu i} - p_\mu \frac{v^i}{c}\left(\partial_i h^{\mu 0} + \partial_i h^{\mu\nu}\gamma^0\gamma_\nu\right)\right.\right.$$
$$\left.\left. + 2\frac{p_\mu p_\nu}{E(\boldsymbol{p})}v^i \partial_i h^{\mu\nu} + \frac{c^2(t-t')}{E(\boldsymbol{p})}\left(\delta^{ij} - \frac{v^i}{c}\frac{v^j}{c}\right)\partial_i\partial_j\left(p_\mu p_\nu h^{\mu\nu}\right)\right]\right\}_{\boldsymbol{p}=\boldsymbol{p}_0} \tag{191}$$

Appendix D: Derivation of the Wave Function Using the Momentum Representation

The method used for the amplitude in the main text can be used to derive the wave function. The spinorial wave function for one–(anti)particle states is

$$\chi(x) = \langle\phi_0|\,\theta(x)\,|\chi(t)\rangle\ . \tag{192}$$

For particles:

$$\langle\phi_0|\,\theta^{(+)}(x)\,|\chi(t)\rangle = \langle\phi_0|\,\theta^{(+)}(x)\,|\chi_i\rangle \tag{193}$$
$$-\frac{i}{\hbar}\frac{1}{(2\pi\hbar)^{3/2}}\sum_{r,r'}\int_{t_0}^{t} dt' \int d^3p \int \frac{d^3k}{(2\pi)^{3/2}}\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}u^{(r)}(\boldsymbol{p}+\hbar\boldsymbol{k})u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})$$
$$\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t')e^{i\boldsymbol{k}\cdot\boldsymbol{x}}e^{\frac{i}{\hbar}[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})](t'-t)}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p})e^{\frac{i}{\hbar}(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)}\langle 1_{\boldsymbol{p},r'}|\chi(t')\rangle$$

[14] See, for instance [24].

and a similar formula for antiparticles. To first order we get

$$\chi^{(1)}(x) = \tag{194}$$
$$-\frac{i}{\hbar}\frac{1}{(2\pi\hbar)^{3/2}}\sum_{r,r'}\int_{t_0}^{t}dt'\int d^3p\int\frac{d^3k}{(2\pi)^{3/2}}\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}u^{(r)}(\boldsymbol{p}+\hbar\boldsymbol{k})u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})$$
$$\widetilde{\mathcal{V}}_G(\boldsymbol{k},\boldsymbol{p},t')e^{i\boldsymbol{k}\cdot\boldsymbol{x}}e^{\frac{i}{\hbar}[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})](t'-t)}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p})e^{\frac{i}{\hbar}(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)}\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle$$

The next idea is to express the propagator which appears in this equation in order to write the outgoing spinor in the form of Dirac matrices multiplying the initial spinor plane wave components.

From

$$\sum_r u^{(r)}(\boldsymbol{p})\overline{u}^{(r)}(\boldsymbol{p}) = \frac{1}{2mc}\left[\gamma^\mu p_\mu + mc\right] \tag{195}$$

one obtains

$$\sum_r \frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}u^{(r)}(\boldsymbol{p}+\hbar\boldsymbol{k})u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k}) =$$
$$\frac{c}{2E(\boldsymbol{p})}\left[\gamma^\mu\left(p_\mu+\hbar\kappa_\mu\right)+mc\right]\left(1-\frac{\hbar\kappa^0 c}{E(\boldsymbol{p})}\right)\gamma^0\,, \tag{196}$$

from which, after some algebra, one finds the spinor

$$\chi^{(1)}(x) = -\frac{i}{\hbar}\frac{1}{(2\pi\hbar)^{3/2}}\sum_{r'}\int_{t_0}^{t}dt'\int(d^3p)\int\frac{d^3k}{(2\pi)^{3/2}}\frac{c^2}{2E(\boldsymbol{p})}$$
$$\left\{\left(p^\mu+\hbar\kappa^\mu\right)\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{E(\boldsymbol{p})}\right)-\frac{i\hbar}{2}\kappa_\rho\sigma^{\rho\nu}\widetilde{h}_{\mu\nu}p^\mu\right\} \tag{197}$$
$$e^{i\boldsymbol{k}.\boldsymbol{x}}e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})](t'-t)/\hbar}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p})e^{i(\boldsymbol{p}\cdot\boldsymbol{x}-E(\boldsymbol{p})t)/\hbar}\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle$$

where the time–dependent exponential can be expanded to any desired order for recoil shift corrections:

$$e^{i[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})](t'-t)/\hbar} = \tag{198}$$
$$e^{i\boldsymbol{k}\cdot\boldsymbol{v}(t'-t)}\left[1+i\hbar\frac{c^2(t'-t)}{2E(\boldsymbol{p})}\left(\boldsymbol{k}^2-\left(\frac{c\boldsymbol{p}\cdot\boldsymbol{k}}{E(\boldsymbol{p})}\right)^2\right)\right]\simeq e^{i\boldsymbol{k}\cdot\boldsymbol{v}(t'-t)}\left[1+i\delta(t'-t)\right]\,.$$

Finally one can check that:

$$\langle\chi_{\rm ref}|\,U^{(1)}(t,t_0)\,|\chi(t_0)\rangle = \int d^3x\;\chi^\dagger_{\rm ref}(x)\chi^{(1)}(x)\,. \tag{199}$$

We obtain indeed

$$\int d^3x\, \chi^{\dagger}_{\rm ref}(x)\chi^{(1)}(x) = -\frac{i}{\hbar}\sum_{r,r'}\int_{t_0}^{t}dt'\int d^3p \int \frac{d^3k}{(2\pi)^{3/2}} e^{\frac{i}{\hbar}[E(\boldsymbol{p}+\hbar\boldsymbol{k})-E(\boldsymbol{p})]t'}$$

$$\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}}u^{(r)\dagger}(\boldsymbol{p}+\hbar\boldsymbol{k})\ \ \langle\chi_{\rm ref}|\, 1_{\boldsymbol{p}+\hbar\boldsymbol{k},r}\rangle \frac{c^2}{2E(\boldsymbol{p})} \tag{200}$$

$$\left\{(p^\mu+\hbar\kappa^\mu)\,\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{E(\boldsymbol{p})}\right)-\frac{i\hbar}{2}\kappa_\rho\sigma^{\rho\nu}\widetilde{h}_{\mu\nu}p^\mu\right\}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p})\langle 1_{\boldsymbol{p},r'}\,|\chi_i\rangle\ ,$$

with

$$\sqrt{\frac{mc^2}{E(\boldsymbol{p}+\hbar\boldsymbol{k})}}u^{\dagger(r)}(\boldsymbol{p}+\hbar\boldsymbol{k})\frac{c^2}{2E(\boldsymbol{p})}$$

$$\left\{(p^\mu+\hbar\kappa^\mu)\,\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{E(\boldsymbol{p})}\right)-i\frac{\hbar}{2}\kappa_\rho\sigma^{\rho\nu}\widetilde{h}_{\mu\nu}p^\mu\right\}\sqrt{\frac{mc^2}{E(\boldsymbol{p})}}u^{(r')}(\boldsymbol{p}) \tag{201}$$

$$= \frac{c^2}{2E(\boldsymbol{p})}\,(p^\mu+\hbar\kappa^\mu)\,\widetilde{h}_{\mu\nu}p^\nu\left(1-\frac{\hbar\kappa^0 c}{2E(\boldsymbol{p})}\right)\delta_{rr'}+\frac{i\hbar}{2m(\gamma+1)}\left[(\boldsymbol{k}\times\boldsymbol{p})\frac{c^2p^\mu\widetilde{h}_{\mu\nu}p^\nu}{2E^2(\boldsymbol{p})}\right]$$

$$\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}-\frac{i\hbar c}{4\gamma}\left[\boldsymbol{k}\times\left(\widetilde{\boldsymbol{h}}-\overset{\widetilde{\Rightarrow}}{h}\cdot\frac{\boldsymbol{p}c}{E(\boldsymbol{p})}\right)\right]\cdot w^{(r)\dagger}\boldsymbol{a}w^{(r')}\,,$$

from which equation (132) is recovered.

Acknowledgement

One of us (Ch.J. B.) would like to thank Dr. C. Lämmerzahl and Prof. Ph. Tourrenc for many stimulating discussions and Prof. Dr. W. Ertmer for his hospitality at the University of Hannover within the Sonderforschungsbereich 407.

References

1. P. Laurent *et al.*: *Cold atom clocks on earth and in space*, in R. Blatt, J. Eschner, D. Leibfried and F. Schmidt–Kaler (eds.): *Laser Spectroscopy*, Proceedings of the 14th International Conference on Laser Spectroscopy (World Scientific, Singapore 1999), p. 41.
2. P. Berman (ed.): *Atom Interferometry* (Academic Press 1997).
3. Ch.J. Bordé: Atomic interferometry and laser spectroscopy, in: *Laser Spectroscopy X* (World Scientific, Singapore 1991), p. 239.
4. U. Sterr *et al.*: Atom interferometry based on separated light fields, in [2] and *Appl. Phys.* **B 54**, 341 (1992).
5. F. Riehle, Th. Kisters, A. Witte, J. Helmcke and Ch.J. Bordé: Optical Ramsey Spectroscopy in a Rotating Frame: Sagnac Effect in a Matter-Wave Interferometer, *Phys. Rev. Lett.* **67**, 177-180 (1991).

6. B.C. Young, M. Kasevich and S. Chu: Precision atom interferometry with light pulses, in [2] and references therein.
7. Ch.J. Bordé: Atomic interferometry with internal state labelling, *Phys. Lett.* **A 140**, 10 (1989).
8. Ch.J. Bordé *et al.*: Optical Ramsey fringes with travelling waves, *Phys. Rev.* **A 30**, 1836 (1984).
9. Ch.J. Bordé *et al.*: Molecular Interferometry Experiments, *Phys. Lett.* **A 188**, 187 (1994).
10. Ch.J. Bordé: Matter–wave interferometers: a synthetic approach, in [2].
11. For an early treatment of a gravitational wave detector using an atom interferometric gradiometer see: Ch.J. Bordé, J. Sharma, Ph. Tourrenc and Th. Damour: Theoretical approaches to laser spectroscopy in the presence of gravitational fields, *J. Physique Lettres* **44**, L-983 (1983).
12. Ch.J. Bordé, A. Karasiewicz and Ph. Tourrenc: General relativistic framework for atomic interferometry, *Int. J. of Mod. Phys.* **D 3**, 157 (1994).
13. F.W. Hehl and Wei–Tou Ni: Inertial effects of a Dirac particle, *Phys. Rev.* **D 42**, 2045 (1990).
14. S. Weinberg: *Gravitation and Cosmology* (John Wiley and Sons, New York 1972).
15. S.N. Gupta: Quantization of Einstein's gravitational field: linear approximation, *Proc. Phys. Soc.* **A 65**, 161 (1952).
16. S.N. Gupta: Quantization of Einstein's gravitational field: general treatment, *Proc. Phys. Soc.* **A 65**, 608 (1952).
17. J.D. Bjorken and S.D. Drell: *Relativistic Quantum Mechanics*, (McGraw–Hill, New York 1964).
18. J.D. Bjorken and S.D. Drell: *Relativistic Quantum Fields* (McGraw–Hill, New York 1965).
19. R.P. Feynman, F.B. Morinigo and W.G. Wagner: *Feynman Lectures on Gravitation,* edited by B. Hatfield (Addison–Wesley, Reading MA 1995).
20. L.D. Landau and E.M. Lifschitz: *The Classical Theory of Fields* (Addison–Wesley, Reading MA 1951).
21. S.S. Schweber: *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York 1961).
22. Ch.J. Bordé: Quantum theory of clocks and of gravitational sensors using atom interferometry, in R. Blatt, J. Eschner, D. Leibfried and F. Schmidt–Kaler (eds.): *Laser Spectroscopy*, Proceedings of the 14th International Conference on Laser Spectroscopy, (World Scientific, Singapore 1999), p. 160.
23. B. Linet et P. Tourrenc: Changement de phase dans un champ de gravitation: possibilité de détection interférentielle, *Can. J. Phys.* **54**, 1129 (1976).
24. M. Born and E. Wolf: *Principles of Optics*, third edition (Pergamon Press, Oxford 1964).
25. L. Hörmander: *The Analysis of Linear Partial Differential Operators*, Vol. 1, (Springer–Verlag, Berlin 1983).
26. V.B. Berestetskii, E.M. Lifschitz and L.P. Pitaevskii: *Relativistic Quantum Theory* (Pergamon Press, Oxford 1971).
27. V. Bargmann and E.P. Wigner: Group theoretical discussion of relativistic wave equations, *Proc. Nat. Acad. Sci.* (USA) **34**, 211 (1948).
28. L. de Broglie: *Théorie Générale des particules à spin (Méthode de fusion)* (Gauthier–Villars, Paris 1943).
29. E. Durand: 16–component theory of the spin–1 particle and its generalization to arbitrary spin, *Phys. Rev.* **D 11**, 3405 (1975).

30. L.H. Thomas: The Kinematics of an Electron with an Axis, *Phil. Mag.* **3**, 1 (1927).
31. J.K. Lubanski: Sur la théorie des particules élémentaires de spin quelconque.I., *Physica* **9**, 310 (1942).
32. J.M. Jauch and F. Rohrlich: *The Theory of Photons and Electrons* (Addison–Wesley, Cambridge MA 1955).
33. Ch.J. Bordé: *A comparison of electromagnetic and weak gravitational interactions in matter–wave interferometry*, to be published.
34. I.N. Bronshtein and K.A. Semendyayev: *Handbook of Mathematics* (Springer, Berlin 1997), p. 770.
35. A. Peters, K. Y. Chung and S. Chu: A measurement of gravitational acceleration by dropping atoms, *Nature* **400**, 849 (1999).
36. A. Landragin, T. L. Gustavson and M. A. Kasevich: Precision atomic gyroscope, in R. Blatt, J. Eschner, D. Leibfried and F. Schmidt–Kaler (eds.): *Laser Spectroscopy*, Proceedings of the 14th International Conference on Laser Spectroscopy (World Scientific, Singapore 1999), p. 170.
37. Ch.J. Bordé: *Propagation of Laser beams and of atomic systems*, in J. Dalibard *et al.* (eds.): *Fundamental Systems in Quantum Optics*,(Elsevier 1991), p. 287.
38. Ch.J. Bordé and C. Lämmerzahl: Atom beam interferometry as two–level particle scattering by a periodic potential, *Ann. Phys.* (Leipzig) **8**, 83 (1999).
39. C. Lämmerzahl and Ch.J. Bordé: Rabi oscillations in gravitational fields: exact solution, *Phys. Lett.* **A 203**, 59 (1995).
40. C. Lämmerzahl and Ch.J. Bordé: Atom interferometry in gravitational fields: influence of gravitation on the beam splitter, *Gen. Rel. Grav.*, **31**, 635 (1999).
41. J. Audretsch and K.–P. Marzlin: Atom interferometry with arbitrary laser configurations: exact phase shift for potentials including inertia and gravitation, *J. Phys. II* (France) **4**, 2073 (1994).
42. J. Audretsch and K.–P. Marzlin: Ramsey fringes in atomic interferometry: measurability of the influence of space–time curvature, *Phys. Rev.* **A 50**, 2080 (1994).
43. C. Lämmerzahl: Relativistic treatment of Raman light–pulse atom beam interferometer with applications in gravity theory, *J. Phys. II* (France) **4**, 2089 (1994).
44. P. Wolf and Ph. Tourrenc: Gravimetry using atom interferometers: some systematic effects, *Phys. Lett.* **A 251**, 241 (1999).
45. B.S. DeWitt: Superconductors and gravitational drag, *Phys. Rev. Lett.* **16**, 1092 (1966)
46. G. Papini: Particle wave functions in weak gravitational fields, *Nuovo Cimento* **52B**, 136 (1967).
47. T.J. Phillips: Measuring the gravitational acceleration of antimatter with an antihydrogen interferometer, *Hyperfine Interactions* **100**, 163 (1996).

Spin in Gravity

Wei–Tou Ni

Center for Gravitation and Cosmology
Department of Physics, National Tsing Hua University,
Hsinchu, Taiwan, 30055 Republic of China

Abstract. In these two talks, we report on the efforts to probe the role of spin and polarization in gravitation. After reviewing the motivation and historical background, we focus the talks on the experimental searches. These experimental searches are mainly of two categories: (i) laboratory searches (torsion–balance experiments, magnetic resonance experiments, SQUID experiments), and (ii) astrophysical and cosmological searches (pulsar observations, radio–galaxy observations, gamma–ray observations). We first discuss experimental searches for photon polarization coupling and then discuss experimental searches for electron spin–coupling. In the discussion of photon polarization coupling, we review the astrophysical and cosmological electromagnetic propagation observations. In the discussion of electron spin–coupling, we review the weak equivalence principle experiments, the finite–range spin coupling experiments, the spin–spin coupling experiments and the cosmic–spin coupling experiments. We discuss two recent laboratory experiments, a SQUID experiment and a torsion–balance experiment in detail to illustrate the experimental techniques. The ultimate searches for the role of spin in gravitation is to measure the gyrogravitational ratio. A discussion of the strategies to perform such experiments conclude these two talks.

1 Introduction

According to our present understanding of physics, particles and fields transform appropriately under inhomogeneous Lorentz transformations. These inhomogeneous Lorentz transformations form a group called the Poincaré group. The only invariants characterizing irreducible representations of the Poincaré group are mass and spin (or helicity in the case of zero mass). Both electroweak and strong interactions are strongly spin–dependent. The question comes whether the gravitational interaction is spin–dependent (polarization–dependent). In this paper, we review the searches for the role of spin in gravitation. The gravitational interaction is the earliest formulated interaction. Both Newtonian gravitation and Einstein's general relativity are universal–interaction theories about masses. There are no (direct) polarization–dependent effects in these theories. Historically, these theories were formulated before spin was discovered. Ever since the existence of spin (quantum spin) was noticed, people started to propose possible polarization–dependent effects in gravitation on various levels. If there are spin–dependent effects in gravitation, Einstein's Equivalence Principle (EEP) would be violated at a certain appropriate level. Since mass and spin (helicity) are two independent invariants of the Poincaré group, there is the question whether the gravitational interaction between masses and the "gravitational" interaction between spins have the same coupling constant. If the strengths of coupling

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 439–456, 2001.

are different, then the question comes whether we shall call the spin–spin or spin–mass interaction gravitational. This question can only be answered if the strengths are determined and a working theory (e.g., superstring theory) is formulated and adopted. From a phenomenological approach, we ask whether there is a long–range (or semi–long–range) spin interaction (in addition to electromagnetic spin interaction) and what is its strength and interaction form. Therefore, in reviewing the experimental searches, we include the related efforts.

In Section 2, we review the theoretical relations among the equivalence principles and the way leading to the axial interaction. In section 3, we discuss spin and gravitation emphasizing torsion and gyrogravitational ratio. In Section 4, we articulate on the inter–relations among spin, equivalence principle and (semi–)long–range forces. In Section 5, we review the experimental searches for photon polarization coupling and related tests of EEP. In Section 6, we review the experimental searches for electron spin–coupling; we discuss weak equivalence principle experiments, finite–range spin–coupling experiments, spin–spin coupling experiments and cosmic spin–coupling experiments in four subsections. In Section 7, we present an outlook.

2 WEP, EEP and the Axial Interaction

Equivalence principles [1,2] are cornerstones in the foundation of gravitation theories. In the theoretical study of the foundation problems, to what extent the Galileo weak equivalence principle {Universality of free–fall trajectories, (WEP I)} implies the validity of the Einstein equivalence principle (EEP) is an important issue. Schiff [3] conjectured that the Galileo weak equivalence principle implies the Einstein equivalence principle. In 1972, I started to investigate this issue and reached a counterexample of Schiff's conjecture [4]. In order to find out to what extent the violation occurs, I followed up using a general framework — the $\chi - g$ framework to study Schiff's conjecture and theoretical relations of various equivalence principles [5,6].

The $\chi - g$ framework can be summarized in the following interaction Lagrangian density

$$L_I = -\frac{1}{16\pi}\chi^{ijkl}F_{ij}F_{kl} - A_k j^k (-g)^{1/2} - \Sigma_I m_I \frac{ds_I}{dt}\delta(\mathbf{x} - \mathbf{x}_I) \tag{1}$$

where $\chi^{ijkl} = \chi^{klij} = -\chi^{klji}$ is a tensor density of the graviational fields (e.g., g_{ij}, ϕ, etc.) or fields to be investigated and j^k, $\mathrm{F}_{ij} \equiv A_{j,i} - A_{i,j}$ have the usual meaning. The gravitation constitutive tensor density χ^{ijkl} dictates the behavior of electromagnetism in a gravitational field and has 21 independent components in general. For a metric theory (when EEP holds), χ^{ijkl} is determined completely by the metric g^{ij} and equals $(-g)^{1/2}(\frac{1}{2}g^{ik}g^{jl} - \frac{1}{2}g^{il}g^{kj})$.

We proved that for a system whose Lagrangian density is given by Eq. (1), WEP I holds if and only if

$$\chi^{ijkl} = (-g)^{1/2}\left[\,\tfrac{1}{2}g^{ik}g^{jl} - \tfrac{1}{2}g^{il}g^{kj} + \phi\epsilon^{ijkl}\,\right], \tag{2}$$

where ϕ is a scalar function of the gravitational fields or fields to be investigated, and $\epsilon^{ijkl} = (-g)^{1/2} e^{ijkl}$ with

$$e^{ijkl} = \begin{cases} 1, & \text{if } (ijkl) \text{ is an even permutation of } (0123) \\ -1, & \text{if } (ijkl) \text{ is an odd permutation of } (0123), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

If $\phi \neq 0$ in (2), the gravitational coupling to electromagnetism is not minimal and EEP is violated. Hence WEP I does not imply EEP and Schiff's conjecture is incorrect [4–6]. However, WEP I does constrain the 21 degrees of freedom of χ to only one degree of freedom (ϕ), and Schiff's conjecture is largely right in spirit.

The theory with $\phi \neq 0$ is an axial theory with important astrophysical and cosmological consequences (Section 5). This is an example that investigations in fundamental physical laws lead to implications in cosmology. Investigations of CP problems in high energy physics leads to a theory with a similar piece of Lagrangian with ϕ the axion field [7–12] (Section 5.2).

In the nonmetric theory with χ^{ijkl} ($\phi \neq 0$) given by Eq. (2) [4–6], there are anomalous torques on electromagnetic–energy–polarized bodies so that different test bodies will change their rotation state differently, like magnets in magnetic fields. Since the motion of a macroscopic test body is determined not only by its trajectory but also by its rotation state, the motion of polarized test bodies will not be the same. We, therefore, have proposed the following stronger weak equivalence principle (WEP II) to be tested by experiments, which states that in a gravitational field, the motion of a test body with a given initial motion state is independent of its internal structure and composition (universality of freefall motion) [5,6]. Therefore, in this framework, the imposition of WEP II guarantees that EEP is valid.

Before mid–seventies, the actual weak equivalence experiments were performed on unpolarized bodies. These experiments constrained only 2 degrees of freedom of χ. Only when experiments are performed on polarized bodies with various electromagnetic energy configurations, can they constrain the other 18 degrees of freedom. This situation motivated us to study other existing and potential experimental and observational evidences for EEP, and to perform experiments on polarized bodies and to search for spin–dependent forces.

Our efforts to test equivalence principle for polarized bodies [13] and to search for spin–dependent forces [14–17] led us to search for axion–like forces also [18,19]. We discuss these searches in Section 6. Since the electromagnetic energy contents of laboratory polarized–bodies and unpolarized–bodies are small, other experimental and observational evidences are crucial in laying the foundation for the Einstein equivalence principle. In Section 5, we discuss the constraints from these evidences and how to generalize the $\chi - g$ framework to give a more general framework for testing the foundation of relativistic gravity including microscopic phenomena.

3 Spin and Gravitation

Ten years after the discovery of general relativity, in 1925-26, Goudsmit and Uhlenbeck [20] introduced our present concept of electron spin as the culmination of a series of studies of doublet and triplet structures in spectra. From the very beginning of its discovery, spin remains a microscopic object. One way to incorporate spin into the classical general relativity is to treat the aggregate of spins as ordinary angular momentum. This is a standard way to extend general relativity.

However, as we know, for the electromagnetic interaction, the gyromagnetic ratios of elementary particles are different from one, and these ratios reveal the inner electromagnetic structures of elementary particles. What would be the gyrogravitatinal ratios of elementary particles? If they differ from one, they will definitely reveal the inner gravitational structures of elementary particles. These will give clues to the microscopic origin of gravity.

The Stanford Orbiting Gyro Relativity (GP-B) experiment to be launched in 2001 is a very precise and difficult experiment [21]. It aims at detecting the Lense–Thirring effect. When this is done, and some of the rotating gyros are replaced by spin–polarized bodies, the gyrogravitational ratio will have a chance to be measured [22]. Microscopic experiments, when developed, may contribute to this very difficult task too (Section 7).

Theoretically, since a spin 1/2 particle is the most fundamental object in the consideration of quantum spin, we look into its inertia effects [23] and curvature effects of a Dirac particle in the standard theory of gravitation as extended by Cartan, Sciama, Kibble, and Hehl et al. [24–29]. In 1921, Eddington [30] mentioned the notion of an asymmetric affine connection in discussing possible extensions of general relativity. In 1922, Cartan [24] introduced torsion as the antisymmetric part of an asymmetric affine connection and laid the foundation of this generalized geometry. Cartan [25] proposed that the torsion of spacetime might be connected with the intrinsic angular momentum of matter. Following this idea, Sciama [27,28] and Kibble [26] developed a theory of gravitation which is commonly called the Einstein–Cartan–Sciama–Kibble theory. Hehl et al. [29] extended it to a dynamical torsion theory. The inertia effects include the Bonse–Wroblewski phase shift due to acceleration, the Sagnac–type effect, the rotation–spin effect, the redshift of the kinetic energy, and the inertial spin–orbit coupling [23]. The torsion effects are analyzed in [31,32] and references therein. The curvature effects give the gyrogravitational ratio.

The study of this problem from more basic points of view involves two approaches:

The first approach starts with the connection. After the formulation of gauge theory by Yang and Mills in 1954, many efforts have been made to bring the gravitation into the present gauge–theoretic framework. Yang [33] proposed his gravitational equation in 1974 along this line. However, there are spurious solutions [34] and the metric is postulated instead of derived. Affine connections correspond to gauge potentials. To be truly analogous to the present gauge theories, the metric ought to be derivable from the affine connection and the

equations of motion. To pursue this approach further, we first obtain the necessary and sufficient conditions for the existence of a metric in an affine manifold [35]. Now the problem comes as how to transform these conditions into equations of motion derivable from a variational principle. Ashtekar's formulation [36] of general relativity is an approach in this general direction.

The second approach starts from an extended framework for Dirac particle in a gravitational field and explores various possible relations between spin and gravity [37,38].

Metric–affine theories treat metric and affine connection more or less on equal footing. For recent research results, see [39,40] and references therein.

4 Spin, Equivalence Principle, and Long–Range Forces

The equivalence principle is an important cornerstone of universal gravitation. The precision of its validity puts an important constraint on gravitation theories and particle theories. Possible deviation from equivalence would give a clue to the microscopic origin of gravity or some new fundamental forces(s). In relation to spins, we look into polarization–dependent deviations from equivalence. In this respect, experiments with polarized entities play an especially important role.

Particles and fields transform appropriately under inhomogeneous Lorentz transformations which form the Poincaré group. The only invariants characterizing irreducible representations of the Poincaré group are mass and spin (or helicity in the case of zero mass). Gravitational interaction is a long–range mass–mass interaction. From a phenomenological approach, we ask whether there is a long–range (or semi–long–range) spin–mass or spin–spin interaction and what is its strength and form of interaction. Experiments on macroscopic spin–polarized bodies are sensitive tools to detect and study these possible interactions to a good precision.

In the new general relativity of Hayashi and Shirafuji [41], the coupling with an antisymmetric field leads to a universal spin–spin interaction. From gauging a sub–group of the Lorentz group, Naik and Pradhan [42,43] proposed a similar interaction. Around 1980, the particle physics community began to realize the possible existence of Goldstone bosons and/or pseudo–Goldstone bosons. These bosons generate (semi–)long–range forces of monopole–monopole type, monopole–dipole(spin) type, and dipole–dipole (spin–spin) type, just like the new general relativity of Hayashi and Shirafuji [41]. The axion [7–12] is such a pseudo–Goldstone boson. The recent issue of the fifth force arises from the existence of a semi–long–range coupling to baryon number/hyper–charge/lepton number. Attempts have been made to construct models of long–range forces in higher dimensional Kaluza–Klein type theories and superstring theories. An attempt to generalize Nambu–Goldstone mechanism shows that the restoration of a spontaneous violation of a "fact", such as the "fact" that translation generators in different directions commute, implies the existence of a massless excitation and, therefore, a long–range force. All the above cases can be explored exper-

imentally by gravitation–type experiments on macroscopic bodies–Eötvös–type experiments, Galileo–type ("free–fall") experiments and cavendish–type experiments [13].

5 Experimental Searches for Photon Polarization Coupling and Tests of EEP

In the $\chi - g$ framework, for a weak gravitational field,

$$\chi^{ijkl} = \chi^{(0)ijkl} + \chi^{(1)ijkl} \tag{4}$$

where

$$\chi^{(0)ijkl} = \frac{1}{2}\eta^{ik}\eta^{jl} - \frac{1}{2}\eta^{il}\eta^{kj} \tag{5}$$

with η^{ij} the Minkowski metric and all $|\chi^{(1)}| \ll 1$. In this field the dispersion relation for ω for a plane–wave propagating in the z–direction is

$$\omega_{\pm} = k\{1 + \frac{1}{4}[(K_1 + K_2) \pm \sqrt{(K_1 - K_2)^2 + 4K^2}]\} \tag{6}$$

where

$$\begin{aligned} K_1 &= \chi^{(1)1010} - 2\chi^{(1)1013} + \chi^{(1)1313}, \\ K_2 &= \chi^{(1)2020} - 2\chi^{(1)2023} + \chi^{(1)2323}, \\ K &= \chi^{(1)1020} - \chi^{(1)1023} - \chi^{(1)1320} + \chi^{(1)1323}. \end{aligned} \tag{7}$$

Photons with two different polarizations propagate with different speeds $V_{\pm} = \omega_{\pm}/k$ and would split in 4–dimensional spacetime. The conditions for no splitting (no retardation) is $\omega_+ = \omega_-$, i.e.,

$$K_1 = K_2\,, \qquad K = 0. \tag{8}$$

Eq. (8) gives two constraints on the $\chi^{(1)}$'s [44–46].

The condition for no splitting (no retardation) in all directions gives ten constraints on the $\chi^{(1)}$'s. With these ten constraints, χ can be written in the following form

$$\chi^{ijkl} = (-H)^{1/2}\left[\,\tfrac{1}{2}H^{ik}H^{jl} - \tfrac{1}{2}H^{il}H^{kj}\,\right]\psi + \phi e^{ijkl} \tag{9}$$

where $\mathrm{H} = \det(\mathrm{H}_{ij})$ [44–46]. With the null–birefringence observations of pulsar pulses and micropulses before 1980, these relations are empirically verified to $10^{-14} - 10^{-16}$ [44–46]. With the present pulsar observations, these limits would be improved. Analyzing the data from polarization measurements of extragalactic radio sources, Haugan and Kauffmann [47] inferred that the resolution for null–birefringence is 0.02 cylce at 5 GHz. This corresponds to a time resolution of 4×10^{-12} s. With a detailed analysis and more extragalactic radio observations, (9) would be tested down to $10^{-28} - 10^{-29}$ at cosmological distances. The

electromagnetic propagation in Moffat's nonsymmetric gravitational theory fits the $\chi - g$ framework. Krisher [48] and Haugan-Kauffmann [47] have used the pulsar data and extragalactic radio observations to constrain it. The effect of ϕ in (9) is to change the phase of two different circular polarizations in gravitation field and gives polarization rotation for linearly polarized light [4,49,50]. Using polarization observations of radio galaxies, Carroll, Field and Jackiw [49,50] put a limit of 0.1 on $\Delta\phi$ over cosmological distances. Using a different analysis of polarization observation of radio galaxies, Nodland and Ralston [51] found indication of anisotropy in electromagnetic propagation over cosmological distances with a birefringence scale of order 10^{25} m (i.e. about 0.1-0.2 Hubble distance). This gives $\Delta\phi \sim 5$ - 10 over Hubble distance. Later analyses [52–56] did not confirm this result and put a limit of $\Delta\phi \leq 1$ over cosmological distance scale. Further observations to test and measure $\Delta\phi$ to 10^{-6} is promising. The natural coupling strength ϕ is of order 1. However, the isotropy of our observable universe to 10^{-5} may leads to a change of $\Delta\phi$ over cosmological distance scale 10^{-5} smaller. Hence, observations to test and measure $\Delta\phi$ to 10^{-6} is very significant.

Eq. (9) is verified empirically to high accuracy from pulsar observations and from polarization measurements of extragalactic radio sources, and we can now look into the empirical constraints for H^{ij} and ψ. In Eq. (1), ds is the line element determined from the metric g_{ij}. From Eq. (9), the gravitational coupling to electromagnetism is determined by the metric H_{ij} and two scalar field ϕ and ψ. If H_{ij} is not proportional to g_{ij}, then the hyperfine levels of the lithium atom, the beryllium atom, the mercury atom and other atoms will have additional shifts. But this is not observed to high accuracy in Hughes–Drever experiments [57,58]. Therefore H_{ij} is proportional to g_{ij} to a certain accuracy. Since a change of H^{ik} to λH^{ij} does not affect χ^{ijkl} in Eq. (9), we can define $H_{11} = g_{11}$ to remove this scale freedom [44,59].

In Hughes–Drever experiments [57,58] $\Delta m/m \leq 0.5 \times 10^{-28}$ or $\Delta m/m_{\mathrm{e.m.}} \leq 0.3 \times 10^{-24}$ where $m_{\mathrm{e.m.}}$ is the electromagnetic binding energy. Using Eq. (9) in Eq. (1), we have three kinds of contributions to $\Delta m/m_{\mathrm{e.m.}}$. These three kinds are of the order of (i) $(H_{\mu\nu} - g_{\mu\nu})$, (ii) $(H_{0\mu} - g_{0\mu})\mathrm{v}$, and (iii) $(H_{00} - g_{00})\mathrm{v}^2$ respectively [44,59]. Here the Greek indices μ, ν denote space indices. Considering the motion of laboratories from earth rotation, in the solar system and in our galaxy, we can set limits on various components of $(H_{ij} - g_{ij})$ from Hughes–Drever experiments as follows:

$$\begin{aligned} |H_{\mu\nu} - g_{\mu\nu}|/U &\leq 10^{-18} \\ |H_{0\mu} - g_{0\mu}|/U &\leq 10^{-13} - 10^{-14}, \\ |H_{00} - g_{00}|/U &\leq 10^{-10}. \end{aligned} \tag{10}$$

where U ($\sim 10^{-6}$) is the galactical gravitational potential.

Eötvös–Dicke [60–63] experiments are performed on unpolarized test bodies. In essence, these experiments show that unpolarized electric and magnetic energies follow the same trajectories as other forms of energy to a certain accuracy. The constraints on Eq. (9) are

$$|1 - \psi|/U < 10^{-10} \tag{11}$$

and

$$|H_{00} - g_{00}|/U < 10^{-6} \tag{12}$$

where U is the solar gravitational potential at the earth.

In 1976, Vessot and Levine [64] used an atomic hydrogen maser clock in a space probe to test and confirm the metric gravitational redshift to an accuracy of $1.4{\times}10^{-4}$ [65]. The space probe attained an altitude of 10,000 km above the earth's surface. With Eq. (11), the constraint on Eq. (9) is

$$|H_{00} - g_{00}|/U \leq 1.4 \times 10^{-4}. \tag{13}$$

Thus, we see that for the constraint on $|H_{00} - g_{00}|/U$, Hughes–Drever experiments give the most stringent limit. However, STEP mission concept [66] proposes to improve the WEP experiment by five orders of magnitude. This will again lead in precision in determining H_{00}.

The theory (1) with χ^{ijkl} given by (2) is studied in [4] and [22]. In (1), particles considered have charges but no spin. To include spin–$\frac{1}{2}$ particles, we can add the Lagrangian for Dirac particles. In the next section, we review the experimental tests of the equivalence principle for polarized bodies.

In the above discussions, we assume $\chi^{(0)ijkl}$ in (4) to be given by the special relativistic value (5). In general, $\chi^{(0)ijkl}$ is determined from cosmological model in a particular theory and provides a framework to test special relativity. From null birefringence of pulsar observations, $\chi^{(0)ijkl}$ is constrained to have the value given in (5) to a precision of 10^{-16}. From the polarization measurements of extragalactic radio sources, the agreement to special relativity is to 10^{-20}.

To include QCD and other gauge interactions, we have generalized the $\chi - g$ framework [68]. Now we are working on a more comprehensive generalization to include a framework to test special relativity, and a framework to test the gravitational interactions of scalar particles and particles with spins together with gauge fields [38]. The relation of this generalized framework with respect to the Mansouri–Sexl framework [69], the Tourrenc–Melliti–Bosredon framework [70] and Lämmerzahl framework [37] is under study.

6 Experimental Searches for Electron Spin–Coupling

In this section, we review and discuss electron spin–coupling experiments — weak equivalence principle experiments, finite–range spin–coupling experiments, spin–spin coupling experiments and cosmic spin–coupling experiments. An important issue is to make a spin–polarized body. Here we describe the strategy and method.

To make a polarized–body with a net spin but without net magnetic moment, we need both the orbital angular momentum contribution and spin contribution of magnetic moment so that these contributions cancel each other, with a net total spin remaining.

For iron–group transition elements in a crystal, the elements are exposed to a noncentral electric field. In a noncentral field the plane of the orbit will

move about; the angular momentum components are no longer constant and may average to zero (quenched). For iron–group elements the orbital magnetic moments are mostly quenched although spins may drag some orbital momenta along with them.

For rare–earth elements the orbital angular momenta in the unfilled 4f shell is not quenched. Therefore, rare–earth Fe, Co, Ni compounds would be good materials for making spin–polarized bodies.

The experiments would be much easier to do at room temperature than at low temperature. For light rare–earth elements and their compounds, the Curie temperatures or Néel temperatures are quite low. Heavy rare–earth compounds, such as Tb, Dy, Ho compounds generally have much higher Curie temperatures. Dy_6Fe_{23} has compensation temperature near room temperature.

In the Dy-Fe compounds, magnetization versus temperature curves indicate artiferromagnetic interaction between iron and dysprosium atoms. From the susceptibility data, the strengths of Fe-Fe, Dy-Fe and Dy-Dy exchange interactions can be derived. Fe-Fe exchange interaction dominates the others. Dy-Fe compounds are ferrimagnetic at room temperature. The effective ordering of the iron lattice and dysprosium lattice have different temperature dependence because the strengths of exchange interactions are different. At the compensation temperature, the magnetic moments of two lattices compensate each other so that there is no net magnetization. Dy^{+++} has $L = 5$ and $S = 5/2$. Half of the Dy magnetization comes from orbital angular momentum, the other half from spin. Most of the iron magnetization comes from spin. So there is a net spin (and net total angular momentum) remaining.

To make samples, Dy_6Fe_{23} was synthesized by melting stoichiometric quantities of metallic iron and metallic dysprosium. The Dy_6Fe_{23} ingots were crushed, pressed into a cylindrical aluminum cup, and magnetized along the axial or a transverse direction direction. The magnetic field was shielded by two halves of pure iron casing, a thin aluminum spacer, and two sets of two fitting μ–metal cups with another thin aluminum spacer between the two sets.

Measuring the magnetization–temperature curve of our sample Dy_6Fe_{23}, comparing with temperature dependence curve in the literature and calculating from the magnetic properties of Dy_6Fe_{23}, there is at least 0.4 net polarized electron per atom.

In addtion to the Dy_6Fe_{23} samples, we also make $DyFe_3$, $HoFe_3$, Ho_6Fe_{23} and Tb_6Fe_{23} samples.

6.1 Weak Equivalence Principle Experiments

To investigate the equivalence principle for spin–polarized bodies or to probe the mass–spin (monopole–spin/baryon–number–to–spin) interactions, we have used both a beam balance [71] and a torsion balance [13] to test a magnetically shielded spin–polarized body of Dy_6Fe_{23}. From these results, we have inferred that, to an accuracy of 5×10^{-3}, the polarized electron falls at the same rate as unpolarized bodies in the earth's gravitational field, and that it falls at the

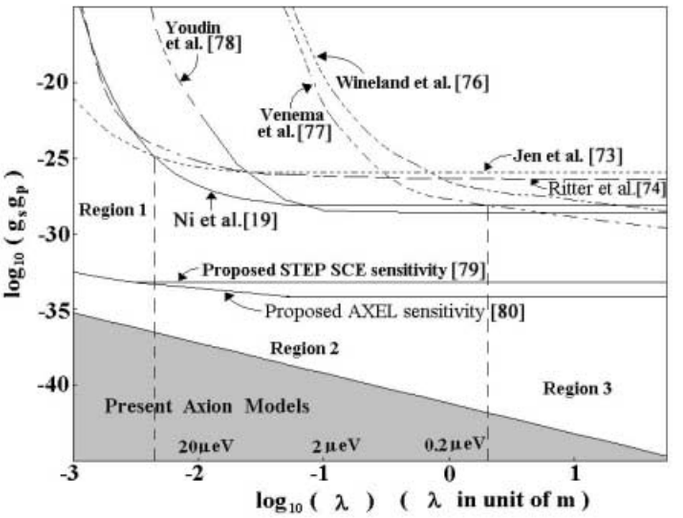


Fig. 1. Limits on $\boldsymbol{\sigma}\cdot\hat{\boldsymbol{r}}$ spin coupling for axionlike interactions from various experiments.

same rate as unpolarized bodies in the solar gravitational field with a deviation from this unity ratio estimated at $(3 \pm 4) \times 10^{-2}$.

At the present, we are working on a rotatable torsion–balance experiment and have improved the solar equivalence principle test by one order of magnitude [72].

6.2 Finite–Range Spin–Coupling Experiments

In [73], we use torsion balance with two cylindrical copper test masses and two cylindrical polarized "attracting" masses to search for finite–range mass–spin interactions with the Hamiltonian of the form: $H_{\rm int} = f(r)\boldsymbol{\sigma}\cdot\hat{\boldsymbol{r}}$ which includes the finite–range Leitner–Okubo–Hari Dass interaction with $f(r) = -Ae^{-\mu r}mU$. This result shows that for the range of 3-5 cm, the upper limit of this interaction for our test mass and polarized mass is below 1% of the gravitational interaction.

Ritter, Winkler and Gillies [74] use spin–polarized Dy_6Fe_{23} masses acting on unpolarized copper masses in a dynamic–mode torsion pendulum, and searched for interaction of the axion [7–12,75] form

$$H_{\rm int} = \frac{\hbar g_s g_p}{8\pi mc}\left(\frac{1}{\lambda r} + \frac{1}{r^2}\right) e^{-r/\lambda}\boldsymbol{\sigma}\cdot\hat{\boldsymbol{r}}\,. \tag{14}$$

In (14), λ is the range of the interaction, g_s and g_p are the coupling constants of vertices at the polarized and unpolarized particles and m is the mass of the polarized particle. Constraints on the coupling $g_s g_p/\hbar c$ with respect to the range from various experiments are plotted in a logarithmic plot (Fig. 1). For

$\lambda < 0.3$ m, references 73 and 74 give more stringent constraints than references 76 and 77, and for $\lambda > 0.3$ m, vice versa. References 76 and 77 investigate the existence of hypothetical anomalous spin–dependent forces by sensing the interaction of polarized trapped ions with fermions in the earth. These experiments are more sensitive to longer range forces, while experiments with laboratory sources are more sensitive to shorter range forces. Reference 8 has the best limit for 0.1 m $< \lambda <$ 8 m. Our works [19,73] have the best limit for $\lambda < 0.1$ m. In figure, we also show the proposed sensitivities of the STEP spin–coupling experiment [79,66] and the AXEL spin–coupling experiment (AXial Experiment at Low temperature) [80,81,18] together with allowed region of present axion models.

Speake's group at the University of Birmingham is working on the development of a new superconducting torsion balance to detect force on the mass for the spin–coupling experiment. H.J. Paik at the University of Maryland proposes to use superconducting accelerometers for a spin–coupling experiment with a high Q. They are also aiming at very significant improvement.

In Fig. 1, there are magnetic resonance experiments, torsion balance experiments and SQUID experiments. In the following, we give a taste of experimental procedure using a SQUID experiment [19]. The experiment measures the effective $\boldsymbol{B}_{\text{eff}}$ field produced by hypothetical axion or axion–like interaction while magnetic field is shielded by two niobium superconducting shields. Equation (14) can be written in the form

$$H_{\text{int}} = -\boldsymbol{m} \cdot \boldsymbol{B}_{\text{eff}} = -\mu_e \boldsymbol{\sigma} \cdot \boldsymbol{B}, \tag{15}$$

with $\mu_e = -\mid \mu_e \mid$ the magnetic moment of the electron. Hence in this experiment, we sensitivly measure $\boldsymbol{B}_{\text{eff}}$ field given by

$$\boldsymbol{B}_{\text{eff}} = -\frac{\hbar}{\mu_e} \frac{g_s g_p}{8\pi m_e c} \left(\frac{1}{\lambda r} + \frac{1}{r^2} \right) e^{-r/\lambda} \hat{\boldsymbol{r}} \,. \tag{16}$$

The scheme of our experimental setup is shown in Fig. 2. Our copper mass is sitting on one side of the turntable underneath the dewar. In the data taking, a laser beam and a chopper–photodetector systm is used to lock the output signal of the dc SQUID to the rotation angle of the polarized bodies. The laser beam is intercepted by the chopper when the copper axis is in line with the axis of the paramagnetic salt. We define this angle to be zero degree, and expect the $\boldsymbol{\sigma} \cdot \boldsymbol{r}$ interaction signal to be proportional to $\cos\Theta$, where Θ is the angular position of the copper mass.

To start the measurement, we set the turntable with copper mass rotating at 0.96 cycle per second with a stepping motor system. The stability of the rotation speed is better than 10^{-4}. The output of voltage of the dc SQUID system for 1 ϕ_o from the most sensitive scale of the dc SQUID controller is 10 V. This output is further amplified 1000 times and low–pass filtered to 10 Hz bandwidth, and then read into a computer with an analog to digital converter. The angular position of the copper mass is simultaneously read into this computer. The typical noise of the SQUID output after 1000 times amplification and 10 Hz low pass filtering

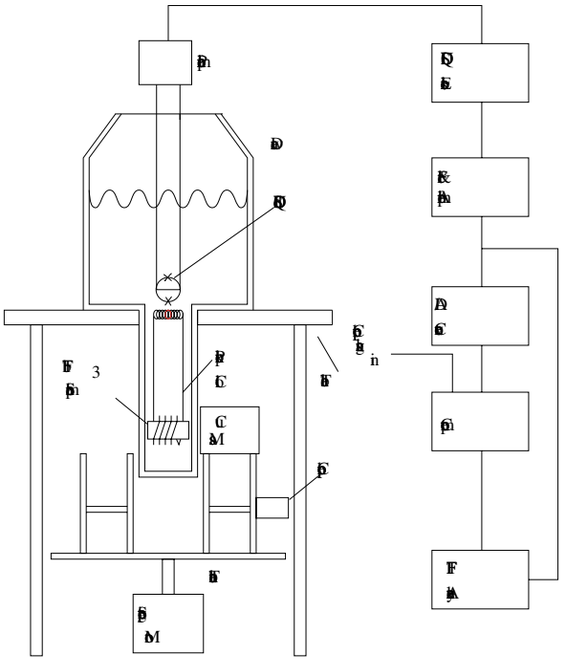


Fig. 2. Schematic for spin–coupling experiment.

as recorded by ADC (Analog–to–Digital Converter) is about ±300 mV. This is consistent with the dc SQUID noise 200 mV/$\sqrt{\text{Hz}}$ after amplification. When we average the data for 400 cycles, the typical output is about ±50 mV and the pattern repeats. To subtract this interference background, we average the data for 4-5 hours, alternatively take away and put back the copper cylinder to average the data for another 4-5 hours and subtract the results to find the net effects.

The weighted average of the six runs for the amplitude of $\cos\Theta$ component is (0.49±2.34) mV. Expressed in terms of flux amplitude, it becomes (0.49±2.34)× 10^{-7} ϕ_o. Converted to $\boldsymbol{B}_{\text{eff}}$, we have $(1.13 \pm 5.38) \times 10^{-12}$ Gauss and the coupling constant $g_s g_p/\hbar c$ is (0.14±0.67)× 10^{-28} for $\lambda > 30$mm. Our experimental constraint on the coupling constant $g_s g_p/\hbar c$ improves over previous results by 2 orders of magnitude at $\lambda = 30$ mm. Further improvement will be implemented. For finite–range Leitner–Okubo–Hari Dass interaction, the dimensionless parameter A is constrained to less than 10 for the range parameter $\lambda = \mu^{-1} > 30$ mm.

6.3 Spin–Spin Coupling Experiments

Usually the dipole–monopole (spin–mass) part of an interaction is larger than its dipole–dipole (spin–spin) part. Monopole–monopole part is sometimes larger

if there is no constraint. However the monopole–monopole part does not change with polarization and is usually harder to detect. Therefore in search for a new interaction, searching for dipole–monopole part is usually more significant. This is true for axion search. However, for axial photon [42,43] search and arion [82] search, the search for anomalous spin–spin interactions give stringent constraints.

Let α_s be the strength of the anomalous spin–spin interaction compared to the magnetic spin–spin interaction. The pioneer work of Graham and Newman [83] used carefully prepared hybrid split toroids of $GdNi_5$/NdNi at superconducting temperatures with a torsion balance of the feedback deflection type. Their experimental result assigns uncertainties in two parts: statistical at the 1 σ level and an estimated systematic uncertainty $\alpha_s = (8.0 \pm 6.3 \pm 1.1) \times 10^{-11}$. The torsion balance experiment led by Ritter at the University of Virginia uses the period method and gives the constraint $\alpha_s < 1.5 \times 10^{-12}$ [84]. Adapting the induced ferromagnetism method of Vorobyov and Gitarts [85], we use a low noise dc SQUID system to search for the interaction of spins in a spin–polarized test mass and those in a paramagnetic salt, separated by a μ–metal shield and a double–layer superconducting shield [15–17]. Our result limit the strength of α_s to $\alpha_s = (1.2 \pm 2.0) \times 10^{-14}$ [17]. This limits the coupling of axial photon and the arion coupling to a level much lower than originally proposed. We are currently working on an improvement of this experiment.

6.4 Cosmic Spin–Coupling Experiments

Hughes–Drever experiments [57,58] test the Cosmic Spatial Isotropy for spin 3/2 particle very precisely. Recently, frequency and clock experiments push this limit even further.

As to the spin 1/2 particle, Phillips [86] used a cryogenic torsion pendulum carrying a transversely polarized magnet with superconducting shields to set a stringent limit of 8.5×10^{-18}eV for the splitting of the spin states of an electron at rest on Earth. In our laboratory we have used a room–temperature torsion balance with a magnetically–compensated polarized–body and set a spin energy level splitting limit of 3×10^{-18}eV [87,88]. Berglund et al. [89] use a magnetic resonance technique and set a limit of 1.8×10^{-18} eV on the energy splitting.

For the analysis of cosmic anisotropy for electrons, we can use the following Hamiltonian:

$$H_{\text{cosmic}} = C_1\sigma_1 + C_2\sigma_2 + C_3\sigma_3 \tag{17}$$

in the cosmic frame of reference. This includes the following two cases, (i) $H_{\text{cosmic}} = g\boldsymbol{\sigma} \cdot \boldsymbol{n}$ where $\boldsymbol{n}$ is a cosmic vector which defines a preferred direction, with $C_1 = gn_1$, $C_2 = gn_2$, $C_3 = gn_3$ as considered in the references [87,88]; here C's are constants, (ii) $H_{\text{cosmic}} = g\boldsymbol{\sigma} \cdot \boldsymbol{v}$ where $\boldsymbol{v}$ is the velocity with respect to the comic background defined by the isotropy of background radiation, with $C_1 = gv_1$, $C_2 = gv_2$, $C_3 = gv_3$ as considered in the context of references [90,91]; in this case, since $\mathbf{v}$ is largely the velocity of our solar system through the cosmic preferred frame, to a first approximation, C's are also constants. For convenience, we use the celestial equatorial coordinate system from the center of earth for our laboratory position,

i.e., the earth rotation axis (North pole direction) as z–axis and the direction of the spring equinox as the positive x–direction. All the above experimental constraints are on C_1 and C_2. The constraints on C_3 are crude.

To improve the precision and to constrain on C_3, we use a rotatable torsion balance carrying a transversely spin–polarized ferrimagnetic Dy_6Fe_{23} mass. With a rotation period of one hour, the period of anisotropy signal is reduced from one sidereal day by about 24 times, and hence the 1/f noise is greatly reduced. Our present experimental results constrain the cosmic anisotropy constant C_1, C_2, C_3 to $\sqrt{C_1^2 + C_2^2} < (1.8 \pm 5.3) \times 10^{-21}$ eV and $\mid C_3 \mid < (1.2 \pm 3.5) \times 10^{-19}$ eV. This improves the previous limits on (C_1, C_2) by 120 times and C_3 by a factor of 800. Our experimental set–up is schematically shown in Fig. 3.

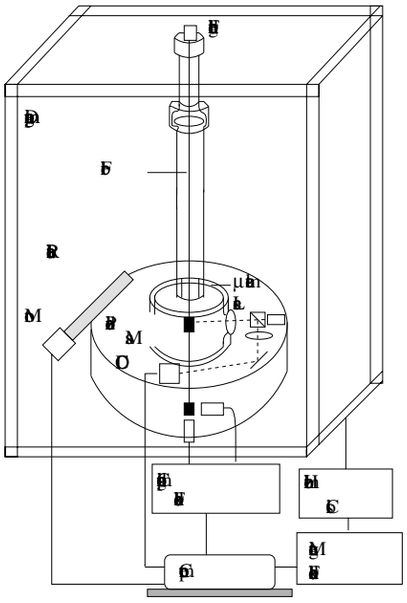


Fig. 3. Experimental setup for spatial isotropy test.

The angular velocity of the cosmic signals is $\Omega+\omega$, $\Omega-\omega$, and ω. By the earth rotation the projection of the electron spin in the $x-y$ plane rotates to opposite direction relative to the neutrino background or cosmos after half of a sidereal day (11 hours 58 min 2 seconds). Adding the two data sets separated by half sidereal day, we can eliminate the $\Omega+\omega$, $\Omega-\omega$ term, and estimate C_3. Subtracting between the same two data sets, we can eliminate the ω term. With 4 sequential data sets (each set's separated by half sidereal day) in opposite rotational direction of rotatable table, the signals with frequencies $\Omega+\omega$, $\Omega-\omega$, ω can be separated. The results of eight such sets of runs gives the limits on C_1, C_2, C_3 just mentioned. This experiment also gives a stringent CPT test [94].

7 Outlook

We have reviewed and discussed theoretical motivations and experimental searches for the role of spin and polarization in gravity. Efforts in this direction will bear fruits both in fundamental physics and cosmology. For example, the study of equivalence principles leads to polarization effects in cosmology to be tested by astrophysical observations, and experiments on spin–polarized bodies constitute strong CPT tests.

The ultimate searches for the role of spin in gravity is to measure the gyrogravitational ratios of particles. For these searches difficult experiments are to be

performed. As we have already seen, the difficulties of the Stanford Gyroscope experiment to be launched in 2002 and the technological achievement already made are very great. However, there is a chance of success by using the following method:

(i) Polarized–body method [22],
(ii) Atom interferometry [95],
(iii) Superfluid ^{3}He [96].

Let us look into the future.

References

1. G. Galilei, 1683, *Discorsi e dimostriazioni matematiche intorno a due nuove scienze*, Elzevir, Leiden. English Translation by H. Crew and A. de Salvio, *Dialogues Concerning Two New Sciences*, Macmillan, New York, 1914; reprinted by Dover, New York, 1954.
2. A. Einstein: *Jahrb. Radioakt. Elektronik* **4**, 411 (1907); Corrections by Einstein in *Jahrb. Radioakt. Elektronik* **5**, 98 (1908); English Translations by H. M. Schwartz in *Am. J. Phys.* **45**, 512, 811, 899 (1977).
3. L. I. Schiff: *Am. J. Phys.* **28**, 340 (1960)
4. W.–T. Ni: A Nonmetric Theory of Gravity, preprint, Montana State University, Bozeman, Montana, USA (1973), http://gravity5.phys.nthu.edu.tw.
5. W.–T. Ni: *Bull. Am. Phys. Soc.* **19**, 655 (1974).
6. W.–T. Ni: *Phys. Rev. Lett.* **38**, 301 (1977).
7. S. Weinberg: *Phys. Rev. Lett.* **40**, 233 (1978).
8. F. Wilczek: *Phys. Rev. Lett.* **40**, 279 (1978).
9. M. Dine *et al.*: *Phys. Let.t* **B 104**, 1999 (1981).
10. M. Shifman *et al.*: *Nucl. Phys.* **B 166**, 493 (1980).
11. J. Kim: *Phys. Rev. Lett.* **43**, 103 (1979).
12. S. L. Cheng, C. Q. Geng and W.–T. Ni: *Phys. Rev.* **D 52** 3132 (1995) and references therein.
13. Y. Chou, W.–T. Ni and S.–L. Wang: *Mod. Phys. Lett.* **A 5**, 2297 (1990); and references therein.
14. S.–S. Pan, W.–T. Ni and S.–C. Chen: *Mod. Phys. Lett.* **A 7**, 1287 (1992); and references therein.
15. T. C. P. Chui and W.–T. Ni: *Phys. Rev. Lett.* **71**, 3247 (1993).
16. W.–T. Ni, S.–S. Pan, T. C. P. Chui and B.–Y. Cheng: *Int. J. Mod. Phys.* **8**, 5153 (1993).
17. W.–T. Ni, T. C. P. Chui, S.–S. Pan and B.–Y. Cheng: *Physica* **B 194-196** 153 (1994).
18. W.–T. Ni: *Class. Quantum Grav.* **13**, A135 (1996); and references therein.
19. W.–T. Ni, S.–S. Pan, H.–C Yeh, L.–S. Hou, J. Wan: *Phys. Rev. Lett.* **88**, 2439 (1999).
20. G. Uhlenbeck and S. Goudsmit: *Naturwiss.* **13**, 953 (1925); *Nature* **117**, 264 (1926).
21. C. W. F. Everitt et al.: this book (2000).
22. W.–T. Ni: "Spin, Torsion and Polarized Test–Body Experiments", in Proceedings of the 1983 International School and Symposium on Precision Measurement and Gravity Experiment, Taipei, Republic of China, January 24-February 2, 1983, ed. by W.-T. Ni (Published by National Tsing Hua University, Hsinchu, Taiwan, Republic of China, June 1983), p. 531.

23. F.W. Hehl and W.–T. Ni: *Phys. Rev.* **D 42**, 2045 (1990); and references therein.
24. É. Cartan: Sur une Généralisation de la notion de courbure de Riemann et les espaces à torsion. *C. R. Acad. Sci.* (Paris) **174**, 593, (1922).
25. É. Cartan: Sur les variétés à connexion affine et la théorie de la relativitée généralisée I, I (suite), II, *Ann. Ec. Norm. Sup.* **40**, 325 (1923); **41**, 1 (1924); **42**, 17 (1925).
26. T.W.B. Kibble: Lorentz invariance and the gravitational field. *J. Math. Phys.* **2**, 212 (1961).
27. D.W. Sciama: On the analogy between charge and spin in general relativity, in *Recent developments in general relativity* (Pergamon+PWN, Oxford, 1962), p. 415.
28. D.W. Sciama: The physical structure of general relativity. *Rev. Mod. Phys.* **36**, 463, 1103 (1964).
29. F.W. Hehl, P. von der Heyde, G.D. Kerlick, and J.M. Nester: *Rev. Mod. Phys.* **48**, 393 (1976).
30. A. S. Eddington: A generalization of Weyl's theory of the electromagnetic and gravitational fields, *Proc. Roy. Soc. Lond. Ser.* **A 99**, 104 (1921).
31. C. Lämmerzahl, *Phys. Lett.* **A 228**, 223 (1997).
32. P. Singh and L.H. Ryder: *Class. Quantum Grav.* **14**, 3513 (1997).
33. C. N. Yang: *Phys. Rev. Lett.* **33**, 445 (1974).
34. W.–T. Ni: *Phys. Rev. Lett.* **35**, 319 (1975).
35. W.–T. Ni: *Math. Proc. Cambridge Phil. Soc.* **90**, 517 (1981); K.–S. Cheng and W.–T. Ni, ibid **87**, 527 (1980).
36. A. Ashtekar: *Phys. Rev. Lett.* **57**, 2244 (1986); *Phys. Rev.* **D 36**, 1587 (1987); Lectures on Non–Perturbative Canonical Gravity (World Scientific, Singapore 1991).
37. C. Lämmerzahl: *Class. Quantum Grav.* **15**, 13 (1998) (also *Gen. Rel. Grav.* **28**, 1043 (1996)).
38. C. Lämmerzahl and W.–T. Ni: work in progress.
39. Gronwald and F. W. Hehl: Metric–affine gauge theory of gravity; I. Foundatoins, *Int. J. Mod. Phys.* **D6**, 263 (1997).
40. F. W. Hehl and A. Macias: Metric–affine gauge theory of gravity: II. Exact solutions, *Int. J. Mod. Phys.* **D 8**, 399 (1999).
41. K. Hayashi and T. Shirafuji: *Phys. Rev.* **D 19**, 3524 (1979).
42. P.C. Naik and T. Pradhan: *J. Phys. A: Math. Gen.* **14**, 2795 (1981).
43. T. Pradhan, R.P. Malik and P.C. Naik: *Pramana* **24**, 77 (1985).
44. W.–T. Ni: "Equivalence Principles and Precision Experiments" p. 647, in B.N. Taylor and W.D. Phillips (eds.): *Precision Measurement and Fundamental Constants II*, Natl. Bur. Stand. (U.S.), Spec. Publ. **617** (1984).
45. W.–T. Ni: "Timing Observations of the Pulsar Propagations in the Galactic Gravitational Field as Precision Tests of the Einstein Equivalence Principle", in B. Hidayat and M.W. Feast (eds.): *Proceedings of the Second Asian-Pacific Regional Meeting of the International Astronomical Union*, (Tira Pustaka, Jakarta 1984), p. 441.
46. W.-T. Ni: "Equivalence Principles, Their Empirical Foundations, and the Role of Precision Experiments to Test Them", in W.-T. Ni (ed.): *Proceedings of the 1983 International School and Symposium on Precision Measurement and Gravity Experiment, Taipei, Republic of China, January 24-February 2, 1983*, (Published by National Tsing Hua University, Hsinchu, Taiwan, Republic of China, June, 1983), p. 491.
47. M.P. Haugan and T.F. Kauffmann: *Phys. Rev.* **D 52**, 3168 (1995).
48. T.P. Krisher: *Phys. Rev.* **D 44**, R2211 (1991).

49. S.M. Carroll, G.B. Field, R. Jackiw: *Phys. Rev.* **D 41**, 1231 (1990).
50. S.M. Carroll and G.B. Field: *Phys. Rev.* **D 43**, 3789 (1991).
51. B. Nodland and J.P. Ralston: *Phys. Rev. Lett.* **78**, 3043 (1997).
52. J.F.C. Wardle, R.A. Perley, and M.H. Cohen: *Phys. Rev. Lett.* **79**, 1801 (1997).
53. D.J. Eisenstein and E.F. Bunn: *Phys. Rev. Lett.* **79**, 1957 (1997).
54. S.M. Carroll and G.B. Field: *Phys. Rev. Lett.* **79**, 2394 (1997).
55. T.J. Loredo, E.A. Flanagan, and I.M. Wasserman: *Phys, Rev.* **D 56**, 7507 (1997).
56. S. M. Carroll: *Phys. Rev. Lett.* **81**, 3067 (1998).
57. V.W. Hughes, H.G. Robinson, and V. Beltran–Lopez: *Phys. Rev. Lett.* **4**, 342 (1960); V. Beltran–Lopez, H.G. Robinson, and V.W. Hughes: *Bull. Am. Phys. Soc.* **6**, 424 (1961); R.W.P. Drever: *Phil. Mag.* **6**, 683 (1962); J.F. Ellena, W.–T. Ni and T.–S. Ueng: *IEEE Transactions on Instrumentation and Measurement* **IM-36**, 175 (1987).
58. T.E. Chupp, R.J. Hoara, R.A. Loveman, E.R. Oteiza, J.M. Richardson, M.E. Wagshul: *Phys. Rev. Lett.* **63**, 1541 (1989).
59. W.–T. Ni: "Implications of Hughes–Drever Experiments", in W.-T. Ni (ed.): *Proceedings of the 1983 International School and Symposium on Precision Measurement and Gravity Experiment, Taipei, Republic of China, January 24-February 2, 1983*, (Published by National Tsing Hua University, Hsinchu, Taiwan, Republic of China, June, 1983), p. 519.
60. R.V. Eötvös, D.Pekar, and E. Fekete: *Ann. Phys.* (Leipzig) **68**, 11 (1922); also R.V. Eötvös: *Math. Naturwiss. Ber. Ungarn* (Budapest) **8**, 65 (1890).
61. P.G. Roll, R.Krotkov, and R.H. Dicke: *Ann. Phys.* (N. Y.) **26**, 442 (1964).
62. V.B. Braginsky and V.I. Panov: *Zh. Eksp. Teor. Fiz.* **61**, 873 (1971) [*Sov. Phys. JETP* **34**, 463 (1972)].
63. Y. Su *et al.*: *Phys. Rev.* **D 50**, 3614 (1994); S. Baessler *et al.*: *Phys. Rev. Lett.* **83**, 3585 (1999).
64. R.F.C. Vessot and M.W. Levine: *Gen. Rel. Grav.* **10**, 181 (1979).
65. R.F.C. Vessot *et al.*: *Phys. Rev. Lett.* **45**, 2081 (1980).
66. ESA SCI(93)4, STEP (Satellite test of the equivalence principle) report on the phase A study (1993).
67. MiniSTEP (A minimal–cost version of a satellite test of the equivalence principle) Stanford, April 1996.
68. W.–T. Ni: *Phys. Lett.* **A 120**, 174 (1987).
69. R. Mansouri and R. U. Sexl: *Gen. Rel. Grav.* **8**, 497 (1977); **8**, 515 (1977); **8**, 809 (1977).
70. Ph. Tourrenc, T. Melliti and J. Bosredon: *Gen. Rel. Grav.* **28**, 1071 (1996).
71. W.–T. Ni, Y. Chou, S.–S. Pan, C.–H. Lin, T.–Y. Hwong, K.–L. Ko and K.–Y. Li: "An Improvement of the Equivalence Principle Test for Spin–Polarized Bodies and the Mass Loss", in Proceddings of the 3rd ROC-ROK Metrology Symposium, Hsinchu, May 22-24, published by the Center of Measurement Standards, I.T.R.I. (1990) p. 107; C.–H. Hsieh, P.–Y. Jen, K.–L. Ko, K.–Y. Li, W.–T. Ni, S.–S. Pan, Y.–H Shi and R.–J. Tyan: *Mod. Phys. Lett.* **A4** (1989) 1597; W.–T. Ni, P.–Y Jen, C.–H. Hsieh, K.–L. Ko, S.–C Chen, S.–S. Pan and M.–H Tu, "Test of the Equivalence Principle for Spin–Polarized Bodies", in J.W. Won and Y.K. Park (eds.): *Proceddings of Second ROK-ROC Metrology Symposium* (Korea Standards Research Institute, 1988) p. VII-2-1.
72. L.–S. Hou and W.–T. Ni: Rotatable–Tosion–Balance Equivalence Principle Experiment for the Spin–Polarized Ho_6Fe_{23}, submitted to *Mod. Phys. Lett.* **A**.

73. T.–H. Jen, W.–T. Ni, S.–S. Pan and S.–L. Wang: "Torsion Balance Experiment Searching for Finite–Range Mass–Spin Interactions", in H. Sato and T. Nakamura (eds.): *Proceedings of the Sixth Marcel Grossmann Meeting on General Relativity* (World Scientific, Singapore 1992), p. 489.
74. R.C. Ritter *et al.*, L.I. Winkler and G.T. Gillies: *Phys. Rev. Lett.* **70**, 701 (1993); the limit on the electron have been multiplied by $8\pi^2$ to be consistent with our notation (R.C. Ritter, private communication).
75. J.E. Moody and F. Wilczek: *Phys. Rev.* **D30**, 130 (1984).
76. D.J. Wineland, J.J. Bollinger, D.J. Heinzen, W.M. Itano, and M.G. Raizen: *Phys. Rev. Lett.* **67,** 1735 (1991).
77. B.J. Venema, P.K. Majumder, S.K. Lamoreaux, B.R. Heckel, and E. N. Fortson: *Phys. Rev. Lett.* **68**, 135 (1992).
78. A.N. Youdin *et al.*: *Phys. Rev. Lett.* **77**, 2170 (1996).
79. D. Shaul *et al.*: *Class. Quantum Grav.* **13**, A107 (1996), and references therein.
80. W.–T. Ni: *Physica* **B 284-8**, 2137 (2000).
81. Y.–C. M. Li and W.–T. Ni: *Physica* **B 284-8**, 2139 (2000).
82. A.A. Ansel'm and N.G. Ural'tsev: *Zh. Eksp. Teor. Fiz.* **82**, 1725 (1982) [*Sov. Phys. JETP* **55**, 997 (1982)]; A.A. Ansel'm: *Pis'ma Zh. Eksp. Teor. Fiz.* **36**, 46 (1982) [*JETP Lett.* **36**, 55 (1982)].
83. D. Graham and R. Newman, in *Proceedings of the Eleventh International Conference on General Relativity and Gravitation*, Stockholm, Sweden, 1986, edited by M. A. H. MacCallum (Cambridge Univ. Press, New York, 1987), p.614; D. Graham, Ph.D. Dissertation (University of California, Irvine 1987).
84. P. V. Vorobyov and Ya. I. Gitarts: *Phys. Lett.* **B 208**, 146 (1988).
85. R. C. Ritter, C. E. Goldblum, W.–T. Ni, G. T. Gillies, and C. C. Speake: *Phys. Rev.* **D 42**, 977 (1990).
86. P. R. Phillips: *Phys. Rev. Lett.* **59**, 1784 (1987).
87. S.–L. Wang, W.–T. Ni and S.–S. Pan: *Mod. Phys. Lett.* **A 8**, 3715 (1993).
88. F.–L. Chang, H.–C. Yeh, W.–T. Ni and S.–S. Pan: "Improved experimental limit on the cosmological spatial anisotropy for polarized electrons", International Workshop on Gravitation and Cosmology (1995), pp. 21.
89. C. J. Berglund *et al.*: *Phys. Rev. Lett.* **75**, 1879 (1995).
90. L. Stodolsky: *Phys. Rev. Lett.* **34**, 110 (1975).
91. H. Nielson and I. Picek: *Nucl. Phys.* **211B**, 269 (1983).
92. L.–S. Hou and W.–T. Ni, "Test of Spatial Anisotropy for Polarized Electrons Using a Rotatable Torsion Balance", CD–ROM Proceedings for the International Workshop on Gravitation and Astrophysics, November 17-19, 1997, Tokyo (Published in August, 1998).
93. L.–S. Hou, W.–T. Ni and Y. C. M. Li: Test of Cosmic Spatial Isotropy for Polarized Electrons Using a Rotatable Torsion Balance, submitted to *Phys. Rev. Lett.*
94. R. Bluhm and V.A. Kostelecky: *Rhys. Rev. Lett.* **84**, 1381 (2000).
95. P.R. Berman (ed.): *Atom Interferometry*, (Academic Press, 1997), and references therein.
96. Y. Mukharsky, O. Avenel and É. Varoquaux: Rotation Measurements with a Superfluid ^{3}He Gyrometer, CD-ROM of the 22nd Low-Temperature Conference, Helsinki, 4-11 August, 1999.

Spin in Special and General Relativity

Lewis H. Ryder

School of Physical Sciences, University of Kent, Canterbury CT2 7NR, UK

Spin is the ultimate gyroscope. The smallest possible amount of angular momentum is $\hbar/4\pi$ – that possessed by a spin $\frac{1}{2}$ particle. When the day comes that it becomes realistic to theorise about and to measure the precession of a spin $\frac{1}{2}$ particle it will be necessary to have to hand the relevant theoretical tools; in other words, to be able to give a description of spin one–half particles which is consistent with Special Relativity, and to generalise that description to General Relativity. In general terms, then, this is an exercise in relativistic quantum mechanics, and in the case of General Relativity, in quantum mechanics in a curved space. This latter is, of course, different from quantum gravity. Quantum gravity is a theory describing the quantum nature of the gravitational field itself, for example in terms of gravitons. For our purposes the gravitational field is treated *classically*, as a curved space–time. The only thing to be quantised is the spin $\frac{1}{2}$ particle.

It may be thought that this exercise has already been performed, since the Dirac equation is nothing other than a relativistic equation for spin $\frac{1}{2}$ particles. It turns out, however, that the Dirac equation itself does not automatically yield a relativistic spin operator. The problems connected with finding such an operator were already identified in 1950 by Foldy and Wouthuysen. After outlining these problems, we shall describe how this operator is constructed. The paper concludes with some remarks about the extent to which it makes sense to talk about spin in the context of general relativity. In particular it will be pointed out that spin precession is *inevitable* in GR; there is no such thing as conserved spin in curved spacetime.

1 The Dirac Equation

The Dirac equation is a first order wave equation for spin one–half particles. In the usual notation, with a metric $(+,-,-,-)$ and $c = 1$, it is

$$(\gamma^0 E + \boldsymbol{\gamma}\cdot\boldsymbol{p})\psi = m\psi \tag{1}$$

To satisfy the Einstein (mass–shell) condition $E^2 - \boldsymbol{p}\cdot\boldsymbol{p} = m^2$ it is found that the coefficients γ^μ cannot be ordinary numbers, but must be matices – in fact, 4×4 matrices. In the "standard representation" (see for example [1])

$$\gamma^0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \tag{2}$$

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 457–462, 2001.

where each of the entries above is a 2×2 matrix and σ^i are the Pauli matrices. Since the γs are 4×4 matrices, the "wave function" ψ must have 4 components. We may write

$$\psi = \begin{pmatrix} u \\ v \end{pmatrix} \tag{3}$$

where the "Pauli spinors" u and v each have two components.

Dirac was setting out to find a wave equation for the electron; that is, for a spin $\frac{1}{2}$ particle. It is reasonable, therefore, to ask what there is in the above to convince us that equation (1) does indeed describe a particle with spin $\frac{1}{2}$. A spin $\frac{1}{2}$ particle has two spin components and the generators of the group SU(2) (the covering group of SO(3)) are proportional to the Pauli matrices; we have

$$\left[\tfrac{\hbar}{2}\sigma^i, \tfrac{\hbar}{2}\sigma^j\right] = i\hbar\epsilon_{ijk}\tfrac{\hbar}{2}\sigma^k \,. \tag{4}$$

It is, of course, very well known that the reason ψ above has 4 components rather than 2 is that the Dirac equation predicts the existence of particles and antiparticles together, and they each have spin $\frac{1}{2}$. It is, then, a simple matter to define

$$\Sigma^i = \begin{pmatrix} \sigma^i & 0 \\ 0 & \sigma^i \end{pmatrix} , \tag{5}$$

and $\frac{\hbar}{2}\Sigma^i$ will clearly also obey the commutation relations (4) above. Moreover, the Dirac equation is a relativistic wave equation so the proposition immediately suggests itself that $\frac{\hbar}{2}\Sigma^i$ is in fact the relativistic spin operator. This proposition runs into a problem, however. The Dirac Hamiltonian is

$$H = \gamma^0(m + \boldsymbol{\gamma}\cdot\boldsymbol{p}) = \begin{pmatrix} m & \boldsymbol{\sigma}\cdot\boldsymbol{p} \\ \boldsymbol{\sigma}\cdot\boldsymbol{p} & -m \end{pmatrix} \tag{6}$$

and the problem is that $[\Sigma^i, H] \neq 0$; spin, given by the operator (5), would not be a conserved quantity. It turns out that the Dirac equation suffers from another problem, which turns out to be related to this one; or, at least, the solution to the one problem is also the solution to the other. This problem is that the velocity operator $\boldsymbol{v} = d\boldsymbol{r}/dt$ has eigenvalues ± 1, i.e. $\pm c$, so Dirac particles should travel at the speed of light. This is clearly inconsistent with relativity.

These problems were addressed and solved by Foldy and Wouthuysen almost half a century ago [2]. Their strategy was to find a transformation of the wave function ψ, $\psi' = e^{iS}\psi$, under which the new Hamiltonian

$$H' = e^{iS}He^{-iS} - ie^{iS}\frac{\partial}{\partial t}e^{-iS} \,, \tag{7}$$

is block–diagonal. Equivalently, it should contain no terms which coupled the two–spinors u and v (see equation (3)) – in the words of FW, no "odd" terms. The required transformation turns out to be

$$e^{iS} = \frac{E + m + i\boldsymbol{\gamma}\cdot\boldsymbol{p}}{\sqrt{2E(E+m)}} \,, \tag{8}$$

and the new Hamiltonian is

$$H' = \begin{pmatrix} \sqrt{m^2 + \boldsymbol{p}^2} & 0 \\ 0 & \sqrt{m^2 + \boldsymbol{p}^2} \end{pmatrix}, \tag{9}$$

a sensible result indeed! Under the FW transformation a general operator A becomes $A' = e^{iS} A e^{-iS}$, so the new spin operator is

$$\boldsymbol{X} = e^{iS} \frac{1}{2} \boldsymbol{\Sigma} e^{-iS} + \frac{\boldsymbol{\Sigma} \cdot \boldsymbol{p}}{2E(E+m)} + \frac{i}{2E} \gamma^5 \gamma^0 \boldsymbol{\Sigma} \times \boldsymbol{p} \tag{10}$$

This highly non–linear operator is called the "mean spin operator" by Foldy and Wouthuysen. It commutes with the Hamiltonian (6) and therefore corresponds to a conserved quantity. When acting on positive energy states (those for which $\boldsymbol{\sigma} \cdot \boldsymbol{p} = \gamma^5 E - \gamma^5 \gamma 0 m$) the mean spin operator becomes

$$\boldsymbol{X}_{\mathrm{FW}} = \frac{1}{2} \boldsymbol{\sigma} \gamma^0 - \frac{\boldsymbol{p}}{2(E+m)} \gamma^5 (1 + \gamma^0) \,. \tag{11}$$

Historically the Foldy–Wouthuysen transformation has been regarded as an intrinsically non–relativistic procedure – as the "correct" way of taking the non–relativistic limit of the Dirac equation. This would seem to be the philosophy of the treatment by Bjorken and Drell [1], for instance, as well as by Hehl and Ni [3]. What I shall show, however, is that the FW operator above, when acting on positive energy states, is in fact the relativistic spin operator for spin $\frac{1}{2}$ particles.

2 Spin and the Poincaré Group

In non–relativistic quantum mechanics spin is consistently treated as being associated with rotations, and therefore, in group theoretic language, to be approached by analysing the rotation group. This is SO(3) in the simplest approach; but SO(3) does not have representations giving spin $\frac{1}{2}$. To treat these particles one has to observe that the covering group of SO(3) is SU(2), and SU(2) has a two–component complex spinor ("Pauli spinor") as the basis state for its fundamental representation. In the non–relativistic domain, then, a satisfactory understanding of spin emerges if one views it as described by the covering group of the rotation group.

On entering the relativistic domain, however, a fresh look has to be taken. It was Wigner who took the freshest look at this problem and whose 1939 paper [4] (1965) gives us the profoundest understanding of spin. Wigner pointed out that what is involved is not the homogeneous, but – surprisingly – the *inhomogeneous* Lorentz group, or Poincaré group; that is, the group of rotations, Lorentz "boost" transformations and translations in spacetime. This group is the isometry group of the Minkowski metric – though Wigner did not describe it in these terms. The surprising thing is that translations should be relevant to the task of describing spin, but Wigner's argument, paraphrased rather drastically, was that spin was the additional "label" necessary to identify quantum states after their mass had

already been specified. Mass and spin are the two quantities by which states can be specified, and they are both kinematic in origin. Mass (rest–mass, that is) is given by the Einstein relation, mentioned above (with $c = 1$)

$$M^2 = P_\mu P^\mu = E^2 - \boldsymbol{P} \cdot \boldsymbol{P} \,. \tag{12}$$

E and $\boldsymbol{P}$ are, however, the generators of translations in time and space and it is because of this that it is the Poincaré group, rather than the homogeneous Lorentz group, which is the group of significance in this problem. In group theoretical language $P_\mu P^\mu$ is the first Casimir operator of the Poincaré group; that is, a quadratic function of the group generators which commutes with all of them. The Poincaré group is, however, a group of rank 2 so there should be two Casimir operators. If the first one gives mass, which it clearly does by equation (12), the second one should yield spin. Wigner was unable to find an explicit expression for the second Casimir operator, but he did pioneering work on elucidating the group structure of the problem. The crucial step to take was to observe that the group which described spin was the *little group* of the Poincaré group; that is, the subgroup of it which leaves invariant a given 4–vector P_μ. Wigner showed that this subgroup depends on M^2. States are classified as timelike, spacelike or lightlike according as $M^2 > 0, < 0$, or $= 0$. There is also the case where $P_\mu = 0$, which is the null case. It is only in the case of timelike states that the little group is the rotation group; in the other cases it is non–compact. This immediately explains, for example, why the photon does not have a longitudinal polarisation state, for that would only exist if spin were desribed by SO(3), which it is not for lightlike states. The problem of identifying specific *spin operators*, however, was left unsolved by Wigner.

It is clear that a necessary property of the spin operators is that they should commute with the translation operators P_μ of the Poincaré group. The other 6 generators, however, usually denoted $J_{\mu\nu}$, do not commute with P_μ. (It may be useful to remark what a "disappointment" this is, since the "obvious" spin operators would be J_{0i}, $(i = 1, 2, 3)$, which generate SU(2).) A crucial step forward, however, was made by Pauli, who defined [5] the operator

$$W_\mu = \frac{1}{2} \epsilon_{\mu\nu\rho\sigma} J^{\rho\sigma} P^\nu \,, \tag{13}$$

generally known as the Pauli–Lubanski pseudovector. It is, of course, a 4–vector, but because of the epsilon symbol only the intrinsic spin part of $J_{\rho\sigma}$, and not the orbital part (which contains P_μ), will contribute to W_μ, so W_μ clearly has something to do with spin. In addition,

$$W = W_\mu W^\mu \tag{14}$$

commutes with all ten generators of the Poincaré group, and is therefore the second Casimir operator of this group. For timelike states with mass M, $W = M^2 s(s+1)$, where s is the spin, so it is clear that M and W are the invariants labelling mass and spin. The question still remains, however: what are the spin operators, which generate, for timelike states, the group SU(2)?

This question was all but answered by Gürsey [6] in 1965; the present author did the mopping up recently [7]. First the commutator

$$W_{\mu\nu} = \frac{1}{M^2}[W_\mu, W_\nu] \tag{15}$$

is defined, as well as its dual

$$W^*_{\kappa\lambda} = \frac{1}{2}\epsilon_{\kappa\lambda\rho\sigma}W^{\rho\sigma}\,. \tag{16}$$

The combinations

$$X_{\mu\nu} = -i\left(W_{\mu\nu} + iW^*_{\mu\nu}\right) \tag{17}$$

$$Y_{\mu\nu} = -i\left(W_{\mu\nu} - iW^*_{\mu\nu}\right) = iX^*_{\mu\nu} \tag{18}$$

$$\tag{19}$$

may be shown to obey the commutation relations

$$[X_{\mu\nu}, X_{\kappa\lambda}] = -i\left(g_{\mu\kappa}X_{\nu\lambda} - g_{\mu\lambda}X_{\nu\kappa} + g_{\nu\lambda}X_{\mu\kappa} - g_{\nu\kappa}X_{\mu\lambda}\right) \tag{20}$$

and similarly for Y. These are the same in form as the commutation relations for the angular momentum operators $J_{\mu\nu}$ which generate the homogeneous Lorentz group, and it therefore follows, by familiar arguments, that the two sets of operators

$$X_i = \frac{1}{2}\epsilon_{ijk}X_{jk}\,, \qquad Y_i = \frac{1}{2}\epsilon_{ijk}Y_{jk} \tag{21}$$

obey the SU(2) commutation relations

$$[X_i, X_j] = \epsilon_{ijk}X_k\,, \qquad [Y_i, Y_j] = \epsilon_{ijk}Y_k\,. \tag{22}$$

They therefore generate two SU(2) groups, and they are both relativistic spin operators. In fact they are the operators appropriate for the left–handed and right–handed states, so that the spin operator appropriate for Dirac particles is given by

$$Z_i = \frac{1}{2}(1-\gamma_5)X_i + \frac{1}{2}(1+\gamma_5)Y_i\,. \tag{23}$$

Substituting the relations

$$W_0 = \frac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{p}\,, \qquad W_i = \frac{1}{2}m\sigma_i + \frac{p_i}{2(E+m)}\boldsymbol{\sigma}\cdot\boldsymbol{p} \tag{24}$$

into the above equations, we find

$$\boldsymbol{Z} = \frac{1}{2}\boldsymbol{\sigma}\gamma^0 - \frac{\boldsymbol{p}}{2(E+m)}\gamma_5(1+\gamma^0) \tag{25}$$

which, on comparison with equation (11), is the same as the Foldy–Wouthuysen operator $\boldsymbol{X}_{\mathrm{FW}}$ when operating on positive energy states. We have therefore shown that a relativistic spin operator can be defined for Dirac particles. The starting point for its discovery is the realisation that the group of motion, or isometry group, of Minkowski space, is the Poincaré group; the spin operator is defined in terms of the generators of this group.

3 General Relativity

The concluding remarks of the last paragraph make clear that in a general Riemannian space, since there is no such simple isometry group as the Poincaré group, there is also no spin operator; spin, that is to say, will not be conserved in a curved space–time. It is therefore inevitable that *spin precession* will take place in a gravitational field. Relativists are, of course, quite familiar with this fact, and understand well the differing contributing factors to spin precession; Thomas precession, geodetic precession and Lense–Thirring precession. The purpose of the observation above is to place these facts in a somewhat more fundamental context, by drawing attention to the connection between the existence of a (conserved) spin operator and that of an isometry group of the metric.

It is interesting to note finally that there is one space besides Minkowski space with a ten parameter isometry group, and that is de Sitter space, a space of constant curvature which, in the Euclideanised version, can be regarded as S^4, a hypersurface embedded in 5–dimensional Euclidean space. The "radius" of this S^4 may be denoted R and the limit $R \to \infty$ corresponds to the transition to Minkowski space. The de Sitter group, like the Poincaré group, has rank 2 and the associated two Casimir operators may be defined [8]. It turns out, hardly surprisingly, that in the limit $R \to \infty$ these two Casimir operators become, in essence, mass and spin, the Casimir operators of the Poincaré group. But this is only in the limit $R \to \infty$; in de Sitter space proper, the operator which is conserved does not correspond with pure spin. And in any case, as just remarked, de Sitter space is a very special case. In a general Riemannian space no conserved spin operator may be defined. If the curvature of the space is small – which in astronomical environments such as our own it is – then spin, as defined with reference to the Poincaré group, is only *approximately* conserved.

References

1. J.D. Bjorken and S.D. Drell: *Relativistic Quantum Mechanics* (McGraw–Hill, New York 1964).
2. L.L. Foldy and S.A. Wouthuysen: *Phys. Rev.* **78**, 29 (1950).
3. F.W. Hehl and W-T. Ni: *Phys. Rev.* **D 42**, 2045 (1990).
4. E.P. Wigner: *Ann. Math.* **40**, 149 (1939).
5. W. Pauli: *Ergebnisse der Exakten Naturwissenschaften* **37**, 85 (1965).
6. F. Gürsey: *Phys. Lett.* **14**, 330 (1965).
7. L.H. Ryder: *Gen. Rel. Grav.* **31**, 775 (1999).
8. F. Gürsey: Introduction to the De Sitter group, in F. Gürsey (ed), *Group Theoretical Concepts and Methods in Elementary Particle Physics* (Gordon and Breach, 1964).

Testing the Dirac Equation

Claus Lämmerzahl[2] and Christian J. Bordé[1]

[1] Fakultät für Physik, Universität Konstanz, 78457 Konstanz, Germany
[2] Laboratoire de Physique des Lasers, Institut Galilée
Université Paris 13, 93430 Villetaneuse, France

Abstract. The dynamical equations which are basic for the description of the dynamics of quantum felds in arbitrary space–time geometries, can be derived from the requirements of a unique deterministic evolution of the quantum fields, the superposition principle, a finite propagation speed, and probability conservation. We suggest and describe observations and experiments which are able to test the unique deterministic evolution and analyze given experimental data from which restrictions of anomalous terms violating this basic principle can be concluded. One important point is, that such anomalous terms are predicted from loop gravity as well as from string theories. Most accurate data can be obtained from future astrophysical observations. Also, laboratory tests like spectroscopy give constraints on the anomalous terms.

1 Introduction

Experimental Quantum Gravity, the experimental search for deviations from Einstein's General Relativity, which includes also Special Relativity, has been an very active area in physics since a few years. All unifying theories or quantum gravity theories predict small modifications from General Relativity, for example deviations from the $1/r$–potential, and violation of the equivalence principle, see e.g. [1], violation of Lorentz–invariance and violations of the universality of the gravitational red shift. Violations of these basic principles underlying General Relativity go together with a modification of the equations of motion for test matter in the gravitational field.

Questions about the structure of the field equations for quantum objects came up very recently in the context of quantum gravity: From loop gravity [2] as well as from string theory [3] there are predictions about quantum gravity induced modifications of the field equations governing the motion of spin–$\frac{1}{2}$–particles. Up to now the predictions consist in directional derivatives and higher–order spatial derivatives added to the usual Dirac equation. However, more general modifications can also be expected. For example, the prediction of second order spatial derivatives, which is worked out in a distinguished frame of reference, makes it reasonable to expect also second order time derivatives. Therefore, it is important to study experimental consequences of general non–standard modifications of the Dirac equation. In order to have a guiding principle at hand which tells us something about the meaning and the physical consequences of certain modifications of the usual Dirac equation, we will study these modifications in the frame of a constructive axiomatic derivation of the Dirac equation.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 463–476, 2001.

In that approach it is possible to derive the Dirac equation from a very few fundamental *operational* principles which can be tested directly in experiments. The derivation of the Dirac equation can be divided into two parts: The first part consists in the derivation of a system of linear hyperbolic field equations, which we call a generalized Dirac equation, from four fundamental principles which the quantum field has to obey, namely (i) the unique deterministic evolution, (ii) the superposition principle, (iii) finite propagation speed, and (iv) conservation of probability:

$$\left.\begin{array}{l}\text{unique deterministic evolution}\\ \text{superposition principle}\\ \text{finite propagation speed}\\ \text{conservation law}\end{array}\right\} \Rightarrow \begin{cases}\text{generalized Dirac–equation}\\ 0 = i\gamma^a(x)\partial_a\varphi(x) - M(x)\varphi(x)\end{cases}$$

Here γ^a $(a = 0, \ldots, 3)$ are some matrices which are not assumed to fulfill a Clifford algebra. Also M is a matrix.

The usual Dirac equation where the matrices γ^a fulfill a Clifford algebra and where M is proportional to the unit matrix, can be derived from the additional assumptions (v) uniquenes of the null cones, (vi) two helicity states only, and (vii) uniqueness of the mass shell, compare [4,5]. From these demands we arrive at a Dirac equation in Riemann–Cartan space–time, where the coupling to torsion consists in the axial part only. All these principles are operational since they can be proven directly by experiments.

Some basic features of quantum theory mentioned above have been questioned previously and subsequently been tested or estimated from some existing data on atoms, for example. One of these features is the linearity of quantum theory which is basic in our understanding of all quantum phenomena. A generalized quantum field equation including a non–linear term has been introduced and discussed by e.g. Shimony [6], and Weinberg [7], and references therein. Shimony himself proposed a neutron interferometry experiment which subsequently was performed by Shull et al [8] giving a strong restriction on the strength of a hypothetical nonlinear term in the Schrödinger equation. Also spectral data of the hydrogen atom have been used for this purpose. Another features which has been discussed was the conservation of probability [9]. The second part of assumptions (v) to (vii) manifests itself in a breaking of local Lorentz– and local position invariance which can be tested by Hughes–Drever, red–shift and atomic interferometry experiments [10].

In this paper we want to question the first of the four basic principles (i) to (iv) underlying our basic understanding of quantum theory, namely the unique deterministic evolution. Here, unique deterministic evolution means that if a quantum state $\psi(x)$ is prepared at a time t_0, then the state is uniquely determined for times $t > t_0$. This implies that the evolution of quantum states is described by an evolution equation which contains a first time derivative only, that is, $\frac{d}{dt}\psi = A\psi$, where A is some operator.

In order to test this principle, we propose a generalization of the usual Dirac equation by adding a second time derivative which violates this principle. This

modified Dirac equation is used to calculate modifications of the propagation of spin–$\frac{1}{2}$–particles as well as the corresponding modifications of the atomic spectrum (The hydrogen spectrum can also be used as justification of the four–dimensionality of space–time on atomic scales [11]). Up to now, all the experimental results are well explained using the standard theory, that is, the usual Dirac equation together with quantum corrections. If everything is well explained using the standard theory, then the modifications of the results due to the modifications in the Dirac equation, can be only smaller than the experimental error. Therefore, all the modifications can be restricted by comparing the calculated effects with the accuracy for the various measured effects. – However, future observation of neutrinos and high energy photons from gamma ray bursts (GRB) may be capable to distinguish between the various models of the Dirac equation. Nevertheless, one has to keep in mind, that on cosmological distances the parameters may depend on the position and thus the effect we are looking for may be cancelled during the propagation over long distances. Therefore, these observations have to be complemented by laboratory experiments like spectroscopy.

2 The Model: A Modification of the Dirac Equation

The unique deterministic evolution implies that the evolution equation for the quantum field is of first order in the time derivative. This means especially, that the evolution equation is an equation without memory. In terms of a system of partial differential equations this means that this system should be of first order in time as it is the case for the Dirac equation $i\hbar\partial_t\psi = -i\hbar c\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}\psi + \beta mc^2\psi$. If an evolution possesses a memory, the time derivative has to be replaced by an operator, for example, an integral expression: $B\psi = -i\hbar c\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}\psi + \beta mc^2\psi$ with, for example, $B\psi = \int B(t,t')\psi(t')dt'$. In the case that the kernel $B(t,t')$ of that kind of equation possesses certain properties (it should be analytic), then one can expand that kernel resulting in a system of partial differential equations with an infinite sum of terms containing time derivatives of arbitrary order, $\sum_0^\infty a_i\partial_t^i\psi$.

Therefore, if the quantum field does not evolve uniquely deterministic or if quantum theory has a memory, then the resulting field equation in these cases contains arbitrary high orders of time derivatives. In a first approximation, this may be modeled by adding to a conventional quantum field equation like the Dirac equation or the Schrödinger equation, a term containing a second time derivative:

$$i\hbar\frac{\partial}{\partial t}\psi = -i\hbar c\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}\psi + \beta mc^2\psi - \epsilon\frac{\hbar^2}{mc^2}\frac{\partial^2}{\partial t^2}\psi\,. \tag{1}$$

In order to make ϵ dimensionless, we introduced a factor $1/mc^2$ in the term containing the last term. Here we assume that ϵ is constant, i.e. does not depend on time or position. For $\epsilon = 0$, the above equation reduces to the usual Dirac equation. It is clear that the last term in (1) violates Lorentz covariance.

This modified Dirac equation can be derived from a Lagrangian:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}i\hbar\left(\psi^+\partial_t\psi - \partial_t\psi^+\psi\right)\\ &\quad + \frac{1}{2}i\hbar c\left(\psi^+\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}\psi - \boldsymbol{\nabla}\psi^+\cdot\boldsymbol{\alpha}\psi\right) - mc^2\psi^+\beta\psi + \epsilon\frac{\hbar^2}{mc^2}\partial_t\psi^+\partial_t\psi \qquad (2)\\ &= \frac{1}{2}i\hbar\left(\bar\psi\gamma^a\partial_a\psi - \partial_a\bar\psi\gamma^a\psi\right) - mc^2\bar\psi\psi - \epsilon\frac{\hbar^2}{mc^2}\partial_t\bar\psi\gamma^0\partial_t\psi\,. \qquad (3)\end{aligned}$$

This implies, in particular, that we have a conservation law

$$0 = \partial_t\rho + \boldsymbol{\nabla}\cdot\boldsymbol{j} \qquad (4)$$

with the probability density

$$\rho = \psi^+\psi - \epsilon\frac{\hbar^2}{mc^2}i\left(\partial_t\psi^+\psi - \psi^+\partial_t\psi\right) \qquad (5)$$

and a current

$$\boldsymbol{j} = \hbar c\psi^+\boldsymbol{\alpha}\psi\,. \qquad (6)$$

The coupling to the electromagnetic field can, as usual, be introduced by means of the minimal coupling procedure:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}i\hbar\left(\psi^+(\partial_t + ie\phi)\psi - (\partial_t - ie\phi)\psi^+\psi\right)\\ &\quad + \frac{1}{2}i\hbar c\left(\psi^+\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\boldsymbol{A}\right)\psi - \left(\boldsymbol{\nabla} + \frac{ie}{\hbar c}\boldsymbol{A}\right)\psi^+\cdot\boldsymbol{\alpha}\psi\right)\\ &\quad - mc^2\psi^+\beta\psi + \epsilon\frac{\hbar^2}{mc^2}(\partial_t - ie\phi)\psi^+(\partial_t + ie\phi)\psi\,. \qquad (7)\end{aligned}$$

The corresponding modified Dirac equation is

$$i\hbar\partial_t\psi = -i\hbar c\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\boldsymbol{A}\right)\psi + \beta mc^2\psi + e\phi + \epsilon\frac{\hbar^2}{mc^2}(\partial_t - ie\phi)^2\psi\,. \qquad (8)$$

3 Plane Wave Solutions and Neutrino Propagation

It is not difficult to present an exact plane wave solution for (1). Inserting the ansatz $\psi = \exp\left(i(Et - \boldsymbol{p}\cdot\boldsymbol{x})\right)a$ into (1) gives $Ea = \left(\boldsymbol{\alpha}\cdot\boldsymbol{p}c + \beta mc^2 + \frac{\epsilon}{mc^2}E^2\right)a$, or[1]

$$0 = \left(\gamma^0\left(E - \frac{\epsilon}{mc^2}E^2\right) + \gamma^{\hat a}p_{\hat a}c - mc^2\right)a\,. \qquad (9)$$

The corresponding dispersion relation reads

$$\left(E - \frac{\epsilon}{mc^2}E^2\right)^2 - \boldsymbol{p}^2c^2 = m^2c^4 \qquad (10)$$

[1] Indices with a hat run from 1 to 3 and, for example, $p_{\hat a}$ is represented by $\boldsymbol{p}$.

which possesses the four solutions

$$\begin{aligned} E_{\pm}^{(1)} &= \frac{mc^2 - \sqrt{m^2c^4 \mp 4\epsilon mc^3\sqrt{\boldsymbol{p}^2c^2 + m^2c^2}}}{2\epsilon} \\ E_{\pm}^{(2)} &= \frac{mc^2 + \sqrt{m^2c^4 \pm 4\epsilon mc^3\sqrt{\boldsymbol{p}^2c^2 + m^2c^2}}}{2\epsilon}\,. \end{aligned} \tag{11}$$

For small ϵ this reduces to

$$\begin{aligned} E_{\pm}^{(1)} &= \pm c\sqrt{\boldsymbol{p}^2 + m^2c^2} + \epsilon\left(mc^2 + \frac{\boldsymbol{p}^2}{m}\right) \\ E_{\pm}^{(2)} &= \frac{mc^2}{\epsilon} \pm c\sqrt{\boldsymbol{p}^2 + m^2c^2} - \epsilon\left(mc^2 + \frac{\boldsymbol{p}^2}{m}\right)\,. \end{aligned} \tag{12}$$

That means that for $\epsilon \to 0$ the solutions $E_{\pm}^{(1)}$ reduce to the well–known solutions. The other two solutions are new and diverge for small ϵ. However, this large quantity will drop out by considering energy differences. The $\pm$ in the solutions for the energy corresponds to positive/negative energies, the (1), (2) corresponds to the two solutions which come up with $\epsilon \neq 0$.

In the high energy limit $m \to 0$ we get ($p = |\boldsymbol{p}|$)

$$\begin{aligned} E_{\pm}^{(1)} &= \pm cp \pm \frac{m^2c^3}{2p} + \epsilon\left(mc^2 + \frac{p^2}{m}\right) \\ E_{\pm}^{(2)} &= \frac{mc^2}{\epsilon} \pm cp \pm \frac{m^2c^3}{2p} - \epsilon\left(mc^2 + \frac{\boldsymbol{p}^2}{m}\right)\,. \end{aligned} \tag{13}$$

In the low energy limit $\boldsymbol{p} \to 0$ we get instead

$$\begin{aligned} E_{\pm}^{(1)} &= \pm mc^2 \pm \frac{\boldsymbol{p}^2}{2m} + \epsilon\left(mc^2 + \frac{\boldsymbol{p}^2}{m}\right) \\ E_{\pm}^{(2)} &= \frac{mc^2}{\epsilon} \pm mc^2 \pm \frac{\boldsymbol{p}^2}{2m} - \epsilon\left(mc^2 + \frac{\boldsymbol{p}^2}{m}\right)\,. \end{aligned} \tag{14}$$

The exact expression for the group velocity reads

$$\begin{aligned} (v_{\pm}^{(1)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}c}{\sqrt{\boldsymbol{p}^2 + m^2c^2}\sqrt{1 \mp 4\frac{\epsilon}{mc}\sqrt{\boldsymbol{p}^2 + m^2c^2}}} \\ (v_{\pm}^{(2)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}c}{\sqrt{\boldsymbol{p}^2 + m^2c^2}\sqrt{1 \pm 4\frac{\epsilon}{mc}\sqrt{\boldsymbol{p}^2 + m^2c^2}}}\,. \end{aligned} \tag{15}$$

Therefore, for positive energies, particles with the same momentum but with different directions propagate with different velocities (the same is true for particles with negative energies). This property can be used for a comparison with data from neutrino propagation.

For small ϵ we get from (15)

$$\begin{aligned}(v_{\pm}^{(1)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}c}{\sqrt{\boldsymbol{p}^2+m^2c^2}} + 2\epsilon\frac{p_{\hat{a}}}{m}\\ (v_{\pm}^{(2)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}c}{\sqrt{\boldsymbol{p}^2+m^2c^2}} - 2\epsilon\frac{p_{\hat{a}}}{m}\,,\end{aligned} \tag{16}$$

for $m \to 0$

$$\begin{aligned}(v_{\pm}^{(1)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}}{p}c \mp \frac{m^2c^3}{2p^3}p_{\hat{a}} + 2\epsilon\frac{p_{\hat{a}}}{m}\\ (v_{\pm}^{(2)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}}{p}c \mp \frac{m^2c^3}{2p^3}p_{\hat{a}} - 2\epsilon\frac{p_{\hat{a}}}{m}\,,\end{aligned} \tag{17}$$

and for $p_{\hat{a}} \to 0$

$$\begin{aligned}(v_{\pm}^{(1)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}}{m} + 2\epsilon\frac{p_{\hat{a}}}{m}\\ (v_{\pm}^{(2)})^{\hat{a}} &= \pm\frac{p_{\hat{a}}}{m} - 2\epsilon\frac{p_{\hat{a}}}{m}\,.\end{aligned} \tag{18}$$

For a comparison with data from the propagation of neutrinos, which may be produced in connection with GRBs, we use (17). We compare the arrival time of neutrinos with the arrival time of light over a distance of $l = 10^{10}$ ly. If the neutrinos and the photons are produced during the same event, and if we take the mass of the neutrinos to be 1 eV and the momentum $p = 10^5$ GeV, then we get as difference of the time–of–arrival

$$\Delta t = \frac{l}{c} - \frac{l}{v_{+}^{(1,2)}} \approx \frac{l}{c}\left(\frac{m^4c^4}{8p^4} \pm 2\epsilon\frac{p}{mc}\right) \approx \left(4\times10^{-9} \pm 6.4\,\epsilon\,10^{31}\right)\ \text{sec}\,. \tag{19}$$

The first term can be neglected compared to the second one, so that we get $|\Delta t| = |\epsilon|6.4\times10^{31}$ sec. Asssuming a temporal structure of the source of about a millisecond [12] and *assuming a null–result*, then we can get from observations of the propagation of neutrinos and of photons the estimate

$$|\epsilon| \le 1.6\times10^{-35}\,. \tag{20}$$

Thus, neutrino observations in the future have the potentiality of high precision determination of the parameter ϵ. Any $|\epsilon|$ larger than that given by (20) should be detectable by this means.

In quantum gravity theories, $|\epsilon|$ is proportional to the ratio of the Planck length and some intermediate length, $\epsilon = \kappa l_p/L$ where κ is assumed to be of the order 1 [38]. If we take $L = \hbar/p$, then, in terms of κ, the above estimate means $|\kappa| \le 1.6\times10^{-21}$ which certainly is in contradiction to the assumption that κ is of the order 1. From this we conclude, that, if the Dirac equation contains an additional quantum gravity induced term with the second time derivative of the neutrino field, then this term should be observable in the future by comparing neutrino propagation with photon propagation.

However, from the derivation of the modifications of the Dirac equation [38], the parameter ϵ or κ may be constant over the scale L of the "weave" states only. Therefore, it may be possible that the effect, as we calculated it, may not occur due to an averaging to zero over larger distances. Consequently, it is necessary also to perform tests which take place on a small scale only. One kind of such tests is atomic spectroscopy what we are going to discuss below.

4 The Non–relativistic Limit

4.1 The Non–relativistic Field Equation

First we calculate the modified Pauli equation corresponding to eqn (1). In order to do so, we first subtract the rest energy from the wave function by means of the substitution

$$\psi = e^{-\frac{i}{\hbar}(1+\epsilon)mc^2 t}\psi' \tag{21}$$

resulting in an elimination of the rest mass in one part of the wave function. This gives

$$\begin{aligned}(1+2\epsilon(1+\epsilon))i\hbar\partial_t\psi' = &-i\hbar c\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla}-\frac{ie}{\hbar c}\boldsymbol{A}\right)\psi' + (\beta-1)mc^2\psi' + e\phi\psi' \\ &-\epsilon\frac{\hbar^2}{mc^2}(\partial_t + ie\phi)^2\psi' \,.\end{aligned} \tag{22}$$

With the projection operators $P_\pm := \frac{1}{2}(1\pm\beta)$ we define the 'large' and 'small' parts of the wave function: $\psi'_\pm = P_\pm\psi'$. Multiplication of (1) with P_+ and P_- gives the two equations

$$(1+2\epsilon(1+\epsilon))i\hbar\partial_t\psi'_+ = -i\hbar c\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla}-\frac{ie}{\hbar c}\boldsymbol{A}\right)\psi'_- + e\phi\psi_+ - \epsilon\frac{\hbar^2}{mc^2}(\partial_t+ie\phi)^2\psi'_+ \tag{23}$$

$$\begin{aligned}(1+2\epsilon)i\hbar\partial_t\psi'_- = &-i\hbar c\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla}-\frac{ie}{\hbar c}\boldsymbol{A}\right)\psi'_+ + e\phi\psi_- - 2mc^2\psi'_- \\ &-\epsilon\frac{\hbar^2}{mc^2}(\partial_t+ie\phi)^2\psi'_- \,.\end{aligned} \tag{24}$$

As usual, we assume that in the second equation the energies $i\hbar\partial_t\psi'_-$ and $e\phi\psi_-$ are small compared with the rest mass term $mc^2\psi'_-$. Therefore, we approximate

$$\psi'_- \approx -\frac{1}{2mc}i\hbar\boldsymbol{\alpha}\cdot\left(\boldsymbol{\nabla}-\frac{ie}{\hbar c}\boldsymbol{A}\right)\psi'_+ \,. \tag{25}$$

Inserting this into the first equation (23) gives

$$\begin{aligned}(1+2\epsilon(1+\epsilon))i\hbar\partial_t\psi'_+ = &-\frac{\hbar^2}{2m}\left(\boldsymbol{\nabla}-\frac{ie}{\hbar c}\boldsymbol{A}\right)^2\psi'_+ + e\phi\psi'_+ \\ &+\frac{e}{mc}\boldsymbol{\Sigma}\cdot\boldsymbol{B} - \epsilon\frac{\hbar^2}{mc^2}(\partial_t+ie\phi)^2\psi'_+ \,,\end{aligned} \tag{26}$$

with $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma} & 0 \\ 0 & \boldsymbol{\sigma} \end{pmatrix}$ and $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$. Here $\boldsymbol{\sigma}$ are the three Pauli matrices. After division through $1 + 2\epsilon(1+\epsilon)$ we get

$$i\hbar\partial_t \psi'_+ = -\frac{\hbar^2}{2m^*}\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\boldsymbol{A}\right)^2 \psi'_+ + e^*\phi\psi'_+ + \frac{e}{m^*c}\boldsymbol{\Sigma}\cdot\boldsymbol{B} - \epsilon\frac{\hbar^2}{m^*c^2}(\partial_t + ie\phi)^2\psi'_+ , \tag{27}$$

where we absorbed the ϵ–factor in a redefinition of mass and charge: $m^* := (1 + 2\epsilon(1+\epsilon))m$, $e^* = e/(1 + 2\epsilon(1+\epsilon))$.

As our result we get a modified Schrödinger equation for a two–spinor ψ

$$i\hbar\frac{\partial}{\partial t}\psi = -\frac{\hbar^2}{2m^*}\left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\boldsymbol{A}\right)^2 \psi + e^*\phi\psi + \frac{e}{mc}\boldsymbol{\sigma}\cdot\boldsymbol{B} - \epsilon\frac{\hbar^2}{m^*c^2}(\partial_t + ie\phi)^2\psi . \tag{28}$$

This equation can also be derived from a Lagrangian

$$\begin{aligned}\mathcal{L} = {} & \frac{i}{2}\hbar(\psi^*(\partial_t - ie\phi)\psi - (\partial_t + ie\phi)\psi^*\psi) \\ & -\frac{\hbar^2}{2m^*}\left(\boldsymbol{\nabla} + \frac{ie}{\hbar c}\boldsymbol{A}\right)\psi^* \cdot \left(\boldsymbol{\nabla} - \frac{ie}{\hbar c}\boldsymbol{A}\right)\psi - \epsilon\frac{\hbar^2}{m^*c^2}(\partial_t - ie\phi)\psi^*(\partial_t + ie\phi)\psi .\end{aligned} \tag{29}$$

Equation (28) also possesses plane wave solutions whose energies are given by (14). We also have a conservation law

$$\frac{d}{dt}\rho + \boldsymbol{\nabla}\boldsymbol{j} = 0 \tag{30}$$

with the probability

$$\rho = \psi^*\psi + i\epsilon\frac{\hbar}{m^*c^2}\left(\frac{\partial}{\partial t}\psi^*\psi - \psi^*\frac{\partial}{\partial t}\psi\right) . \tag{31}$$

and the current

$$\boldsymbol{j} = \frac{i\hbar}{2m^*}\left(\nabla\psi^*\psi - \psi^*\nabla\psi\right) . \tag{32}$$

Like in the Klein–Gordon equation there seems to exist the possibility to get negative probablities. However, using the Schrödinger equation, we get for the probability

$$\begin{aligned}\rho &= \psi^*\psi - \epsilon\frac{1}{mc^2}\left(-i\hbar\frac{\partial}{\partial t}\psi^*\psi + \psi^* i\hbar\frac{\partial}{\partial t}\psi\right) \\ &= \psi^*\psi + \epsilon\frac{1}{m^*c^2}\left(\frac{\hbar^2}{2m^*}\Delta\psi\psi + \psi^*\frac{\hbar^2}{2m^*}\Delta\psi\right) .\end{aligned} \tag{33}$$

This quantity is strictly positive if $\epsilon < 0$, and for $\epsilon > 0$ this is positive if $|\epsilon|$ is small enough.

4.2 Modifications of the Energy Levels

It is possible to calculate the energy levels of the hydrogen atom exactly. In order to do so we choose $\boldsymbol{A} = 0$ and $\phi = \phi(r)$ as the usual spherically symmetric electrostatic potential:

$$i\hbar\frac{\partial}{\partial t}\psi = -\frac{\hbar^2}{2m^*}\Delta\psi + e^*\phi\psi + \epsilon\frac{1}{m^*c^2}\left(i\hbar\frac{\partial}{\partial t} - e\phi\right)^2\psi\,. \tag{34}$$

We asume a stationary solution, then $i\hbar\partial_t\psi = E\psi$, and we get from the above Hamiltonian

$$\left(E - \epsilon\frac{1}{m^*c^2}E^2\right)\psi = -\frac{\hbar^2}{2m^*}\Delta\psi + \left(e^* - 2\epsilon\frac{e}{m^*c^2}E\right)\phi\psi + \epsilon\frac{1}{m^*c^2}e^2\phi^2\psi\,. \tag{35}$$

With $\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) - \frac{1}{r^2}\widehat{\boldsymbol{L}}^2$, where $\widehat{\boldsymbol{L}}$ is the angular momentum operator with the eigenvalue equation $\widehat{\boldsymbol{L}}^2 Y_l^m = l(l+1)Y_l^m$ with $l = 1, 2, 3, \ldots$. With a splitting of the wave function into a radial and an angular part $\psi = R(r)Y_l^m(\vartheta, \varphi)$ we get the radial part of the wave equation

$$\begin{aligned}\left(E - \epsilon\frac{1}{m^*c^2}E^2\right)R = &-\frac{\hbar^2}{2m^*}\left(\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}R\right) - \frac{1}{r^2}l(l+1)R\right)\\ &+\left(e^* - 2\epsilon\frac{e}{m^*c^2}E\right)\phi R + \epsilon\frac{1}{m^*c^2}e^2\phi^2 R\,.\end{aligned} \tag{36}$$

With the explicit expression for the electrostatic potential $\phi = -e/r$ we get by multiplication with $2m^*/\hbar^2$

$$\begin{aligned}0 = &\frac{d^2}{dr^2}R + \frac{2}{r}\frac{d}{dr}R + \frac{2m^*}{\hbar^2}\left(E - \epsilon\frac{1}{m^*c^2}E^2\right)R\\ &+\frac{2m^*}{\hbar^2}\left(e^* - 2\epsilon\frac{e}{m^*c^2}E\right)\frac{e}{r}R - \frac{1}{r^2}\left(l(l+1) + \epsilon\frac{2e^4}{\hbar^2c^2}\right)R\,.\end{aligned} \tag{37}$$

Since the r–dependence is the same as for the usual hydrogen atom, this equation can be solved using the standard scheme: With

$$\mathcal{E} := \frac{2m^*}{\hbar^2}\left(E - \frac{\epsilon}{m^*c^2}E^2\right), \quad q := \frac{2m^*e}{\hbar^2}\left(e^* - \frac{2\epsilon e}{m^*c^2}E\right), \quad \ell := l(l+1) + 2\epsilon\alpha^2 \tag{38}$$

where $\alpha = e^2/\hbar c$ is the fine structure constant, we get for the radial part of the wave function

$$\left(\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} + \mathcal{E} + \frac{q}{r} - \frac{\ell}{r^2}\right)R = 0\,. \tag{39}$$

We introduce dimensionless coordinates by

$$r' = \frac{r}{r_0}, \quad -\mathcal{E} = \frac{1}{4r_0^2}, \quad q' = qr_0 \tag{40}$$

and get with $R'(r') := R(r(r'))$

$$\left(\frac{d^2}{dr'^2} + \frac{2}{r'}\frac{d}{dr'} - \frac{1}{4} + \frac{q'}{r'} - \frac{1}{r'^2}\ell\right) R' = 0\,. \tag{41}$$

We also introduce a new variable $f(r')$ through

$$R = e^{-\frac{1}{2}r'} r'^{\gamma} f(r') \tag{42}$$

with a parameter γ which will be specified later. We get the following equation for the function f:

$$0 = r\frac{d^2 f(r')}{dr'^2} + (2\gamma + 2 - r)\frac{df(r')}{dr'} + \left(\frac{\gamma(\gamma+1) - \ell}{r'} + q' - \gamma - 1\right) f(r')\,. \tag{43}$$

In order to solve this equation we specify the value of γ by the requirement that the term $\sim 1/r'$ should vanish:

$$\gamma\,(\gamma + 1) - \ell = 0\,. \tag{44}$$

This gives the two possibilities

$$\gamma_\pm = -\frac{1}{2} \pm \sqrt{\ell + \frac{1}{4}} = -\frac{1}{2} \pm \sqrt{\left(l + \frac{1}{2}\right)^2 + 2\epsilon\alpha^2}\,, \tag{45}$$

and from (43) the differential equations

$$zf'' + (\vartheta_\pm - z)f' - \beta_\pm f = 0\,, \tag{46}$$

with

$$\vartheta_\pm = 2\gamma_\pm + 2 = 1 \pm 2\sqrt{\ell + \frac{1}{4}} \tag{47}$$

$$\beta_\pm = \gamma_\pm + 1 - q' = \frac{1}{2} \pm \sqrt{\ell + \frac{1}{4}} - q'\,. \tag{48}$$

Eqn. (46) is the confluent hypergeometric differential equation with the solution [13]

$$f(\beta_\pm, \vartheta_\pm, z) = \sum_{\nu=0}^{\infty} \frac{(\beta_\pm + \nu)!\vartheta_\pm!}{\beta_\pm!(\vartheta_\pm + \nu)!}\frac{z^\nu}{\nu!}\,. \tag{49}$$

which is appropriate for our problem.

It is clear that, in order to get no infinite terms, $\vartheta_\pm$ is not allowed to be a negative integer: $\vartheta_\pm \neq -1, -2, \ldots$, which is fulfilled if $\epsilon \neq 0$ and $|\epsilon| < 1$. For $\epsilon = 0$ we cannot use the solution ϑ_-. In addition, if the sum in (49) does not terminate, then the solution diverges for large r faster than $\exp\left(\frac{1}{2}r'\right)$ which leads to non–normalizable solutions. The condition for a termination of the sum is $\beta_\pm \in \mathbb{Z}^-$, or

$$\beta_\pm = -k, \qquad k \in \mathbb{N}\,. \tag{50}$$

With (48), (40), and (38) we get four energy eigenvalues E

$$E_{\pm}^{(1)} = \frac{m^*c^2}{2\epsilon}\frac{K_{\pm}^2 - 2\epsilon\alpha\alpha^*\frac{e}{e^*}}{K_{\pm}^2 - 2\epsilon\alpha^2} + \frac{m^*c^2}{2\epsilon}\frac{K_{\pm}\sqrt{K_{\pm}^2 + 2\epsilon\alpha\alpha^*(1-2\frac{e}{e^*})}}{K_{\pm}^2 - 2\epsilon\alpha^2} \tag{51}$$

$$E_{\pm}^{(2)} = \frac{m^*c^2}{2\epsilon}\frac{K_{\pm}^2 - 2\epsilon\alpha\alpha^*\frac{e}{e^*}}{K_{\pm}^2 - 2\epsilon\alpha^2} - \frac{m^*c^2}{2\epsilon}\frac{K_{\pm}\sqrt{K_{\pm}^2 + 2\epsilon\alpha\alpha^*(1-2\frac{e}{e^*})}}{K_{\pm}^2 - 2\epsilon\alpha^2} \tag{52}$$

with $K_{\pm} = k - \frac{1}{2} \mp \sqrt{(l+\frac{1}{2})^2 + 2\epsilon\alpha^2}$ and $\alpha^* = {e^*}^2/\hbar c$. To first order in ϵ we get

$$E_{+}^{(1)} = m^*c^2\left(\frac{1}{\epsilon} + \frac{\alpha^2}{2}\frac{1}{(1-k+l)^2} - \alpha^4\epsilon\frac{5-8k+2l}{4(1-k+l)^4(1+2l)} + \mathcal{O}(\epsilon^2)\right) \tag{53}$$

$$E_{-}^{(1)} = m^*c^2\left(\frac{1}{\epsilon} + \frac{\alpha^2}{2}\frac{1}{(k+l)^2} + \epsilon\alpha^4\frac{3-8k-2l}{4(k+l)^4(1+2l)} + \mathcal{O}(\epsilon^2)\right) \tag{54}$$

$$E_{+}^{(2)} = -m^*c^2\left(\frac{\alpha^2}{2}\frac{1}{(1-k+l)^2} - \epsilon\alpha^4\frac{5-8k+2l}{4(1-k+l)^4(1+2l)} + \mathcal{O}(\epsilon^2)\right) \tag{55}$$

$$E_{-}^{(2)} = -m^*c^2\left(\frac{\alpha^2}{2}\frac{1}{(k+l)^2} + \epsilon\alpha^4\frac{3-8k-2l}{4(k+l)^4(1+2l)} + \mathcal{O}(\epsilon^2)\right). \tag{56}$$

The solution is then given by

$$R(r) = R(r(r')) = R'(r') = Ae^{-\frac{1}{2}r'}r'^{\gamma}f(r') = Ae^{-\frac{1}{2}r'}r'^{\gamma_{\pm}}f(-k,\vartheta_{\pm},r') \tag{57}$$

where A is a normalization constant and where all parameters depend on l.

For $\epsilon \neq 0$ all these energy values are well defined. Even for very small ϵ the first two energies (53,54) are valuable solutions, too, since only energy differences are observable and thus the first term drops out. However, there are two reasons which justify to drop the first two solutions: (i) Except the first term mc^2/ϵ, the upper two sets of energy levels (53,54) are the same as the lower two sets (55,56) up to sign. For the first two sets of energy levels the continuum is below the discrete spectrum. By postulating that all particles fall into the lowest energy level, then all atoms will fall into the continuous part of the spectrum which has never been observed. (One also can postulate that all particles want to go to the highest energy level. But due to the symmetry of the two sets of spectra, this will give the same answer. Therefore we do not consider the upper two sets by convention.) (ii) We want to describe small modifications of the known energy levels given for $\epsilon = 0$. Owing to these reasons, we keep the last two sets of energy levels.

Defining for $E_{+}^{(2)}$ the principal quantum number $n := 1 - k + l$, then the energy levels are given by

$$E_{+}^{(2)} = -m^*c^2\frac{\alpha^2}{2}\left(\frac{1}{n^2} + \epsilon\alpha^2\left(\frac{3}{2n^4} - \frac{4}{n^3(1+2l)}\right) + \mathcal{O}(\epsilon^2)\right) \tag{58}$$

and if we define for $E_-^{(2)}$ the principal quantum number $n := k + l$, then the energy levels are

$$E_-^{(2)} = -m^* c^2 \frac{\alpha^2}{2} \left(\frac{1}{n^2} + \epsilon\alpha^2 \left(\frac{3}{2n^4} - \frac{4}{n^3(1+2l)} \right) + \mathcal{O}(\epsilon^2) \right) , \tag{59}$$

where $mc^2\alpha^2/2$ is the Rydberg constant Ry. It is remarkable, that in both cases we get the same scheme of energy differences. Therefore, at this order we cannot distinguish between energies E_+ and E_-.

There are two differences between this result and the usual spectrum of the hydrogen atom: (i) No degeneracy with respect to l. (ii) Additional $1/n^4$– and $1/n^3$–terms. From (58) we get for the Lyman–series with $n = 1$ and $l = 0$ as ground state, for example, the energy differences

$$\Delta E_{\text{Lyman}} = Ry \left(1 - \frac{1}{n^2} + \epsilon\alpha^2 \left(\frac{3}{2} \left(1 - \frac{1}{n^4} \right) - 4 \left(1 - \frac{1}{n^3(1+2l)} \right) \right) + \mathcal{O}(\epsilon^2) \right) . \tag{60}$$

The additional terms $1/n^4$ and $n^3(1 + 2l)$ modify the structure of this series. The ionization energy is given by $n \to \infty$ and is

$$E_{\text{ionization}} = Ry \left(1 - \frac{5}{2}\epsilon\alpha^2 + \mathcal{O}(\epsilon^2) \right) . \tag{61}$$

From numbers related to ΔE_{Lyman} or $E_{\text{ionization}}$ one can draw estimates on the value of ϵ.

Since the accuracy $\delta\Delta E/\Delta E$ for recent measurements, see for example [14], is of the order of 10^{-13} which agrees completely with the conventional theory, we can conclude that the relative deviation from the usually calculated energy has to be smaller than this uncertainty:

$$\frac{\delta\Delta E}{\Delta E} = \epsilon\alpha^2 \frac{\left| -\frac{5}{2} - \frac{3}{2n^4} + \frac{4}{n^3} \right|}{1 - \frac{1}{n^2}} \leq 10^{-13} . \tag{62}$$

This means that

$$\epsilon \leq 10^{-13} \frac{1}{\alpha^2} \frac{1 - \frac{1}{n^2}}{\left| -\frac{5}{2} - \frac{3}{2n^4} + \frac{4}{n^3} \right|} = 10^{-13} \frac{1}{\alpha^2} \frac{3}{8 + \frac{3}{8}} \approx 7 \times 10^{-10} . \tag{63}$$

In terms of a κ as introduced after Eqn. (20) we have the estimate $\kappa \leq 7 \times 10^4$ which is outside the assumption that κ is of the order unity. Therefore, the accuracy of atomic spectroscopy is still at least five orders of magnitude too small in order to be able to detect any influence of quantum gravity on atomic levels.

5 Conclusion

We have shown that an additional term in the Dirac equation containing a second time derivative, as it is motivated from quantum gravity, influences neutrino

propagation and atomic spectroscopy. While neutrino propagation may be capable to 'see' this additional term, the accuracy of spectroscopy has still to be improved by five orders of magnitude in order to be sensitive to this term.

A clear difficulty with the hydrogen spectrum is that, precisely, it cannot be calculated exactly with an arbitrary accuracy with the present state–of–the–art. There is a number of corrections to the Dirac solutions: recoil, QED, finite nucleus size, see [16] for a review. These corrections scale with $1/n^3$ and amount to a few kilohertz. They are very difficult to calculate and there are still some discrepancies between theoretical results. Fortunately, some combinations of frequencies are independent of these corrections at their leading order and can be used for higher accuracies. Concerning the measurements themselves, there is presently a very rapid evolution towards much higher accuracies. For example, it was shown recently that, because of the extreme regularity of the frequency comb generated by femtosecond lasers over a very wide spectrum [17], they could be used to compare frequencies of oscillators which differ by several orders of magnitude. This provides a way to compare many transition frequencies of the hydrogen atom between themselves and with microwave clocks, with the potential accuracy of the cesium fountain clock which is presently 10^{-15} and should improve quickly by another order of magnitude. Also, the techniques to interrogate narrow transitions of cold atom hydrogen by sub–Doppler methods or atom interferometry have improved very significantly either in cold thermal beams (Hänsch and coworkers in Garching, Biraben and coworkers in Paris) or in clouds generated from Bose–Einstein condensates (Kleppner and coworkers at MIT). A subkilohertz linewidth is presently achieved for the 1S-2S two–photon transition [17] and could still be reduced by one or two orders of magnitude in the near future. The hydrogen atom is thus potentially a universal clock by itself covering the full spectrum from UV to microwaves (hydrogen maser).

Acknowledgement

Ch.B. thanks C. Lämmerzahl and J. Mlynek for their hospitality at the University of Konstanz. We acknowledge financial support of the Optik–Zentrum of the University of Konstanz.

References

1. E. Fischbach and C.L. Talmadge. *The Search for Non–Newtonian Gravity.* Springer–Verlag, New York, 1999.
2. J. Alfaro, H.A. Morales-Tecotl, and L.F. Urrutia. Quantum gravity corrections to neutrino propagation. *Phys. Rev. Lett.*, 84:to appear, 2000.
3. J. Ellis, N.E. Mavromatos, D.V. Nanopoulos, and G. Volkov. Gravitational–recoil effects on fermion propagation in space-time foam. gr-qc/9911055.
4. J. Audretsch and C. Lämmerzahl. A new constructive axiomatic scheme for the geometry of space–time. In Majer U. and Schmidt H.–J., editors, *Semantical Aspects of Space–Time Geometry*, page 21. BI Verlag, Mannheim, 1993.

5. C. Lämmerzahl and U. Bleyer. A quantum test theory for basic principles of special and general relativity with applications to atomic interferometry and Hughes–Drever type experiments, 1999. preprint, University of Konstanz.
6. A. Shimony. Proposed neutron interferometer test of some nonlinear variants of wave mechanics. *Phys. Rev.*, A 20:394, 1979.
7. S. Weinberg. Testing quantum mechanics. *Ann. Phys.*, 194:336, 1989.
8. C.G. Shull, D.K. Atwood, J. Arthur, and M.A. Horne. Search for a nonlinear variant of the Schrödinger equation by neutron interferometry. *Phys. Rev. Lett.*, 44:765, 1980.
9. J. Ellis, S. Hagelin, D.V. Nanopoulos, and M. Srednicki. Search for violations of quantum mechanics. *Nucl. Phys.*, B 241:381, 1984.
10. C. Lämmerzahl. Quantum Tests of Foundations of General Relativity. *Class. Quantum Grav.*, 14:13, 1998.
11. F. Burgbacher, C. Lämmerzahl, and A. Macias. Is there a stable hydrogen atom in higher dimensions? *J. Math. Phys.*, 40:625, 1999.
12. P.N. Bhat, G.J. Fishman, C.A. Meegan, R.B. Wilson, M.N. Brock, and W.S. Paciesas. Evidence for sub–millisecond structure in a γ–ray burst. *Nature*, 359:217, 1992.
13. I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products.* Academic Press, Orlando, 1983.
14. Th. Udem, A. Huber, B. Gross, J. Reichert, M. Prevedelli, M. Weitz, and T.W. Hänsch. Phase–coherent measurement of the Hydrogen 1s–2s transition frequency with an optical frequency interval divider chain. *Phys. Rev. Lett.*, 79:2646, 1997.
15. C. Kiefer and T.P. Singh. Quantum gravitational corrections to the functional Schrödinger equation. *Phys. Rev.* **D 44**, 1067 (1991).
16. K. Pachucki, D. Leibfried, M. Weitz, A. Huber, W. Konig and T. W. Hansch. Theory of the energy levels and precise two-photon spectroscopy of atomic hydrogen and deuterium. *J. Phys. B: At. Mol. Opt. Phys.* **29**, 177 (1996).
17. T. Udem, J. Reichert, R. Holzwarth, T. W. Hänsch. Accurate measurement of large optical frequency differences with a mode–locked laser. *Opt. Lett.* **24**, 881 (1999).

How Does the Electromagnetic Field Couple to Gravity, in Particular to Metric, Nonmetricity, Torsion, and Curvature?

Friedrich W. Hehl[1,2] and Yuri N. Obukhov[3]

[1] School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA
[2] Institute for Theoretical Physics, University of Cologne, 50923 Köln, Germany
[3] Department of Theoretical Physics, Moscow State University, 117234 Moscow, Russia

Abstract. The coupling of the electromagnetic field to gravity is an age-old problem. Presently, there is a resurgence of interest in it, mainly for two reasons: (i) Experimental investigations are under way with ever increasing precision, be it in the laboratory or by observing outer space. (ii) One desires to test out alternatives to Einstein's gravitational theory, in particular those of a gauge-theoretical nature, like Einstein-Cartan theory or metric-affine gravity.— A clean discussion requires a reflection on the foundations of electrodynamics. If one bases electrodynamics on the conservation laws of electric charge and magnetic flux, one finds Maxwell's equations expressed in terms of the excitation $H = (\mathcal{D}, \mathcal{H})$ and the field strength $F = (E, B)$ without any intervention of the metric or the linear connection of spacetime. In other words, there is still no coupling to gravity. Only the constitutive law $H = \text{functional}(F)$ mediates such a coupling. We discuss the different ways of how metric, nonmetricity, torsion, and curvature can come into play here. Along the way, we touch on non-local laws (Mashhoon), non-linear ones (Born-Infeld, Heisenberg-Euler, Plebański), linear ones, including the Abelian axion (Ni), and find a method for *deriving* the metric from linear electrodynamics (Toupin, Schönberg). Finally, we discuss possible non-minimal coupling schemes.

1 Introduction

General relativity was proposed in 1915. One of its predictions was the bending of light rays of stars in the gravitational field of the Sun. This effect was verified observationally soon afterwards by Dyson et al. in 1920 and put, as a result, Einstein's theory in the forefront of gravitational research.

Within the framework of general relativity, a light ray can be extracted from classical electrodynamics in its *geometrical optics* limit, i.e., for wavelengths much smaller than the local curvature radius of space. Accordingly, the bending of light can be understood as a result of a nontrivial refractive index of spacetime, see Skrotskii et al. [81,89], due to the coupling of the electromagnetic field F to the gravitational field g. Classically, we have in nature just these two fundamental fields F and g, the weak and the strong fields being confined to microphysical dimensions of 10^{-19}m or 10^{-15}m, respectively. Therefore, the coupling of F and g is of foremost importance in classical physics.

C. Lämmerzahl, C.W.F. Francis, and F.W. Hehl (Eds.): LNP 562, pp. 479–504, 2001.

The conventional way that coupling is achieved is to display the Maxwell-Lorentz equations of vacuum electrodynamics in the (flat) Minkowski world of special relativity theory in Cartesian coordinates. For this purpose, usually the formalism of tensor analysis (Ricci calculus) is used, see [78]:

$$F^{ij}{}_{,j} = I^i\,, \qquad F_{ij,k} + F_{jk,i} + F_{ki,j} = 0\,. \tag{1}$$

Here $F_{ij} = -F_{ji} = (F_{01}, F_{02}, F_{03}, F_{23}, F_{31}, F_{12}) = (\boldsymbol{E}, \boldsymbol{B})$ is the electromagnetic field strength, I^i the electric 4-vector current, and

$$F^{ij} := g^{ik} g^{jl}\, F_{kl}\,, \tag{2}$$

with g^{ij} as the contravariant components of the metric. The commas in (1) denote partial differentiation with respect to the local spacetime coordinates x^i.

If we switch on *gravity*, the flat Minkowski world becomes curved, the spacetime geometry now being a Riemannian geometry with a variable metric $g_{ij}(x)$ of Minkowskian signature $(+---)$. The coupling of the Maxwell-Lorentz set (1) to gravity is now brought about by the *comma goes to semicolon rule* , $\to$; (see [52]), where the semicolon represents the covariant derivative $\nabla_i \equiv {}_{;i}$ with respect to the Riemannian connection ("Levi-Civita connection"):

$$F^{ij}{}_{;j} = I^i\,, \qquad F_{ij;k} + F_{jk;i} + F_{ki;j} = 0\,. \tag{3}$$

This translation rule from special to general relativity is also alluded to as *minimal coupling* with the additional understanding that the components of the metric in (2) become spacetime dependent fields.

The metric field $g_{ij}(x)$, entering (3) via (2) and via the covariant derivatives, i.e., via the semicolons, has to fulfill the Einstein field equation,

$$Ric_{ij} - \frac{1}{2}\, g_{ij}\, Ric_k{}^k = \kappa \left(\overset{\text{Max}}{T}{}_{ij} + \overset{\text{mat}}{T}{}_{ij} \right)\,, \tag{4}$$

with

$$Ric_{ij} := R_{kij}{}^k\,, \qquad \overset{\text{Max}}{T}{}_i{}^j := \sqrt{\frac{\varepsilon_0}{\mu_0}} \left(-\, F_{ik} F^{jk} + \frac{1}{4}\, \delta_i^j\, F_{kl} F^{kl} \right)\,. \tag{5}$$

Here $R_{ijl}{}^k$ is the curvature and $\overset{\text{mat}}{T}{}_{ij}$ the material energy-momentum tensor. The coupled Einstein-Maxwell system describes correctly a wealth of experiments, in particular the gravitational bending of light, the gravitational redshift, the time delay of radar pulses in the gravitational field of the Sun, and the gravitational lensing and microlensing of starlight in the gravitational field of galaxies.

But in all these experiments, we study the propagation of light along null-geodesics in a prescribed (and perhaps slowly varying) gravitational field which is a solution of the Einstein vacuum equation – and not of the *electro*-vacuum equation. We could call this the *non* self-consistent Einstein-Maxwell theory. In the solar system, e.g., the Schwarzschild metric is taken as solution of the Einstein vacuum equation and the motion of a "photon" is described by the null geodesic

equation on this background. A true novel effect of the Einstein-Maxwell theory would be, e.g., the generation of electromagnetic waves by gravitational waves. Because of their smallness, no such effects were ever observed. Accordingly, the interaction of a classical electromagnetic field F_{ij} in the form of a lightray with a prescribed gravitational field $g_{ij}(x)$ is well described by means of Eqs.(3). Nevertheless, further consequences of these equations need to be compared with experiment as soon as more sensitive measuring methods are available.

2 On the Equivalence Principle

According to Einstein's equivalence principle, see [20], gravity can be *locally simulated* in a gravity-free region of spacetime by going over from the Cartesian coordinates, anchored in an inertial frame of reference (including an intertial clock) and used in (1), to arbitrary curvilinear coordinates yielding a non-inertial frame in general, as in (3). In this context, the metric g_{ij}, occurring in (2) and in the semicolons of (3), is understood as a flat metric in curvilinear coordinates. Thus, the minimal coupling can be interpreted, in a first step, just as a coordinate transformation from Cartesian to curvilinear coordinates. And, moreover, it *identifies* the metric as the gravitational potential.

On the other hand, let us assume that we are in a region *with* gravity and (2) and (3) are valid together with the Einstein equation for the metric. Then, also according to Einstein's equivalence principle, we must be able to pick suitable coordinates such that locally the equations look like in special relativity in Cartesian coordinates. In Riemannian geometry, the local coordinates are called Riemannian normal (hence geodesic) coordinates at one point P, if the Christoffel symbols

$$\Gamma_{ij}{}^{k} := \frac{1}{2}\, g^{kl}\, (g_{il,j} + g_{jl,i} - g_{ij,l}) \tag{6}$$

vanish at P and and the metric becomes Minkowskian:

$$\Gamma_{ij}{}^{k}|_{\mathrm{P}} \stackrel{*}{=} 0\,, \qquad g_{ij}|_{\mathrm{P}} \stackrel{*}{=} \mathrm{diag}\,(+1,-1,-1,-1)\,. \tag{7}$$

Accordingly, the semicolon becomes a comma and the metric in (2), at one given point, looks flat.

Still, the curvature is non-vanishing, of course: $R_{ijk}{}^{l}|_{\mathrm{P}} \neq 0$. The equations look flat since they contain only *first* derivatives. If they contained second derivatives, then the semicolons goes to comma rule and its reverse would *not* work since on that level not only the Christoffels enter but potentially also the curvature which, in contrast to the Christoffels, is a tensor and cannot be nullified by means of a suitable choice of coordinates. For that reason, the minimal coupling procedure, being in this context an expression of the equivalence principle, must be applied only to first order differential equations. The safest thing is then to apply it, as a rule, only on the level of a *Lagrangian*, since there ordinarily only first-order expressions are allowed for. Non-minimal couplings of the gravitational field to electromagnetism have also been investigated, see Prasanna [74],

Buchdahl [8], Goenner [23], and Müller-Hoissen [58–60], for example, or for light rays in non-minimally coupled theories, see Drummond and Hathrell [17], but the price one has to pay is to introduce a new constant of nature; and there is no evidence for such a constant in nature – unless one takes the Planck length itself. We will come back to these questions in Sec.7.

Therefore we can conclude that the equivalence principle and minimal coupling work well for the Maxwell-Lorentz equations (1) and that they lead to experimentally established equations.

Wave Equation for the Electromagnetic Field Strength. We hasten to add that, within the framework of the minimally coupled Maxwell-Lorentz equations, we find 2nd derivatives if we derive the wave equation for the electromagnetic field strength F — and then also curvature terms are expected to emerge. This is exactly what happens, as already found by Gordon [24] and Eddington [18].

In the framework of exterior calculus (see Frankel [21]), let us consider the electromagnetic field strength 2-form $F = \frac{1}{2} F_{ij}\, dx^i \wedge dx^j$. In Maxwell-Lorentz vacuum electrodynamics, it satisfies $dF = 0$, $\varepsilon_0 d\, {}^\star F = \frac{1}{c} J$. We denote the codifferential by $\delta := {}^\star d\, {}^\star$. Then we find, with the wave operator (d'Alembertian)

$$\Box := \delta\, d + d\, \delta\,, \tag{8}$$

and by using the Maxwell-Lorentz equations, the wave equation

$$\Box F = \frac{1}{\varepsilon_0 c}\, d\, {}^\star J\,, \tag{9}$$

see [53]. The left hand side of this equation, in terms of components, can be determined by substituting (8):

$$\Box F = \frac{1}{2} \left(\nabla^k \nabla_k\, F_{ij} + 2\, Ric_{[i}{}^k\, F_{j]k} - R^{kl}{}_{ij}\, F_{kl} \right) dx^i \wedge dx^j\,. \tag{10}$$

Accordingly, minimal coupling can lead to curvature terms of a prescribed form.[1]

3 A Caveat

Soon after general relativity had been proposed, it became clear, see Einstein [19], that one can introduce as *auxiliary variables* the densities

$$\mathcal{F}^{ij} := \sqrt{-g(x)}\, g^{ik}(x)\, g^{jl}(x)\, F_{kl}\,, \qquad \mathcal{I}^i := \sqrt{-g(x)}\, I^i\,, \tag{11}$$

with $g(x) := \det g_{ij}(x)$, in terms of which the Maxwell-Lorentz equations (1) can be rewritten in a metric-free way as

$$\mathcal{F}^{ij}{}_{,j} = \mathcal{I}^i\,, \qquad F_{ij,k} + F_{jk,i} + F_{ki,j} = 0\,. \tag{12}$$

[1] Of course, we could have non-minimal coupling as, e.g., in $\Box F + \gamma \left(e_\alpha \rfloor e_\beta \rfloor R^{\alpha\beta} \right) \wedge F = \frac{1}{\varepsilon_0 c}\, d\, {}^\star J$, see Sec.7.

Similarly, the charge conservation law $I^i{}_{;i} = 0$ can be put in the form

$$\mathcal{I}^i{}_{,i} = 0\,. \tag{13}$$

The metric enters only via the densities defined in (11). In fact, if we started from the set (12) in special relativity right away, then no comma goes to semicolon rule would have been necessary: These equations are generally covariant, they are valid in arbitrary curvilinear coordinates, be it in the framework of special or general relativity theory.

In the calculus of *exterior differential forms* (Cartan calculus), see Frankel [21], these equations can be formulated very succinctly. We introduce the electric current as odd 3-form,

$$J := \rho - j \wedge dt = \frac{1}{3!}\, J_{ijk}\, dx^i \wedge dx^j \wedge dx^k\,, \tag{14}$$

the electromagnetic excitation as odd 2-form

$$H = \mathcal{D} - \mathcal{H} \wedge dt = \frac{1}{2!}\, H_{ij}\, dx^i \wedge dx^j\,, \tag{15}$$

and the electromagnetic field strength as even 2-form

$$F = B + E \wedge dt = \frac{1}{2!}\, F_{ij}\, dx^i \wedge dx^j\,. \tag{16}$$

Then (12) reads

$$dH = J\,, \qquad\qquad dF = 0\,, \tag{17}$$

with

$$dJ = 0\,. \tag{18}$$

The set (17) represents the Maxwell equations. They are independent of metric and connection. The constitutive relation for the vacuum reads

$$H = \sqrt{\frac{\varepsilon_0}{\mu_0}}\, {}^\star F\,, \tag{19}$$

where the star $\star$ represents the metric-dependent and odd Hodge duality operator. Eq.(19) corresponds to $(11)_1$ and $\mathcal{I}^i$ can be related to the components of J,

$$\mathcal{F}^{ij} = \frac{1}{2!}\sqrt{\frac{\mu_0}{\varepsilon_0}}\, \epsilon^{ijkl} H_{kl}\,, \qquad\qquad \mathcal{I}^i = \frac{1}{3!}\, \epsilon^{ijkl} J_{jkl}\,, \tag{20}$$

with $\epsilon^{ijkl} = \pm 1, 0$, the totally antisymmetric Levi-Civita tensor density.

Now the equivalence principle looks empty: Since the Maxwell equations (17) are formulated in a coordinate and frame independent way, they are valid in this form in arbitrary coordinate systems and frames, be it in a flat or in a curved spacetime. Only the constitutive relation (19) "feels", up to a conformal factor, the presence of a flat or a non-flat metric, i.e., the constitutive relation couples to

the conformally invariant part of the metric. The coupling of electromagnetism to gravity becomes almost trivial. Is all this just a mathematical trick, which distracts from the physical content of Maxwell's theory, or is it more?

One further observation hints also at the need for clarification. The *Einstein-Cartan theory* of gravity is a *viable* gravitational theory, see [30,51,86]. It is the simplest model of the metric-affine *gauge* theory of gravity, see [31,25]. In the Einstein-Cartan theory, spacetime is described by means of a Riemann-Cartan geometry with torsion and curvature.[2] If we couple (3) to gravity, do we have to use the semicolons as covariant derivatives with respect to the Riemann-Cartan connection or still with respect to the Christoffels, see [14]? In the context of (17), this question cannot even be posed, since the exterior derivative d is all what is needed. Are then the equations (17) misleading as a starting point for coupling to gauge gravity? What could be the appropriate starting point?

Provided one formulates Maxwell's theory and its coupling to gravity in terms of a Lagrangian with the electromagnetic potential A as variable, gauge covariance of the formalism results in (17) cum (19), as was pointed out by Benn, Dereli, and Tucker [5]. However, we would like to have some more immediate insight into the structure of electromagnetism as induced by experiment even without having a variational formulation at our disposal.

4 Electric Charge and Magnetic Flux Conservation

The metric is a quantity which allows to define *lengths* and *angles* in spacetime. There are, however, laws in physics which don't require the knowledge of a metric. Take the conservation law of electric charge as an example. Mark a 3-dimensional simply connected submanifold Ω_3. We know from experiment that a possible electric charge inside Ω_3 is composed of charge "quanta", i.e., there is an integer number of elementary charges in Ω_3. Recent advances in technology made it possible, see [13,44], to trap and to count single electrons and protons. Thus, as soon as we have such quanta available, we can rely on *counting procedures*, see Post [73], the use of a meter stick or a chronometer is superfluous under such circumstances.

Electric charge conservations is experimentally well-established and is one of the pillars electromagnetism rests on. We formulate it, following Kottler-Cartan-van Dantzig, see [72,87] and also [32], most appropriately as an integral law. According to (14), we assume the existence of the odd electric current 3-form J. We take charge conservation as axiom 1, that is, J integrated over a closed 3-dimensional hypersurface Ω_3 has to vanish, if this hypersurface is the boundary of a connected 4-volume Ω_4:

$$\int\limits_{\partial\Omega_4} J = \int\limits_{\Omega_4} dJ = 0\,, \qquad \text{or} \qquad dJ = 0\,. \tag{21}$$

[2] A proper discussion of the equivalence principle in the context of Einstein-Cartan theory requires the introduction of local coframes, see [25,27,38]. Being concerned here only with electromagnetism, it is sufficient to use natural, i.e., holonomic coframes.

Here we applied the Stokes theorem.

If (21) is assumed to be valid $\int_{C_3} J = 0$ for all three-cycles $\partial C_3 = 0$, $C_3 \neq \partial\Omega_4$, then, according to a theorem of de Rham, J is exact, see [73]. Thus the inhomogeneous Maxwell equation is a consequence,

$$J = dH\,, \tag{22}$$

with the odd 2-form H of the electromagnetic excitation, see (15). The excitation is only determined up to an exact form. Nevertheless, the electric excitation $\mathcal{D}$ can be *measured* by means of Maxwellian double plates as charge per unit area, the magnetic excitation $\mathcal{H}$ by means of a small test coil, which compensates the $\mathcal{H}$-field to be measured, as current per unit length. This is possible since in theses null experiments vanishing field strength F implies vanishing excitation H, see [32]. In other words, the extensive quantities $\mathcal{D}$ and $\mathcal{H}$ – and thus the 4-dimensional excitation H – have an operationally significance of their own, since they are related to charge at rest or in motion, respectively. Accordingly, the somewhat formalistic introduction of the densities in (11) has now been legitimized as a transition to operationally meaningful additive quantities. Note that up to now only the differential structure of the spacetime was needed, a metric has not been involved.

Let us choose a field of 4-frames e_α and consider the motion of a point particle with respect to the reference frame thus defined. As axiom 2 one can take an operational definition of the electromagnetic field strength F via the Lorentz force density

$$f_\alpha = (e_\alpha \rfloor F) \wedge J\,. \tag{23}$$

The interior product (contraction) is denoted by $\rfloor$. The force density f_α is a notion from classical mechanics. It is an odd covector-valued 4-form. Accordingly, Eq.(23) can be read as a definition of the even 2-form F, see (16). Again, we don't need a metric. And we know the recipe of how to proceed in the same manner.

That $\int_{\Omega_2} F$ can be interpreted as magnetic flux is obvious if we choose Ω_2 as a 'spacelike' surface (strictly, at this point we don't know what spacelike means; we will come back to this later). In superconductors under suitable circumstances we can count (in an Abrikosov flux line lattice) quantized magnetic flux lines. This suggests that magnetic flux is a conserved quantity (axiom 3):

$$\int\limits_{\partial\Omega_3} F = \int\limits_{\Omega_3} dF = 0 \qquad \text{or} \qquad dF = 0\,. \tag{24}$$

In this way, by means of the axioms (21), (23), and (24), we recovered the fundamental structure of Maxwell's theory: $dH = J$, $dF = 0$. This is what had been called *metric-free electrodynamics*.[3] What is missing so far is the relation between the excitation H and the field strength F, and it is exactly there where the metric, i.e., the gravitational potential comes in.

[3] Stachel [83] calls it *generalized* electrodynamics. We don't follow this suggestion, since Maxwell's equations were originally given in terms of $(\mathcal{D}, \mathcal{H})$ and $(E, \mu\mathcal{H})$ in

5 No Interaction of Charge and Flux "Substrata" with Gravity

We now understand that the inhomogeneous Maxwell equation $dH = J$, as an expression of electric charge conservation, cannot be influenced by gravity, i.e., by the metric tensor g, or, in the case of metric-affine gravity or its specific subcases, such as Einstein-Cartan theory, by the connection Γ of spacetime. The electric charge substratum of spacetime has rules of its own. Spacetime can be deformed by the presence of metric and connection, but the charge substratum and the net electric charge to be attributed to a prescribed 3-dimensional (3D) volume won't change. Thus the additivity 3D volume-wise of the charge lays at the foundation of the Maxwellian framework. And it translates into the 2D additivity of the integrated excitation $\int_{\Omega_2} H$ – this being the reason why one uses this integral for the operational interpretation of H.

Similar arguments can be advanced for the homogeneous Maxwell equation $dF = 0$. However, first of all it should be stressed that the axiomatics we are using strongly suggests the non-existence of magnetic charges. If there were magnetic charges, then we would have no reason to believe in electric charge conservation either; compare for this argument axiom 1, Eq.(21), with axiom 3, Eq.(24). Conventionally, the inhomogeneous equation $dH = J$ is seen in analogy to the homogeneous one $dF = 0$. But not so in the framework of our axiomatics which has a firm empirical basis. We put $dJ = 0$ in analogy to $dF = 0$. The whole historical development of electromagnetism, starting with Ørsted and Ampère, points to the elimination of the phenomenologically introduced magnetic charges. Most recent experiments, see [1,29], exclude magnetic charges with very good precision. Thus theoretical as well as experimental evidence speak against the existence of magnetic charges.

Having said this, we hasten to add that, nevertheless, there is some kind of magnetic substratum in spacetime, namely the magnetic flux $\int_{\Omega_2} F$. It is a substratum of its own right. The fluxoids, the quantized magnetic flux lines in superconductors, see [84], do convey a clear message. Besides electric charge[4], magnetic flux[5] (and *not* magnetic charge) has an independent standing in electromagnetism, too. Thus rightfully, it is governed by an own axiom, namely axiom 3.

Axiom 3 is again a conservation theorem. In contrast to axiom 1, which has a fermionic smell, axiom 3 is more of a bosonic nature. Moreover, magnetic flux adds up 2D area-wise. For this reason, magnetic flux is represented by a 2-form and not, like the charge, by a 3-form. Accordingly, there are essential differences between these two conservation laws which express the peculiarities of the

a form 'isomorphic' to the $(1 + 3)$-decomposition of $dH = J$ and $dF = 0$, see [50]. Therefore, the "generalized" Maxwell equations, $dH = J$ and $dF = 0$, correspond in actual fact, just to *Maxwell's equations* (modulo the substitution $B \to \mu\mathcal{H}$). And this is how we will name them.

[4] SI-unit Coulomb, elementary charge $e = 1.60217733 \times 10^{-19}$ C.

[5] SI-unit Weber, elementary fluxoid $h/(2e) = 2.06783461 \times 10^{-15}$ Wb.

electromagnetic phenomena. Electric and magnetic effects enter the Maxwellian framework in an asymmetric way, in spite of all that talk about a duality between electricity and magnetism. But there is also a similarity in that both axioms are formulated as integral conservation laws. The possibility to count the fluxoids assures us that axiom 3 has to be again a law free of metric and connection.

Incidentally, there is a nice visualization of the fundamental quantities entering electrodynamics. If one describes the quantum Hall effect for low lying Landau levels, then the concept of a *composite fermion* is very helpful: it consists of one electron and an even number of fluxoids is attached to it, see Jain [41,42] and [43]. Isn't that a very clear indication of what the fundamental quantities are in electrodynamics? Namely, electric charge (see axiom 1) and magnetic flux (see axiom 3), see also Nambu [61] in this general context.

Our conclusion is then that, as long as we opt for electric charge and magnetic flux conservation, the Maxwell equations in gravity-free regions, i.e., in the Minkowski spacetime of special relativity, read $dH = J$ and $dF = 0$; they *remain the same* irrespective of the switching on of gravity, be it in Einstein's theory, in metric-affine gravity (see [76]), or in any other geometrical theory of gravity.

6 Constitutive Law of Electrodynamics and Its Relation to Gravity

After having discussed extensively that gravity does not influence the Maxwell equations, we eventually turn to the constitutive law via which gravity does influence electrodynamics. It is true, the charge substratum and the flux substratum themselves do not couple to gravity, as we have shown in the last section. However, the interrelationship between both substrata is affected by gravity. Metaphorically speaking, the "flow" of *each* of the substrata is ruled by a particular gravity-independent conservation theorem, but the flows of electric charge and magnetic flux are coupled via a gravity-dependent constitutive law since, in the end, magnetism has to be expressed in terms of electricity.

Let us choose arbitrary local spacetime coordinates x^i. Then we have,

$$H = \frac{1}{2}\, H_{ij}\, dx^i \wedge dx^j, \qquad F = \frac{1}{2}\, F_{ij}\, dx^i \wedge dx^j. \tag{25}$$

We will turn first to the electrodynamics of material media in order to develop some intuition on the concept involved, but eventually, it will be the *vacuum*, be it in inertial or non-inertial frames, which will occupy the center of our interest.

6.1 Non-local

Moving *macroscopic* matter defines a $(1+3)$-splitting of spacetime specified by a well defined average 4-vector velocity field u which describes the congruence of worldlines of the flow of the medium. Such a vector field can be defined operationally from the motion of matter as follows.

Let a 3–dimensional arithmetic space R^3 be equipped with the coordinates $\xi^a, a = 1, 2, 3$. We will use these coordinates (known as *Lagrange* coordinates in continuum mechanics) as labels which enumerate elements of a material medium. A smooth mapping $x_{(0)} : R^3 \to X_4$ into the spacetime defines a 3–dimensional space domain (hypersurface) V which represents the initial distribution of matter. In local spacetime coordinates, this mapping (or labeling) is given by the four functions $x^i_{(0)}(\xi^a)$. It should be preserved at any time, i.e. along any worldline of a particular element its labels ξ^a are constant.

Given the initial configuration V of matter, we parameterize dynamics of the medium by the "time" coordinate τ which is defined as the proper time measured along an element's worldline from the original hypersurface V. The resulting local coordinates (τ, ξ^a) are usually called the normalized comoving coordinates. Thus finally, the motion of matter is described by the functions $x^i(\tau, \xi^a)$. Subsequently, we define the 4–velocity vector field by

$$u := \partial_\tau = \left(\frac{dx^i}{d\tau}\right)_{\xi^a = \text{const}} \partial_i \,. \tag{26}$$

Evidently, a family of observers comoving with the matter is characterized by the same timelike congruence $x^i(\tau, \xi^a)$. They are making physical (in particular, electrodynamical) measurements in their local reference frames which drift with the material motion.

One says that a medium, moving in general, has *dispersion* properties when the electromagnetic fields produce non-instantaneous polarization and magnetization effects. The most general *linear* constitutive law is then given, in the comoving system, by means of the integral

$$H_{ij}(\tau, \xi) = \frac{1}{2} \int d\tau' K_{ij}{}^{kl}(\tau, \tau')\, F_{kl}(\tau', \xi) \,. \tag{27}$$

The coefficients of the kernel $K_{ij}{}^{kl}(\tau, \tau')$ are called the response functions. We expect the metric to be involved in their set-up. Their form is defined by the internal properties of matter and by the motion of a medium.

Mashhoon [49] has proposed a physically very interesting example of such a non-local electrodynamics in which non-locality comes as a direct consequence of the *non-inertial* dynamics of observers. In this case, instead of (25), one should use the field expansions

$$H = \frac{1}{2} H_{\alpha\beta}\, \vartheta^\alpha \wedge \vartheta^\beta, \qquad F = \frac{1}{2} F_{\alpha\beta}\, \vartheta^\alpha \wedge \vartheta^\beta \tag{28}$$

with respect to the coframe of a non-inertial observer $\vartheta^\alpha = e_i{}^\alpha\, dx^i$. The constitutive law is then replaced by

$$H_{\alpha\beta}(\tau, \xi) = \frac{1}{2} \int d\tau' K_{\alpha\beta}{}^{\gamma\delta}(\tau, \tau')\, F_{\gamma\delta}(\tau', \xi) \,, \tag{29}$$

and the response kernel in (29) is now defined by the acceleration and rotation of the observer's reference system. It is a constitutive law for the vacuum as viewed from a non-inertial frame of reference.

Mashhoon imposes an addititional *assumption* that the kernel is of *convolution* type, i.e., $K_{\alpha\beta}{}^{\gamma\delta}(\tau,\tau') = K_{\alpha\beta}{}^{\gamma\delta}(\tau-\tau')$. Then the kernel can be uniquely determined by means of the Volterra technique, and often it is possible to use the Laplace transformation in order to simplify the computations. Unfortunately, although Mashhoon's kernel is always calculable in principle, in actual practice one normally cannot obtain K explicitly in terms of the observer's acceleration and rotation.

Preserving the main ideas of Mashhoon's approach, one can abandon the convolution condition. Then the general form of the kernel can be worked out explicitly (u is the observer's 4-velocity):

$$K_{\alpha\beta}{}^{\gamma\delta}(\tau,\tau') = \frac{1}{2}\,\epsilon_{\alpha\beta}{}^{\lambda[\delta}\left(\delta_\lambda^{\gamma]}\,\delta(\tau-\tau') - u\rfloor\Gamma_\lambda{}^{\gamma]}(\tau')\right). \tag{30}$$

The influence of non-inertiality is manifest in the presence of the connection 1-form. The kernel (30) coincides with the original Mashhoon kernel in the case of constant acceleration and rotation, but in general the two kernels are different [55]. Perhaps, only the direct observations would establish the true form of the non-local constitutive law. However, such a non-local effect has not been confirmed experimentally as yet.

6.2 Non-linear

But the constitutive law can also be non-linear (or non-local and non-linear at the same time). In the local and non-linear *Born-Infeld* electrodynamics [6], with the dimensionfull parameter $f_{\rm e}$ as maximal attainable electric field strength, we have

$$H = -\,\frac{\partial V_{\rm BI}}{\partial F} \sim \frac{\partial\sqrt{-\det\left|g_{kl}+\frac{1}{f_{\rm e}}F_{kl}\right|}}{\partial F}\,. \tag{31}$$

The metric as symmetric second rank tensor enters here in a very natural way. It adds up with the antisymmetric electromagnetic field to an asymmetric tensor – much in the way Einstein had hoped to find for his unified field theories of gravity and electromagnetism. By differentiation, we find

$$H = \sqrt{\frac{\varepsilon_0}{\mu_0}}\,\frac{{}^\star F + \frac{1}{2f_{\rm e}^2}\,{}^\star(F\wedge F)\,F}{\sqrt{1-\frac{1}{f_{\rm e}^2}\,{}^\star(F\wedge{}^\star F)-\frac{1}{4f_{\rm e}^4}\,[{}^\star(F\wedge F)]^2}}\,, \tag{32}$$

now the metric being absorbed in the (odd) *Hodge star* operator, see [21]. For $f_{\rm e}\to\infty$, we recover the conventional local and linear Maxwell-Lorentz theory for vacuum with $H = \sqrt{\frac{\varepsilon_0}{\mu_0}}\,{}^\star F$. The Born-Infeld electrodynamics is presently used as a toy model in string theories, see [22]. The problem with Born-Infeld electrodynamics is that, in contrast to Maxwell's theory, it defies quantization. It is an interesting model, but nothing like an established theory.

A similar example is the non-linear *Heisenberg-Euler* electrodynamics [34]. Quantum electrodynamical vacuum fluctuations yield corrections to Maxwell's

theory that can be accounted for by an effective constitutive law constructed by Heisenberg and Euler. To the first order in the fine structure constant $\alpha_{\rm f} = \frac{e^2}{4\pi\varepsilon_0\hbar c}$, it is given by (see also [39,35])

$$H = \sqrt{\frac{\varepsilon_0}{\mu_0}} \left\{ \left[1 + \frac{8\,\alpha_{\rm f}}{45\,B_k^2} {}^\star\big(F \wedge {}^\star F\big) \right] {}^\star F + \frac{14\,\alpha_{\rm f}}{45\,B_k^2} {}^\star\big(F \wedge F\big)\, F \right\}, \qquad (33)$$

where $B_k = \frac{m^2c^2}{e\hbar} \approx 4.4 \times 10^9\,\mathrm{T}$, with the mass of the electron m. The metric is again hidden in the Hodge star and the Maxwell-Lorentz limit results analogously for $m \to \infty$. The Casimir force between two uncharged electrically conducting plates, also an effect of vacuum fluctuations, has been experimentally verified as have been non-linear effects in the "superposition" of strong laser beams. Accordingly, the non-linear constitutive law (33) is a valid post-classical approximation of vacuum electrodynamics and as such experimentally confirmed.

Note that these variants of classical electrodynamics respect charge and flux conservation. This underlines the fact that our axiomatics clearly points to that structure of electrodynamics, namely the constitutive law, which can be changed without giving up the essentials of electrodynamics.

Both, Eqs.(32) and (33) are special cases of Plebański's more general non-linear electrodynamics [71]. Let the quadratic invariants of the electromagnetic field strength be denoted by

$$I_1 := \frac{1}{2}{}^\star(F \wedge {}^\star F) = \frac{1}{2}(\boldsymbol{E}^2 - \boldsymbol{B}^2) \quad \text{and} \quad I_2 := \frac{1}{2}{}^\star(F \wedge F) = \boldsymbol{E} \cdot \boldsymbol{B}\,, \qquad (34)$$

where I_1 is an even and I_2 is an odd scalar (the Hodge operator is odd). Then Plebański postulated a non-linear electrodynamics with the constitutive law[6]

$$H = U(I_1, I_2)\, {}^\star F + V(I_1, I_2)\, F\,, \qquad (35)$$

where U and V are functions of the two invariants. Note that in the Born-Infeld case U and V depend on both invariants whereas in the Heisenberg-Euler case we have $U_{\rm HE}(I_1)$ and $V_{\rm HE}(I_2)$. Nevertheless, in both cases U is required as well as V. And in both cases, see (32) and (33), U is an even function and V and odd one such as to preserve parity invariance.

If one chose V to be an even function, e.g., then parity violating terms would emerge. Such terms were most recently discussed by Majumdar, Mukhopadhyaya, and SenGupta [48,56]; for the experimental situation (there seem no signatures for parity violations) compare Lue et al. [47].

Singularity-Free Electro-Gravitodynamics. Recently, Ayón-Beato & García [2], for earlier work see Shikin [79], have proposed a constitutive law

$$H = U(I_1)\, {}^\star F\,, \qquad (36)$$

[6] Strictly, Plebański assumed a Lagrangian which yields (17) together with the *structural relations* $F = u(I_1, I_2)\, {}^\star H + v(I_1, I_2)\, H$. The latter law, apart from singular cases, is equivalent to (35).

which, as subcases, does neither encompass (32) nor (33) and thus makes it appear as rather academic. The explicit form of U is *defined* by the requirement of obtaining completely singularity-free solutions of the coupled system of the gravitational field (Einstein) and the electromagnetic field (non-linear Maxwell). Examples of suitable functions $U(I_1)$ are given in [79,2].

In terms of the local time and space coordinates (t, r, θ, ϕ), the general spherically symmetric ansatz for the coframe can be written as

$$\vartheta^{\hat{0}} = f(r)\, d\,t\,, \quad \vartheta^{\hat{1}} = \frac{1}{f(r)}\, d\,r\,, \quad \vartheta^{\hat{2}} = r\, d\,\theta\,, \quad \vartheta^{\hat{3}} = r\, \sin\theta\, d\,\phi\,, \tag{37}$$

whereas, for the electromagnetic field, we have

$$F = \varphi(r)\, \vartheta^{\hat{0}} \wedge \vartheta^{\hat{1}}\,. \tag{38}$$

The exact solution of the coupled system of gravitational and electromagnetic field equations, i.e., of Einstein's equation (4) and Maxwell's equations $dF = 0,\ dH = 0$, reads

$$\varphi = \frac{q}{U(I_1)\, r^2}, \qquad f^2 = 1 - \frac{2m}{r} + \frac{Q(r)}{r^2}, \tag{39}$$

where q, m are integration constants and ('Tolman's integral')

$$Q(r) = \kappa\, r \int_r^\infty dr'\, \mathcal{K}(r')\, r'^2, \qquad \mathcal{K} = 2I_1\, U(I_1) - \int^{I_1} dI_1'\, U(I_1'). \tag{40}$$

In the last function one should substitute the explicit form of the quadratic invariant I_1 computed on the spherically symmetric configuration (37) with (38).

It is shown in [79,2–4] that the constitutive function $U(I_1)$ can be chosen in such a way that the functions in (39) describe a completely regular, i.e., singularity-free configuration.

6.3 Linear: Abelian Axion, inter Alia

A very important case is that of a linear constitutive law between the components of the two-forms H and F. It postulates the existence of the $6 \times 6 = 36$ constitutive functions $\kappa_{ij}{}^{kl}(t, x) = -\kappa_{ji}{}^{kl} = -\kappa_{ij}{}^{lk}$ such that

$$H_{ij} = \frac{1}{2}\, \kappa_{ij}{}^{kl}\, F_{kl}. \tag{41}$$

This kind of an ansatz we know from the physics of anisotropic crystals. The factor $1/2$ is chosen in order to have a smooth transition to the conventional $\boldsymbol{D} = \varepsilon_0\, \varepsilon\, \boldsymbol{E}$ etc. relations, cf. [72] p.127.

The choice of the local coordinates is clearly unimportant. In a different coordinate system the linear constitutive law preserves its form due to the tensorial

transformation properties of $\kappa_{ij}{}^{kl}$. Alternatively, instead of the local coordinates, one may choose an anholonomic frame and may then decompose the two-forms H and F with respect to it.

Since H is an odd and F an even form, the constitutive functions $\kappa_{ij}{}^{kl}(t,x)$ are *odd*. Taking the Levi-Civita symbol, we can split off the odd piece according to

$$\kappa_{ij}{}^{kl} =: \frac{1}{2}\,\epsilon_{ijmn}\,\chi^{mnkl} \qquad \text{or} \qquad \chi^{ijkl} = \frac{1}{2}\,\epsilon^{ijmn}\kappa_{mn}{}^{kl}\,. \tag{42}$$

Because of the corresponding properties of the Levi-Civita symbol, the χ^{ijkl} are even scalar densities of weight $+1$. For the Levi-Civita symbols with upper and lower indices, we have $\epsilon^{ijkl}\,\epsilon_{mnpq} = \delta^{\,i\,j\,k\,l}_{mnpq}$.

With the linear constitutive law (as with more general laws), we can set up a Lagrangian 4-form; here we call it $V_{\rm lin}$. Because of $H = -\,\partial V_{\rm lin}/\partial F$, the Lagrangian must be quadratic in F. Thus we find

$$\begin{aligned} V_{\rm lin} &= -\,\frac{1}{2}\,H\wedge F = -\,\frac{1}{8}\,H_{ij}F_{pq}\,dx^i\wedge dx^j\wedge dx^p\wedge dx^q \\ &= -\,\frac{1}{32}\,\left(\epsilon^{pqij}\epsilon_{ijmn}\chi^{mnkl}\right)F_{kl}F_{pq}\,dx^0\wedge dx^1\wedge dx^2\wedge dx^3\,. \end{aligned} \tag{43}$$

The components of the field strength F enter in a *symmetric* way. Therefore, without loss of generality, we can impose the symmetry condition $\chi^{ijkl} = \chi^{klij}$ on the constitutive functions reducing them to 21 independent functions at this stage.

The $\kappa_{ij}{}^{kl}$ carry the dimension $[\kappa] = [\chi] = e^2/\hbar$. Therefore, still before introducing the metric, we can split off the totally antisymmetric part of χ^{ijkl} and define the dimensionless constitutive functions according to

$$\chi^{ijkl} = f\ \overset{\circ}{\chi}{}^{ijkl} + \alpha\,\epsilon^{ijkl}\,, \qquad \text{with} \qquad \overset{\circ}{\chi}{}^{[ijkl]} = 0\,. \tag{44}$$

Here $[f] = [\alpha] = \hbar/e^2$, and $f = f(t,x)$ and $\alpha = \alpha(t,x)$ represent one scalar and one pseudo-scalar constitutive function, respectively. Thus the *linearity ansatz* eventually reads

$$H_{ij} = \frac{1}{4}\,\epsilon_{ijmn}\,\chi^{mnkl}\,F_{kl} = \frac{f}{4}\epsilon_{ijmn}\ \overset{\circ}{\chi}{}^{mnkl}\,F_{kl} + \alpha\,F_{ij}\,, \tag{45}$$

with

$$\overset{\circ}{\chi}{}^{mnkl} = -\,\overset{\circ}{\chi}{}^{nmkl} = -\,\overset{\circ}{\chi}{}^{mnlk} = \overset{\circ}{\chi}{}^{klmn} \qquad \text{and} \qquad \overset{\circ}{\chi}{}^{[mnkl]} = 0\,, \tag{46}$$

i.e., besides α, we have 20 independent constitutive functions. Thus $\overset{\circ}{\chi}{}^{nmkl}$ has the same algebraic symmetries and the same number of independent components as a curvature tensor in a Riemannian spacetime.

Pseudo-scalars are also called axial scalars. So far, our axial scalar $\alpha(x)$ is some kind of permittivity/permeability field. If one adds a kinetic term of the α-field to the electromagnetic Lagrangian (43), then $\alpha(x)$ becomes propagating and one can call it legitimately an Abelian[7] axion. Ni [62] was the first to introduce

[7] In contrast to the axions related to *non*-Abelian gauge theories, see [91,92,54] and the reviews in [45] and [80].

such an axion field α in the context of the coupling of electromagnetism to gravity, see also deSabbata & Sivaram [16] and the references given there.

The *Abelian* axion has the following properties:

- Pseudoscalar field, i.e., spin $= 0$, parity $= -1$.
- Couples to Maxwell's field in the Lagrangian according to $\alpha\, F \wedge F = 2\alpha\, E \wedge B \wedge dt$, see (45) and (43). Here E is the 3-dimensional electric field 1-form and B the corresponding magnetic field 2-form. This term in the total Lagrangian can be written as

$$\alpha\, F \wedge F = -\, d\alpha \wedge A \wedge F\,, \tag{47}$$

  dropping, as usual, the total derivative. This contributes to the excitation $H = -\partial L/\partial F$ a term

$$\sim d\alpha \wedge A\,. \tag{48}$$

- Since it arises on the same level as the metric, see Eq.(64) below, it is a field of a similar universality as the gravitational field.

As yet, the Abelian axion has not been found experimentally, see the discussion of Cooper & Stedman [12] on corresponding ring laser experiments.

6.4 Isotropic

The linearity ansatz (45) can be further constrained in order to arrive eventually at an isotropic constitutive tensor. We will proceed here somewhat unconventional in that we don't assume a metric of spacetime beforehand but rather derive it in the following way:

Duality Operator, Electric and Magnetic Reciprocity. The constitutive tensor $\overset{\circ}{\chi}{}^{klmn}$ of (45) defines a new duality operator which acts on 2-forms on X. In components, an arbitrary 2-form $\Theta = \frac{1}{2}\Theta_{ij}\, dx^i \wedge dx^j$ is mapped into the 2-form ${}^{\#}\Theta$ by

$${}^{\#}\Theta_{ij} := \frac{1}{4}\epsilon_{ijkl}\, \overset{\circ}{\chi}{}^{klmn}\, \Theta_{mn}\,, \tag{49}$$

see [67,33]. No metric is involved in this process. Now the linear material law (45) can be written as

$$H = \left(f\,{}^{\#} + \alpha\right) F. \tag{50}$$

We postulate that the duality operator, applied twice, should, up to a sign, lead back to the identity. Such a closure relation or the "electric and magnetic reciprocity" [85] reduces the number of independent components of $\overset{\circ}{\chi}$ to 9 (without using a metric). One can demonstrate that this is a sufficient condition for the *non*existence of *birefringence* in vacuum, see [62,63,46,64,28]. Then the

fourth-order general Fresnel equation degenerates to the second-order light cone equation. Therefore, we impose

$$^{\#\#} = -1\,. \tag{51}$$

The minus sign yields Minkowskian signature[8], whereas the condition $^{\#\#} = +1$ would lead to Euclidean or to the mixed signature $(+,+,-,-)$.

Seemingly Toupin [85] and Schönberg [77] were the first to deduce the conformally invariant part of a spacetime metric from duality operators and relations like (49) and (51). This was later rediscovered by Jadczyk [40], whereas Wang [90] gave a revised presentation of Toupin's results. A forerunner was Peres [69], see in this context also the more recent papers by Piron and Moore [70]. Brans [7] and subsequently numerous authors discussed such structures in the framework of general relativity theory, see, e.g., [9,37,26,68] and the references given there.

It is convenient to adopt a more compact *bivector* notation by defining the indices $I, J, \cdots = (01, 02, 03, 23, 31, 12)$. Then $\overset{\circ}{\chi}{}^{ijkl}$ becomes the 6×6 matrix $\overset{\circ}{\chi}{}^{IK}$ and (51) goes over into

$$\overset{\circ}{\chi}{}^{IJ}\,\epsilon_{JK}\,\overset{\circ}{\chi}{}^{KL}\,\epsilon_{LM} = -\delta^I_M\,. \tag{52}$$

In terms of 3×3-constituents an arbitrary symmetric $\overset{\circ}{\chi}{}^{IK} = \overset{\circ}{\chi}{}^{KI}$ constitutive matrix reads

$$\overset{\circ}{\chi}{}^{IJ} = \overset{\circ}{\chi}{}^{JI} = \begin{pmatrix} A & C \\ C^{\mathrm{T}} & B \end{pmatrix}, \quad \epsilon^{IJ} = \epsilon^{JI} = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \tag{53}$$

where $A = A^{\mathrm{T}}, B = B^{\mathrm{T}}$, and the superscript $^{\mathrm{T}}$ denotes transposition. The algebraic condition $\epsilon_{IJ}\,\overset{\circ}{\chi}{}^{IJ} \equiv 0$ is provided by $\mathrm{tr}\, C = 0$.

The general non-trivial solution of the closure relation (52) can be written in the form

$$\overset{\circ}{\chi}{}^{IJ} = \begin{pmatrix} pB^{-1} + qN & B^{-1}K \\ -KB^{-1} & B \end{pmatrix}. \tag{54}$$

Here B is a nondegenerate arbitrary *symmetric* 3×3 matrix (6 independent components B_{ab}), K an arbitrary *antisymmetric* matrix (3 independent components $K_{ab} =: \epsilon_{abc}\,k^c$), N the symmetric matrix with components $N^{ab} := k^a k^b$, and $q := -1/\det B$, $p := [\mathrm{tr}(NB)/\det B] - 1$. Thus, Eq.(54) subsumes 9 independent components.

Triplet of Self-Dual 2-Forms. The duality operator $^{\#}$ induces a decomposition of the 6-dimensional space of 2-forms into two 3-dimensional invariant subspaces corresponding to the eigenvalues $\pm i$. Writing the 2-form basis

[8] One could define a different duality operator by $^{\widehat{\#}}\Theta_{ij} = \frac{f}{4}\epsilon_{ijkl}\,\overset{\circ}{\chi}{}^{klmn}\,\Theta_{mn}$ such that $^{\widehat{\#}\widehat{\#}} = -f^2$.

$\Theta^I = dx^i \wedge dx^j$ in terms of the two 3-dimensional column vectors

$$\Theta^I = \begin{pmatrix} \beta^a \\ \gamma_b \end{pmatrix}, \quad a, b, \cdots = 1, 2, 3, \tag{55}$$

one can construct the corresponding self-dual basis $\overset{(s)}{\Theta}{}^I := \frac{1}{2}(\Theta^I - i\,{}^{\#}\Theta^I)$. In the 3-vector representation,

$$\overset{(s)}{\Theta}{}^I = \begin{pmatrix} \overset{(s)}{\beta}{}^a \\ \overset{(s)}{\gamma}{}_b \end{pmatrix}, \tag{56}$$

one of the 3-dimensional invariant subspaces can be spanned either by the upper or by the lower components which are related to each other by a linear transformation. For example, $\overset{(s)}{\beta}$ can be expressed in terms of $\overset{(s)}{\gamma}$ according to $\overset{(s)}{\beta} = (i + B^{-1}K)B^{-1}\overset{(s)}{\gamma}$. Therefore $\overset{(s)}{\gamma}$ or, equivalently, the *triplet of 2-forms*

$$\begin{aligned} S^{(a)} &:= -(B^{-1})^{ab}\ \overset{(s)}{\gamma}{}_b \\ &= \frac{i}{2}\,(dx^0 \wedge dx^a - (\det B)^{-1}\, k_b\, dx^b \wedge dx^a \\ &\qquad + i\,(B^{-1})^{ab}\,\epsilon_{bcd}\, dx^c \wedge dx^d). \end{aligned} \tag{57}$$

subsume the properties of this invariant subspace. Each of the 2-forms carry 3 independent components, i.e., they add up to 9 components.

The information of the constitutive matrix $\overset{\circ}{\chi}{}^{IJ}$ is now encoded into the triplet of 2-forms $S^{(a)}$. One can verify that the latter satisfies the completeness relation

$$S^{(a)} \wedge S^{(b)} = \frac{1}{3}\,(B^{-1})^{ab}\,(B)_{cd}\, S^{(c)} \wedge S^{(d)}\,. \tag{58}$$

Extracting the Metric. Within the context of $SU(2)$ Yang-Mills theory, Urbantke [88] was able to derive a 4-dimensional spacetime metric g_{ij} from a triplet of 2-forms satisfying a completeness condition of the type (58). Explicitly, the Urbantke formulas read

$$\sqrt{\det g}\; g_{ij} = -\frac{2}{3}\sqrt{\det B}\,\epsilon_{abc}\,\epsilon^{klmn}\,S^{(a)}_{ik}\,S^{(b)}_{lm}\,S^{(c)}_{nj}\,, \tag{59}$$

$$\sqrt{\det g} \quad = -\frac{1}{6}\,\epsilon^{klmn}\,B_{cd}\,S^{(c)}_{kl}\,S^{(d)}_{mn}\,. \tag{60}$$

The $S^{(a)}_{ij}$ are the components of the 2-form triplet $S^{(a)} = S^{(a)}_{ij}dx^i \wedge dx^i/2$. If we substitute the forms (57) into (59) and (60), we can display the metric explicitly in terms of the constitutive coefficients:

$$g_{ij} = \frac{1}{\sqrt{\det B}}\left(\begin{array}{c|c} \det B & -\,k_a \\ \hline -\,k_b & -\,B_{ab} + (\det B)^{-1}\,k_a\,k_b \end{array}\right). \tag{61}$$

Here $k_a := B_{ab}k^b = B_{ab}\,\epsilon^{bcd}K_{cd}/2$. One can verify that the metric in (61) has Minkowskian signature. Since the triplet $S^{(a)}$ is defined up to an arbitrary scalar factor, we obtain a *conformal* class of metrics.

Given a metric, we can now define eventually the notion of local isotropy. Let $T^{i_1\dots i_p}$ be the contravariant coordinate components of a tensor field and $T^{\alpha_1\dots\alpha_p} := e_{i_1}{}^{\alpha_1}\cdots e_{i_p}{}^{\alpha_p}\,T^{i_1\dots i_p}$ its frame components with respect to an orthonormal frame $e_\alpha = e^i{}_\alpha\,\partial_i$. A tensor is said to be locally isotropic at a given point, if its frame components are invariant under a Lorentz rotation of the orthonormal frame. Similar considerations extend to tensor densities.

There are only two geometrical objects which are numerically invariant under local Lorentz transformations: the Minkowski metric $o_{\alpha\beta} = \mathrm{diag}(+1,-1,-1,-1)$ and the Levi-Civita tensor density $\epsilon_{\alpha\beta\gamma\delta}$. Thus

$$\mathcal{T}^{ijkl} = \phi(x)\,\sqrt{-g}\,\left(g^{ik}g^{jl} - g^{jk}g^{il}\right) + \varphi(x)\,\epsilon^{ijkl} \tag{62}$$

is the most general locally isotropic contravariant fourth rank tensor density of weight +1 with the symmetries $\mathcal{T}^{ijkl} = -\mathcal{T}^{jikl} = -\mathcal{T}^{ijlk} = \mathcal{T}^{klij}$. Here ϕ and φ are scalar and pseudo-scalar fields, respectively.

One can prove that the constitutive tensor in (45) with the closure property (51) is *locally isotropic with respect to the metric* (61), see also [63]. Accordingly, for the constitutive tensor, we finally have

$$\overset{\circ}{\chi}{}^{ijkl} = 2\sqrt{-g}\,g^{k[i}g^{j]l}\,. \tag{63}$$

Thus, the isotropic law reads

$$H_{ij} = \frac{f}{2}\,\epsilon_{ijmn}\,\sqrt{-g}\,g^{km}g^{ln}\,F_{kl} + \alpha\,F_{ij} \tag{64}$$

or, if written with the help of the Hodge star operator belonging to the metric (61),

$$H = (f\,{}^\star + \alpha)\,F\,. \tag{65}$$

6.5 Centrosymmetric

If we want the constitutive tensor to be reflection symmetric at each point of spacetime, i.e., if we require centrosymmetry, then we have to kill the Abelian axion and arrive, provided F is chosen in accordance with the SI-conventions, at the usual law for Maxwell-Lorentz vacuum electrodynamics[9]

$$H = f\,{}^\star F = \sqrt{\frac{\varepsilon_0}{\mu_0}}\,{}^\star F\,. \tag{66}$$

[9] Remember that in Ricci calculus the excitation is defined according to $\mathcal{H}^{ij}|_{Ric} = \epsilon^{ijkl}\,H_{kl}/2$. Then $\mathcal{H}^{ij}|_{Ric} = -\,f\sqrt{-g}\,F^{ij}$, see $(11)_1$.

7 Non-minimal Coupling Involving Curvature, Nonmetricity and Torsion?

7.1 Non-minimal Coupling Violating Charge and/or Flux Conservation

The Maxwell equation $dH = J$ reflects (and comes from) the electric current conservation, $dJ = 0$, see Sec.4. By modifying the left hand side of $dH = J$, one can arrive at a model violating charge conservation. Such a modification can typically originate from a non-minimal coupling of the electromagnetic to the gravitational field. Given the torsion 2-form T^α, one can consider, for example, the field equation

$$dH + \alpha\ (e_\alpha \rfloor T^\alpha) \wedge H + \beta\, {}^\star(\vartheta_\alpha \wedge T^\alpha) \wedge H = J\,, \tag{67}$$

or, with the nonmetricity 1-form $Q_{\alpha\beta} := -Dg_{\alpha\beta}$ and the Weyl 1-form $Q := Q_\alpha{}^\alpha/4$,

$$dH + \gamma\, Q \wedge H + \delta\, {}^\star(\vartheta^\alpha \wedge \vartheta^\beta \wedge Q_{\alpha\beta}) \wedge H = J\,. \tag{68}$$

Similar non-minimal terms could emerge in $dF = 0$. However, curvature dependent terms cannot be accommodated at the level of the Maxwell equations, since the contraction of the indices produces always a form of even rank whereas the Maxwell equations are represented by 3-forms, i.e., by forms of odd rank. In any case, violating charge or flux conservation is not possible without giving up most of the experimentally established structure of the theory of electromagnetism. Therefore we will not follow this path.

Incidentally, there are some papers in the literature in which the conventional vacuum constitutive law $H \sim {}^\star F$ is uphold, but the Maxwell equations are coupled to torsion in an inconsistent way. A closer inspection of the papers [75,15,82] shows that the proposed "non-minimal" coupling of torsion to the electromagnetic field is void of physical contents. In fact, torsion drops out if the algebra is done correctly.

Another procedure comes to mind if we talk about the violation of the conservation laws. Hojman et al. [36] introduced a new scalar field $\varphi(x)$, the *tlaplon*, see also Mukku & Sayed [57]. The gradient $d\varphi$ of the tlaplon was put proportional to the trace part ${}^{(2)}T^\alpha := \vartheta^\alpha \wedge T$ of the torsion T^α, see [31]; here $T := e_\beta \rfloor T^\beta$. In fact, we have $T = \frac{3}{2}\, d\varphi$.

Superficially, the axial scalar $\alpha(x)$ (Abelian axion) and the scalar tlaplon $\varphi(x)$ may look similar. However, the axion already emerges from spacetime viewed as a differential manifold as soon as a linear constitutive law is assumed for electromagnetism, whereas the tlaplon can only be introduced if the differential manifold in equipped with a linear connection. In other words, the axion is a pre-metric and a pre-connection animal, the tlaplon, in contrast, needs to be 'housed' in a linear connection.

Moreover, as we saw, the axion respects the conservation laws (and pleases us thereby), whereas the tlaplon defies these rules and appears as an anti-

electromagnetic creature. The electromagnetic field is not defined in the conventional way, namely by $F = dA$. Instead, Hojman et al. define it via a "covariant" derivative according to

$$\widehat{F} = F - T^\alpha A_\alpha = F + \frac{1}{3}\, T \wedge A = F + \frac{1}{2}\, d\varphi \wedge A. \tag{69}$$

Thus the "electromagnetic" Lagrangian becomes

$$\widehat{F} \wedge {}^\star \widehat{F} = F \wedge {}^\star F + d\varphi \wedge A \wedge {}^\star F + \frac{1}{4}\, T \wedge A \wedge {}^\star (T \wedge A)\,. \tag{70}$$

The last term does not contribute to the excitation $H = -\partial L/\partial F$, but the second term produces a contribution

$$\sim {}^\star (d\varphi \wedge A). \tag{71}$$

We can compare now the two contributions from the axion (48) and the tlaplon (71). They are reminiscent of each other since one is equal to the Hodge dual of the other. Therefore, in the tlaplon case, besides a connection, we need additionally a metric. But the most decisive difference is, as can be read off from (70), that the Maxwell equations get amended and axiom 1 and axiom 3 are no longer valid.

7.2 "Admissible" Non-minimal Coupling

The message is then that a change of the Maxwell equations $dH = J\,,\ dF = 0$ is to be avoided, unless one allows for a violation of electric charge or magnetic flux conservation. By introducing the metric g^{ij} into the constitutive law, one gets a smooth and natural transition from special relativity to general relativity and to gauge theories of gravity. In the constitutive law for vacuum one could imagine, along with the contributions depending only on the metric, couplings like [23]

$$\begin{aligned} \chi^{ijkl} = {} & A_1\, R^{ijkl} + A_2\, R^{*\,ijkl} + A_3\, {}^\star R^{*\,ijkl} \\ & + A_4\, (Ric^{i[k} g^{l]j} - Ric^{j[k} g^{l]i}) + A_5\, R\, g^{i[k} g^{l]j} + A_6\, R\, \epsilon^{ijkl} \end{aligned} \tag{72}$$

without violating the conservation laws. Here $R := e^\alpha \rfloor e_\beta \rfloor R_\alpha{}^\beta$ is the curvature scalar, and we denote the *right* or *Lie* dual of an $so(1,3)$-valued form $\psi_{\alpha\beta}$ by $\psi^*_{\alpha\beta} := \frac{1}{2}\, \epsilon_{\alpha\beta\mu\nu}\, \psi^{\mu\nu}$.

Choose, for example, $\chi^{ijkl} = 2\sqrt{-g} f_0 (1 + \beta^2\, R)\, g^{i[k} g^{l]j}$, then the inhomogeneous Maxwell equation would read

$$d\, {}^\star F + \beta^2 d\, (R\, {}^\star F) = \frac{1}{f_0}\, J\,, \tag{73}$$

i.e., a coupling of curvature and electromagnetic field strength would be possible. However, one had to introduce a new natural constant with the dimension of

$[\beta] = 1/\text{length}$. On the level of the Lagrangian, this coupling would be non-minimal,

$$V_{\text{non}-\text{m}} = -\frac{f_0}{2}\left(1+\beta^2 R\right) {}^\star F \wedge F\,, \tag{74}$$

but such an ansatz would not spoil the fundamental principles of electrodynamics; it would seem to be the most natural way of achieving an RF-coupling. Goenner [23], see also the literature given there, derived this non-minimal Lagrangian from some fundamental principles, like the existence of a decent Newtonian limit. However, in his view, such a model violates charge conservation. We disagree with him on this point.

We want to stress that one cannot achieve a similar non-minimal coupling to the torsion T^α of spacetime. First of all, the Maxwell equations are independent of torsion. By means of the constitutive law one maps the 2-form F to the 2-form H. The curvature $R_\alpha{}^\beta$ is a 2-form of type $(1,1)$, i.e., it carries two $GL(4,R)$ indices, whereas the torsion 2-form carries only one index. Therefore, by contraction, we cannot get a scalar out of the torsion. A coupling like $(e_\alpha \rfloor T^\alpha) \wedge {}^\star F$ is not possible, since this is a 3-form. However, higher powers in T^α would be possible such as

$$H = f_0\left[1+\gamma^2\,{}^\star(T^\alpha \wedge T_\alpha)\right] \wedge {}^\star F \tag{75}$$

or

$$H = f_0\left[1+\delta^4\,{}^\star\big((e_\alpha\rfloor T^\alpha)\wedge(e_\beta\rfloor T^\beta)\wedge(e_\gamma\rfloor T^\gamma)\wedge(e_\delta\rfloor T^\delta)\big)\right] \wedge {}^\star F\,. \tag{76}$$

Here $[\gamma] = [\delta] = 1/\mathit{length}$. Accordingly, it is not too difficult to introduce a coupling of torsion to electromagnetism. However, the price one has to pay is the introduction of new natural constants γ and δ. In other words, even if possible, we don't take such models too seriously.

Also non-minimally coupled nonmetricity could be installed by additional quadratic pieces such as

$$H = f_0\left[1+\xi^2(e_\alpha\rfloor Q^{\alpha\gamma})(e_\beta\rfloor Q^\beta{}_\gamma)\right] {}^\star F\,. \tag{77}$$

Therefore, there are quite a number of different "admissible" options available as soon as we allow non-minimal couplings to arise.

8 Outlook

Using astronomical observations on the propagation of light, the upper bounds for non-minimal coupling effects should be determined in a systematic way, as is done, for example, by Haugan and Lämmerzahl [28]. For such a purpose, we will develop [66] the geometrical optics limit of the Maxwell equations $dH = J$, $dF = 0$ and will use particular constitutive laws, as, e.g., the linear law. Possible couplings to curvature, torsion, and nonmetricity should come under sharper focus in this way. Non-linear effect à la Ayón-Beato & García should be investigated in the context of, say, the metric-affine gauge theory of gravity.

Acknowledgments

The authors are indebted to Steve Adler (Princeton), Tetsuo Fukui (Nishinomiya), Hubert Goenner (Göttingen), Yakov Itin (Jerusalem), Claus Lämmerzahl (Konstanz), Bahram Mashhoon (Columbia, MO), and Guillermo Rubilar (Cologne) for helpful remarks and/or interesting discussions. FWH is grateful to the Institute for Advanced Study for hospitality and to the Volkswagen-Stiftung, Hannover for support.

References

1. B. Abbott et al. (D0 Collaboration), *A search for heavy pointlike Dirac monopoles*, *Phys. Rev. Lett.* **81** (1998) 524-529.
2. E. Ayón-Beato and A. García, *Regular black hole in general relativity coupled to nonlinear electrodynamics*, *Phys. Rev. Lett.* **80** (1998) 5056-5059.
3. E. Ayón-Beato and A. García, *Non-singular charged black hole solution for nonlinear source*, *Gen. Rel. Grav. J.* **31** (1999) 629-633.
4. E. Ayón-Beato and A. García, *New regular black hole solution from nonlinear electrodynamics*, *Phys. Lett.* **B464** (1999) 25-29.
5. I.M. Benn, T. Dereli, and R.W. Tucker, *Gauge field interactions in spaces with arbitrary torsion*, *Phys. Lett.* **B96** (1980) 100-104.
6. M. Born and L. Infeld, *Foundations of the new field theory*, *Proc. Roy. Soc. (London)* **A144** (1934) 425-451.
7. C.H. Brans, *Complex 2-forms representation of the Einstein equations: The Petrov Type III solutions*, *J. Math. Phys.* **12** (1971) 1616-1619.
8. H.A. Buchdahl, *On a Lagrangian for non-minimally coupled gravitational and electromagnetic fields*, *J. Phys.* **A12** (1979) 1037-1043.
9. R. Capovilla, T. Jacobson, and J. Dell, *General relativity without the metric*, *Phys. Rev. Lett.* **63** (1989) 2325-2328.
10. S.M. Carroll, G.B. Field, and R. Jackiw, *Limits on a Lorentz- and parity-violating modification of electrodynamics*, *Phys. Rev.* **D41** (1990) 1231-1240.
11. C. Caso et al. (Particle Data Group), *European Phys. J.* **C3** (1998) 1-794 and 1999 partial update for edition 2000 (ULR: http://pdg.lbl.gov), here: *Axions and other Very Light Bosons.*
12. L. Cooper and G.E. Stedman, *Axion detection by ring lasers*, *Phys. Lett.* **B357** (1995) 464-468.
13. H. Dehmelt, R. Mittleman, R.S. van Dyck, Jr., and P. Schwinberg, *Past electron positron $g-2$ experiments yielded sharpest bound on CPT violation*, *Phys. Rev. Lett.* **83** (1999) 4694-4696.
14. V.C. de Andrade, J.G. Pereira, *Torsion and the electromagnetic field*, *Int. J. Mod. Phys.* **D8** (1999) 141-151.
15. R. de Ritis, M. Lavorgna, and C. Stornaiolo, *Geometric optics in a Riemann-Cartan space-time*, *Phys. Lett.* **A98** (1983) 411-413.
16. V. de Sabbata and C. Sivaram, *Spin and Torsion in Gravitation.* Singapore, World Scientific (1994).
17. I.T. Drummond and S.J. Hathrell, *QED vacuum polarization in a background gravitational field and its effect on the velocity of photons*, *Phys. Rev.* **D22** (1980) 343-355.

18. A.S. Eddington, *The Mathematical Theory of Relativity.* Cambridge University Press, Cambridge (1924) Sec.74(b).
19. A. Einstein, *Eine neue formale Deutung der Maxwellschen Feldgleichungen der Elektrodynamik*, *Sitzungsber. Königl. Preuss. Akad. Wiss. (Berlin)* (1916) pp. 184-188; see also *The Collected Papers of Albert Einstein.* Vol.6, A.J. Kox et al., eds. (1996) pp. 263-269.
20. A. Einstein, *The Meaning of Relativity,* 5th edition. Princeton University Press, Princeton (1955).
21. T. Frankel, *The Geometry of Physics* – An Introduction. Cambridge University Press, Cambridge (1997).
22. G.W. Gibbons and D.A. Rasheed, *Magnetic duality rotations in nonlinear electrodynamics, Nucl. Phys.* **B454** (1995) 185-206.
23. H. Goenner, *Theories of gravitation with nonminimal coupling of matter and the gravitational field*, *Found. Phys.* **14** (1984) 865-881.
24. W. Gordon, *Zur Lichtfortpflanzung nach der Relativitätstheorie, Ann. Phys. (Leipzig)* **72** (1923) 421-456.
25. F. Gronwald and F. W. Hehl, *On the gauge aspects of gravity*, in: *International School of Cosmology and Gravitation:* 14$^{\text{th}}$ Course: Quantum Gravity, held May 1995 in Erice, Italy. Proceedings. P.G. Bergmann et al. (eds.). World Scientific, Singapore (1996) pp. 148–198. Los Alamos Eprint Archive, gr-qc/9602013.
26. G. Harnett, *Metrics and dual operators*, *J. Math. Phys.* **32** (1991) 84-91.
27. D. Hartley: *Normal frames for non-Riemannian connections*, *Class. Quantum Grav.* **12** (1995) L103-L105.
28. M. Haugan and C. Lämmerzahl, *On the experimental foundations of the Maxwell equations*, *Ann. Physik (Leipzig)*, to be published (2000).
29. Y.D. He, *Search for a Dirac magnetic monopole in high energy nucleus-nucleus collisions*, *Phys. Rev. Lett.* **79** (1997) 3134-3137.
30. F.W. Hehl, P. von der Heyde, G.D. Kerlick, and J.M. Nester, *General relativity with spin and torsion: Foundations and prospects*, *Rev. Mod. Phys.* **48** (1976) 393-416.
31. F.W. Hehl, J.D. McCrea, E.W. Mielke, and Y. Ne'eman, *Metric-affine gauge theory of gravity: Field equations, Noether identities, world spinors, and breaking of dilation invariance*, *Phys. Rep.* **258** (1995) 1-171.
32. F.W. Hehl, Yu.N. Obukhov, and G.F. Rubilar, *Classical electrodynamics: A Tutorial on its Foundations*, in: *Quo vadis geodesia...? Festschrift for Erik W. Grafarend*, F. Krumm and V.S. Schwarze (eds.) Univ. Stuttgart, ISSN 0933-2839 (1999) pp. 171-184. Los Alamos Eprint Archive, physics/9907046.
33. F.W. Hehl, Yu.N. Obukhov, and G.F. Rubilar, *Spacetime metric from linear electrodynamics II*, talk given at International European Conference on Gravitation: *Journées Relativistes 99, Weimar, Germany, 12-17 Sep 1999. Ann. Phys. (Leipzig)* **9** (2000), in print; Los Alamos Eprint Archive, gr-qc/9911096.
34. W. Heisenberg and H. Euler, *Consequences of Dirac's theory of positrons*, *Z. Phys.* **98** (1936) 714-732 (in German).
35. J.S. Heyl and L. Hernquist, *Birefringence and dichroism of the QED vacuum*, *J. Phys.* **A30** (1997) 6485-6492.
36. S. Hojman, M. Rosenbaum, M.P. Ryan, and L.C. Shepley, *Gauge invariance, minimal coupling, and torsion*, *Phys. Rev.* **D17** (1978) 3141-3146.
37. G. 't Hooft, *A chiral alternative to the vierbein field in general relativity*, *Nucl. Phys.* **B357** (1991) 211-221.
38. B.Z. Iliev, *Normal frames and the validity of the equivalence principle. 1. Cases in a neighborhood and at a point*, *J. Phys.* **A29** (1996) 6895-6902.

39. C. Itzykson and J.-B. Zuber, *Quantum Field Theory*. McGraw Hill, New York (1985).
40. A.Z. Jadczyk, *Electromagnetic permeability of the vacuum and light-cone structure*, *Bull. Acad. Pol. Sci., Sér. sci. phys. et astr.* **27** (1979) 91-94.
41. J.K. Jain, *Composite-fermion approach for the fractional quantum Hall effect*, *Phys. Rev. Lett.* **63** (1989) 199-202.
42. J.K. Jain, *Composite fermion theory of fractional quantum Hall effect*, *Acta Phys. Polon.* **B26** (1995) 2149-2166.
43. B.L. Johnson and G. Kirczenow, *Composite fermions in the quantum Hall effect*, *Rep. Prog. Phys.* **60** (1997) 889-939.
44. M.W. Keller, A.L. Eichenberger, J.M. Martinis, and N.M. Zimmerman, *A capacitance standard based on counting electrons*, *Science* **285** (1999) 1706-1709.
45. E.W. Kolb and M.S. Turner, *The Early Universe*. Addison-Wesley, Redwood City (1990) Chapter 10: Axions.
46. C. Lämmerzahl, R.A.Puntigam, and F.W.Hehl, *Can the electromagnetic field couple to post-Riemannian structures?* In: *Proceedings of the 8th Marcel Grossmann Meeting on General Relativity, Jerusalem 1997*, T. Piran and R. Ruffini, eds.. World Scientific, Singapore (1999).
47. A. Lue, L. Wang, and M. Kamionkowski, *Cosmological signature of new parity-violating interactions*, *Phys. Rev. Lett.* **83** (1999) 1506-1509.
48. P. Majumdar and S. SenGupta, *Parity violating gravitational coupling of electromagnetic fields, Class. Quantum Grav.* **16** (1999) L89-L94.
49. B. Mashhoon, *Nonlocal electrodynamics*, in: *Cosmology and Gravitation*, Proc. VII Brasilian School of Cosmology and Gravitation, Rio de Janeiro, August 1993, M. Novello, editor. Editions Frontieres, Gif-sur-Yvette (1994) pp. 245-295.
50. J.C. Maxwell, *A dynamical theory of the electromagnetic field,* Ref.[65], Vol.1, pp. 526-597, see, in particular, Part III of this article.
51. E.W. Mielke, *Geometrodynamics of Gauge Fields* – On the geometry of Yang-Mills and gravitational gauge theories. Akademie–Verlag, Berlin (1987).
52. C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation*. Freeman, San Francisco (1973).
53. S. Mohanty and A.R. Prasanna, *Photon propagation in Einstein and higher derivative gravity*, *Nucl. Phys.* **B526** (1998) 501-508.
54. J.E. Moody and F. Wilczek, *New macroscopic forces? Phys. Rev.* **D30** (1984) 130-138.
55. U. Muench, F.W. Hehl, and B. Mashhoon, *Acceleration-induced nonlocal electrodynamics in Minkowski spacetime*, *Preprint Univ. Missouri-Columbia (March 2000)*, 14 pp.; Los Alamos Eprint Archive, gr-qc/0003093.
56. B. Mukhopadhyaya and S. Sengupta, *A geometrical interpretation of parity violation in gravity with torsion, Phys. Lett.* **B458** (1999) 8-12.
57. C. Mukku and W.A. Sayed, *Torsion without torsion*, *Phys. Lett.* **B82** (1979) 382-386.
58. F. Müller-Hoissen, *Non-minimal coupling from dimensional reduction of the Gauss-Bonnet action*, *Phys. Lett.* **B201** (1988) 325-327.
59. F. Müller-Hoissen, *Modification of Einstein Yang-Mills theory from dimensional reduction of the Gauss-Bonnet action, Class. Quantum Grav.* **5** (1988) L35-L40.
60. F. Müller-Hoissen and R. Sippel, *Spherically symmetric solutions of the nonminimally coupled Einstein-Maxwell equations, Class. Quantum Grav.* **5** (1988) 1473-1488.
61. Y. Nambu, *The Aharonov-Bohm problem revisited,* Los Alamos Eprint archive: hep-th/9810182.

62. W.-T. Ni, *A non-metric theory of gravity*. Dept. Physics, Montana State University, Bozeman. Preprint December 1973. [This paper is referred to by W.-T. Ni in Bull. Amer. Phys. Soc. 19 (1974) 655.] The paper is available via http://gravity5.phys.nthu.edu.tw/ .
63. W.-T. Ni, *Equivalence principles and electromagnetism*, *Phys. Rev. Lett.* **38** (1977) 301-304.
64. W.-T. Ni, S.-S. Pan, H.-C. Yeh, L.-S. Hou, and J.-L. Wan, *Search for an axionlike spin coupling using a paramagnetic salt with a dc SQUID*, *Phys. Rev. Lett.* **82** (1999) 2439-2442.
65. W.D. Niven (ed.), *The Scientific Papers of James Clerk Maxwell*, 2 Volumes. Dover, New York (1965).
66. Yu.N. Obukhov, T. Fukui, and G.F. Rubilar, *Wave propagation in linear electrodynamics*, draft, University of Cologne (Feb. 2000).
67. Yu.N. Obukhov and F.W. Hehl, *Space-time metric from linear electrodynamics*, *Phys. Lett.* **B458** (1999) 466-470.
68. Yu.N. Obukhov and S.I. Tertychniy, *Vacuum Einstein equations in terms of curvature forms*, *Class. Quantum Grav.* **13** (1996) 1623-1640.
69. A. Peres, *Electromagnetism, geometry, and the equivalence principle*, *Ann. Phys. (NY)* **19** (1962) 279-286.
70. C. Piron and D.J. Moore, *New aspects of field theory*, *Turk. J. Phys.* **19** (1995) 202-216.
71. J. Plebański, *Non-Linear Electrodynamics – a Study*, Nordita (1968). Our copy is undated and stems from the CINVESTAV Library, Mexico City (courtesy A. Macías).
72. E.J. Post, *Formal Structure of Electromagnetics* – General Covariance and Electromagnetics. North Holland, Amsterdam (1962) and Dover, Mineola, New York (1997).
73. E.J. Post, *Quantum Reprogramming – Ensembles and Single Systems: A Two-Tier Approach to Quantum Mechanics*. Kluwer, Dordrecht (1995).
74. A.R. Prasanna, *A new invariant for electromagnetic fields in curved space-time*, *Phys. Lett.* **A37** (1971) 331-332.
75. A.R. Prasanna, *Maxwell's equations in Riemann-Cartan space U_4*, *Phys. Lett.* **A54** (1975) 17-18.
76. R.A. Puntigam, C. Lämmerzahl, and F.W. Hehl, *Maxwell's theory on a post-Riemannian spacetime and the equivalence principle*, *Class. Quantum Grav.* **14** (1997) 1347-1356.
77. M. Schönberg, *Electromagnetism and gravitation*, *Rivista Brasileira de Fisica* **1** (1971) 91-122.
78. J.A. Schouten, *Tensor Analysis for Physicists*. 2nd edition. Dover, Mineola, New York (1989).
79. G.N. Shikin, *Static spherically symmetric solutions of the system of Einstein equations and non-linear electrodynamics equations with an arbitrary Lagrangian*, in: *"Relativity theory and gravitation"*, V.I. Rodichev, ed. Nauka, Moscow (1976) pp. 129-132 (in Russian).
80. P. Sikivie, ed., *Axions '98*. Proceedings of the 5th IFT Workshop on Axions, Gainesville, Florida, USA. *Nucl. Phys.* **B** (Proc. Suppl.) **72** (1999) 1-240.
81. G.V. Skrotskii, *The influence of gravitation on the propagation of light*, *Sov. Phys. Doklady* **2** (1957) 226-229.
82. L.L. Smalley, *On the extension of geometric optics from Riemannian to Riemann-Cartan spacetime*, *Phys. Lett.* **A117** (1986) 267-269.

83. J. Stachel, *Covariant formulation of the Cauchy problem in generalized electrodynamics and general relativity*, *Acta Phys. Polon.* **35** (1969) 689-709.
84. M. Tinkham, *Introduction to Superconductivity.* 2nd ed.. McGraw-Hill, New York (1996).
85. R.A. Toupin, *Elasticity and electro-magnetics*, in: *Non-Linear Continuum Theories, C.I.M.E. Conference, Bressanone, Italy 1965.* C. Truesdell and G. Grioli coordinators. Pp.203-342.
86. A. Trautman, *Gauge and optical aspects of gravitation*, *Class. Quantum Grav.* **16** (1999) A157-A175.
87. C. Truesdell and R.A. Toupin, *The Classical Field Theories*, in: *Handbuch der Physik,* Vol. III/1, S. Flügge ed.. Springer, Berlin (1960) pp. 226-793.
88. H. Urbantke, *A quasi-metric associated with $SU(2)$ Yang-Mills field*, *Acta Phys. Austriaca Suppl.* **XIX** (1978) 875-816.
89. A.M. Volkov, A.A. Izmest'ev, and G.V. Skrotskii, *The propagation of electromagnetic waves in a Riemannian space*, *Sov. Phys. JETP* **32** (1971) 686-689 [*ZhETF* **59** (1970) 1254-1261 (in Russian)].
90. C. Wang, *Mathematical Principles of Mechanics and Electromagnetism, Part B: Electromagnetism and Gravitation.* Plenum Press, New York (1979).
91. S. Weinberg, *A new light boson? Phys. Rev. Lett.* **40** (1978) 223-226.
92. F. Wilczek, *Problem of strong P and T invariance in the presence of instantons*, *Phys. Rev. Lett.* **40** (1978) 279-282.

Subject Index